



DATA NOTE

The genome sequence of lesser burdock, *Arctium minus* (Hill)

Bernh. (Asteraceae)

[version 1; peer review: 3 approved]

Maarten J. M. Christenhusz^{1,2}, Claudia A. Martin^{3,4},
 Royal Botanic Gardens Kew Genome Acquisition Lab,
 Plant Genome Sizing collective, Darwin Tree of Life Barcoding collective,
 Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory
 team,
 Wellcome Sanger Institute Scientific Operations: Sequencing Operations,
 Wellcome Sanger Institute Tree of Life Core Informatics team,
 Tree of Life Core Informatics collective, Darwin Tree of Life Consortium

¹Curtin University, Perth, Western Australia, Australia²Hortus Botanicus, University of Technology Delft, Delft, The Netherlands³Royal Botanic Garden Edinburgh Library, Edinburgh, Scotland, UK⁴The University of Edinburgh, Edinburgh, Scotland, UK

V1 First published: 15 Oct 2024, 9:589
<https://doi.org/10.12688/wellcomeopenres.23160.1>

Latest published: 15 Oct 2024, 9:589
<https://doi.org/10.12688/wellcomeopenres.23160.1>

Abstract

We present a genome assembly of a diploid specimen of *Arctium minus* (lesser burdock; Tracheophyta; Magnoliopsida; Asterales; Asteraceae). The genome sequence is 1,903.1 megabases in span. Most of the assembly is scaffolded into 18 chromosomal pseudomolecules. The mitochondrial and plastid genome assemblies have lengths of 312.58 kilobases and 152.71 kilobases, respectively. Gene annotation of this assembly on Ensembl identified 27,734 protein-coding genes.

Keywords

Arctium minus, lesser burdock, genome sequence, chromosomal, Asterales



This article is included in the [Tree of Life gateway](#).

Open Peer Review

Approval Status ✓ ✓ ✓

	1	2	3
version 1 15 Oct 2024	✓ view	✓ view	✓ view

- Norma Paniego** Instituto de Agrobiotecnología y Biología Molecular, Hurlingham, Argentina
- Ayoob Obaid Alfalahi** University of Anbar, Ramadi, Iraq
- Eric González-Segovia** University of British Columbia, British Columbia, Canada

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Darwin Tree of Life Consortium (mark.blaxter@sanger.ac.uk)

Author roles: **Christenhusz MJM:** Investigation, Resources, Writing – Review & Editing; **Martin CA:** Writing – Original Draft Preparation, Writing – Review & Editing;

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome through core funding to the Wellcome Sanger Institute [206194, <https://doi.org/10.35802/206194>] and the Darwin Tree of Life Discretionary Award [218328, <https://doi.org/10.35802/218328>]. *The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

Copyright: © 2024 Christenhusz MJM *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.


How to cite this article: Christenhusz MJM, Martin CA, Royal Botanic Gardens Kew Genome Acquisition Lab *et al.* **The genome sequence of lesser burdock, *Arctium minus* (Hill) Bernh. (Asteraceae) [version 1; peer review: 3 approved]** Wellcome Open Research 2024, 9:589 <https://doi.org/10.12688/wellcomeopenres.23160.1>

First published: 15 Oct 2024, 9:589 <https://doi.org/10.12688/wellcomeopenres.23160.1>



DATA NOTE

The genome sequence of lesser burdock, *Arctium minus* (Hill) Bernh. (Asteraceae)

Maarten J. M. Christenhusz^{1,2}, Claudia A. Martin ^{3,4},
Royal Botanic Gardens Kew Genome Acquisition Lab,
Plant Genome Sizing collective, Darwin Tree of Life Barcoding collective,
Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory
team,
Wellcome Sanger Institute Scientific Operations: Sequencing Operations,
Wellcome Sanger Institute Tree of Life Core Informatics team,
Tree of Life Core Informatics collective, Darwin Tree of Life Consortium

¹Curtin University, Perth, Western Australia, Australia

²Hortus Botanicus, University of Technology Delft, Delft, The Netherlands

³Royal Botanic Garden Edinburgh Library, Edinburgh, Scotland, UK

⁴The University of Edinburgh, Edinburgh, Scotland, UK

V1 First published: N/A, N/A: N/A N/A
Latest published: N/A, N/A: N/A N/A

Open Peer Review

Approval Status AWAITING PEER REVIEW

Any reports and responses or comments on the article can be found at the end of the article.

Abstract

We present a genome assembly of a diploid specimen of *Arctium minus* (lesser burdock; Tracheophyta; Magnoliopsida; Asterales; Asteraceae). The genome sequence is 1,903.1 megabases in span. Most of the assembly is scaffolded into 18 chromosomal pseudomolecules. The mitochondrial and plastid genome assemblies have lengths of 312.58 kilobases and 152.71 kilobases, respectively. Gene annotation of this assembly on Ensembl identified 27,734 protein-coding genes.

Keywords

Arctium minus, lesser burdock, genome sequence, chromosomal, Asterales



This article is included in the [Tree of Life gateway](#).

Corresponding author: Darwin Tree of Life Consortium (mark.blaxter@sanger.ac.uk)

Author roles: **Christenhusz MJM:** Investigation, Resources, Writing – Review & Editing; **Martin CA:** Writing – Original Draft Preparation, Writing – Review & Editing;

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome through core funding to the Wellcome Sanger Institute [206194, <https://doi.org/10.35802/206194>] and the Darwin Tree of Life Discretionary Award [218328, <https://doi.org/10.35802/218328>]. *The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

Copyright: © 2024 Christenhusz MJM *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Christenhusz MJM, Martin CA, Royal Botanic Gardens Kew Genome Acquisition Lab *et al.* **The genome sequence of lesser burdock, *Arctium minus* (Hill) Bernh. (Asteraceae)** Wellcome Open Research , : <https://doi.org/>

First published: N/A, N/A: N/A N/A

Species taxonomy

Eukaryota; Viridiplantae; Streptophyta; Streptophytina; Embryophyta; Tracheophyta; Euphyllophyta; Spermatophyta; Magnoliopsida; Mesangiospermae; eudicotyledons; Gunneridae; Pentapetales; asterids; campanulids; Asterales; Asteraceae; Carduoideae; Cardueae; Arctiinae; *Arctium*; *Arctium minus* (Hill) Bernh., 1800 (NCBI:txid143172).

Background

Lesser burdock (also known as little burdock, louse-bur, common burdock, button-bur, cuckoo-button, or wild rhubarb), of the genus *Arctium* L., is a biennial herbaceous plant belonging to the daisy family, Asteraceae. This species, commonly known for its burr-like seed heads, is widespread and abundant across most of Britain where it is commonly perceived as a persistent weed (Preston *et al.*, 2002). Its distribution ranges across lowland and upland landscapes, although it is less common in the far north and west of Scotland. The adaptability of *A. minus* to various soil types and climatic conditions has facilitated its spread and establishment in diverse geographic regions, and it thrives in a variety of urban and rural environments. It is particularly successful in areas with disturbed habitats such as roadsides, field margins, woodland edges and waste grounds, where there is minimal competition with other plant species. *A. minus* is native to all of Europe and western Asia, extending south to Morocco and east to Afghanistan, and it has been widely introduced across North America, southeastern Brazil, southeastern Australia and the North Island of New Zealand (Gross *et al.*, 1979; Hultén & Fries, 1986; POWO, 2024).

