DATA NOTE

The genome sequence of the Roundleaf geranium, *Geranium rotundifolium* L. (Geraniales: Geraniaceae)

[version 1; peer review: 2 approved]

Maarten J. M. Christenhusz[1,2], Alex D. Twyford[3,4],
Royal Botanic Gardens Kew Genome Acquisition Lab,
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 Gardens Kew, Richmond, England, UK
[2]Curtin University, Perth, Western Australia, Australia
[3]Royal Botanic Garden Edinburgh, Edinburgh, Scotland, UK
[4]The University of Edinburgh, Edinburgh, Scotland, UK

v1 **First published:** 18 Mar 2026, **11**:168
https://doi.org/10.12688/wellcomeopenres.26116.1
Latest published: 18 Mar 2026, **11**:168
https://doi.org/10.12688/wellcomeopenres.26116.1

Abstract

We present a genome assembly of *Geranium rotundifolium* (Roundleaf geranium; Streptophyta; Magnoliopsida; Geraniales; Geraniaceae). The genome sequence has a total length of 497.00 megabases. Most of the assembly (97.57%) is scaffolded into 13 chromosomal pseudomolecules. The mitochondrial sequence has a length of 335.81 kilobases and the plastid genome assembly has a length of 169.49 kilobases. Gene annotation of this assembly on Ensembl identified 29 331 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

Geranium rotundifolium; Roundleaf geranium; genome sequence; chromosomal; Geraniales

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ✓ ✓

	1	2
version 1 18 Mar 2026	✓ view	✓ view

1. **Nils Stein**, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Seeland, Germany

2. **Margaret E Staton**, The University of Tennessee, Knoxville, USA

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; **Twyford AD**: 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: Christenhusz MJM, Twyford AD, Royal Botanic Gardens Kew Genome Acquisition Lab *et al.* **The genome sequence of the Roundleaf geranium, *Geranium rotundifolium* L. (Geraniales: Geraniaceae) [version 1; peer review: 2 approved]** Wellcome Open Research 2026, **11**:168 https://doi.org/10.12688/wellcomeopenres.26116.1

First published: 18 Mar 2026, **11**:168 https://doi.org/10.12688/wellcomeopenres.26116.1

Species taxonomy

Eukaryota; Viridiplantae; Streptophyta; Streptophytina; Embryophyta; Tracheophyta; Euphyllophyta; Spermatophyta; Magnoliopsida; Mesangiospermae; eudicotyledons; Gunneridae; Pentapetalae; rosids; malvids; Geraniales; Geraniaceae; *Geranium*; *Geranium rotundifolium* L. (NCBI:txid379955).

Background

The round-leaved crane's-bill, *G. rotundifolium* L. (Figure 1), is an annual herbaceous plant species that possesses abundant glandular hairs, has small pink flowers, and leaves that are kidney shaped in outline with relatively shallow lobes (Stace *et al.*, 2019). It is locally abundant in the south of England and south Wales, while it is virtually absent from northern Scotland and much of Ireland (Stroh *et al.*, 2023). However, the species has experienced a marked northwards expansion in the past ~50 years, perhaps associated with climate change. Outside of Britain and Ireland, its native range includes much of Europe, North Africa and Western Asia, and it has naturalised in many locations including in North America, South America and South Africa (Royal Botanic Gardens, Kew, 2025). *G. rotundifolium* is typically found growing in walls and dry roadside banks, as well as a weed in disturbed sites and waste ground.

The species belongs to the Geraniaceae, and is a diploid with $2n = 26$ (Henniges *et al.*, 2022). Although numerous genomes exist for the family Geraniaceae, this assembly provides the first chromosomally complete sequence for *G. rotundifolium*, enabling comparative analyses. This genome was assembled using the Tree of Life pipeline from a specimen collected in Teddington Lock, Richmond, Surrey, UK.

