DATA NOTE

The genome sequence of the Marsh Pennywort, *Hydrocotyle vulgaris* L. (Apiales: Araliaceae)

[version 1; peer review: 3 approved]

Markus Ruhsam[1], Peter M Hollingsworth[1],
Royal Botanic Garden Edinburgh Genome Acquisition Lab,
Darwin Tree of Life Barcoding collective, Plant Genome Sizing 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

[1]Royal Botanic Garden Edinburgh, Edinburgh, Scotland, UK

v1 **First published:** 28 Jul 2025, **10**:370
https://doi.org/10.12688/wellcomeopenres.24582.1
Latest published: 28 Jul 2025, **10**:370
https://doi.org/10.12688/wellcomeopenres.24582.1

Abstract

We present a genome assembly from a specimen of *Hydrocotyle vulgaris* (Marsh Pennywort; Streptophyta; Magnoliopsida; Apiales; Araliaceae). The assembly contains two haplotypes with total lengths of 870.40 megabases and 862.86 megabases. Most of haplotype 1 (97.78%) is scaffolded into 48 chromosomal pseudomolecules. Haplotype 2 was assembled to scaffold level. The mitochondrial and plastid genome assemblies have lengths of 562.52 kilobases and 153.03 kilobases, respectively.

Keywords

Hydrocotyle vulgaris, Marsh Pennywort, genome sequence, chromosomal, Apiales

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ✓ ✓ ✓

	1	2	3
version 1 28 Jul 2025	✓ view	✓ view	✓ view

1. **Sunil Kumar Sahu**, BGI Research, Shenzhen, China
2. **Peng-Fei Ma**, Chinese Academy of Sciences, Kunming, China
3. **Bimal K Chetri**, Indian Institute of Technology, Guwahati, India

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: Ruhsam M: Investigation, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Hollingsworth PM**: Investigation, Resources, 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 (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.

How to cite this article: Ruhsam M, Hollingsworth PM, Royal Botanic Garden Edinburgh Genome Acquisition Lab *et al.* **The genome sequence of the Marsh Pennywort, *Hydrocotyle vulgaris* L. (Apiales: Araliaceae) [version 1; peer review: 3 approved]** Wellcome Open Research 2025, **10**:370 https://doi.org/10.12688/wellcomeopenres.24582.1

First published: 28 Jul 2025, **10**:370 https://doi.org/10.12688/wellcomeopenres.24582.1

Species taxonomy

Eukaryota; Viridiplantae; Streptophyta; Streptophytina; Embryophyta; Tracheophyta; Euphyllophyta; Spermatophyta; Magnoliopsida; Mesangiospermae; eudicotyledons; Gunneridae; Pentapetalae; asterids; campanulids; Apiales; Apiineae; Araliaceae; *Hydrocotyle*; *Hydrocotyle vulgaris* L. (NCBI: txid82091)

Background

Hydrocotyle vulgaris is a low growing (5–15 cm) aquatic perennial with peltate leaves on permanently or seasonally inundated substrates such as bogs, fens, marshes and lakesides (Stace *et al.*, 2019). It is native to Europe and Northern Africa. In Britain and Ireland it is widespread and locally common at altitudes from sea level to 635m, but has experienced some declines associated with drainage and development in the English Midlands and parts of Scotland (Stroh *et al.*, 2023).

Marsh Pennywort was formerly included in the Apiaceae subfamily Hydrocotyloideae but has now been transferred to Araliaceae (Plunkett *et al.*, 2004). However, it is sometimes treated as member of its own family Hydrocotylaceae (Stace *et al.*, 2019).

Hydrocotyle vulgaris is a stoloniferous plant, which roots at all nodes forming extensive clonal patches. Only emergent plants produce flowers, which are easily overlooked due to their small sizes (~ 1 mm) and their position below the leaves (Schou *et al.*, 2023). Small insects may act as pollinators, however, the flowers are considered to be mainly self-pollinated (Schou *et al.*, 2023) and it is likely that Marsh Pennywort's main mode of reproduction is via clonal growth. For example, although seed production of *H. vulgaris* has been observed in China, where it was introduced as an ornamental in the 1990s, there is currently no evidence for the establishment of individuals via seeds (Zhao *et al.*, 2024). Clonal spread has been so successful in China that *H. vulgaris* is now considered to be invasive there with numerous studies investigating its mechanism of clonal reproduction (Liu *et al.*, 2014; Miao LiHua *et al.*, 2011; Si *et al.*, 2020; Wang *et al.*, 2018; Zhao *et al.*, 2024).

Marsh pennywort has also been studied for its potential medicinal properties (Ureta *et al.*, 2018) as well as its role in phytoremediation (Ni *et al.*, 2018; Vafaei *et al.*, 2013).

Hydrocotyle vulgaris is an octoploid ($2n$ = 96) based on Dutch (Gadella & Kliphuis, 1966), British (Al-Bermani *et al.*, 1993; Hollingsworth *et al.*, 1992), Italian (Amadei *et al.*, 1982) and Portuguese material (Queiros, 1972) with a likely basic chromosome number of x = 12 (Yi *et al.*, 2004).

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 – we present a chromosomally complete genome sequence for *Hydrocotyle vulgaris*. This genome was assembled using the Tree of Life pipeline from a specimen collected in Aberlady, Scotland, UK (Figure 1).

Sample acquisition, flow cytometry and DNA barcoding

A specimen of *Hydrocotyle vulgaris* (specimen ID EDTOL01496, ToLID drHydVulg1; Figure 1) was used for genome sequencing. It was collected from Aberlady, UK (latitude 56.0161, longitude –2.8529) on 2021-06-14. The specimen was collected and identified by Markus Ruhsam (Royal Botanic Garden Edinburgh). Metadata collection followed the recommended standards of the Darwin Tree of Life project (Lawniczak *et al.*, 2022).