The life cycle of the plant spans two years; in the first year, it forms a basal rosette of leaves, while in the second year, it produces tall, branched flowering stems reaching heights of 1 to 2 metres (Stace *et al.*, 2019). The flowering period extends from July to September, during which purple flowers, similar to that of thistles, appear (Figure 1). When dry, these flowers turn brown and form ‘burrs’ that cling to animal fur and clothing. This mechanism of seed dispersal inspired technological innovations such as Velcro, which utilises the hook and loop fastening mechanism observed and invented by George de Mestral, when he saw seed heads of burdock entangled in his dog’s fur (Christenhusz *et al.*, 2017). The species forms an important food source for many species of Lepidoptera in Britain.

Arctium minus has been valued for both its culinary and medicinal uses. The roots of the plant, rich in inulin, have been consumed as a vegetable in some cultures (tasting like a cross between sweet chestnut and parsnip). Inulin is a soluble dietary fibre found in a variety of plants, and it is known for its prebiotic effects, promoting the growth of beneficial gut bacteria (Moro & Clerici, 2021). In traditional medicine, the plant was utilised for its purported detoxifying properties (Chevallier, 1996), and its extracted seed oil is rich in fatty acids and phytosterols. In modern-day Britain, most people will only be aware of the use of this species in the production of ‘burdock beer’, a traditional British soft drink, made from

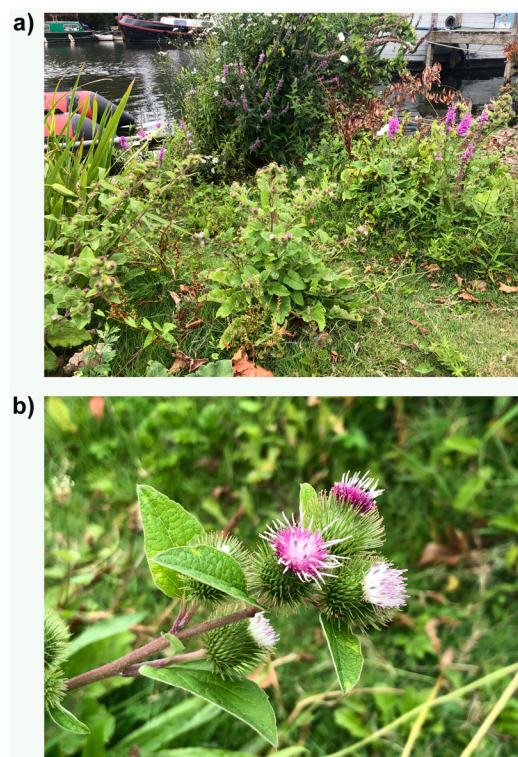


Figure 1. Photographs of the *Arctium minus* (daArcMinu1) specimen used for genome sequencing showing (a) the whole plant, and (b) a close up of one of the flowering stems.

the roots of this and related species, often mixed with dandelion (*Taraxacum* L.) roots, to create the well-known beverage ‘dandelion and burdock’, consumed since the Middle Ages. Originally a type of mead, it is now a carbonated soft drink (Lewis-Stempel, 2010).

The genus name *Arctium* is derived from the Ancient Greek ἄρκτος (romanised as Arctus), meaning bear, but it was also the name of a centaur in Greek mythology. It likely refers to the rough, bristly appearance of the burrs and perhaps the toughness of the plant. The Latin species epithet “minus”, small, denotes the relatively small size compared to the greater burdock (*A. lappa* L.). The taxonomic history of *A. minus* has been chequered, and it has sometimes been treated as a variety or subspecies of *A. lappa* or *A. nemorosum* Lej. The species is variable, and identification can sometimes be difficult as it can closely resemble other species of the genus, especially wood burdock (*A. nemorosum*; Stace *et al.*, 2019).

While *A. minus* has been reported to be a diploid with either a chromosome number of $2n = 2x = 32$ or 36 , all UK material counted to date shows it to be $2n = 36$ (Stace *et al.*, 2019), and the previous reports of $2n = 32$ are now considered to erroneous (Gross *et al.*, 1980). Species of the genus *Arctium* are known to hybridise (Wang *et al.*, 2019), and this has been documented from Britain and Ireland. As a result, polyploidy is

frequent in related species, contributing to further identification challenges and taxonomic debates. Hybrids such as *Arctium* × *nothum* (Ruhmer) J.Weiss (a hybrid between *A. minus* and *A. lappa*) and *Arctium* × *mixtum* (Simonk.) Nyman (a hybrid between *A. minus* and *A. tomentosum* Mill., woolly burdock) exhibit triploid chromosome counts ($2n = 3x = 54$), likely arising from unreduced gametes from one parent. These hybrid zones, primarily areas where the ranges of the parent species overlap, are typically in disturbed habitats that provide suitable conditions for both species to co-occur. The presence of hybrids indicates active gene flow between the species, resulting in intermediate forms, but this typically leads to reduced fertility resulting in low persistence of these cytotypes.

Here, we present the first chromosome-level *A. minus* genome, which we anticipate will help in understanding the taxonomic diversity of the genus, including hybrid evolution, and facilitate comparative genomic studies to uncover evolutionary and functional genomic insights. Secondly, as a highly tolerant species, this genome can help to enhance our understanding of the genetic basis of adaptability and resilience in diverse environments. In addition, this provides the opportunity to

explore genes involved in the synthesis of medicinal compounds like inulin and antimicrobial agents.

Genome sequence report

The genome of a specimen of *Arctium minus* (Figure 1) was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating a total of 72.51 Gb (gigabases) from 5.53 million reads, providing approximately 24-fold coverage. Using flow cytometry, the genome size (1C-value) was estimated to be 2.11 pg, equivalent to 2,070 Mb. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data, which produced 213.23 Gb from 1,412.15 million reads, yielding an approximate coverage of 112-fold. Specimen and sequencing information is summarised in Table 1.

Manual assembly curation corrected two missing joins or mis-joins and two haplotypic duplications. The final assembly has a total length of 1,903.10 Mb in 30 sequence scaffolds with a scaffold N50 of 103.5 Mb (Table 2) with 14 gaps. The snail plot in Figure 2 summarises the assembly statistics, while the blob plot in Figure 3 shows the distribution of assembly scaffolds by GC proportion and coverage. The cumulative assembly plot in Figure 4 shows curves for subsets of scaffolds

Table 1. Specimen and sequencing data for *Arctium minus*.

Project information			
Study title	Arctium minus		
Umbrella BioProject	PRJEB53860		
Species	<i>Arctium minus</i>		
BioSample	SAMEA7521931		
NCBI taxonomy ID	143172		
Specimen information			
Technology	ToLID	BioSample accession	Organism part
PacBio long read sequencing	daArcMinu1	SAMEA7521964	leaf
Hi-C sequencing	daArcMinu1	SAMEA7521962	leaf
RNA sequencing	daArcMinu1	SAMEA7521959	flower
Sequencing information			
Platform	Run accession	Read count	Base count (Gb)
Hi-C Illumina NovaSeq 6000	ERR9881701	1.41e+09	213.23
PacBio Sequel Iie	ERR9902008	9.86e+05	14.05
PacBio Sequel Iie	ERR9902011	1.90e+06	23.34
PacBio Sequel Iie	ERR9902009	9.75e+05	13.81
PacBio Sequel Iie	ERR9902010	1.67e+06	21.32
RNA Illumina NovaSeq 6000	ERR10378020	5.62e+07	8.48
RNA Illumina NovaSeq 6000	ERR10378019	5.89e+07	8.9

Table 2. Genome assembly data for *Arctium minus*, daArcMinu1.1.