Methods

Sample acquisition, flow cytometry and DNA barcoding

A specimen of *G. rotundifolium* (specimen ID KDTOL10167, ToLID drGerRotu1) was used for genome sequencing. It was collected from Teddington Lock, Richmond, Surrey, UK (latitude 51.4303, longitude −0.319) on 2021-04-19. The specimen was collected and identified by Maarten J. M. Christenhusz (Royal Botanic Gardens Kew). The same specimen was used for RNA sequencing. 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. 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

Figure 1. Photograph of *Geranium rotundifolium* by Alvesgaspar.

was *Petroselinum crispum* (Mill) Nyman ex A.E.Hill '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 Life Core Laboratory are available on protocols.io (Howard *et al.*, 2025). The drGerRotu1 sample was weighed and triaged to determine the appropriate extraction protocol. Tissue from the leaf was homogenised by cryogenic disruption using the Covaris cryoPREP® Automated Dry Pulverizer. HMW DNA was extracted using the Plant Organic Extraction protocol. We used centrifuge-mediated fragmentation to produce DNA fragments in the 8–10 kb range, following the Covaris g-TUBE protocol for ultra-low input (ULI). 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 1.95 ng/μL and a yield of 760.50 ng.

RNA was extracted from leaf tissue of drGerRotu1 in the Tree of Life Laboratory at the WSI using the RNA Extraction: Automated MagMax™ *mir*Vana protocol. 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.

PacBio HiFi library preparation and sequencing

Library preparation and sequencing were performed at the WSI Scientific Operations core. Prior to library preparation, the DNA was fragmented to ~10 kb. Ultra-low-input (ULI) libraries were prepared using the PacBio SMRTbell® Express Template Prep Kit 2.0 and gDNA Sample Amplification Kit. Samples were normalised to 20 ng DNA. Single-strand overhang removal, DNA damage repair, and end-repair/A-tailing were performed according to the manufacturer's instructions, followed by adapter ligation. A 0.85× pre-PCR clean-up was carried out with Promega ProNex beads.

The DNA was evenly divided into two aliquots for dual PCR (reactions A and B), both following the manufacturer's protocol. A 0.85× post-PCR clean-up was performed with ProNex beads. DNA concentration was measured using a Qubit Fluorometer v4.0 (Thermo Fisher Scientific) with the Qubit HS Assay Kit, and fragment size was assessed on an Agilent Femto Pulse Automated Pulsed Field CE Instrument (Agilent Technologies) using the gDNA 55 kb BAC analysis kit. PCR reactions A and B were then pooled, ensuring a total mass of ≥500 ng in 47.4 μl.

The pooled sample underwent another round of DNA damage repair, end-repair/A-tailing, and hairpin adapter ligation. A 1× clean-up was performed with ProNex beads, followed by DNA quantification using the Qubit and fragment size analysis using the Agilent Femto Pulse. Size selection was performed on the Sage Sciences PippinHT system, with target fragment size determined by Femto Pulse analysis (typically 4–9 kb). Size-selected libraries were cleaned with 1.0× ProNex beads and normalised to 2 nM before sequencing.

The sample was sequenced using the Sequel IIe system (Pacific Biosciences, California, USA). The concentration of the library loaded onto the Sequel IIe was in the range 40–135 pM. The SMRT link software, a PacBio web-based end-to-end workflow manager, was used to set-up and monitor the run, and to perform primary and secondary analysis of the data upon completion.

Hi-C

Sample preparation and crosslinking

Hi-C data were generated from the leaf tissue of drGerRotu1 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 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. Sequencing was performed using paired-end 150 bp reads on the Illumina NovaSeq 6000.

RNA library preparation and sequencing

Libraries were prepared using the NEBNext® Ultra™ II Directional RNA Library Prep Kit for Illumina (New England Biolabs), following the manufacturer's instructions. Poly(A) mRNA in the total RNA solution was isolated using oligo (dT) beads, converted to cDNA, and uniquely indexed; 14 PCR cycles were performed. Libraries were size-selected to produce fragments between 100–300 bp. Libraries were quantified, normalised, pooled to a final concentration of 2.8 nM, and diluted to 150 pM for loading. Sequencing was carried out on the Illumina NovaSeq X to generate 150-bp 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. 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 (Cheng *et al.*, 2021) with the --primary option. The Hi-C reads (Rao *et al.*, 2014) were mapped to the primary contigs using bwa-mem2 (Vasimuddin *et al.*, 2019), and the contigs were scaffolded in YaHS (Zhou *et al.*, 2023) with 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 [zhou2025Oatk].

