



GENOME NOTE

The genome sequence of meadow thistle, *Cirsium dissectum*

(L.) Hill (Asterales: Asteraceae)

[version 1; peer review: 1 approved]

Michael F. Fay ¹, Ilia J. Leitch ¹, Sahr Mian ¹, José Ignacio Márquez-Corro ¹, Colin T. Kelleher ², 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

¹Royal Botanic Gardens Kew, Richmond, Surrey, England, UK

²National Botanic Gardens of Ireland, Glasnevin, Dublin, Ireland

V1 First published: 28 Sep 2026, 11:763
<https://doi.org/10.12688/wellcomeopenres.27952.1>

Latest published: 28 Sep 2026, 11:763
<https://doi.org/10.12688/wellcomeopenres.27952.1>

Abstract

We present a genome assembly of *Cirsium dissectum* (meadow thistle; Streptophyta; Magnoliopsida; Asterales; Asteraceae). The assembly consists of two haplotypes with total lengths of 1,069.23 megabases and 1,020.16 megabases. Most of the haplotype 1 assembly (94.46%) was scaffolded into 17 chromosomal pseudomolecules. Haplotype 2 was assembled to scaffold level. The mitochondrial genome assembly has a length of 316.65 kilobases and the plastid genome assembly has a length of 152.52 kilobases. Gene annotation of the haplotype 1 assembly by Ensembl identified 32,454 protein-coding genes. This assembly was generated as part of the Darwin Tree of Life project, which produces reference genomes for eukaryotic species found in Britain and Ireland.

Keywords

Cirsium dissectum, meadow thistle, genome sequence, chromosomal, Asterales

Open Peer Review

Approval Status

1

version 1

28 Sep 2026



[view](#)

1. **Automated Benchmarking**, Wellcome Sanger Institute, Hinxton, UK

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



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

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

Author roles: **Fay MF:** Investigation, Resources; **Leitch IJ:** Investigation, Resources; **Mian S:** Investigation, Resources; **Márquez-Corro JI:** Investigation, Resources; **Kelleher CT:** Investigation, Resources;

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome through core funding to the Wellcome Sanger Institute [220540, <https://doi.org/10.35802/220540>] 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: © 2026 Fay MF *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: Fay MF, Leitch IJ, Mian S *et al.* **The genome sequence of meadow thistle, *Cirsium dissectum* (L.) Hill (Asterales: Asteraceae) [version 1; peer review: 1 approved]** Wellcome Open Research 2026, 11:763
<https://doi.org/10.12688/wellcomeopenres.27952.1>

First published: 28 Sep 2026, 11:763 <https://doi.org/10.12688/wellcomeopenres.27952.1>

Species taxonomy

Eukaryota; Viridiplantae; Streptophyta; Streptophytina; Embryophyta; Tracheophyta; Euphyllophyta; Spermatophyta; Magnoliopsida; Mesangiospermae; eudicotyledons; Gunneridae; Pentapetalae; asterids; campanulids; Asterales; Asteraceae; Carduoideae; Cardueae; Carduinae; *Cirsium*; *Cirsium dissectum* Hill (NCBI:txid1155260)

Background

Cirsium dissectum (L.) Hill, the meadow thistle, is a long-lived perennial plant. Its tall, cottony stems usually bear a single magenta-purple flower head (Figure 1), while rosettes of long-stalked leaves are whitish beneath and edged with weak spines (Harrap, 2025; Rimes & Smith, 2017).

This Oceanic Temperate species is restricted to the milder, wetter parts of Europe. It is widespread but localised in northern and western France and also occurs in Belgium, while populations have declined greatly in north-western Germany and the Netherlands. In Britain, meadow thistle is concentrated in south Wales and south-west England, with more scattered populations elsewhere. It is widely distributed in Ireland, especially in western counties, but is very rare in Scotland (Rimes & Smith, 2017).

Meadow thistle grows on nutrient-poor, peaty or waterlogged soils that remain damp throughout the year and receive base-rich water. In Britain it is a defining species of *Molinia caerulea*–*Cirsium dissectum* fen-meadows. It reproduces mainly by clonal spread through rhizomes that can reach 40 cm, although seed production is also important. Flowering is mostly in June and July. Its scented, nectar-rich florets attract butterflies, bumblebees and long-tongued flies (Rimes & Smith, 2017).

Cirsium dissectum has declined through drainage and agricultural improvement of wetlands, insufficient management, nutrient enrichment and habitat fragmentation. Studies using microsatellite markers have linked smaller populations with lower genetic diversity and reproductive fitness (Rimes & Smith, 2017).

Cirsium dissectum is diploid, with $2n = 2x = 34$ (Bureš *et al.*, 2010; Morton, 1977; Rice *et al.*, 2015).

We sequenced the genome of *Cirsium dissectum* as part of the Darwin Tree of Life Project, which aims to generate high-quality reference genomes for all named eukaryotic species in Britain and Ireland to support research, conservation, and the sustainable use of biodiversity. The sequenced specimen was collected from Dublin, Ireland.



Figure 1. Illustrative photograph of *Cirsium dissectum*. This is not the sequenced specimen. Photograph by [Natasha de Vere](#), CC BY 2.0, via Flickr.

Methods

Sample acquisition, flow cytometry and DNA barcoding

A specimen of *C. dissectum* (specimen ID KDTOL10835, ToLID daCirDiss1) was used for genome sequencing. It was collected from Dublin, Dublin, Ireland (latitude 53.3694, longitude -6.2706) on 2023-06-08. The specimen was collected by Mike Fay, Ilia

Leitch, Sahr Mian and José Ignacio Márquez-Corro and identified by Colin Kelleher. The daCirDiss1 specimen was also used for Hi-C and RNA sequencing. The herbarium voucher for the sequenced plant (collector number MC9544) is deposited at Kew ([K001527174](#)).

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

The initial identification was verified by an additional DNA barcoding process according to the framework developed by [Twyford *et al.* \(2024\)](#). Part of the plant specimen was preserved in silica gel desiccant ([Chase & Hills, 1991](#)). DNA extracted from the dried plant was amplified by PCR for standard barcode markers, with the amplicons sequenced and compared to public sequence databases including GenBank and the Barcode of Life Database (BOLD) ([Ratnasingham & Hebert, 2007](#)). Following whole genome sequence generation, the relevant DNA barcode region was also used alongside the initial barcoding data for sample tracking at the WSI ([Twyford *et al.*, 2024](#)). The standard operating procedures for Darwin Tree of Life barcoding are available on [protocols.io](#).

