Contents lists available at ScienceDirect

Data in Brief

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

Data Article

Whole genome sequencing data of greater burdock (*Arctium lappa*) naturalized in the United States of America

Danae Kala Rodriguez Bardaji, Girish Kumar, Samantha Tran, Gabriella Fedus, Michael A. Savka, Dawn Carter, André O. Hudson*

Thomas H. Gosnell School of Life Sciences, College of Science, Rochester Institute of Technology, 85 Lomb Memorial Drive, Rochester NY 14623, United States

ARTICLE INFO

Article history:
Received 14 January 2025
Revised 14 February 2025
Accepted 19 March 2025
Available online 21 March 2025

Dataset link: Dataset for Arctium lappa genome (Original data)

Keywords:
Illumina
Greater burdock
Genome annotation
Antimicrobial compound

ABSTRACT

This dataset comprises whole genome sequencing burdock (*Arctium lappa*) naturalized in a residential yard in Rochester, New York, USA. Total DNA was extracted from a leaf sample and processed using the Illumina Nextera XT DNA library preparation kit. Sequencing on the NextSeq 2000 platform produced 127.4 GB of raw data, yielding 125.8 GB of high-quality reads after filtering, with an average genome coverage of 75x. The genome was assembled de novo into 792,817 contigs, achieving a total genome length of 1,075,454,921 base pairs with a GC content of 37.03 %. Scaffolding against a Chinese *A. lappa* reference genome improved genome completeness from 49.1 % to 94.93 %, successfully recovering the majority of protein-coding genes.

Variant analysis identified approximately 20.8 million Single Nucleotide Polymorphisms (SNPs) and 1.3 million indels, including functionally significant mutations. The Internal Transcribed Spacer 2 (ITS2) ribosomal region was isolated and compared with global references, revealing significant genetic differentiation between the U.S.A and Chinese populations. This comprehensive genomic dataset has been deposited in publicly accessible repositories, including National Center for Biotechnology Information (NC and Zenodo. The sequencing of this sample provides a valuable resource

* Corresponding author.
E-mail address: aohsbi@rit.edu (A.O. Hudson).
Social media: @HudsonLabRIT (A.O. Hudson)

https://doi.org/10.1016/j.dib.2025.111508

for comparative genomics, population genetics, and investigations into bioactive compounds with antimicrobial properties, supporting agricultural and pharmaceutical applications. Direct access to the dataset is available at 10.5281/zenodo.14607136

Specifications Table

Subject	Biology
Specific subject area	Plant genomics, population genetics
Type of data	Sequencing raw reads, variant calls, genome sequence, Table and Figure
Data collection	Total DNA was extracted from a leaf sample of *Arctium lappa* individual. Approximately 100 ng of unfragmented genomic DNA was processed using the Illumina Nextera XT DNA library preparation kit followed by sequencing on the NextSeq2000 (2 × 150 bp read configuration) generating 127.4 gb of data, representing an average genome coverage of 75x
Data source location	The *Arctium lappa* sample was collected from a residential garden in Rochester, New York, USA in August 2024
Data accessibility	Raw Sequencing data has been deposited in NCBI under the Bioproject Number PRJNA1167729 and BioSample Number SAMN44012260. The draft genome, scaffolded genome (fasta), variant calls (vcf) and Internal Transcriped Spacer 2 (ITS2) region (fasta) have been deposited in the zenodo database. Direct URL to data: 10.5281/zenodo.14607136
Related research article	None

1. Value of the Data

- This study presents the first genomic representation of naturalized wild type burdock (*Arctium lappa*) in the United States, filling a critical gap in the genomic resources available for this species. The data provide a foundation for understanding its genetic makeup and evolutionary history.
- Comparative genomic analysis reveals insights into the genetic variation present in the U.S.-planted greater burdock when contrasted with its reference genome from China. This comparison highlights significant genetic differentiation, including approximately 20.8 million identified SNPs, reflecting potential adaptations or genetic drift in the U.S. population.
- Recovery and analysis of the ITS2 region, a high-copy-number ribosomal marker, enable us to examine the genetic similarity between the U.S. and Chinese populations of greater burdock. The ITS2 sequence from the U.S. population displayed 100 % identity with an ITS2 sequence from *A. lappa* sample from Ontario, Canada, while showing slight divergence (98.16 % identity, at least 4 mismatches) from Chinese samples.
- Understanding the genomics of *A. lappa* provides a valuable framework for identifying genes as well as their gene variants involved in the biosynthesis of bioactive compounds with antimicrobial properties. These compounds, which are integral to the plant's defense mechanisms, hold significant potential for pharmaceutical and agricultural applications.