Genome assembly		
Assembly name	daArcMinu1.1	
Assembly accession	GCA_954870635.1	
Accession of alternate haplotype	GCA_954871535.1	
Span (Mb)	1,903.10	
Number of contigs	46	
Contig N50 length (Mb)	80.9	
Number of scaffolds	30	
Scaffold N50 length (Mb)	103.5	
Longest scaffold (Mb)	199.74	
Assembly metrics*		Benchmark
Consensus quality (QV)	60.7	≥ 50
k-mer completeness	100.0%	≥ 95%
BUSCO**	C:98.1%[S:92.0%,D:6.1%], F:0.6%,M:1.4%,n:2,326	C ≥ 95%
Percentage of assembly mapped to chromosomes	99.92%	≥ 95%
Organelles	Mitochondrial genome: 312.58 kb; plastid genome: 152.71 kb	complete single alleles
Genome annotation at Ensembl		
Number of protein-coding genes	27,734	
Number of non-coding genes	10,938	
Number of gene transcripts	52,022	

* Assembly metric benchmarks are adapted from column VGP-2020 of “Table 1: Proposed standards and metrics for defining genome assembly quality” from [Rhie et al. \(2021\)](#).

** BUSCO scores based on the eudicots_odb10 BUSCO set using version 5.4.3. C = complete [S = single copy, D = duplicated], F = fragmented, M = missing, n = number of orthologues in comparison. A full set of BUSCO scores is available at https://blobtoolkit.genomehubs.org/view/daArcMinu1_1/dataset/daArcMinu1_1/busco.

assigned to different phyla. Most (99.92%) of the assembly sequence was assigned to 18 chromosomal-level scaffolds. Chromosome-scale scaffolds confirmed by the Hi-C data are named in order of size ([Figure 5](#); [Table 3](#)). While not fully phased, the assembly deposited is of one haplotype. Contigs corresponding to the second haplotype have also been deposited. The mitochondrial and plastid genomes were also assembled and can be found as contigs within the multifasta file of the genome submission.

The estimated Quality Value (QV) of the final assembly is 60.7 with k-mer completeness of 100.0%, and the assembly has a BUSCO v5.4.3 completeness of 98.1% (single = 92.0%, duplicated = 6.1%), using the eudicots_odb10 reference set ($n = 2,326$).

Metadata for specimens, BOLD barcode results, spectra estimates, sequencing runs, contaminants and pre-curation assembly

statistics are given at <https://links.tol.sanger.ac.uk/species/143172>.

Genome annotation report

The *Arctium minus* genome assembly (GCA_954870635.1) was annotated at the European Bioinformatics Institute (EBI) on Ensembl Rapid Release. The resulting annotation includes 52,022 transcribed mRNAs from 27,734 protein-coding and 10,938 non-coding genes ([Table 2](#); https://rapid.ensembl.org/Arctium_minus_GCA_954870635.1/Info/Index). The average transcript length is 3,940.95. There are 1.35 coding transcripts per gene and 4.83 exons per transcript.

Methods

Sample acquisition, DNA barcoding and genome size estimation

A specimen of *Arctium minus* (specimen ID KDTOL10022, ToLID daArcMinu1) was collected from Canbury Gardens,

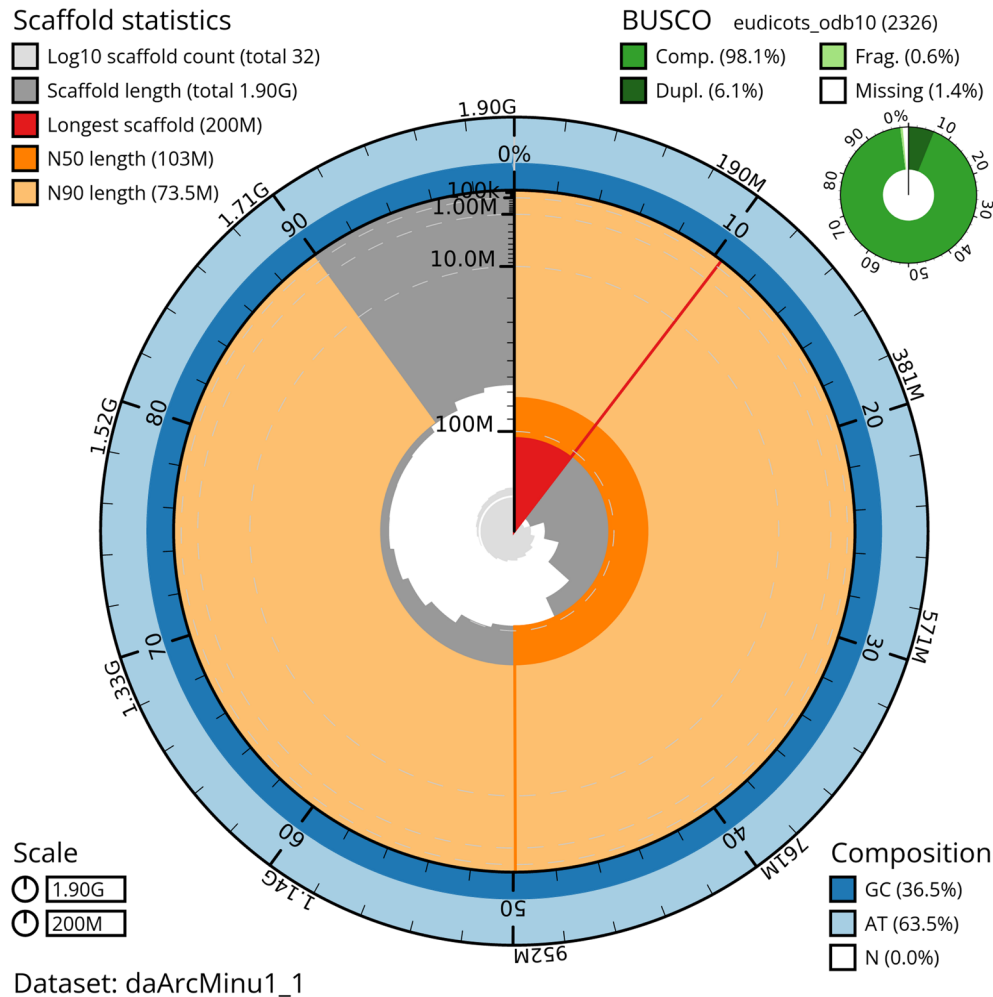


Figure 2. Genome assembly of *Arctium minus*, daArcMinu1.1: metrics. The BlobToolKit snail plot shows N50 metrics and BUSCO gene completeness. The main plot is divided into 1,000 size-ordered bins around the circumference with each bin representing 0.1% of the 1,903,528,901 bp assembly. The distribution of scaffold lengths is shown in dark grey with the plot radius scaled to the longest scaffold present in the assembly (199,739,593 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 scaffold lengths (103,461,716 and 73,472,875 bp), respectively. The pale grey spiral shows the cumulative scaffold count on a log scale with white scale lines showing successive orders of magnitude. The blue and pale-blue area around the outside of the plot shows the distribution of GC, AT and N percentages in the same bins as the inner plot. A summary of complete, fragmented, duplicated and missing BUSCO genes in the eudicots_odb10 set is shown in the top right. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/daArcMinu1_1/dataset/daArcMinu1_1/snail.

Kingston Upon Thames, Surrey, UK (latitude 51.42, longitude -0.31) on 2020-08-06. The specimen was collected and identified by Maarten Christenhusz (Royal Botanic Gardens, Kew) and preserved by freezing at -80°C . The herbarium voucher associated with the sequenced plant is M. Christenhusz 9019 and is deposited in the herbarium of RBG Kew (K) (K001400639).