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. CyStain™ PI OxProtect Staining Buffer (Sysmex UK Ltd) was used for isolation of nuclei (Loureiro *et al.*, 2007), and the internal calibration standard was *Petroselinum crispum* 'Champion Moss Curled' with an assumed 1C-value of 2 200 Mb (Obermayer *et al.*, 2002).

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

Protocols for high molecular weight (HMW) DNA extraction developed at the Wellcome Sanger Institute (WSI) Tree of

Figure 1. Photograph of the *Hydrocotyle vulgaris* (drHydVulg1) specimen from which samples were taken for genome sequencing.

Life Core Laboratory are available on protocols.io (Howard *et al.*, 2025). The drHydVulg1 sample was weighed and triaged to determine the appropriate extraction protocol. Tissue from the whole plant was homogenised by cryogenic bead beating. HMW DNA was extracted using the Automated Plant MagAttract v4 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 using a Nanodrop spectrophotometer and Qubit Fluorometer using the Qubit dsDNA High Sensitivity Assay kit. Fragment size distribution was evaluated by running the sample on the FemtoPulse system. For this sample, the final post-shearing DNA had a Qubit concentration of 6.14 ng/μL and a yield of 798.20 ng. The 260/280 spectrophotometric ratio was 1.87, and the 260/230 ratio was 1.25.

PacBio HiFi library preparation and sequencing

Library preparation and sequencing were performed at the WSI Scientific Operations core. Libraries were prepared using 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 using a Qubit Fluorometer v4.0 (ThermoFisher 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 sample was sequenced on a Revio instrument (Pacific Biosciences). The prepared library was normalised to 2 nM, and 15 μL was used for making complexes. Primers were annealed and polymerases bound to generate circularised complexes, following the manufacturer's instructions. Complexes were purified using 1.2X SMRTbell beads, then diluted to the Revio loading concentration (200–300 pM) and spiked with a Revio sequencing internal control. The sample was sequenced on a Revio 25M SMRT cell. The SMRT Link software (Pacific Biosciences), a web-based workflow manager, was used to configure and monitor the run and to carry out primary and secondary data analysis.

Hi-C

Sample preparation and crosslinking

Hi-C data were generated from the whole plant tissue of drHydVulg1 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. Library amplification was performed using 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 and 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. Sequencing was performed using paired-end 150 bp reads on the Illumina NovaSeq 6000.

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. GenomeScope2 (Ranallo-Benavidez *et al.*, 2020) was used to analyse the *k*-mer frequency distributions, providing estimates of genome size, heterozygosity, and repeat content.

The HiFi reads were assembled using Hifiasm in Hi-C phasing mode (Cheng *et al.*, 2021; Cheng *et al.*, 2022), producing two haplotypes. Hi-C reads (Rao *et al.*, 2014) were mapped to the primary 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 MERQURY.FK (Rhie *et al.*, 2020). The organelle genomes were assembled using OATK (Zhou *et al.*, 2024).

Assembly curation

The assembly was decontaminated using the Assembly Screen for Cobionts and Contaminants (ASCC) pipeline. TreeVal was used to generate the flat files and maps for use in curation.

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). Manual corrections included 4 breaks, 11 joins, and removal of 7 haplotypic duplications. The curation process is documented at https://gitlab.com/wtsi-grit/rapid-curation. PretextSnapshot was used to generate a Hi-C contact map of the final assembly.

Assembly quality assessment

The Merqury.FK 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 databases (k = 31) computed prior to genome assembly. The analysis outputs included assembly QV scores and completeness statistics.

The genome was analysed using the BlobToolKit pipeline, a Nextflow implementation of the earlier Snakemake version (Challis *et al.*, 2020). The pipeline aligns PacBio reads using minimap2 (Li, 2018) and SAMtools (Danecek *et al.*, 2021) to generate coverage tracks. It runs BUSCO (Manni *et al.*, 2021) using lineages identified by querying the GoaT database (Challis *et al.*, 2023). For the three domain-level lineages, BUSCO genes are aligned to the UniProt Reference Proteomes database (Bateman *et al.*, 2023) using DIAMOND blastp (Buchfink *et al.*, 2021). The genome is divided into chunks based on the density of BUSCO genes from the closest taxonomic lineage, and each chunk is aligned to the UniProt Reference Proteomes database with DIAMOND blastx. Sequences without hits are chunked using seqtk and aligned to the NT database with blastn (Altschul *et al.*, 1990). The BlobToolKit suite consolidates all outputs into a blobdir for visualisation. The BlobToolKit pipeline was developed using nf-core tooling (Ewels *et al.*, 2020) and MultiQC (Ewels *et al.*, 2016), with package management via Conda and Bioconda (Grüning *et al.*, 2018), and containerisation through Docker (Merkel, 2014) and Singularity (Kurtzer *et al.*, 2017).