Nucleic acid extraction

High molecular weight (HMW) DNA extraction protocols developed by the Tree of Life Core Laboratory at the Wellcome Sanger Institute (WSI) are available on [protocols.io](#) ([Howard *et al.*, 2025](#)). Tissue from the daCirDiss1 sample was weighed and [triaged](#) to select the appropriate extraction workflow. HMW DNA was prepared for two PacBio HiFi libraries: DTOL15069480 and ToL_PacBio_HiFi015:D2. Leaf tissue was homogenised by [cryogenic bead beating](#). HMW DNA was extracted using the [Automated Sorbitol Washing + Plant Organic Extraction](#) protocol. DNA was sheared into an average fragment size of 12–20 kb following the [MegaRuptor®3 for LI PacBio](#) protocol. Sheared DNA was purified by [automated SPRI](#) (solid-phase reversible immobilisation), using AMPure PB beads (Pacific Biosciences) and the Thermo Fisher KingFisher™ Apex to eliminate shorter fragments and concentrate the DNA.

The concentration of the sheared and purified DNA was assessed with a NanoDrop spectrophotometer and a Qubit Fluorometer using the Qubit dsDNA High Sensitivity Assay kit. Fragment size distribution was evaluated by running the sample on the Femto Pulse system.

RNA was extracted from leaf tissue of daCirDiss1 in the Tree of Life Laboratory at the WSI using the [RNA Extraction: Automated MagMax™ mirVana protocol](#). The RNA concentration was assessed with a NanoDrop spectrophotometer and a Qubit Fluorometer using the Qubit RNA Broad-Range Assay kit. RNA integrity was assessed using the Agilent RNA 6000 Pico Kit and Eukaryotic Total RNA assay.

PacBio HiFi library preparation and sequencing

Library preparation and sequencing took place in the WSI Scientific Operations core. Libraries DTOL15069480 and ToL_PacBio_HiFi015:D2 were prepared with the SMRTbell Prep Kit 3.0 (Pacific Biosciences) according to the manufacturer’s instructions. The kit includes reagents for end repair/A-tailing, adapter ligation, post-ligation SMRTbell bead clean-up, and nuclease treatment. Size selection and clean-up were performed using diluted AMPure PB beads (Pacific Biosciences). DNA concentration was quantified with a Qubit Fluorometer v4.0 (Thermo Fisher Scientific) and the Qubit 1X dsDNA HS assay kit. Final library fragment size was assessed with the Agilent Femto Pulse Automated Pulsed Field CE Instrument (Agilent Technologies) using the gDNA 55 kb BAC analysis kit.

The libraries were sequenced in two runs on a Revo instrument (Pacific Biosciences, California, USA). Prepared libraries with a molarity greater than 2 nM were normalised to 2 nM, and 15 µL was used for making complexes. For libraries with a molarity below 2 nM, the available 10 µL library volume was used without normalisation. Primers were annealed and polymerases were bound to generate circularised complexes, following the manufacturer’s instructions. Complexes were purified using 1.2× SMRTbell beads, diluted to the Revo loading concentration of 200–300 pM, and spiked with a Revo sequencing internal control. Sequencing used Revo 25M SMRT cells. SMRT Link software (Pacific Biosciences), a web-based workflow manager, was used to configure and monitor the runs and to carry out primary and secondary data analysis.

Hi-C

Sample preparation and crosslinking

Hi-C data were generated from the daCirDiss1 leaf sample using the Arima-HiC v2 kit (Arima Genomics). Tissue was finely ground using the Covaris cryoPREP Dry Pulverizer (Covaris), and then subjected to nuclei isolation. Nuclei were isolated using a modified protocol based on the Qiagen QProteome Cell Compartment Kit (Qiagen), in which only the Lysis and CE2 buffers were used, with QIAshredder spin columns. After isolation, nuclei were fixed using formaldehyde to a final concentration of 2% to crosslink the DNA. The crosslinked DNA was then digested and biotinylated according to the manufacturer's instructions. A clean-up step was performed with SPRIselect beads before library preparation. DNA concentration was quantified using the Qubit Fluorometer v4.0 (Thermo Fisher Scientific) and the Qubit HS Assay Kit, following the manufacturer's instructions.

Hi-C library preparation and sequencing

Biotinylated DNA constructs were fragmented using a Covaris E220 sonicator and size-selected to 400–600 bp using SPRIselect beads. DNA was enriched with Arima-HiC v2 kit Enrichment beads. End repair, A-tailing, and adapter ligation were carried out with the NEBNext Ultra II DNA Library Prep Kit (New England Biolabs), following a modified protocol where library preparation occurs while DNA remains bound to the Enrichment beads. Libraries were amplified with KAPA HiFi HotStart mix and a custom Unique Dual Index (UDI) barcode set (Integrated DNA Technologies). Depending on sample concentration and biotinylation percentage determined at the crosslinking stage, libraries were amplified with 10–16 PCR cycles. Post-PCR clean-up was performed with SPRIselect beads. Libraries were quantified using the AccuClear Ultra High Sensitivity dsDNA Standards Assay Kit (Biotium) and a FLUOstar Omega plate reader (BMG Labtech).

Prior to sequencing, libraries were normalised to 10 ng/μL. Normalised libraries were quantified again to create equimolar and/or weighted 2.8 nM pools. Pool concentrations were checked using the Agilent 4200 TapeStation (Agilent) with High Sensitivity D500 reagents before sequencing. Libraries were sequenced on the Illumina NovaSeq X, generating paired-end 150-bp reads.

RNA library preparation and sequencing

Libraries were prepared with the NEBNext® Ultra II Directional RNA Library Prep Kit for Illumina (New England Biolabs), following the manufacturer's instructions. Poly(A) mRNA was isolated from 100 ng of total RNA using oligo(dT) beads, converted to cDNA, and uniquely indexed; 14 PCR cycles were performed. Libraries were prepared with an intended average fragment size of 100–300 bp. Libraries were quantified, normalised, and pooled equimolarly to a final concentration of 2.8 nM before dilution to 150 pM for loading. Sequencing was carried out on the Illumina NovaSeq X, generating paired-end reads.

Genome assembly

Prior to assembly of the PacBio HiFi reads, a database of k -mer counts ($k = 31$) was generated from the filtered reads using [FastK](#). The k -mer frequency distribution was analysed with GenomeScope (version 2.1.0) ([Ranallo-Benavidez et al., 2020](#)), providing fitted point estimates of haploid genome size, heterozygosity and repeat content.