The initial species identification was verified by an additional DNA barcoding process following the framework developed by Twyford *et al.* (2024). Part of the plant specimen was preserved in silica gel desiccant (Chase & Hills, 1991). DNA

was extracted from the dried specimen, then PCR was used to amplify standard barcode regions. The resulting amplicons were sequenced and compared to public sequence databases including GenBank and the Barcode of Life Database (BOLD). The barcode sequences for this specimen are available on BOLD (Ratnasingham & Hebert, 2007). Following whole genome sequence generation, DNA barcodes were also used alongside the initial barcoding data for sample tracking through the genome production pipeline at the Wellcome Sanger Institute (Twyford *et al.*, 2024). The standard operating procedures for the Darwin Tree of Life barcoding have been deposited on protocols.io (Beasley *et al.*, 2023).

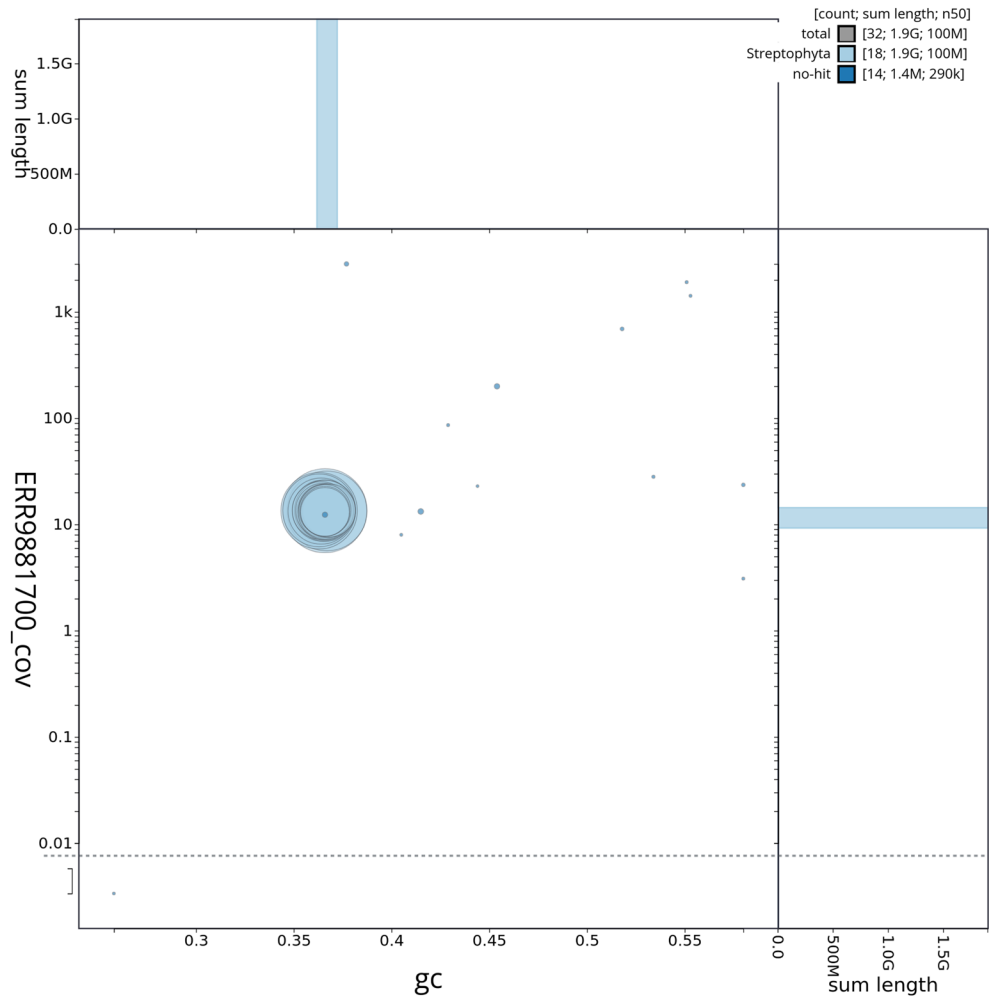


Figure 3. Blob plot of base coverage against GC proportion for sequences in the assembly daArcMinu1.1. Sequences are coloured by phylum. Circles are sized in proportion to sequence length. Histograms show the distribution of sequence length sum along each axis. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/daArcMinu1_1/dataset/daArcMinu1_1/blob.

The genome size was estimated by flow cytometry using the fluorochrome propidium iodide and following the ‘one-step’ method as outlined in Pellicer *et al.* (2021). For this species, the General Purpose Buffer (GPB) supplemented with 3% PVP and 0.08% (v/v) beta-mercaptoethanol was used for isolation of nuclei (Loureiro *et al.*, 2007), and the internal calibration standard was *Solanum lycopersicum* ‘Stupiké polní rané’ with an assumed 1C-value of 968 Mb (Doležel *et al.*, 2007).

Nucleic acid extraction

The workflow for high molecular weight (HMW) DNA extraction at the Wellcome Sanger Institute (WSI) Tree of Life Core Laboratory includes a sequence of core procedures: sample preparation; sample homogenisation, DNA extraction, fragmentation, and clean-up. Detailed protocols are available on protocols.io (Denton *et al.*, 2023). The daArcMinu1

sample was weighed and dissected on dry ice (Jay *et al.*, 2023). For sample homogenisation, leaf tissue was cryogenically disrupted using the Covaris cryoPREP® Automated Dry Pulverizer (Narváez-Gómez *et al.*, 2023).

HMW DNA was extracted using the Automated Plant MagAttract v2 protocol (Todorovic *et al.*, 2023). HMW DNA was sheared into an average fragment size of 12–20 kb in a Megaruptor 3 system (Bates *et al.*, 2023). Sheared DNA was purified by solid-phase reversible immobilisation, using AMPure PB beads to eliminate shorter fragments and concentrate the DNA (Strickland *et al.*, 2023). The concentration of the sheared and purified DNA was assessed using a Nanodrop spectrophotometer and Qubit Fluorometer and Qubit dsDNA High Sensitivity Assay kit. Fragment size distribution was evaluated by running the sample on the FemtoPulse system.