Genome sequence report

Sequence data

The genome of a specimen of *Hydrocotyle vulgaris* was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating 58.72 Gb (gigabases) from 5.47 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 467.23 Mb, with a heterozygosity of 37.71% and repeat content of 39.29% (Figure 2). Using flow cytometry, the genome size (1C-value) of the sample was estimated to be 1.07 pg, equivalent to 1 050.00 Mb. These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 118× coverage. Hi-C sequencing produced 89.30 Gb from 591.38 million reads, which were

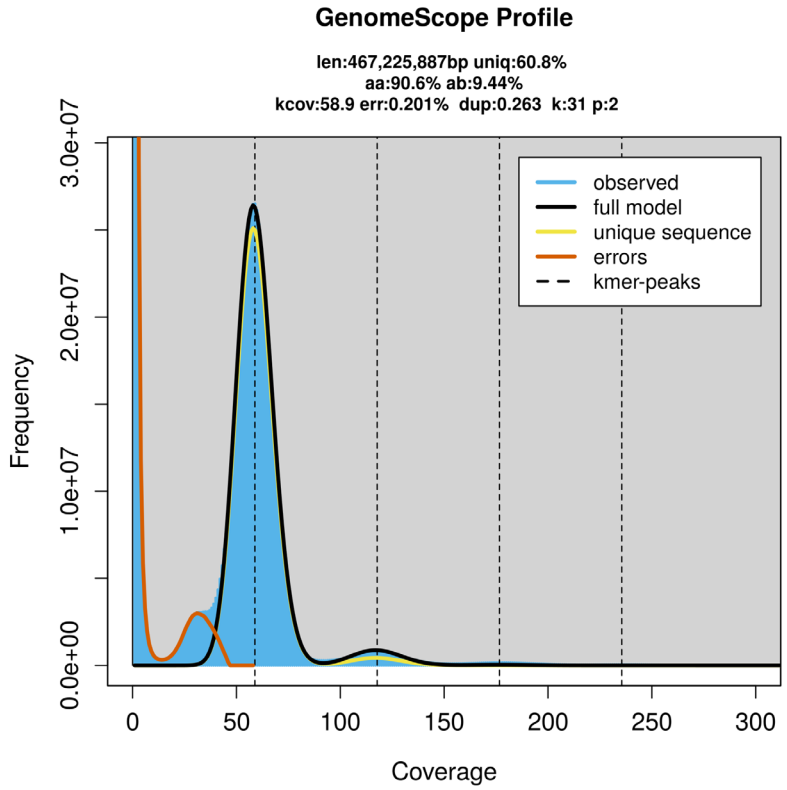

Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2. The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

used to scaffold the assembly. Table 1 summarises the specimen and sequencing details.

Assembly statistics

The genome was assembled into two haplotypes using Hi-C phasing. Haplotype 1 was curated to chromosome level, while haplotype 2 was assembled to scaffold level. The final assembly has a total length of 870.40 Mb in 331 scaffolds, with 261 gaps, and a scaffold N50 of 17.8 Mb (Table 2).

Most of the assembly sequence (98.73%) was assigned to 48 chromosomal-level scaffolds. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 3; Table 3). The mitochondrial and plastid genomes were also assembled. These sequences are included as contigs in the multifasta file of the genome submission and as standalone records.

Assembly quality metrics

For haplotype 1, the estimated QV is 58.3, and for haplotype 2, 58.3. When the two haplotypes are combined, the assembly achieves an estimated QV of 58.3. The *k*-mer completeness is 96.17% for haplotype 1, 96.07% for haplotype 2, and 98.22% for the combined haplotypes (Figure 4).

BUSCO analysis using the eudicots_odb10 reference set (n = 2 326) identified 98.6% of the expected gene set (single = 41.1%, duplicated = 57.6%) for haplotype 1. The snail plot in Figure 5 summarises the scaffold length distribution and other assembly statistics for haplotype 1. The blob plot in Figure 6 shows the distribution of scaffolds by GC proportion and coverage for haplotype 1.

Table 4 lists the assembly metric benchmarks adapted from Rhie *et al.* (2021) the Earth BioGenome Project Report on

Table 1. Specimen and sequencing data for BioProject PRJEB74171.

Platform	PacBio HiFi	Hi-C
ToLID	drHydVulg1	drHydVulg1
Specimen ID	EDTOL01496	EDTOL01496
BioSample (source individual)	SAMEA10241384	SAMEA10241384
BioSample (tissue)	SAMEA10241462	SAMEA10241462
Tissue	whole plant	whole plant
Instrument	Revio	Illumina NovaSeq 6000
Run accessions	ERR12779287	ERR12791519
Read count total	5.47 million	591.38 million
Base count total	58.72 Gb	89.30 Gb

Table 2. Genome assembly statistics.

Assembly name	drHydVulg1.hap1.1	drHydVulg1.hap2.1
Assembly accession	GCA_964106935.1	GCA_964107055.1
Assembly level	chromosome	scaffold
Span (Mb)	870.40	862.86
Number of chromosomes	48	N/A
Number of contigs	592	521
Contig N50	5.33 Mb	5.5 Mb
Number of scaffolds	331	256
Scaffold N50	17.8 Mb	17.76 Mb
Longest scaffold length (Mb)	26.66	N/A
Organelles	Mitochondrion: 562.52 kb; Plastid: 153.03 kb	N/A