2. Background

Arctium lappa (greater burdock) has been traditionally used in various cultures for its antimicrobial effects, primarily attributed to its bioactive compounds such as flavonoids, tannins, terpenes and phenolic acid [1]. Given its rich pharmacological profile, *A. lappa* holds great potential for use in agricultural, pharmaceutical, and cosmetic industries, especially as a natural alternative to synthetic antimicrobial agents [2]. The original motivation for compiling this dataset stems from the need to bridge a significant knowledge gap in the genomic resources available for this plant, particularly for naturalized populations in the United States. The dataset was generated to provide comprehensive genomic insights into the plant's genetic diversity, evolutionary history, and potential adaptations, offering a basis for future comparative and functional studies. This research aligns with theoretical frameworks in plant genomics and evolutionary biology, emphasizing the importance of understanding genetic variation and its implications for species adaptation. Methodologically, the study leveraged state-of-the-art next-generation sequencing and bioinformatics pipelines to ensure high-quality genome assembly and annotation, enabling a robust foundation for further exploration of the species' biosynthetic pathways, particularly those linked to antimicrobial compound production. By presenting the first U.S.-based genomic representation of *A. lappa*, the dataset enriches global genomic databases, creating opportunities for cross-population genetic studies. If this data article complements a related research article, it adds significant value by providing raw and processed genomic data critical for replication, validation, and extended analyses.

3. Data Description

A total of 127.4 GB of sequencing data, comprising 846 million reads, was generated. After filtering, 125.8 GB of high-quality data (836 million paired-end reads) remained. The draft genome of greater burdock (*Arctium lappa*) was assembled into 792,817 contigs, with a total length of 1,075,454,921 base pairs, an N50 of 1.7 kb, and a GC content of 37.03 %. Initial genome completeness, assessed based on single-copy genes, was 49.1 %, but this increased to 94.93 % after scaffolding the draft genome to the *A. lappa* reference genome. This improvement indicates that the majority of protein-coding genes were successfully captured in the draft assembly, though they were fragmented due to repetitive regions. The complete Internal Transcribed Spacer 2 (ITS2) region (217 bp, GC content 61.75 %) was recovered and showed 100 % sequence identity to the ITS2 sequence of *A. lappa* from Ontario, Canada (GenBank Accession Number: MG217926). The next closest matches were to samples from China (Fig. 1). Alignment of the filtered reads to the Chinese reference genome achieved a 96 % alignment rate, revealing approximately 20.8 million SNPs and 1.3 million indel variants. This corresponds to a variant rate of ∼1 per 77 base pairs (1.29 %), indicating significant genetic differentiation from the available *A. lappa* genome from China. Among the identified SNPs, 249,085 were missense mutations, 8637 were nonsense mutations, and 179,699 were silent mutations.

4. Experimental Design, Materials and Methods

4.1. Materials

A leaf sample of *A. lappa* was obtained from a residential garden at Rochester, New York in August 2024. The plant was matured and going through senescence when the tissue was harvested for DNA extraction.

4.2. DNA Extraction, DNA Quantification and Quality Control

Genomic DNA was extracted from *A. lappa* leaf using DNeasy plant pro mini kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. DNA concentration was measured using a Qubit 4.0 fluorometer, while DNA quality was evaluated with a NanoDrop spectrophotometer. DNA integrity was assessed by standard gel electrophoresis on a 1 % (w/v) agarose gel.