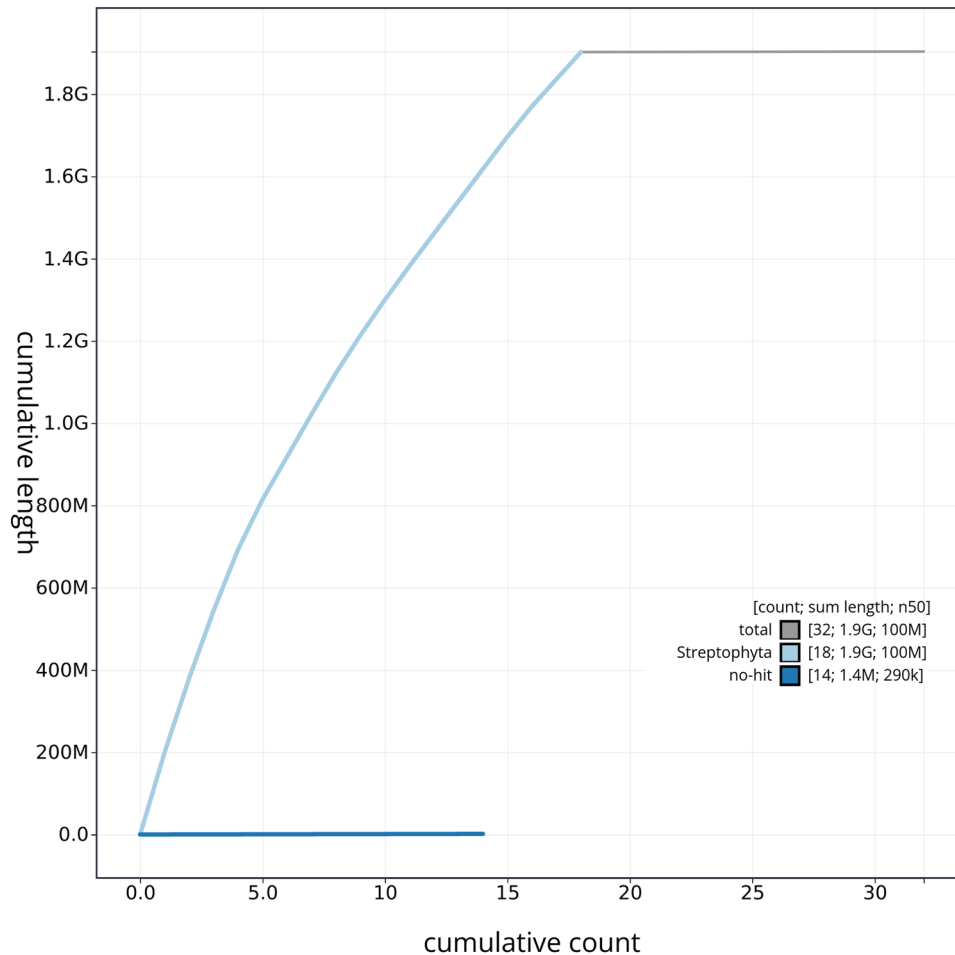


Figure 4. Genome assembly of *Arctium minus* daArcMinu1.1: BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all sequences. Coloured lines show cumulative lengths of sequences assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/daArcMinu1_1/dataset/daArcMinu1_1/cumulative.

RNA was extracted from flower tissue of daArcMinu1 in the Tree of Life Laboratory at the WSI using the RNA Extraction: Automated MagMax™ mirVana protocol (do Amaral *et al.*, 2023). The RNA concentration was assessed using a Nanodrop spectrophotometer and a Qubit Fluorometer using the Qubit RNA Broad-Range Assay kit. Analysis of the integrity of the RNA was done using the Agilent RNA 6000 Pico Kit and Eukaryotic Total RNA assay.

Hi-C preparation

Leaf tissue of daArcMinu1 was processed at the WSI Scientific Operations core, using the Arima-HiC v2 kit. Tissue was finely ground using cryoPREP and then subjected to nuclei isolation using a modified protocol of the Qiagen QProteome Kit. After isolation, the nuclei were fixed, and the DNA crosslinked using 37% formaldehyde solution. The crosslinked DNA was then digested using the restriction enzyme master mix. The 5'-overhangs were then filled in and labelled with biotinylated nucleotides and proximally ligated. An overnight incubation was carried out for enzymes

to digest remaining proteins and for crosslinks to reverse. A clean up was performed with SPRIselect beads prior to library preparation. DNA concentration was quantified using the Qubit Fluorometer v2.0 and Qubit HS Assay Kit according to the manufacturer's instructions.

Library preparation and sequencing

Library preparation and sequencing was performed at the WSI Scientific Operations core. Pacific Biosciences HiFi circular consensus DNA sequencing libraries were prepared using the PacBio Express Template Preparation Kit v2.0 (Pacific Biosciences, California, USA) as per the manufacturer's instructions. The kit includes the reagents required for removal of single-strand overhangs, DNA damage repair, end repair/A-tailing, adapter ligation, and nuclease treatment. Library preparation also included a library purification step using 0.8X AMPure PB beads (Pacific Biosciences, California, USA) and a size selection step to remove templates <3 kb using AMPure PB modified SPRI. Samples were sequenced using the Sequel IIe system (Pacific Biosciences, California, USA). The

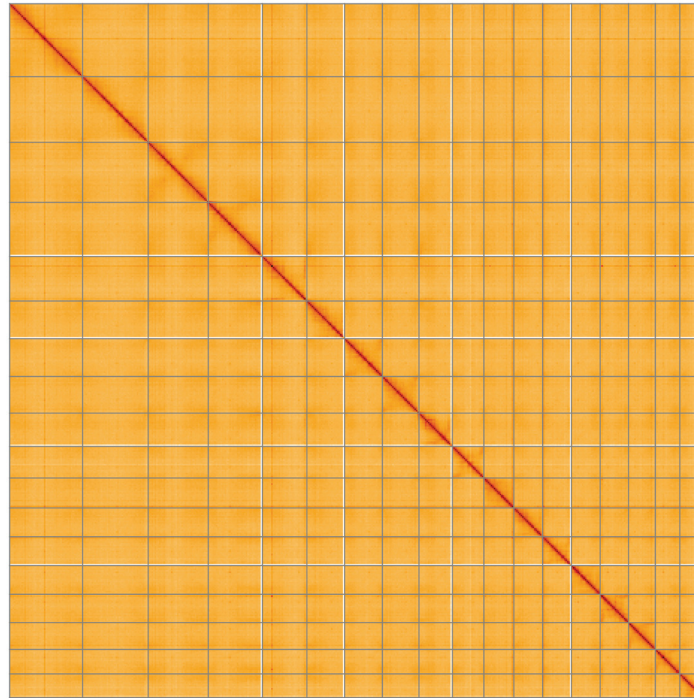


Figure 5. Genome assembly of *Arctium minus*, daArcMinu1.1: Hi-C contact map of the daArcMinu1.1 assembly, visualised using HiGlass. Chromosomes are shown in order of size from left to right and top to bottom. An interactive version of this figure may be viewed at <https://genome-note-higlass.tol.sanger.ac.uk/l/?d=Ej0r1nfwSweTrM1Vac65yg>.

Table 3. Chromosomal pseudomolecules in the genome assembly of *Arctium minus*, daArcMinu1.

INSDC accession	Name	Length (Mb)	GC%
OX941080.1	1	199.74	36.5
OX941081.1	2	179.92	36.5
OX941082.1	3	164.54	36.5
OX941083.1	4	148.83	36.5
OX941084.1	5	121.84	36.5
OX941085.1	6	103.58	36.5
OX941086.1	7	103.46	36.5
OX941087.1	8	100.14	36.5
OX941088.1	9	91.98	36.5
OX941089.1	10	86.24	36.5
OX941090.1	11	81.9	36.5
OX941091.1	12	79.15	36.5
OX941092.1	13	78.86	36.5

INSDC accession	Name	Length (Mb)	GC%
OX941093.1	14	78.65	36.5
OX941094.1	15	77.81	37.0
OX941095.1	16	73.47	36.5
OX941096.1	17	67.44	36.5
OX941097.1	18	64.56	36.5
OX941098.1	MT	0.31	45.5
OX941099.1	Pltd	0.15	37.5

concentration of the library loaded onto the Sequel IIe was within the manufacturer’s recommended loading concentration range of 40–100 pM. The SMRT link software, a PacBio web-based end-to-end workflow manager, was used to set-up and monitor the run, as well as perform primary and secondary analyses of the data upon completion.

Poly(A) RNA-Seq libraries were constructed using the NEB Ultra II RNA Library Prep kit following manufacturer’s instructions. RNA sequencing was performed on the Illumina NovaSeq 6000 instrument.

For Hi-C library preparation, DNA was fragmented to a size of 400 to 600 bp using a Covaris E220 sonicator. The DNA was then enriched, barcoded, and amplified using the NEBNext Ultra II DNA Library Prep Kit, following manufacturers' instructions. The Hi-C sequencing was performed using paired-end sequencing with a read length of 150 bp on an Illumina NovaSeq 6000.

Genome assembly, curation and evaluation

Assembly

The original assembly of HiFi reads was performed using Hifiasm (Cheng *et al.*, 2021) with the --primary option. Haplotypic duplications were identified and removed with purge_dups (Guan *et al.*, 2020). Hi-C reads were further mapped with bwa-mem2 (Vasimuddin *et al.*, 2019) to the primary contigs, which were further scaffolded using the provided Hi-C data (Rao *et al.*, 2014) in YaHS (Zhou *et al.*, 2023) using the --break option. Scaffolded assemblies were evaluated using Gfastats (Formenti *et al.*, 2022), BUSCO (Manni *et al.*, 2021) and MERQURY.FK (Rhie *et al.*, 2020).