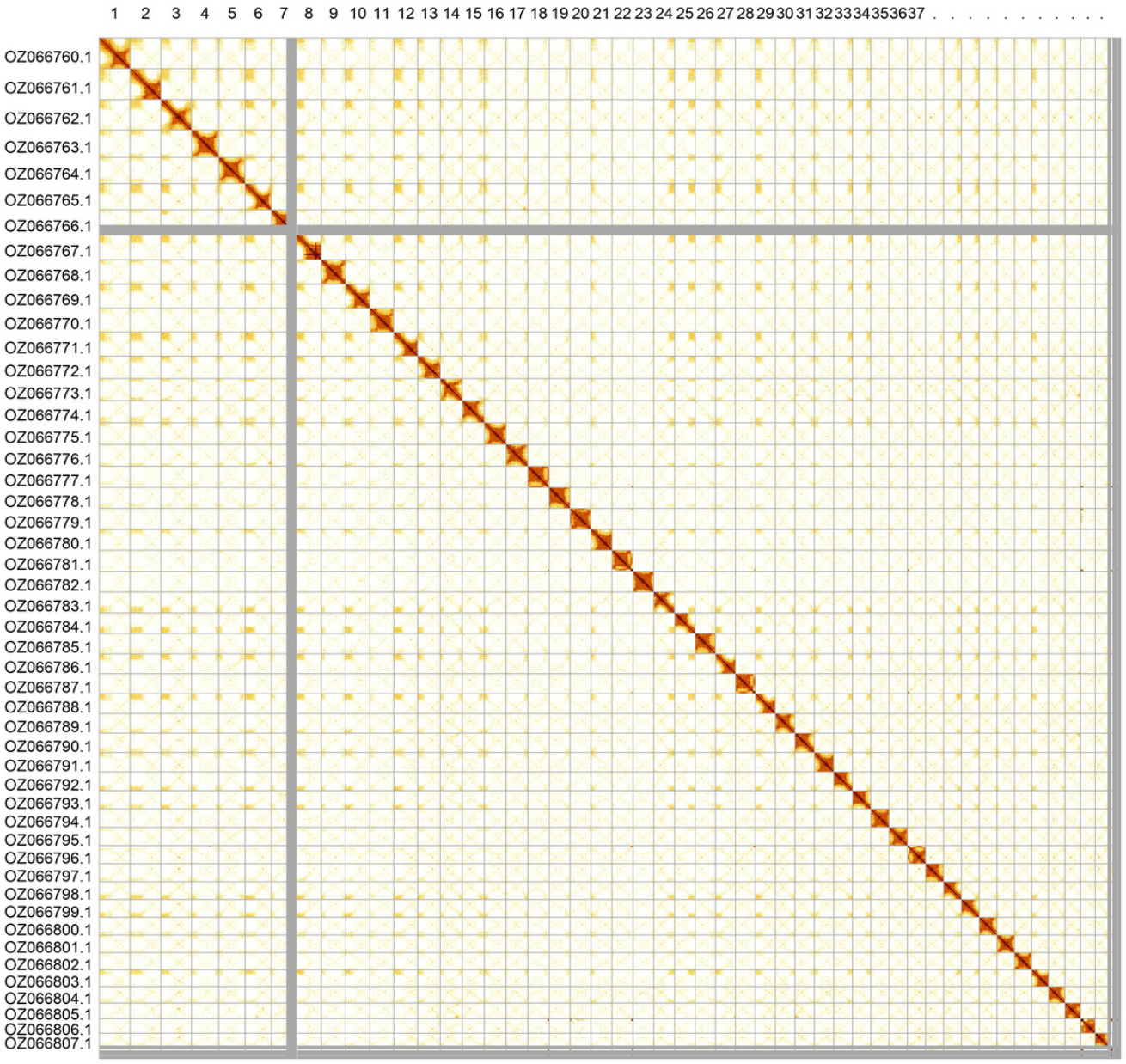

Figure 3. Hi-C contact map of the *Hydrocotyle vulgaris* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes. The plot was generated using PretextSnapshot.

Assembly Standards September 2024. The EBP metric calculated for the haplotype 1 is **6.C.Q58**, meeting the recommended reference standard.

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

Table 3. Chromosomal pseudomolecules in the haplotype 1 genome assembly of *Hydrocotyle vulgaris* drHydVulg1.

INSDC accession	Molecule	Length (Mb)	GC%
OZ066760.1	1	26.66	33
OZ066761.1	2	26.17	33.50
OZ066762.1	3	25.87	33.50
OZ066763.1	4	23.42	34
OZ066764.1	5	22.38	33.50
OZ066765.1	6	22.29	33.50
OZ066766.1	7	21.38	34
OZ066767.1	8	21.16	34.50
OZ066768.1	9	20.77	34
OZ066769.1	10	20.55	33.50
OZ066770.1	11	20.44	33.50
OZ066771.1	12	20.40	33.50
OZ066772.1	13	19.15	33
OZ066773.1	14	18.78	33.50
OZ066774.1	15	18.72	33.50
OZ066775.1	16	18.66	33.50
OZ066776.1	17	18.51	33.50
OZ066777.1	18	17.98	33.50
OZ066778.1	19	17.88	34
OZ066779.1	20	17.88	33.50
OZ066780.1	21	17.81	34
OZ066781.1	22	17.81	33.50
OZ066782.1	23	17.80	34
OZ066783.1	24	17.80	33.50
OZ066784.1	25	17.50	34
OZ066785.1	26	17.36	33.50
OZ066786.1	27	17.01	33
OZ066787.1	28	16.91	33
OZ066788.1	29	16.85	33.50
OZ066789.1	30	16.67	33.50
OZ066790.1	31	16.56	33.50
OZ066791.1	32	16.55	34
OZ066792.1	33	16.09	34
OZ066793.1	34	15.57	34
OZ066794.1	35	15.56	34
OZ066795.1	36	15.53	33.50
OZ066796.1	37	15.49	33
OZ066797.1	38	15.30	33.50
OZ066798.1	39	15.30	33.50
OZ066799.1	40	15.12	33.50
OZ066800.1	41	15	33.50
OZ066801.1	42	14.90	34
OZ066802.1	43	14.70	34
OZ066803.1	44	14.46	34
OZ066804.1	45	13.93	33.50
OZ066805.1	46	13.33	33.50
OZ066806.1	47	12.42	32
OZ066807.1	48	10.93	33.50