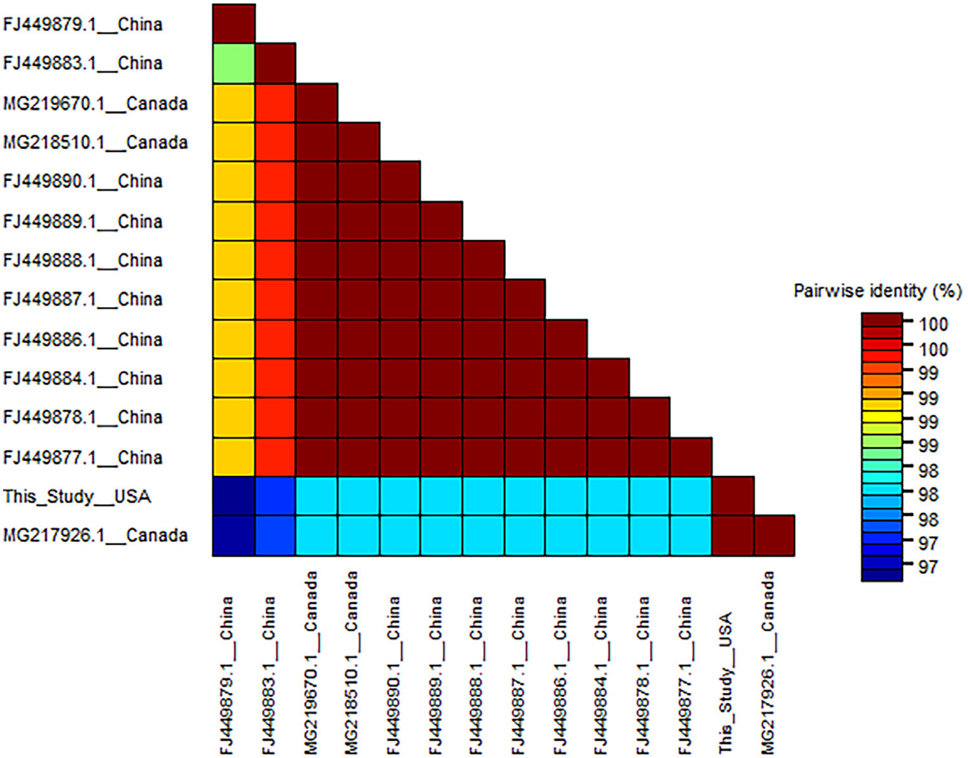

Fig. 1. Sequence identity heatmap showing pairwise percent identity in nucleotide composition between the ITS2 region of selected *A. lappa* samples. Heatmaps were drawn using SDT v1.4 software [3].

4.3. DNA Library Preparation and Sequencing

Sequencing library was prepared using standard procedure outlined in the Nextera XT library preparation kit (Illumina, San Diego, CA). Following library construction, the fragment size distribution was assessed via D1000 screen tape on an Agilent TapeStation 4200. The libraries were then quantified using a Qubit 4.0 fluorometer and diluted to 2 nM prior to sequencing on an Illumina NextSeq 2000 (P2 (2 × 150 cycles) at the Genomics Lab, Rochester Institute of Technology.

4.4. De Novo Assembly

Raw reads were quality and poly-G trimmed using fastp v0.20.1 [4] retaining paired-end reads that are longer than 100 bp. The filtered reads were assembled de novo using megahit v1.2.9 (default settings) [5]. Contigs larger than 500 bp were subsequently used for further analysis. Genome statistics was calculated using seqkit v2.8.2 [6] while genome completeness used compleasm [7] based on the eudicots_odb10 database. Scaffolding of the assembled contigs to the reference genome (Accession Code: GCA_023525745) [8] used RagTag v2.1.0 (default settings) [9]. To recover the high-copy-number ITS2 region, the filtered reads were subsampled to retain 1 million reads. These subsampled reads were subjected to de novo assembly, and the resulting contigs were screened for the ITS2 region using BLASTN against publicly available ITS2 homologs. ITS2 sequences showing significant homologies were downloaded and used as the input for pairwise identification calculation and visualization using SDT v1.4 software (default setting) [3].

4.5. Variant Calling

Filtered reads were aligned to the *Arctium lappa* reference genome using BWA-MEM2 v2.2.1 with default parameters [10]. The resulting BAM file was used as input for variant calling with BCFtools v1.13 [11]. Variants were filtered to retain variants with a quality score >30, read depth ≥10, and mapping quality ≥20. Annotation of the identified variants was performed using SnpEff v5 [12].

Limitations

Whole genome sequencing was conducted on a single individual, which may not fully capture the genetic diversity of burdock populations currently cultivated in the U.S.. This limitation is particularly significant when considering bioactive compounds with antimicrobial properties. The genetic pathways associated with the biosynthesis of these compounds can vary between individuals and populations due to environmental pressures, genetic drift, and selective breeding.