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 eight breaks, 14 joins, and removal of one haplotypic duplication. This reduced the scaffold count by 1.0%. 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 the primary and alternate 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 from NCBI Taxonomy (Schoch *et al.*, 2020). 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 *G. rotundifolium* was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating 44.27 Gb (gigabases) from 4.79 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 503.13 Mb, with a heterozygosity of 0.33% and repeat content of 37.85% (Figure 2). Using flow cytometry, the genome size (1C-value) of the sample was estimated to be 0.65 pg, equivalent to 630.00 Mb. These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 80× coverage. Hi-C sequencing produced 106.75 Gb from 706.93 million reads, which were used to scaffold the assembly. RNA sequencing data were also generated and are available in public sequence repositories. Table 1 summarises the specimen and sequencing details.

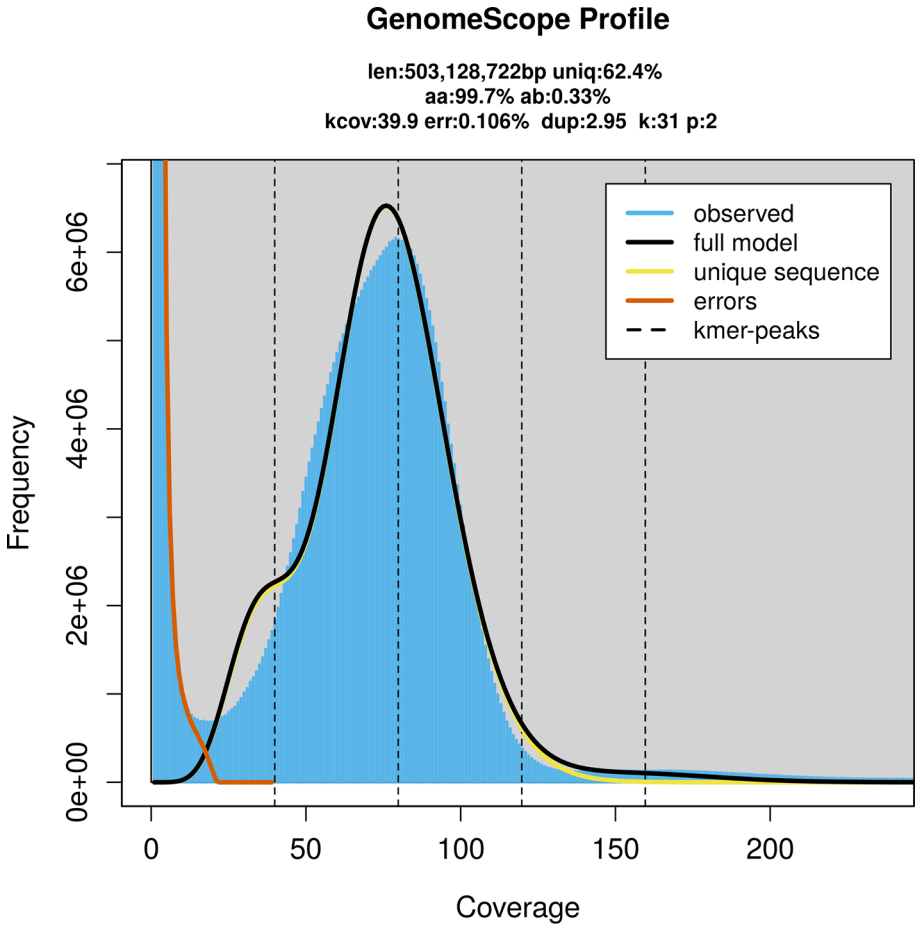

Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2. The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

Table 1. **Specimen and sequencing data for BioProject PRJEB69504.**

Platform	PacBio HiFi	Hi-C	RNA-seq
ToLID	drGerRotu1	drGerRotu1	drGerRotu1
Specimen ID	KDTOL10167	KDTOL10167	KDTOL10167
BioSample (source individual)	SAMEA9143062	SAMEA9143062	SAMEA9143062
BioSample (tissue)	SAMEA9143830	SAMEA9143830	SAMEA9143830
Tissue	leaf	leaf	leaf
Instrument	Sequel IIe	Illumina NovaSeq 6000	Illumina NovaSeq X
Run accessions	ERR12303938; ERR12303939	ERR12318584	ERR13493917
Read count total	4.79 million	706.93 million	97.54 million
Base count total	44.27 Gb	106.75 Gb	14.73 Gb

Table 2. **Genome assembly statistics.**

Assembly name	drGerRotu1.1
Assembly accession	GCA_963920655.1
Alternate haplotype accession	GCA_963920715.1
Assembly level	chromosome
Span (Mb)	497.00
Number of chromosomes	13
Number of contigs	1 194
Contig N50	1.1 Mb
Number of scaffolds	496
Scaffold N50	39.7 Mb
Organelles	Mitochondrion: 335.81 kb; Plastid: 169.49 kb

Assembly statistics

The primary haplotype was assembled, and contigs corresponding to an alternate haplotype were also deposited in INSDC databases. The final assembly has a total length of 497.00 Mb in 496 scaffolds, with 698 gaps, and a scaffold N50 of 39.7 Mb (Table 2).