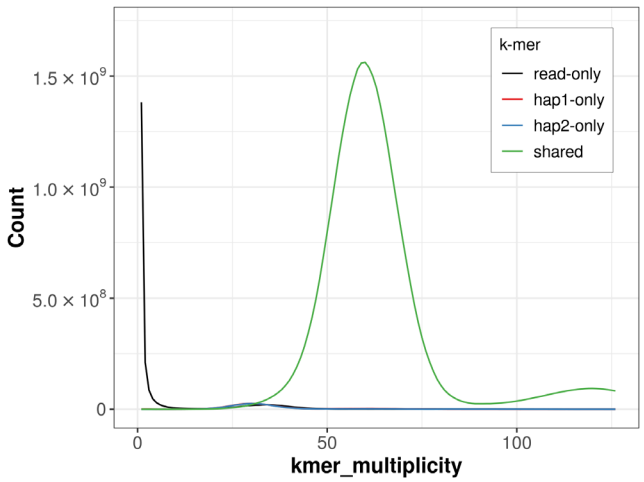

Figure 4. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers that appear in the reads but are not assembled. The green curve (the homozygous peak) corresponds to *k*-mers shared by both haplotypes and the red and blue curves (the heterozygous peaks) show *k*-mers found only in one of the haplotypes.

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

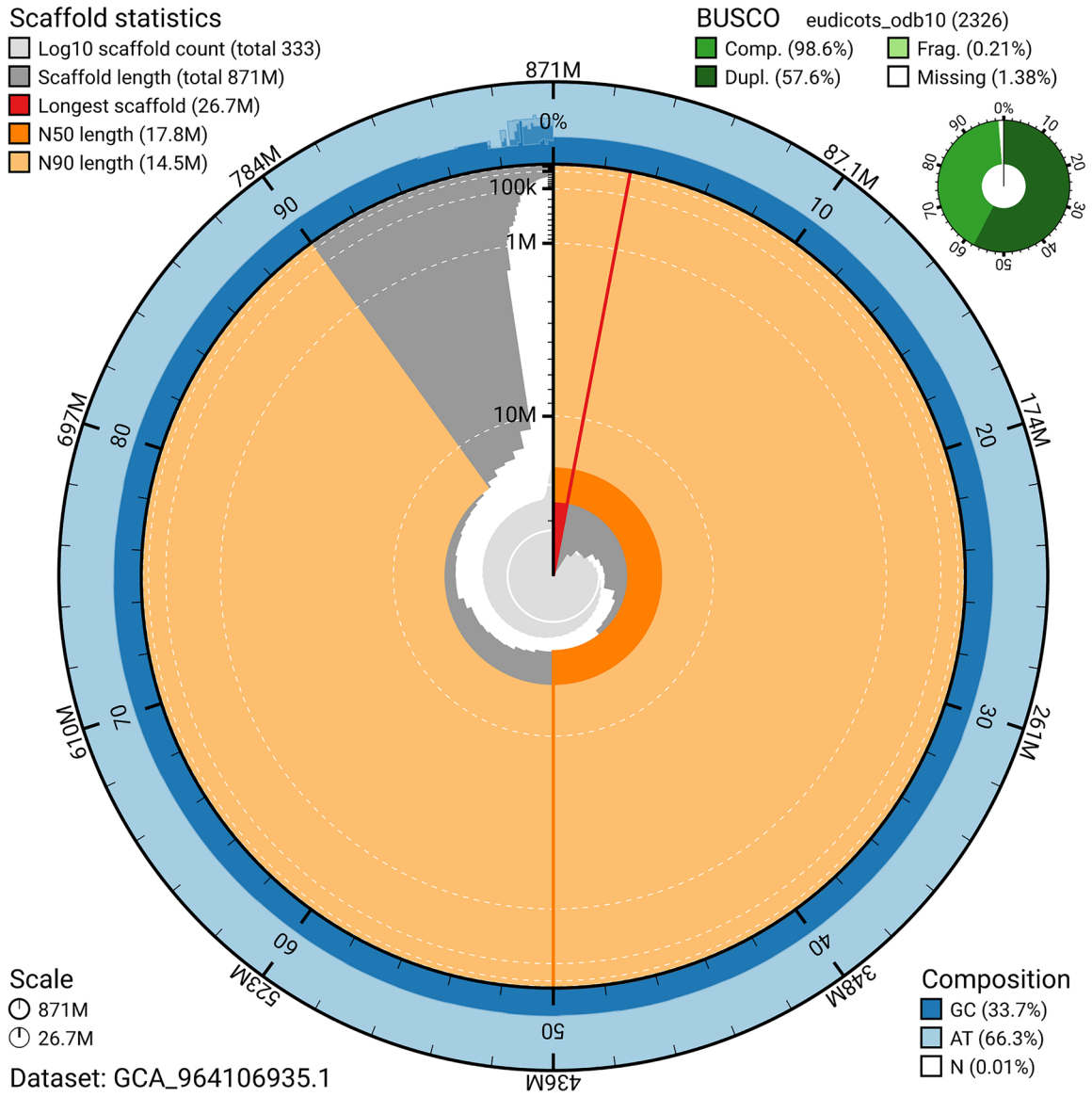

Figure 5. Assembly metrics for drHydVulg1.hap1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the circumference. The outermost 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. The longest scaffold is indicated by the red arc, and the deeper orange and pale orange arcs represent the N50 and N90 lengths. A 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 set is presented at the top right. An interactive version of this figure can be accessed on the BlobToolKit viewer.

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.