The HiFi reads were assembled using Hifiasm in Hi-C phasing mode ([Cheng et al., 2021, 2022](#)), producing two haplotypes. Hi-C reads ([Rao et al., 2014](#)) were mapped to contigs using bwa-mem2 ([Vasimuddin et al., 2019](#)). Contigs were further scaffolded with Hi-C data in YaHS ([Zhou et al., 2023](#)), using the --break option for handling potential misassemblies. The scaffolded assemblies were evaluated using Gfastats ([Formenti et al., 2022](#)), BUSCO ([Manni et al., 2021](#)) and MerquryFK ([Rhie et al., 2020](#)). The organelle genomes were assembled using OATK ([Zhou et al., 2025](#)). The comparator for the plastid assembly was the chloroplast genome of *Cirsium setidens* (NC_063550.1).

Assembly curation

The assembly was decontaminated using the Assembly Screen for Cobionts and Contaminants (ASCC) pipeline. The flat files and maps for curation were generated with [TreeVal](#). Manual curation was conducted primarily in [PretextView](#) and HiGlass ([Kerpedjiev et al., 2018](#)). Scaffolds were visually inspected and corrected as described by [Howe et al. \(2021\)](#). The two haplotype assemblies were combined for curation. Manual corrections included 48 breaks and 101 joins. This reduced the scaffold count by 3.1%, reduced the scaffold N50 by 12.4%, and reduced the total assembly length by 0.7%. The curation process is described at [sanger-tol/curation-resources](#). A Hi-C contact map of the haplotype 1 assembly was generated with PretextViewSnapshot.

Assembly quality assessment

The MerquryFK tool ([Rhie et al., 2020](#)) was run in a Singularity container ([Kurtzer et al., 2017](#)) to evaluate k -mer completeness and assembly quality for both haplotypes using the k -mer database ($k = 31$) computed prior to genome assembly. The analysis outputs included assembly QV scores and completeness statistics.

The haplotype 1 assembly was assessed using BlobToolKit (Challis *et al.*, 2020). PacBio reads were mapped to the assembly to derive sequence coverage, and BUSCO (Manni *et al.*, 2021) was used to assess gene-set completeness. DIAMOND BLASTp searches of extracted BUSCO proteins and DIAMOND BLASTx searches of assembly sequences against UniProt Reference Proteomes (Buchfink *et al.*, 2021), followed by BLASTn searches of sequences without DIAMOND BLASTx hits against the NCBI nt database (Altschul *et al.*, 1990), supported taxonomic assignment. Coverage, sequence composition, taxonomic assignment and BUSCO results were integrated into a BlobDir for interactive inspection and visualisation. The snail and blob plots were rendered locally from this hosted BlobDir using BlobTk v0.8.3. The scaffold-length radial scale function for the snail plot was set to linear (Challis & Blaxter, 2026). The blob plot used circular markers with area proportional to scaffold length (Challis *et al.*, 2020).

Genome sequence report

Sequence data

The genome of a specimen of *C. dissectum* was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating 54.31 Gb (gigabases) from 4.95 million reads. These reads were used to assemble the genome. GenomeScope analysis using a $p = 2$ model gave a k -mer-based haploid genome size estimate of 1,197.07 Mb, with heterozygosity of 1.13% and repeat content of 61.49%; the full-model fit was 97.89% (Figure 2). Using flow cytometry, the genome size (1C-value) of the sample was estimated to be 1.2 pg, equivalent to 1,170.00 Mb. These estimates guided expectations for the assembly. The NCBI assembly record reports 44× genome coverage. Hi-C sequencing produced 97.37 Gb from 322.42 million read pairs, which were used to scaffold the haplotype 1 assembly. RNA sequencing data were also generated and are available in public sequence repositories. Table 1 summarises the specimen and sequencing details.

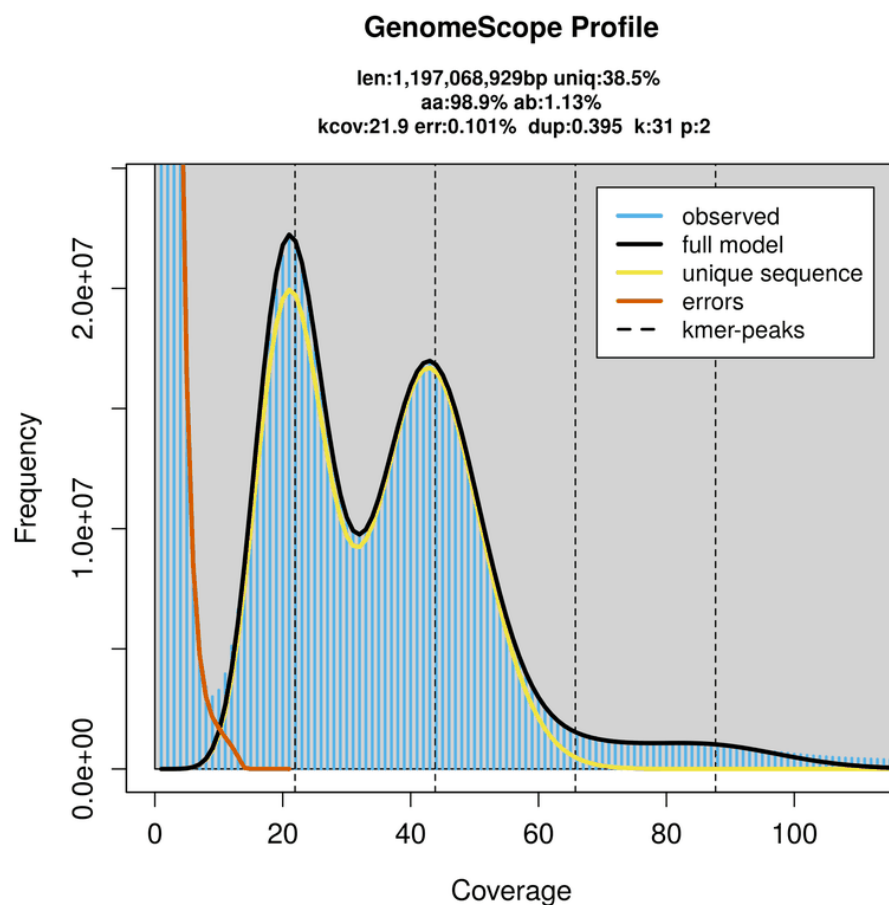


Figure 2. GenomeScope model fitted to the k -mer frequency distribution. The plot shows the observed and modelled k -mer spectra used to estimate haploid genome size, heterozygosity and repeat content from unassembled sequencing reads.