Most of the assembly sequence (97.57%) was assigned to 13 chromosomal-level scaffolds. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 3; Table 3).

The mitochondrial genome (length 335.81 kb, OY987324.1) and plastid genome (length 169.49 kb, OY987325.1) were also assembled. These sequences are included as contigs in the multifasta file of the genome submission and as standalone records.

Assembly quality metrics

The combined primary and alternate assemblies achieve an estimated QV of 53.4. The *k*-mer completeness is 99.20% for the primary assembly, 5.12% for the alternate haplotype, and 99.23% for the combined assemblies (Figure 4).

BUSCO v.5.5.0 analysis using the eudicots_odb10 reference set (n = 2 326) identified 94.8% of the expected gene set (single = 80.0%, duplicated = 14.8%). The snail plot in Figure 5 summarises the scaffold length distribution and other assembly statistics for the primary assembly. The blob plot in Figure 6 shows the distribution of scaffolds by GC proportion and coverage.

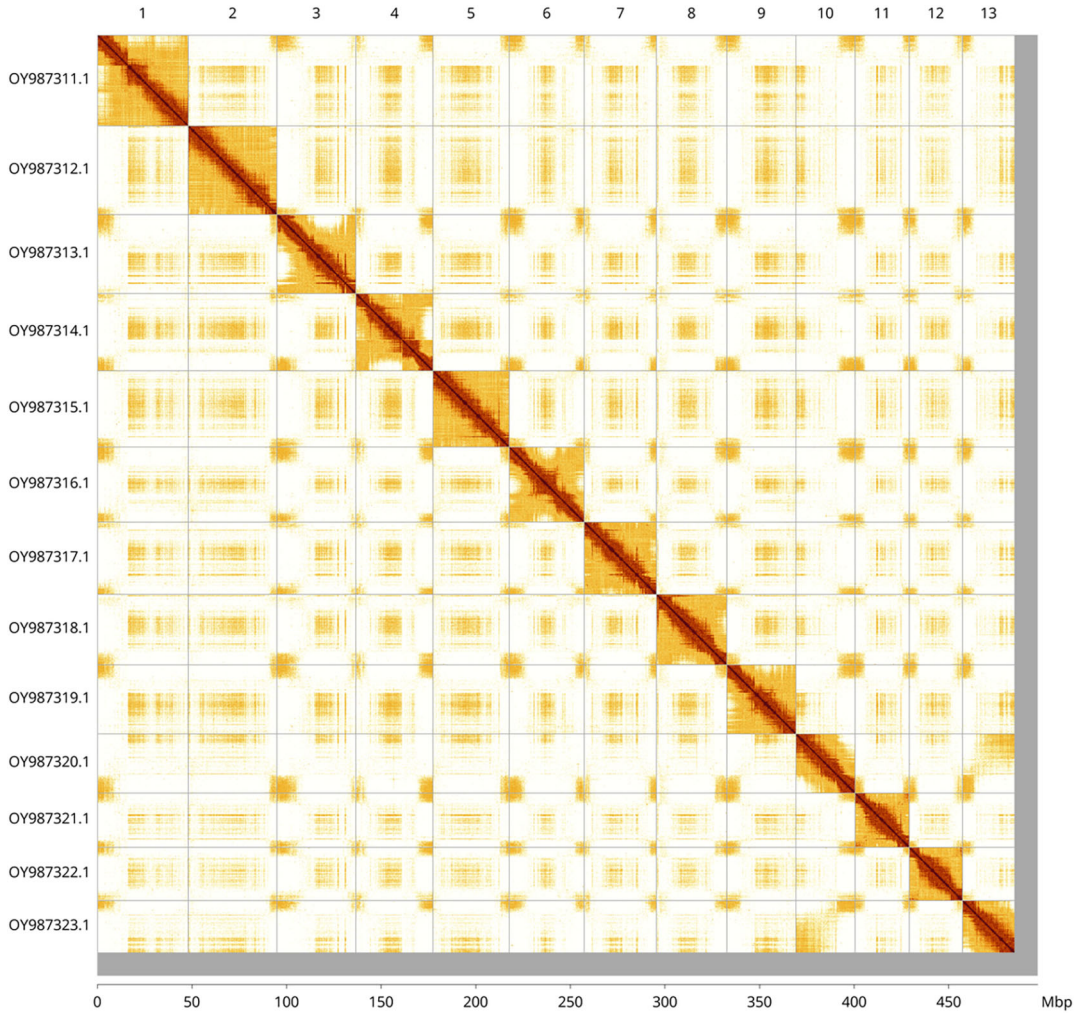

Figure 3. Hi-C contact map of the *Geranium rotundifolium* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes, with a megabase scale shown below. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in the primary genome assembly of *Geranium rotundifolium* drGerRotu1.