The organelle genomes were assembled using MBG (Rautiainen & Marschall, 2021) from PacBio HiFi reads mapping to related genomes. A representative circular sequence was selected for each from the graph based on read coverage.

Curation

The assembly was decontaminated using the Assembly Screen for Cobionts and Contaminants (ASCC) pipeline (article in preparation). Manual curation was primarily conducted using PretextView (Harry, 2022), with additional insights provided by JBrowse2 (Diesh *et al.*, 2023) and HiGlass (Kerpedjiev *et al.*, 2018). Scaffolds were visually inspected and corrected as

described by Howe *et al.* (2021). Any identified contamination, missed joins, and mis-joins were corrected, and duplicate sequences were tagged and removed. The process is documented at <https://gitlab.com/wtsi-grit/rapid-curation> (article in preparation).

Evaluation of final assembly

A Hi-C map for the final assembly was produced using bwa-mem2 (Vasimuddin *et al.*, 2019) in the Cooler file format (Abdennur & Mirny, 2020). To assess the assembly metrics, the *k*-mer completeness and QV consensus quality values were calculated in Merqury (Rhie *et al.*, 2020). This work was done using the "sanger-tol/readmapping" (Surana *et al.*, 2023a) and "sanger-tol/genomenote" (Surana *et al.*, 2023b) pipelines. The genome evaluation pipelines were developed using nf-core tooling (Ewels *et al.*, 2020) and MultiQC (Ewels *et al.*, 2016), relying on the Conda package manager, the Bioconda initiative (Grüning *et al.*, 2018), the Biocontainers infrastructure (da Veiga Leprevost *et al.*, 2017), as well as the Docker (Merkel, 2014) and Singularity (Kurtzer *et al.*, 2017) containerisation solutions.

The genome was also analysed within the BlobToolKit environment (Challis *et al.*, 2020) and BUSCO scores (Manni *et al.*, 2021) were calculated.

Table 4 contains a list of relevant software tool versions and sources.

Wellcome Sanger Institute – Legal and Governance

The materials that have contributed to this genome note have been supplied by a Darwin Tree of Life Partner. The submission of materials by a Darwin Tree of Life Partner is subject

Table 4. Software tools: versions and sources.

Software tool	Version	Source
BlobToolKit	4.1.7	https://github.com/blobtoolkit/blobtoolkit
BUSCO	5.3.2	https://gitlab.com/ezlab/busco
bwa-mem2	2.2.1	https://github.com/bwa-mem2/bwa-mem2
Cooler	0.8.11	https://github.com/open2c/cooler
Gfastats	1.3.6	https://github.com/vgl-hub/gfastats
Hifiasm	0.16.1-r375	https://github.com/chhy1p123/hifiasm
HiGlass	1.11.6	https://github.com/higlass/higlass
MBG	-	https://github.com/maickrau/MBG
Merqury	MerquryFK	https://github.com/thegenemyers/MERQURY.FK
PretextView	0.2	https://github.com/wtsi-hpag/PretextView
purge_dups	1.2.3	https://github.com/dfguan/purge_dups
sanger-tol/genomenote	v1.0	https://github.com/sanger-tol/genomenote
sanger-tol/readmapping	1.1.0	https://github.com/sanger-tol/readmapping/tree/1.1.0
YaHS	yahs-1.1.91eebc2	https://github.com/c-zhou/yahs

to the ‘**Darwin Tree of Life Project Sampling Code of Practice**’, which can be found in full on the Darwin Tree of Life website [here](#). By agreeing with and signing up to the Sampling Code of Practice, the Darwin Tree of Life Partner agrees they will meet the legal and ethical requirements and standards set out within this document in respect of all samples acquired for, and supplied to, the Darwin Tree of Life Project.

Further, the Wellcome Sanger Institute employs a process whereby due diligence is carried out proportionate to the nature of the materials themselves, and the circumstances under which they have been/are to be collected and provided for use. The purpose of this is to address and mitigate any potential legal and/or ethical implications of receipt and use of the materials as part of the research project, and to ensure that in doing so we align with best practice wherever possible. The overarching areas of consideration are:

- Ethical review of provenance and sourcing of the material
- Legality of collection, transfer and use (national and international)

Each transfer of samples is further undertaken according to a Research Collaboration Agreement or Material Transfer Agreement entered into by the Darwin Tree of Life Partner, Genome Research Limited (operating as the Wellcome Sanger Institute), and in some circumstances other Darwin Tree of Life collaborators.

Data availability

European Nucleotide Archive: *Arctium minus*. Accession number PRJEB53860; <https://identifiers.org/ena.embl/PRJEB53860> (Wellcome Sanger Institute, 2023). The genome

sequence is released openly for reuse. The *Arctium minus* genome sequencing initiative is part of the Darwin Tree of Life (DTOL) project. All raw sequence data and the assembly have been deposited in INSDC databases. Raw data and assembly accession identifiers are reported in [Table 1](#).

Author information

Members of the Royal Botanic Gardens Kew Genome Acquisition Lab are listed here: <https://doi.org/10.5281/zenodo.12625079>.

Members of the Plant Genome Sizing collective are listed here: <https://doi.org/10.5281/zenodo.7994306>.

Members of the Darwin Tree of Life Barcoding collective are listed here: <https://doi.org/10.5281/zenodo.12158331>.

Members of the Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team are listed here: <https://doi.org/10.5281/zenodo.12162482>.

Members of Wellcome Sanger Institute Scientific Operations: Sequencing Operations are listed here: <https://doi.org/10.5281/zenodo.12165051>.

Members of the Wellcome Sanger Institute Tree of Life Core Informatics team are listed here: <https://doi.org/10.5281/zenodo.12160324>.

Members of the Tree of Life Core Informatics collective are listed here: <https://doi.org/10.5281/zenodo.12205391>.

Members of the Darwin Tree of Life Consortium are listed here: <https://doi.org/10.5281/zenodo.4783558>.

References

- Abdennur N, Mirny LA: **Cooler: scalable storage for Hi-C data and other genomically labeled arrays**. *Bioinformatics*. 2020; **36**(1): 311–316. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Bates A, Clayton-Lucey I, Howard C: **Sanger Tree of Life HMW DNA fragmentation: diagenode Megaruptor³ for LI PacBio**. *protocols.io*. 2023. [Publisher Full Text](#)
- Beasley J, Uhl R, Forrest LL, et al.: **DNA barcoding SOPs for the Darwin Tree of Life project**. *protocols.io*. 2023; [Accessed 25 June 2024]. [Publisher Full Text](#)
- Challis R, Richards E, Rajan J, et al.: **BlobToolKit – interactive quality assessment of genome assemblies**. *G3 (Bethesda)*. 2020; **10**(4): 1361–1374. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Chase MW, Hills HH: **Silica gel: an ideal material for field preservation of leaf samples for DNA studies**. *Taxon*. 1991; **40**(2): 215–220. [Publisher Full Text](#)
- Cheng H, Concepcion GT, Feng X, et al.: **Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm**. *Nat Methods*. 2021; **18**(2): 170–175. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Chevallier A: **The encyclopedia of medicinal plants: a practical reference guide to over 550 key herbs and their medicinal uses**. Dorling Kindersley, 1996. [Reference Source](#)
- Christenhusz MJM, Fay MF, Chase MW: **Plants of the world an illustrated encyclopedia of vascular plants**. The University of Chicago Press, 2017. [Publisher Full Text](#)
- da Veiga Leprevost F, Grüning BA, Alves Aflitos S, et al.: **BioContainers: an open-source and community-driven framework for software standardization**. *Bioinformatics*. 2017; **33**(16): 2580–2582. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Denton A, Yatsenko H, Jay J, et al.: **Sanger Tree of Life wet laboratory protocol collection V.1**. *protocols.io*. 2023. [Publisher Full Text](#)
- Diesh C, Stevens GJ, Xie P, et al.: **JBrowse 2: a modular genome browser with views of synteny and structural variation**. *Genome Biol*. 2023; **24**(1): 74. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- do Amaral RJV, Bates A, Denton A, et al.: **Sanger Tree of Life RNA extraction: automated MagMax[™] mirVana**. *protocols.io*. 2023. [Publisher Full Text](#)
- Dolezel J, Greilhuber J, Suda J: **Estimation of nuclear DNA content in plants**