Table 1. Specimen and sequencing data for *C. dissectum* (BioProject [PRJEB85034](#))

Platform	PacBio HiFi	Hi-C	RNA-seq
ToLID	daCirDiss1	daCirDiss1	daCirDiss1
Specimen ID	KDTOL10835	KDTOL10835	KDTOL10835
BioSample (source individual)	SAMEA114563213	SAMEA114563213	SAMEA114563213
BioSample (tissue)	SAMEA114563312	SAMEA114563312	SAMEA114563312
Tissue	leaf	leaf	leaf
Instrument	Revio	Illumina NovaSeq X	Illumina NovaSeq X
Library ID & run accession(s)	DTOL15069480: ERR14209144 ; ToL_PacBio_HiFi015:D2: ERR14209145	ERR14224624	ERR14224625
Read count total	4.95 million reads	322.42 million read pairs	49.70 million read pairs
Base count total	54.31 Gb	97.37 Gb	15.01 Gb

Assembly statistics

The genome was assembled into two haplotypes using Hi-C phasing. The haplotype 1 assembly was curated to chromosome level, while haplotype 2 was assembled to scaffold level. The haplotype 1 assembly, which is proposed as the reference assembly for this species, has a total length of 1,069.23 Mb in 955 scaffolds (excluding organelle sequences), with 664 gaps, and its scaffold N50 is 59.47 Mb (Table 2). The scaffold auN for the haplotype 1 assembly is 59.1 Mb.

Table 2. Genome assembly statistics for *C. dissectum*

Genome assembly	Haplotype 1	Haplotype 2
Assembly name	daCirDiss1.hap1.1	daCirDiss1.hap2.1
Assembly accession	GCA_965276805.1	GCA_965276745.1
Assembly level	chromosome	scaffold
Span (Mb)	1,069.23	1,020.16
Number of chromosomes	17	-
Number of contigs	1,619	905
Contig N50 (Mb)	3.15	2.96
Number of scaffolds (excluding organelle sequences)	955	317
Scaffold N50 (Mb)	59.47	59.32
Scaffold auN (Mb)	59.1	59.9
Longest scaffold length (Mb)	91.53	86.47
Organelles	Mitochondrial genome: 316.65 kb; Plastid genome: 152.52 kb	-

Most of the haplotype 1 assembly sequence (94.46%) was assigned to 17 chromosome-level scaffolds. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 3; Table 3). Contigs in the following regions have uncertain order and orientation: chromosome 4 (22,224–25,498 kbp), chromosome 9 (32,498–34,433 kbp), chromosome 10 (29,659–32,127

kbp), and chromosome 17 (19,000–20,450 kbp). During curation, haplotypic inversions were observed on chromosome 1 (55,700–59,900 kbp), chromosome 6 (18,200–28,700 kbp), and chromosome 7 (34,900–40,300 kbp).

The mitochondrial genome (length 316.65 kb, OZ255697.1) and plastid genome (length 152.52 kb, OZ255698.1) were also assembled. These sequences are included as contigs in the multifasta file of the haplotype 1 genome submission and as standalone records.

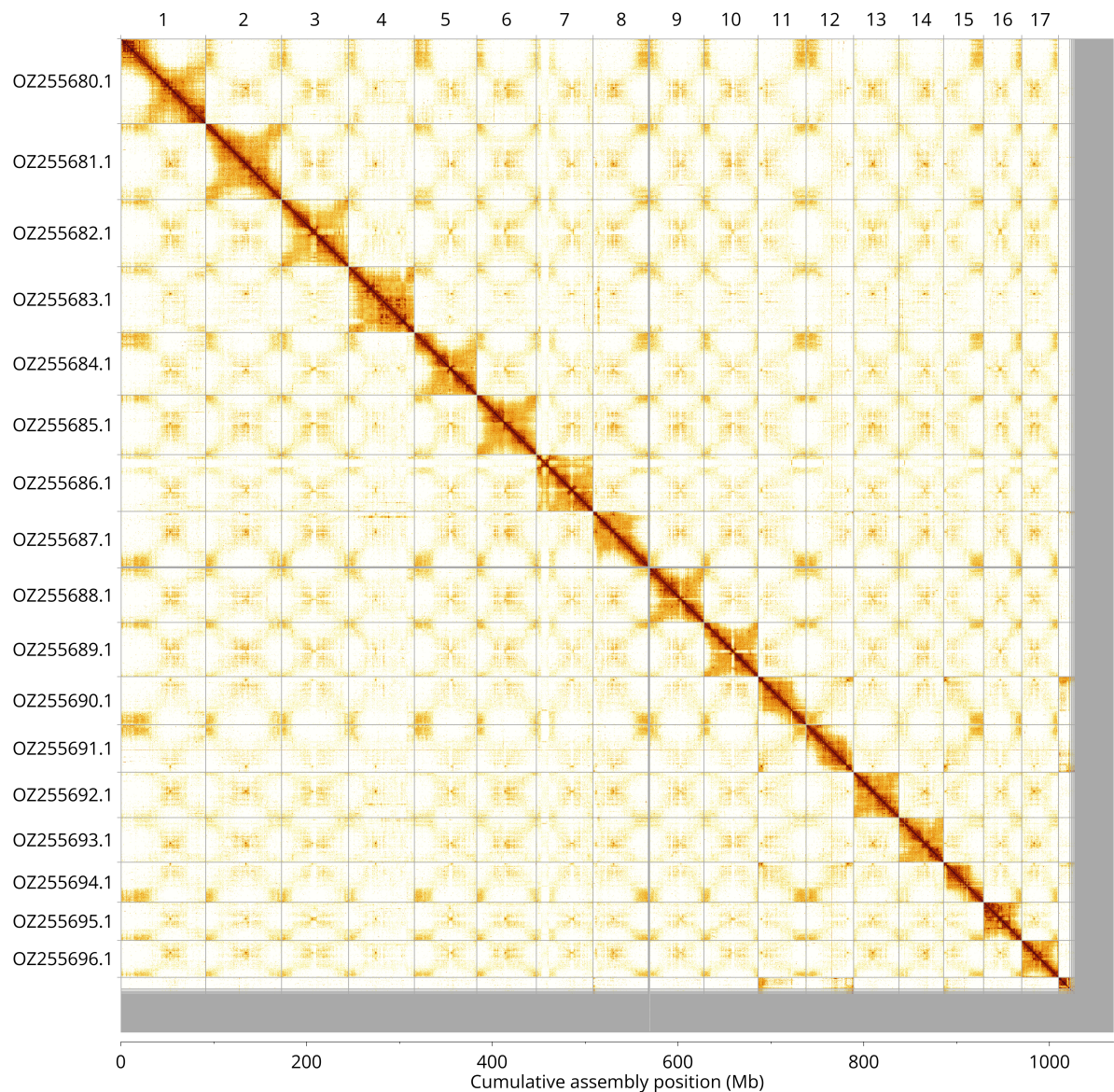


Figure 3. Hi-C contact map of the *C. dissectum* haplotype 1 assembly. Assembled chromosomes are shown in order of size and labelled along the axes, with a megabase scale shown below. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in the haplotype 1 genome assembly of *C. dissectum* daCirDiss1