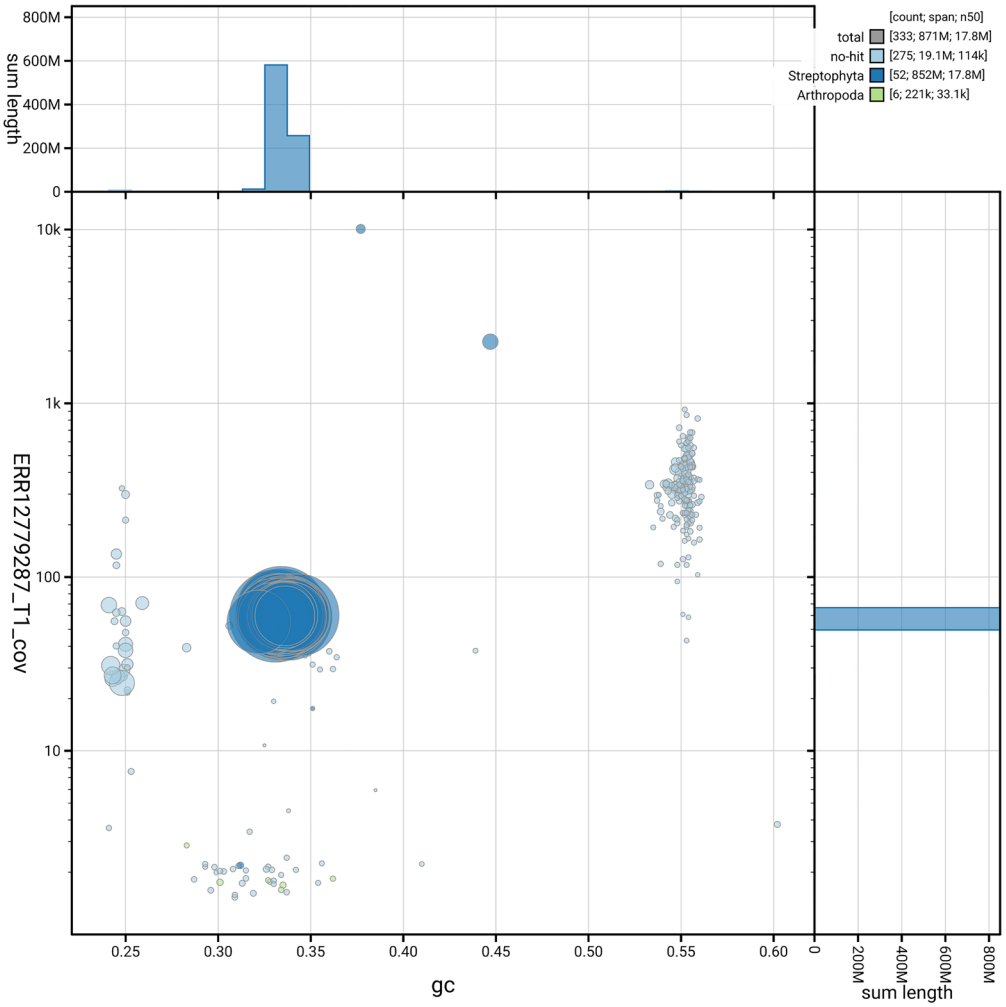

Figure 6. BlobToolKit GC-coverage plot for drHydVulg1.hap1.1. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes display the total length of sequences distributed across different levels of coverage and GC content. An interactive version of this figure is available on the BlobToolKit viewer.

Table 4. Earth Biogenome Project summary metrics for the *Hydrocotyle vulgaris* assembly.

Measure	Value	Benchmark
EBP summary (haplotype 1)	6.C.Q58	6.C.Q40
Contig N50 length	5.33 Mb	≥ 1 Mb
Scaffold N50 length	17.80 Mb	= chromosome N50
Consensus quality (QV)	Haplotype 1: 58.3; haplotype 2: 58.3; combined: 58.3	≥ 40
k-mer completeness	Haplotype 1: 96.17%; Haplotype 2: 96.07%; combined: 98.22%	≥ 95%
BUSCO	C:98.6% [S:41.1%; D:57.6%]; F:0.2%; M:1.2%; n:2 326	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	98.73%	≥ 90%

Data availability

European Nucleotide Archive: Hydrocotyle vulgaris. Accession number PRJEB74171. The genome sequence is released openly for reuse. The *Hydrocotyle vulgaris* genome sequencing initiative is part of the Darwin Tree of Life Project (PRJEB40665) and Sanger Institute Tree of Life Programme (PRJEB43745). All raw sequence data and the assembly have been deposited in INSDC databases. The genome will be annotated using available RNA-Seq data and presented through the Ensembl pipeline at the European Bioinformatics Institute. 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
BEDTools	2.30.0	https://github.com/arq5x/bedtools2
BLAST	2.14.0	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/
BlobToolKit	4.3.9	https://github.com/blobtoolkit/blobtoolkit
BUSCO	5.5.0	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
DIAMOND	2.1.8	https://github.com/bbuchfink/diamond
fasta_windows	0.2.4	https://github.com/tolkit/fasta_windows
FastK	1.1	https://github.com/thegenemyers/FASTK
GenomeScope2.0	2.0.1	https://github.com/tbenavi1/genomescope2.0
Gfastats	1.3.6	https://github.com/vgl-hub/gfastats
GoaT CLI	0.2.5	https://github.com/genomehubs/goat-cli
Hifiasm	0.19.8-r603	https://github.com/chhylp123/hifiasm
HiGlass	1.13.4	https://github.com/higlass/higlass
MerquryFK	1.1.2	https://github.com/thegenemyers/MERQURY.FK
Minimap2	2.24-r1122	https://github.com/lh3/minimap2
Oatk	0.9	https://github.com/c-zhou/oatk
MultiQC	1.14; 1.17 and 1.18	https://github.com/MultiQC/MultiQC
Nextflow	23.10.0	https://github.com/nextflow-io/nextflow
PretextSnapshot	N/A	https://github.com/sanger-tol/PretextSnapshot
PretextView	0.2.5	https://github.com/sanger-tol/PretextView
samtools	1.19.2	https://github.com/samtools/samtools
sanger-tol/ascc	0.1.0	https://github.com/sanger-tol/ascc
sanger-tol/blobtoolkit	0.6.0	https://github.com/sanger-tol/blobtoolkit
Seqtk	1.3	https://github.com/lh3/seqtk
Singularity	3.9.0	https://github.com/sylabs/singularity
TreeVal	1.2.0	https://github.com/sanger-tol/treeval
YaHS	1.2a.2	https://github.com/c-zhou/yahs