Ethics Statement

This work does not involve human subjects or animal subjects. The manuscript is an original work and has not been published elsewhere.

Data Availability

Dataset for Arctium lappa genome (Original data) (Zenodo).

CRediT Author Statement

Danae Kala Rodriguez Bardaji: Conceptualization, Investigation, Methodology, Writing – review & editing; **Girish Kumar:** Investigation, Methodology, Data curation, Software, Formal analysis; **Samantha Tran:** Investigation, Methodology; **Gabriella Fedus:** Investigation, Methodology; **Michael A. Savka:** Conceptualization, Investigation, Writing – review & editing; **Dawn Carter:** Investigation, Writing – review & editing; **André O. Hudson:** Conceptualization, Supervision, Funding acquisition, Writing – original draft, Writing – review & editing.

Acknowledgments

The authors would like to thank the College of Science and The Thomas H. Gosnell School of Life Sciences at the Rochester Institute of Technology for ongoing support.

Declaration of Competing Interest

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

References

[1] M. Shyam, E.P. Sabina, Harnessing the power of *Arctium lappa* root: a review of its pharmacological properties and therapeutic applications, Nat. Prod. Bioprospect. 14 (2024) 49, doi:10.1007/s13659-024-00466-8.

[2] N. Yosri, S.M. Alsharif, J. Xiao, S.G. Musharraf, C. Zhao, A. Saeed, R. Gao, N.S. Said, A. Di Minno, M. Daglia, Z. Guo, S.A.M. Khalifa, H.R. El-Seedi, *Arctium lappa* (Burdock): insights from ethnopharmacology potential, chemical constituents, clinical studies, pharmacological utility and nanomedicine, Biomed. Pharmacother. 158 (2023) 114104, doi:10.1016/j.biopha.2022.114104.

[3] B.M. Muhire, A. Varsani, D.P. Martin, SDT: a virus classification tool based on pairwise sequence alignment and identity calculation, PLoS One 9 (2014) e108277, doi:10.1371/journal.pone.0108277.

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

[5] D. Li, C.-M. Liu, R. Luo, K. Sadakane, T.W. Lam, MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph, Bioinformatics 31 (2015) 1674–1676, doi:10.1093/bioinformatics/btv033.

[6] W. Shen, S. Le, Y. Li, F. Hu, SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation, PLoS One 11 (2016) e0163962, doi:10.1371/journal.pone.0163962.

[7] N. Huang, H. Li, compleasm: a faster and more accurate reimplementation of BUSCO, Bioinformatics 39 (2023), doi:10.1093/bioinformatics/btad595.

[8] Y. Yang, S. Li, Y. Xing, Z. Zhang, T. Liu, W. Ao, G. Bao, Z. Zhan, R. Zhao, T. Zhang, D. Zhang, Y. Song, C. Bian, L. Xu, T. Kang, The first high-quality chromosomal genome assembly of a medicinal and edible plant Arctium lappa, Mol. Ecol. Resour. 22 (2022) 1493–1507, doi:10.1111/1755-0998.13547.

[9] M. Alonge, S. Soyk, S. Ramakrishnan, X. Wang, S. Goodwin, F.J. Sedlazeck, Z.B. Lippman, M.C. Schatz, RaGOO: fast and accurate reference-guided scaffolding of draft genomes, Genome Biol. 20 (2019) 224, doi:10.1186/s13059-019-1829-6.

[10] Md. Vasimuddin, S. Misra, H. Li, S. Aluru, Efficient architecture-aware acceleration of BWA-MEM for multicore systems, in: Proceedings of the 2019 IEEE International Parallel and Distributed Processing Symposium. IPDPS, Rio de Janeiro, Brazil, IEEE, 2019, pp. 314–324, doi:10.1109/IPDPS.2019.00041.

[11] P. Danecek, J.K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M.O. Pollard, A. Whitwham, T. Keane, S.A. McCarthy, R.M. Davies, H. Li, Twelve years of SAMtools and BCFtools, GigaScience 10 (2021) giab008, doi:10.1093/gigascience/giab008.

[12] P. Cingolani, A. Platts, L.L. Wang, M. Coon, T. Nguyen, L. Wang, S.J. Land, X. Lu, D.M. Ruden, A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff, Fly 6 (2012) 80–92 (Austin), doi:10.4161/fly.19695.