INSDC accession	Molecule	Length (Mb)	GC%
OZ255680.1	1	91.53	35
OZ255681.1	2	81.74	35.50

INSDC accession	Molecule	Length (Mb)	GC%
OZ255682.1	3	72.03	35.50
OZ255683.1	4	70.94	36
OZ255684.1	5	67.14	35
OZ255685.1	6	64.09	35.50
OZ255686.1	7	61.08	36
OZ255687.1	8	60.87	36
OZ255688.1	9	58.60	35.50
OZ255689.1	10	58.47	35.50
OZ255690.1	11	51.47	35.50
OZ255691.1	12	51.33	36.50
OZ255692.1	13	48.68	36
OZ255693.1	14	48.05	36
OZ255694.1	15	43.26	35.50
OZ255695.1	16	41.06	35.50
OZ255696.1	17	39.64	36.50

Assembly quality metrics

For the haplotype 1 assembly, the estimated QV is 61.3, and for the haplotype 2 assembly, 61.3. When the two haplotypes are combined, the assembly achieves an estimated QV of 61.3. The *k*-mer completeness is 78.48% for the haplotype 1 assembly, 77.45% for the haplotype 2 assembly, and 98.62% for the combined haplotypes (Figure 4).

BUSCO analysis using the eudicots_odb10 reference set ($n = 2,326$) identified 97.7% of the expected gene set (single = 90.8%, duplicated = 6.9%) in the haplotype 1 assembly. For the haplotype 2 assembly, BUSCO v.5.7.1 analysis identified 97.8% of the expected gene set (single = 90.3%, duplicated = 7.5%).

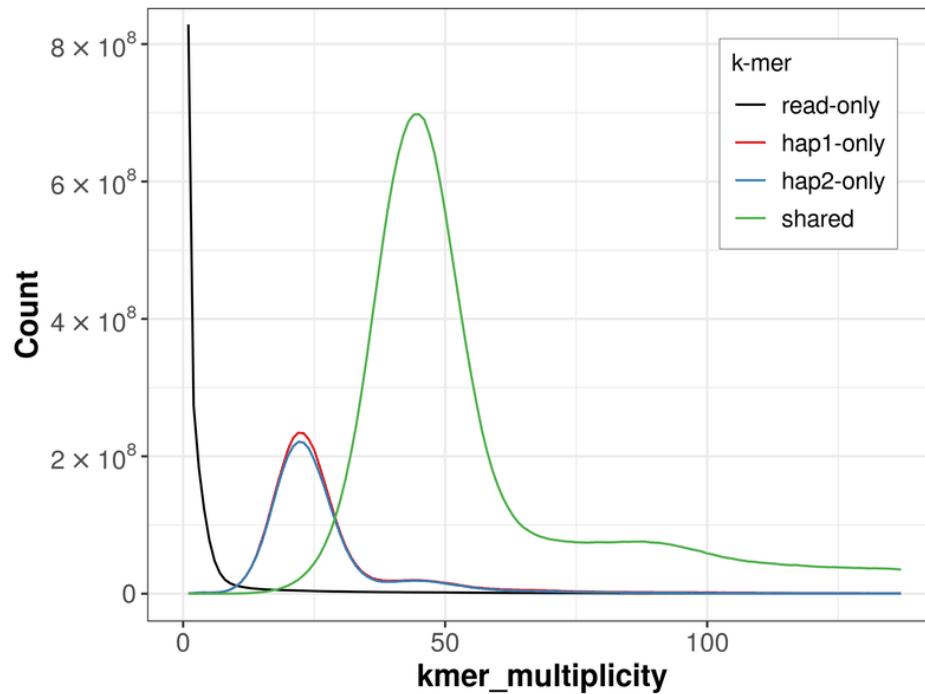


Figure 4. Evaluation of k -mer completeness using MerquryFK. This plot illustrates the recovery of k -mers from the original read data in the haplotype 1 assembly, the haplotype 2 assembly and the combined haplotypes. The horizontal axis represents k -mer multiplicity, and the vertical axis shows the number of k -mers. The black curve represents k -mers that appear in the reads but are not assembled, while the remaining curves correspond to k -mers recovered in the assemblies shown.

The snail plot in Figure 5 provides a summary of assembly contiguity, cumulative scaffold count, sequence composition and BUSCO completeness for the haplotype 1 assembly (Challis & Blaxter, 2026). The relative sizes of the longest scaffold, scaffold N50 and N90, together with the dark-grey scaffold-length distribution, show how much of the assembly is contained in long scaffolds. The outer composition track shows variation in GC, AT and N content, while the central track shows the cumulative scaffold count and the BUSCO inset summarises gene-set completeness.

The blob plot in Figure 6 displays scaffolds from the haplotype 1 assembly by GC proportion and base coverage. Marker area is proportional to scaffold length, and colour indicates taxonomic assignment (Challis *et al.*, 2020).