Author information

- Members of the Royal Botanic Garden Edinburgh 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

References

Al-Bermani AK, Al-Shammary KIA, Gornall RJ, *et al.*: **Contribution to a cytological catalogue of the British and Irish flora, 3.** *Watsonia.* 1993; **19**: 269–271.
Reference Source

Altschul SF, Gish W, Miller W, *et al.*: **Basic Local Alignment Search Tool.** *J Mol Biol.* 1990; **215**(3): 403–410.
PubMed Abstract | Publisher Full Text

Amadei L, Giordani A, Tomei PE: **Numeri cromosomici per la Flora Italiana: 882–888.** *Informatore Botanico Italia.* 1982; **14**: 243–47.
Reference Source

Bateman A, Martin MJ, Orchard S, *et al.*: **UniProt: the Universal Protein Knowledgebase in 2023.** *Nucleic Acids Res.* 2023; **51**(D1): D523–D531.
PubMed Abstract | Publisher Full Text | Free Full Text

Buchfink B, Reuter K, Drost HG: **Sensitive protein alignments at Tree-of-Life scale using DIAMOND.** *Nat Methods.* 2021; **18**(4): 366–368.
PubMed Abstract | Publisher Full Text | Free Full Text

Challis R, Kumar S, Sotero-Caio C, *et al.*: **Genomes on a Tree (GoaT): a versatile, scalable search engine for genomic and sequencing project metadata across the eukaryotic Tree of Life [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2023; **8**: 24.
PubMed Abstract | Publisher Full Text | Free 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

Cheng H, Jarvis ED, Fedrigo O, *et al.*: **Haplotype-resolved assembly of diploid genomes without parental data.** *Nat Biotechnol.* 2022; **40**(9): 1332–1335.
PubMed Abstract | Publisher Full Text | Free Full Text

Danecek P, Bonfield JK, Liddle J, *et al.*: **Twelve years of SAMtools and BCFtools.** *GigaScience.* 2021; **10**(2): giab008.
PubMed Abstract | Publisher Full Text | Free 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

Gadella TWJ, Kliphuis E: **Chromosome numbers of flowering plants in the Netherlands II.** *Proceedings of the Royal Netherlands Academy of Arts and Sciences, Series C: Biological and Medical Sciences.* 1966; **69**: 541–556.
Reference Source

Grüning B, Dale R, Sjödin A, *et al.*: **Bioconda: sustainable and comrehensive software distribution for the life sciences.** *Nat Methods.* 2018; **15**(7): 475–476.
PubMed Abstract | Publisher Full Text | Free Full Text

Hollingsworth PM, Gornall RJ, Bailey JP: **Contribution to a cytological catalogue of the British and Irish flora, 2.** *Watsonia.* 1992; **19**: 134–37.

Howard C, Denton A, Jackson B, *et al.*: **On the path to reference genomes for all biodiversity: lessons learned and laboratory protocols created in the Sanger Tree of Life core laboratory over the first 2000 species.** *bioRxiv.* 2025.
Publisher Full Text

Howe K, Chow W, Collins J, *et al.*: **Significantly improving the quality of genome assemblies through curation.** *GigaScience.* 2021; **10**(1): giaa153.
PubMed Abstract | Publisher Full Text | Free 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

Lawniczak MKN, Davey RP, Rajan J, *et al.*: **Specimen and sample metadata standards for biodiversity genomics: a proposal from the Darwin Tree of Life project [version 1; peer review: 2 approved with reservations].** *Wellcome Open Res.* 2022; **7**: 187.
Publisher Full Text

Li H: **Minimap2: pairwise alignment for nucleotide sequences.** *Bioinformatics.* 2018; **34**(18): 3094–3100.
PubMed Abstract | Publisher Full Text | Free Full Text

Liu R, Chen Q, Dong B, *et al.*: **Effects of vegetative propagule pressure on the establishment of an introduced clonal plant, *Hydrocotyle vulgaris.*** *Sci Rep.* 2014; **4**(1): 5507.
PubMed Abstract | Publisher Full Text | Free Full Text

Loureiro J, Rodriguez 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, Seppey 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

Miao LiHua ML, Ji MengCheng JM, Wang YingYing WY, *et al.*: **Study on invasion risk of *Hydrocotyle vulgaris* as an alien species in wetlands.** *J Zhejiang Univ Agr Life Sci.* 2011; **37**(4): 425–431.