- using flow cytometry.** *Nat Protoc.* 2007; **2**(9): 2233–2244.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Ewels P, Magnusson M, Lundin S, *et al.*: **MultiQC: summarize analysis results for multiple tools and samples in a single report.** *Bioinformatics.* 2016; **32**(19): 3047–3048.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Ewels PA, Peltzer A, Fillinger S, *et al.*: **The nf-core framework for community-curated bioinformatics pipelines.** *Nat Biotechnol.* 2020; **38**(3): 276–278.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Formenti G, Abueg L, Brajuka A, *et al.*: **Gfastats: conversion, evaluation and manipulation of genome sequences using assembly graphs.** *Bioinformatics.* 2022; **38**(17): 4214–4216.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Gross RS, Werner PA, Hawthorn WR: **The biology of Canadian weeds. 38. *Arctium minus* (Hill) Bernh. and *A. lappa* L.** *Can J Plant Sci.* 1979; **59**: 401–413.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Gross RS, Werner PA, Hawthorn WR: **The biology of Canadian weeds. 38. *Arctium minus* (Hill) Bernh. and *A. lappa* L.** *Can J Plant Sci.* 1980; **60**(2): 621–634.
[Publisher Full Text](#)
- Grüning B, Dale R, Sjödin A, *et al.*: **Bioconda: sustainable and comprehensive software distribution for the life sciences.** *Nat Methods.* 2018; **15**(7): 475–476.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Guan D, McCarthy SA, Wood J, *et al.*: **Identifying and removing haplotypic duplication in primary genome assemblies.** *Bioinformatics.* 2020; **36**(9): 2896–2898.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Harry E: **PretextView (Paired REad TEXTure Viewer): a desktop application for viewing pretext contact maps.** 2022.
[Reference Source](#)
- Howe K, Chow W, Collins J, *et al.*: **Significantly improving the quality of genome assemblies through curation.** *GigaScience.* 2021; **10**(1): g1aa153.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Hultén E, Fries M: **Atlas of North European vascular plants north of the tropic of cancer.** Koeltz Scientific Books, 1986.
[Reference Source](#)
- Jay J, Yatsenko H, Narváez-Gómez JP, *et al.*: **Sanger Tree of Life sample preparation: triage and dissection.** *protocols.io.* 2023.
[Publisher Full Text](#)
- Kerpedjiev P, Abdennur N, Lekschas F, *et al.*: **HIGlass: web-based visual exploration and analysis of genome interaction maps.** *Genome Biol.* 2018; **19**(1): 125.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Kurtzer GM, Sochat V, Bauer MW: **Singularity: scientific containers for mobility of compute.** *PLoS One.* 2017; **12**(5): e0177459.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Lewis-Stempel J: **The wild life.** Black Swan, 2010.
- Loureiro J, Rodríguez E, Dolezel J, *et al.*: **Two new nuclear isolation buffers for plant DNA flow cytometry: a test with 37 species.** *Ann Bot.* 2007; **100**(4): 875–888.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Manni M, Berkeley MR, Seppely M, *et al.*: **BUSCO update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes.** *Mol Biol Evol.* 2021; **38**(10): 4647–4654.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Merkel D: **Docker: lightweight Linux containers for consistent development and deployment.** *Linux J.* 2014; **2014**(239): 2, [Accessed 2 April 2024].
[Reference Source](#)
- Moro TMA, Clerici MTPS: **Burdock (*Arctium lappa* L.) roots as a source of inulin-type fructans and other bioactive compounds: current knowledge and future perspectives for food and non-food applications.** *Food Res Int.* 2021; **141**: 109889.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Narváez-Gómez JP, Mbye H, Oatley G, *et al.*: **Sanger Tree of Life sample homogenisation: covaris cryoPREP® automated dry Pulverizer V.1.** *protocols.io.* 2023.
[Publisher Full Text](#)
- Pellicer J, Powell RF, Leitch IJ: **The application of flow cytometry for estimating genome size, ploidy level endopolyploidy, and reproductive modes in plants.** In: Besse, P. (ed.) *Methods Mol Biol.* New York, NY: Humana, 2021; **2222**: 325–361.
[PubMed Abstract](#) | [Publisher Full Text](#)
- POWO: **Plants of the World Online.** Royal Botanic Gardens, Kew, 2024.
[Reference Source](#)
- Preston CD, Pearman D, Trevor DD: **New atlas of the British & Irish flora.** Oxford: Oxford University Press, 2002.
[Reference Source](#)
- Rao SSP, Huntley MH, Durand NC, *et al.*: **A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping.** *Cell.* 2014; **159**(7): 1665–1680.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Ratnasingham S, Hebert PDN: **bold: the Barcode of Life Data system (<http://www.barcodinglife.org>).** *Mol Ecol Notes.* 2007; **7**(3): 355–364.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Rautiainen M, Marschall T: **MBG: Minimizer-based sparse de Bruijn Graph construction.** *Bioinformatics.* 2021; **37**(16): 2476–2478.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Rhie A, McCarthy SA, Fedrigo O, *et al.*: **Towards complete and error-free genome assemblies of all vertebrate species.** *Nature.* 2021; **592**(7856): 737–746.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Rhie A, Walenz BP, Koren S, *et al.*: **Mercury: Reference-free quality, completeness, and phasing assessment for genome assemblies.** *Genome Biol.* 2020; **21**(1): 245.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Stace CA, Thompson H, Stace M: **New flora of the British Isles.** 4th ed. C&M Floristics, 2019.
[Reference Source](#)
- Strickland M, Cornwell C, Howard C: **Sanger Tree of Life fragmented DNA clean up: manual SPRI.** *protocols.io.* 2023.
[Publisher Full Text](#)
- Surana P, Muffato M, Qi G: **sanger-tol/readmapping: sanger-tol/readmapping v1.1.0 - Hebridean Black (1.1.0).** *Zenodo.* 2023a.
[Publisher Full Text](#)
- Surana P, Muffato M, Sadasivan Baby C: **sanger-tol/genomenote (v1.0.dev).** *Zenodo.* 2023b.
[Publisher Full Text](#)
- Todorovic M, Oatley G, Howard C: **Sanger Tree of Life HMW DNA extraction: automated plant MagAttract v.2.** *protocols.io.* 2023.
[Publisher Full Text](#)
- Twyford AD, Beasley J, Barnes I, *et al.*: **A DNA barcoding framework for taxonomic verification in the Darwin Tree of Life project [version 1; peer review: 1 approved].** *Wellcome Open Res.* 2024; **9**: 339.
[Publisher Full Text](#)
- Vasimuddin M, Misra S, Li H, *et al.*: **Efficient architecture-aware acceleration of BWA-MEM for multicore systems.** In: *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS).* IEEE, 2019; 314–324.
[Publisher Full Text](#)
- Wang D, Bädärau AS, Swamy MK, *et al.*: ***Arctium* species secondary metabolites chemodiversity and bioactivities.** *Front Plant Sci.* 2019; **10**: 834.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Wellcome Sanger Institute: **The genome sequence of lesser burdock, *Arctium minus* (Hill) Bernh.** European Nucleotide Archive. [dataset], accession number PRJEB53860, 2023.
- Zhou C, McCarthy SA, Durbin R: **YaHS: yet another Hi-C scaffolding tool.** *Bioinformatics.* 2023; **39**(1): btac808.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Open Peer Review

Current Peer Review Status:   

Version 1

Reviewer Report 28 November 2024

<https://doi.org/10.21956/wellcomeopenres.25504.r109244>

© 2024 González-Segovia E. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Eric González-Segovia 

University of British Columbia, British Columbia, Canada

The article discusses the *Arctium minus* genome assembly, providing a brief overview of the species' biology and its uses. The methods for the specimen collection, DNA extraction, HiFi and Hi-C sequencing as well the genome assembly are well detailed and clearly described.