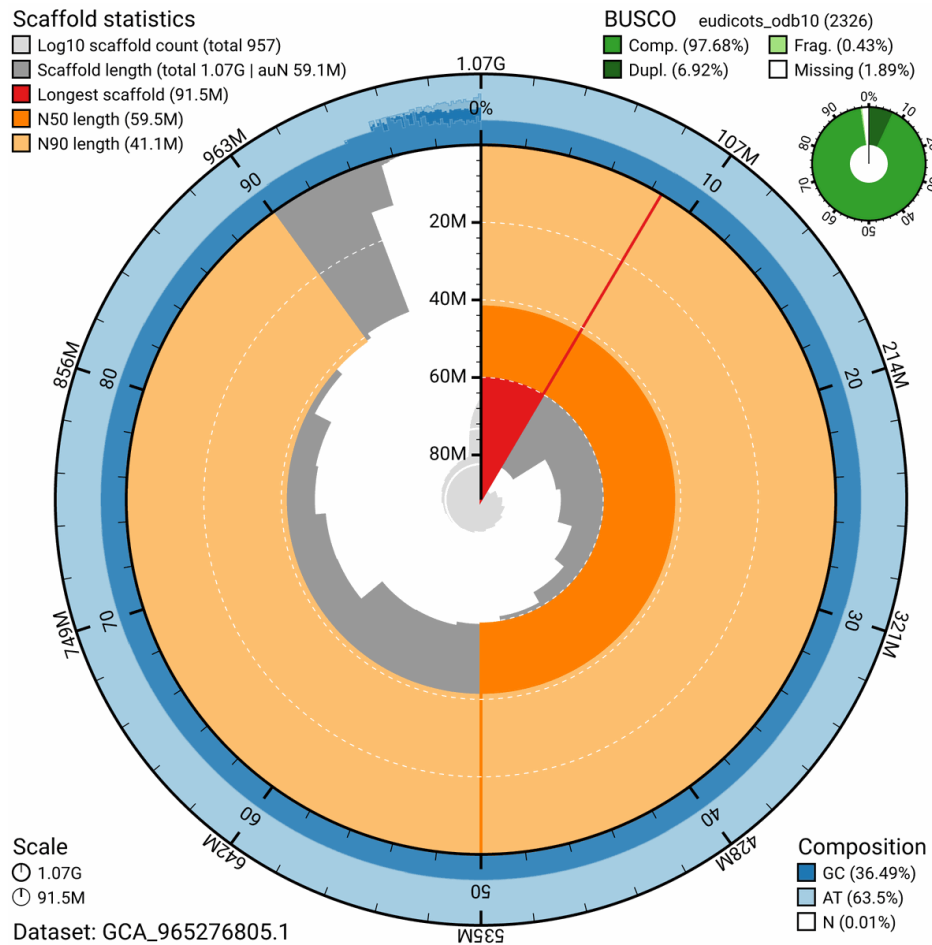


Figure 5. Assembly metrics for daCirDiss1.hap1.1, the haplotype 1 assembly. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the haplotype 1 assembly, and the main plot is divided into 1,000 bins around the circumference. The outermost blue tracks display the distribution of GC, AT, and N percentages across the bins. Scaffolds are arranged clockwise from longest to shortest and are depicted in dark grey. Scaffold length is shown on a linear radial scale, scaled to the longest scaffold. Red shading indicates the longest scaffold, and the deeper orange and pale orange shading represent the N50 and N90 lengths. The summary statistics include scaffold auN, the area under the scaffold Nx curve. The light grey spiral at the centre shows the cumulative scaffold count on a logarithmic scale. A summary of complete, fragmented, duplicated, and missing BUSCO genes in the eudicots_odb10 set is presented at the top right. The assembly can be explored interactively in the [BlobToolKit viewer](#).

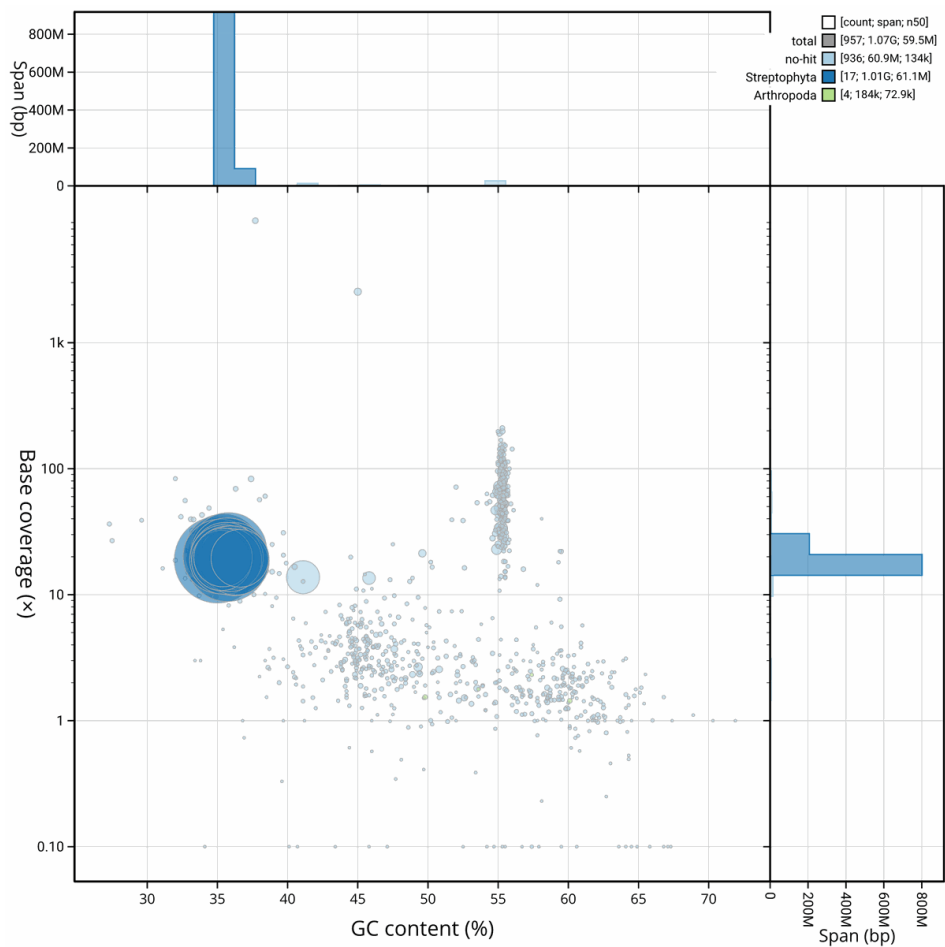


Figure 6. BlobToolKit blob plot for daCirDiss1.hap1.1, the haplotype 1 assembly. The plot shows base coverage (vertical axis) and GC content (horizontal axis). Circles represent scaffolds; marker area is proportional to scaffold length, and colour indicates taxonomic assignment. The histograms along the axes display the total length of sequences distributed across different levels of coverage and GC content. The assembly can be explored interactively in the [BlobToolKit viewer](#).

Table 4 lists the assembly metric benchmarks adapted from [Rhie et al.. \(2021\)](#) and the Earth BioGenome Project (EBP) [Report on Assembly Standards](#). The three-part EBP metric calculated for the haplotype 1 assembly is **6.C.Q61**.

Table 4. Earth BioGenome Project summary metrics for the *C. dissectum* assembly

Measure	Value	Benchmark
EBP summary (haplotype 1)	6.C.Q61	6.C.Q40
Contig N50 length	3.15 Mb	> 1 Mb
Scaffold N50 length	59.47 Mb	= chromosome N50
Consensus quality (QV)	Haplotype 1: 61.3; haplotype 2: 61.3; combined: 61.3	# 40
k-mer completeness	Haplotype 1: 78.48%; Haplotype 2: 77.45%; combined: 98.62%	> 90%
BUSCO	Haplotype 1: C:97.7% [S:90.8%, D:6.9%], F:0.4%, M:1.9%, n:2,326;	S > 90%; D < 5%

Measure	Value	Benchmark
	Haplotype 2: C:97.8% [S:90.3%, D:7.5%], F:0.5%, M:1.7%, n:2,326	
Percentage of assembly assigned to chromosomes	94.46%	# 90%

Notes: The EBP summary uses log10(Contig N50); chromosome-level (C) or log10(Scaffold N50); Q (Mercury QV). BUSCO v5.7.1 using the eudicots_odb10 lineage dataset: C=complete; S=single-copy; D=duplicated; F=fragmented; M=missing; n=orthologues.