Ni J, Sun SX, Zheng Y, *et al.*: **Removal of prometryn from hydroponic media using marsh pennywort (*Hydrocotyle vulgaris* L.).** *Int J Phytoremediation.* 2018; **20**(9): 909–913.
PubMed Abstract | Publisher Full Text

Obermayer R, Leitch IJ, Hanson L, *et al.*: **Nuclear DNA C-values in 30 species double the familial representation in pteridophytes.** *Ann Bot.* 2002; **90**(2): 209–217.
PubMed Abstract | Publisher Full Text | Free 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 (Clifton, N.J.).* New York, NY: Humana, 2021; **2222**: 325–361.
PubMed Abstract | Publisher Full Text

Plunkett GM, Wen J, Lowry II PP: **Infrafamilial classifications and characters in Araliaceae: insights from the phylogenetic analysis of nuclear (ITS) and plastid (*trn*L–*trn*F) sequence data.** *Plant Syst Evol.* 2004; **245**: 1–39.
Publisher Full Text

Queiros M: **Contribuicao ao conhecimento citotaxonomico das spermatophyta de Portugal. VII Umbelliferae.** *Anuário da Sociedade Broteriana.* 1972; **38**: 293–314.

Ranallo-Benavidez TR, Jaron KS, Schatz MC: **GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes.** *Nat Commun.* 2020; **11**(1): 1432.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

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 PD: **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**

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.*: **Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies.** *Genome Biol.* 2020; **21**(1): 245.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Schou JC, Moeslund B, Lansdown RV, *et al.*: **Aquatic plants of Northern and Central Europe including Britain and Ireland.** Oxford: Princeton University Press, 2023.
Publisher Full Text

Si C, Alpert P, Zhang JF, *et al.*: **Capacity for clonal integration in introduced versus native clones of the invasive plant *Hydrocotyle vulgaris*.** *Sci Total Environ.* 2020; **745**: 141056.
PubMed Abstract | **Publisher Full Text**

Stace CA, Thompson H, Stace M: **New flora of the British Isles.** C&M Floristics, 2019.

Stroh PA, Walker KJ, Humphrey TA, *et al.*: **Plant Atlas 2020: mapping changes in the distribution of the British and Irish Flora.** Durham UK; Oxford UK: Botanical Society of Britain and Ireland; Princeton University Press, 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: 2 approved].** *Wellcome Open Res.* 2024; **9**: 339.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Ureta R, Mejico S, Maranan Y: **Free radical scavenging activity and antioxidants of *Hydrocotyle vulgaris* L. (Pennywort): baseline study in developing biocosmetic-antidote for pathological aging.** *Int J Pharm Phytochem Ethnomed.* 2018; **10**(1): 2297–6922.
Reference Source

Vafaei F, Movafeghi A, Khataee AR, *et al.*: **Potential of *Hydrocotyle vulgaris* for phytoremediation of a textile dye: inducing antioxidant response in roots and leaves.** *Ecotoxicol Environ Saf.* 2013; **93**: 128–34.
PubMed Abstract | **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 MZ, Bu XQ, Li L, *et al.*: **Constraints on the evolution of phenotypic plasticity in the clonal plant *Hydrocotyle vulgaris*.** *J Evol Biol.* 2018; **31**(7): 1006–1017.
PubMed Abstract | **Publisher Full Text**

Yi T, Lowry PP, Plunkett GM, *et al.*: **Chromosomal evolution in Araliaceae and close relatives.** *Taxon.* 2004; **53**(4): 987–1005.
Publisher Full Text

Zhao BN, Chen ZH, Liu ZH, *et al.*: **Clonal integration facilitates the expansion of *Hydrocotyle vulgaris* from a limited space to a larger area.** *Plant Spec Biol.* 2024; **39**(5): 286–96.
Publisher Full Text

Zhou C, Brown M, Blaxter M, *et al.*: **Oatk: a de novo assembly tool for complex plant organelle genomes.** *bioRxiv.* 2024.
Publisher Full Text

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 09 September 2025

https://doi.org/10.21956/wellcomeopenres.27091.r130599

✔ **Bimal K Chetri**
Indian Institute of Technology, Guwahati, Assam, India

The assembly is of high quality and clearly valuable, but the manuscript would benefit from an explicit clarification of how the haploid genome size estimate (~467 Mb) and the flow cytometry estimate (~1,050 Mb) relate to the final haplotype assemblies (~870 Mb and ~863 Mb). Since *Hydrocotyle vulgaris* is an octoploid (2n = 96), a short statement connecting polyploidy to these size differences would make the Data Note more transparent and easier for downstream users to interpret.

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

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 09 September 2025

https://doi.org/10.21956/wellcomeopenres.27091.r130600

Peng-Fei Ma
Chinese Academy of Sciences, Kunming, China

This study reported a high-quality genome assembly of a small aquatic perennial, Hydrocotyle vulgaris. This species is native to Europe and Northern Africa, and particularly widespread in Britain and Ireland. It even becomes an invasive plant in China. The genome sequences reported here may be helpful for understanding the clonal reproduction of this species and potentially control of it. I have two questions. In the results, a heterozygosity rate of 37.71% was reported and please check this number as it is so high. The other one is that the assembled quality of Haplotype 2 seems better than Haplotype 1, and why Haplotype 1 was used for Hi-C scaffolding?

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

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 30 August 2025

https://doi.org/10.21956/wellcomeopenres.27091.r128829

Sunil Kumar Sahu
BGI Research, Shenzhen, China

This manuscript presents a high-quality, chromosome-scale genome assembly for *Potentilla sterilis* (Barren Strawberry), generated as part of Darwin Tree of Life (DToL) project. The assembly is robust, utilizing modern long-read and Hi-C scaffolding technologies, and meets high-quality standards. The resource will be a valuable addition to the genomic resources available for the Rosaceae family and for studying polyploidy in plants. The manuscript is generally well-written and follows a standard format for data notes.

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 and evolution

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.