I have just a minor comment regarding these lines: "*While not fully phased, the assembly deposited is of one haplotype. Contigs corresponding to the second haplotype have also been deposited.*"

From my understanding, the authors employed a primary/alternate assembly approach. In this approach, the primary assembly consists of a mixture of both haplotypes, where the best and longer contigs are selected. The alternate assembly contains contigs for heterozygous regions not included in the primary assembly. Given this, I believe it is inaccurate to state that "*the assembly deposited is of one haplotype*" since the primary assembly is a mixture of haplotypes rather than a single haplotype. Similarly, the alternate assembly does not consist solely of contigs from a single haplotype. Instead, they could simply state that the chromosome-level assembly was created using the primary assembly and that the alternate assembly is also available.

Other than this minor comment, the article is well written, and everything was clear to me. However, I do not possess expertise in this species to verify the accuracy of the biological details provided.

References

1. de-Dios T, Fontserè C, Renom P, Stiller J, et al.: Whole genomes from the extinct Xerces Blue butterfly can help identify declining insect species. *Elife*. 2024; **12**. [PubMed Abstract](#) | [Publisher Full Text](#)

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Yes

Are sufficient details of methods and materials provided to allow replication by others?

Yes

Are the datasets clearly presented in a useable and accessible format?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Genetics, genomics and local adaptation.


I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 18 November 2024

<https://doi.org/10.21956/wellcomeopenres.25504.r109240>

© 2024 Alfalahi A. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Ayoob Obaid Alfalahi 

University of Anbar, Ramadi, Al Anbar Governorate, Iraq

In this article “The genome sequence of lesser burdock, *Arctium minus* (Hill) Bernh. (Asteraceae)”, authors produced high-quality genome assembly for diploid *Arctium minus* (Hill) Bernh using DNA barcoding, HiFi, and Hi-C map techniques. The results revealed a genome size of 2,070 mb scaffolded into 18 chromosomal pseudomolecules, along with mitochondrial and plastid genome assemblies. The results are really interesting and will provide insights into chromosomal evolution of this biennial species, in addition to better understanding on adaptation, genomic, and proteomic levels.

The following notes may be considered:

1. In Figure 1, a) more than one plant species are presented in the provided photo which is a bit confusing, hence a close shot indicating only the targeted plant species (*Arctium minus*) may be provided. Also, it would be better to include photo showing all the plant parts (root, stem, leaves, flowers and maybe fruits), this will be more descriptive and informative, especially in such specialized studies.
2. The collection stage of plant sample should be determined, however, full blooming stage is preferred. Furthermore, in Table 1, authors pointed to different organism parts used for molecular analysis (leaf and flower), are these different parts from one plant sample!
3. Authors missed some relevant up-to-date references that can really support the presented information about *Arctium minus* (e.g <https://doi.org/10.1007/s11130-024-01175-w>).

4. The introduction included long stories without relevant references supporting these claims, especially the last section describing the polyploidy phenomenon in *Arctium*.
5. Links for the used software should be provided. The manufactured company, address and origin should be stated in all methodology part not only for Library preparation and sequencing work.

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Yes

Are sufficient details of methods and materials provided to allow replication by others?

Yes

Are the datasets clearly presented in a useable and accessible format?

Yes

Competing Interests: No competing interests were disclosed.**Reviewer Expertise:** Plant Biotechnology**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Reviewer Report 05 November 2024

<https://doi.org/10.21956/wellcomeopenres.25504.r105938>

© 2024 Paniego N. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**Norma Paniego** 

Instituto Nacional de Tecnología Agropecuaria; Consejo Nacional de Investigaciones Científicas y Técnicas, Instituto de Agrobiotecnología y Biología Molecular, Hurlingham, Buenos Aires, Argentina

In this Data Note, the authors present a chromosome-level genome assembly of *Arctium minus*, a biennial herbaceous plant of the Asteraceae family, using a combination of PACBIO HiFi and HiC Illumina sequencing technologies. *A. minus*, like other members of the genus *Arctium*, is valued for its edible and medicinal properties. This species is native to Europe and Western Asia and is distributed in different regions worldwide due to its high adaptability to various soils and environments.

This work provides information on the genome size of *A. minus*, estimated at 2,070 Mb. The results are clearly described and are complemented by complete tables and figures. The figures have interactive versions that can be accessed via web links to understand the results better. The study provides a high-quality assembly of the nuclear genome into 18 chromosomes, together with genome annotations. The mitochondrial and plastid genomes are also presented. The methods applied are described in detail and linked to standardized protocols available on the protocol.io platform. In addition, the software used and the quality control metrics applied are clearly outlined and referenced. All data and metadata, including sequences, annotations, and herbarium specimens, are accessible in public repositories.

The work is original, well written, and provides valuable information for the plant community interested in the Arctium family, as well as for groups studying genes involved in the synthesis of medicinal compounds. Furthermore, being a highly resilient species, this genome can contribute to the understanding of the genetic basis of plant adaptability and resilience in different environments. Finally, from a technological perspective, this work is interesting for research groups involved in whole genome sequencing, as it presents relevant methodological aspects, including robust quality control metrics applied at distinct stages of the process, and efficient forms for graphical visualization of the results.

Minor comments:

I suggest adding in the Background section, the references to genome assembly of *A lappa* (1, 2) and the multi-omic database for *Arctium sp.* (3). In the Results section, I note an inconsistency in the reporting of genome coverage for PACBIO and HiC sequencing. I recommend revising these values in relation to the total gigabases generated and the estimated genome size in this note.

References

1. Zhang D, Xing Y, Xu L, Zhao R, et al.: The complete mitochondrial genome of *Arctium lappa* (Campanulales, Asteraceae). *Mitochondrial DNA Part B*. 2020; **5** (2): 1722-1723 [Publisher Full Text](#)
2. Yang Y, Li S, Xing Y, Zhang Z, et al.: The first high-quality chromosomal genome assembly of a medicinal and edible plant *Arctium lappa*. *Mol Ecol Resour*. 2022; **22** (4): 1493-1507 [PubMed Abstract](#) | [Publisher Full Text](#)
3. Song Y, Yang Y, Xu L, Bian C, et al.: The burdock database: a multi-omic database for *Arctium lappa*, a food and medicinal plant. *BMC Plant Biol*. 2023; **23** (1): 86 [PubMed Abstract](#) | [Publisher Full Text](#)

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Yes

Are sufficient details of methods and materials provided to allow replication by others?

Yes

Are the datasets clearly presented in a useable and accessible format?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Plant genomics, genomics of plant fungal disease resistance, plant genomics breeding

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.