Genome annotation report

The *C. dissectum* genome assembly (GCA_965276805.1) was annotated by Ensembl at the European Bioinformatics Institute (EBI). The annotation comprises 70,044 transcripts from 32,454 protein-coding genes and 24,413 non-coding genes. The average transcript length is 3,138.67 bp, with an average of 1.23 coding transcripts per gene and 4.02 exons per transcript. For further information about the annotation, please refer to the [annotation page](#) on Ensembl.

Data availability

European Nucleotide Archive: *Cirsium dissectum*. Accession number [PRJEB85034](#); <https://identifiers.org/ena.embl/PRJEB85034>. The genome sequence is released openly for reuse. The *C. dissectum* genome sequencing initiative is part of the Darwin Tree of Life Project ([PRJEB40665](#)) and the Sanger Institute Tree of Life Programme ([PRJEB43745](#)). All raw sequence data and the haplotype assemblies have been deposited in INSDC databases. Raw data and assembly accession identifiers are reported in Table 1 and Table 2.

Pipelines used for genome assembly at the WSI Tree of Life are available at <https://pipelines.tol.sanger.ac.uk/pipelines>. Table 5 lists software versions used in this study.

Table 5. Software versions and sources

Software	Version	Source
BLAST	2.14.0	https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/
BlobToolKit	4.4.5	https://github.com/genomehubs/blobtoolkit
BlobTk (hosted BlobDir)	0.5.1	https://github.com/genomehubs/blobtk
BlobTk (snail and blob plot renderer)	0.8.3	https://github.com/genomehubs/blobtk
BUSCO	5.7.1; 6.1.0	https://gitlab.com/ezyab/busco
bwa-mem2	2.2.1	https://github.com/bwa-mem2/bwa-mem2
DIAMOND	2.1.8	https://github.com/bbuchfink/diamond
fasta_windows	0.2.4	https://github.com/tolkit/fasta_windows
FastK	1.2	https://github.com/thegenemyers/FASTK
GenomeScope	2.1.0	https://github.com/tbenavi1/genomescope2.0
Gfastats	1.3.6	https://github.com/vgl-hub/gfastats
Hifiasm	0.19.8-r603	https://github.com/chhyllp123/hifiasm
HiGlass	1.13.4	https://github.com/higlass/higlass
MercuryFK	1.1.0-c1	https://github.com/thegenemyers/MERQURY.FK
Minimap2	2.28-r1209	https://github.com/lh3/minimap2

Software	Version	Source
Oatk	0.9	https://github.com/c-zhou/oatk
MultiQC	1.14; 1.17 and 1.18	https://github.com/MultiQC/MultiQC
Nextflow	24.10.4	https://github.com/nextflow-io/nextflow
PretextSnapshot	0.0.6	https://github.com/sanger-tol/PretextSnapshot
PretextView	1.0.3	https://github.com/sanger-tol/PretextView
samtools	1.21	https://github.com/samtools/samtools
sanger-tol/ascc	0.1.0	https://github.com/sanger-tol/ascc
sanger-tol/blobtoolkit	v0.7.1	https://github.com/sanger-tol/blobtoolkit
sanger-tol/curationpretext	1.4.2	https://github.com/sanger-tol/curationpretext
Seqtk	1.4-r122	https://github.com/lh3/seqtk
Singularity	3.9.0	https://github.com/sylabs/singularity
TreeVal	1.2.0	https://github.com/sanger-tol/treval
YaHS	1.2.2	https://github.com/c-zhou/yahs

Author information

- Members of the [Royal Botanic Gardens Kew Genome Acquisition Lab](#)
- Members of the [Plant Genome Sizing Collective](#)
- Members of the [Darwin Tree of Life Barcoding collective](#)
- Members of the [Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team](#)
- Members of [Wellcome Sanger Institute Scientific Operations – Sequencing Operations](#)
- Members of the [Wellcome Sanger Institute Tree of Life Core Informatics team](#)
- Members of the [Tree of Life Core Informatics collective](#)
- Members of the [Darwin Tree of Life Consortium](#)

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 to the ‘**Darwin Tree of Life Project Sampling Code of Practice**’, which can be found in full on the [Darwin Tree of Life website](#). 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.

References

- Altschul S. F., Gish W., Miller W.: **Basic Local Alignment Search Tool**. *Journal of Molecular Biology*; 1990; **215**(3): 403–410. [PubMed Abstract](#) | [Publisher Full Text](#) |
- Buchfink B., Reuter K., Drost H.-G.: **Sensitive protein alignments at tree-of-life scale using DIAMOND**. *Nature Methods*; 2021; **18**(4): 366–368. [PubMed Abstract](#) | [Publisher Full Text](#) |
- Bureš Petr, Šmarda Petr, Rotreklová Olga: **Pollen viability and natural hybridization of Central European species of *Cirsium***. *Preslia*; 2010. Accessed: 2026-09-25. **82** 391–422. [Reference Source](#)
- Challis R., Richards E., Rajan J.: **BlobToolKit – interactive quality assessment of genome assemblies**. *G3: Genes, Genomes, Genetics*; 2020; **10**(4): 1361–1374. [PubMed Abstract](#) | [Publisher Full Text](#) |
- Challis Richard, Blaxter Mark: **Snail plots are badges of genome assembly quality**. *G3 Genes | Genomes | Genetics*; 2026; **16**(6): jkag074. [PubMed Abstract](#) | [Publisher Full Text](#) |
- Chase M. W., Hills H. H.: **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., Jarvis E. D., Fedrigo O.: **Haplotype-resolved assembly of diploid genomes without parental data**. *Nature Biotechnology*; 2022; **40**(9): 1332–1335. [PubMed Abstract](#) | [Publisher Full Text](#) |
- Cheng H., Concepcion G. T., Feng X.: **Haplotype-resolved *de novo* assembly using phased assembly graphs with Hifiasm**. *Nature Methods*; 2021; **18**(2): 170–175. [PubMed Abstract](#) | [Publisher Full Text](#) |
- Doležal J., Greilhuber J., Suda J.: **Estimation of nuclear DNA content in plants using flow cytometry**. *Nature Protocols*; 2007; **2**(9): 2233–2244. [PubMed Abstract](#) | [Publisher Full Text](#) |
- Formenti G., Abueg L., Brajuka A.: **Gfastats: Conversion, evaluation and manipulation of genome sequences using assembly graphs**. *Bioinformatics*; 2022; **38**(17): 4214–4216. [PubMed Abstract](#) | [Publisher Full Text](#) |
- Harrap Simon: *Harrap's wild flowers: A field guide to the wild flowers of Britain & Ireland*. Bloomsbury Publishing Plc; 2025. 2 ed.
- Howard Caroline, Denton Amy, Jackson Benjamin W.: **On the path to reference genomes for all biodiversity: Laboratory protocols and lessons learned from processing over 2,000 species in the Sanger Tree of Life**. *GigaScience*; 2025; **14** gjaf119. [Reference Source](#) [Publisher Full Text](#) |
- Howe K., Chow W., Collins J.: **Significantly improving the quality of genome assemblies through curation**. *GigaScience*; 2021; **10**(1): gjaa153. [PubMed Abstract](#) | [Publisher Full Text](#) |
- Kerpedjiev P., Abdennur N., Lekschas F.: **HiGlass: Web-based visual exploration and analysis of genome interaction maps**. *Genome Biology*; 2018; **19**(1): 125. [PubMed Abstract](#) | [Publisher Full Text](#) |
- Kurtzer G. M., Sochat V., Bauer M. W.: **Singularity: Scientific containers for mobility of compute**. *PLOS ONE*; 2017; **12**(5): e0177459. [PubMed Abstract](#) | [Publisher Full Text](#) |
- Loureiro J., Rodriguez E., Doležal J.: **Two new nuclear isolation buffers for plant DNA flow cytometry: A test with 37 species**. *Annals of Botany*; 2007; **100**(4): 875–888. [PubMed Abstract](#) | [Publisher Full Text](#) |
- Manni Mosè, Berkeley Matthew R, Seppey Mathieu: **BUSCO update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes**. *Molecular Biology and Evolution*; 2021; **38**(10): 4647–4654. [PubMed Abstract](#) | [Publisher Full Text](#) |
- Morton J. K.: **A cytological study of the Compositae (excluding *Hieracium* and *Taraxacum*) of the British Isles**. *Watsonia*; 1977. Accessed: 2026-09-25. **11** 211–223. [Reference Source](#)
- Pellicer J., Powell R. F., Leitch I. J.: **The application of flow cytometry for estimating genome size, ploidy level endopolyploidy, and reproductive modes in plants**. *Methods in Molecular Biology* Besse P.: New York, NY. 2021. **2222**: 325–361. [PubMed Abstract](#) |
- Ranallo-Benavidez Timothy R., Jaron Kai S., Schatz Michael C.: **GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes**. *Nature Communications*; 2020; **11**(1): 1432. [PubMed Abstract](#) | [Publisher Full Text](#) |
- Rao S. S. P., Huntley M. H., Durand N. C.: **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](#) |
- Ratnasingham S., Hebert P. D. N.: **BOLD: The Barcode of Life Data System (<http://www.barcodinglife.org>)**. *Molecular Ecology Notes*; 2007; **7**(3): 355–364. [PubMed Abstract](#) | [Publisher Full Text](#) |
- Rhie A., Walenz B. P., Koren S.: **Merquy: Reference-free quality, completeness, and phasing assessment for genome assemblies**. *Genome Biology*; 2020; **21**(1): 245. [PubMed Abstract](#) | [Publisher Full Text](#) |
- Rhie A., McCarthy S. A., Fedrigo O.: **Towards complete and error-free genome assemblies of all vertebrate species**. *Nature*; 2021; **592**(7856): 737–746. [PubMed Abstract](#) | [Publisher Full Text](#) |
- Rice Anna, Glick Lior, Abadi Shiran: **The Chromosome Counts Database (CCDB) – a community resource of plant chromosome numbers**. *New Phytologist*; 2015; **206**(1): 19–26. [Publisher Full Text](#) |

Rimes C., Smith S.: **Species account: *Cirsium dissectum***. Botanical Society of Britain and Ireland. 2017. Accessed: 2026-09-24. [Reference Source](#)

Twyford A. D., Beasley J., Barnes I.: **A DNA barcoding framework for taxonomic verification in the Darwin Tree of Life Project**. *Wellcome Open Research*; 2024; **9**: 339. [PubMed Abstract](#) | [Publisher Full Text](#) |

Vasimuddin Md., Misra S., Li H.: **Efficient architecture-aware acceleration of BWA-MEM for multicore systems**. *2019 IEEE international parallel and distributed processing symposium (IPDPS)*. IEEE. 2019. 314–324. [Publisher Full Text](#) |

Zhou C., McCarthy S. A., Durbin R.: **YaHS: Yet another Hi-C scaffolding tool**. *Bioinformatics*; 2023; **39**(1): btac808. [PubMed Abstract](#) | [Publisher Full Text](#) |

Zhou Chenxi, Brown Max, Blaxter Mark: **Oatk: A de novo assembly tool for complex plant organelle genomes**. *Genome Biology*; 2025; **26**(1): 235. [Publisher Full Text](#) |

Open Peer Review

Current Peer Review Status: 

Version 1

Reviewer Report 28 September 2026

<https://doi.org/>

© 2026 **Benchmarking 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.



Automated Benchmarking

Wellcome Sanger Institute, Hinxton, UK

As this article is a Genome Note, Automated Benchmarking has been used to determine whether the reported assembly passes four defined checks established by the Earth BioGenome Project: the raw and assembled data are publicly available, the Earth BioGenome Project three-part metric is met, the k-mer completeness exceeds 90%, and the BUSCO score passes the stated threshold.

For *Cirsium dissectum* (daCirDiss1), all four checks passed.

Data available: pass. ENA study PRJEB85034, raw-read accessions ERR14209144, ERR14209145, ERR14224624, and ERR14224625, and assembly accessions GCA_965276805.1 and GCA_965276745.1 are public.

Earth BioGenome Project three-part metric: pass. Haplotype 1 is reported as 6.C.Q61, meeting the 6.C.Q40 reference standard.

k-mer completeness: pass. The combined haplotype assemblies have 98.62% k-mer completeness, exceeding the Earth BioGenome Project target of 90%.

BUSCO: pass. Haplotype 1, which is proposed as the reference assembly for this species, has 97.7% complete BUSCOs (90.8% single-copy and 6.9% duplicated). BUSCO software version: 5.7.1; lineage dataset: eudicots_odb10. It meets the plant-specific target of greater than 90% complete BUSCOs.

This decision uses values reported in the Genome Note and public accession records.

Competing Interests: This Automated Benchmarking peer-review report was generated automatically and relies on analyses performed by the Tree of Life programme, which also

produced the Genome Note under review.

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.