INSDC accession	Molecule	Length (Mb)	GC%
OY987311.1	1	48.13	39
OY987312.1	2	46.77	39.50
OY987313.1	3	41.76	38
OY987314.1	4	40.76	38.50
OY987315.1	5	40.25	39.50
OY987316.1	6	39.70	38
OY987317.1	7	38.16	38.50
OY987318.1	8	37.16	39

Table 3. *Continued*

INSDC accession	Molecule	Length (Mb)	GC%
OY987319.1	9	36.52	39
OY987320.1	10	31.32	38
OY987321.1	11	28.59	38
OY987322.1	12	28.13	38.50
OY987323.1	13	27.66	38.50

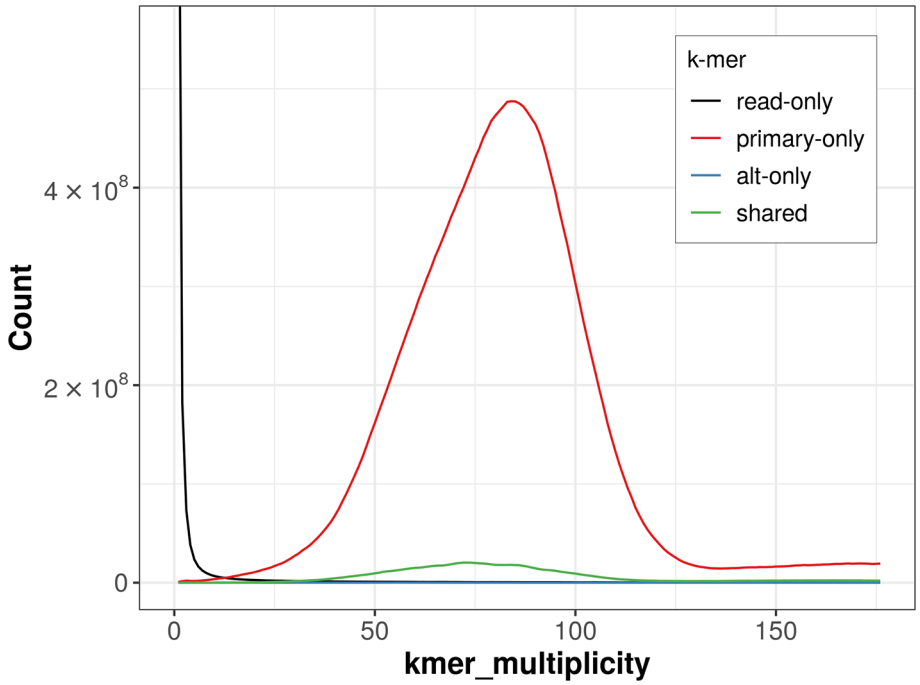

Figure 4. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers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

Table 4 lists the assembly metric benchmarks adapted from Rhie *et al.* (2021) and the Earth BioGenome Project Report on Assembly Standards September 2024. The EBP metric calculated for the primary assembly is **6.C.Q54**, meeting the recommended reference standard.

Genome annotation report

The *G. rotundifolium* genome assembly (GCA_963920655.1) was annotated by Ensembl at the European Bioinformatics Institute (EBI). This annotation includes 48 755 transcribed mRNAs from 29 331 protein-coding and 9 228 non-coding genes. The average transcript length is 2 941.21 bp, with an average of 1.26 coding transcripts per gene and 5.03 exons per transcript. For further information about the annotation, please refer to the annotation page on Ensembl.

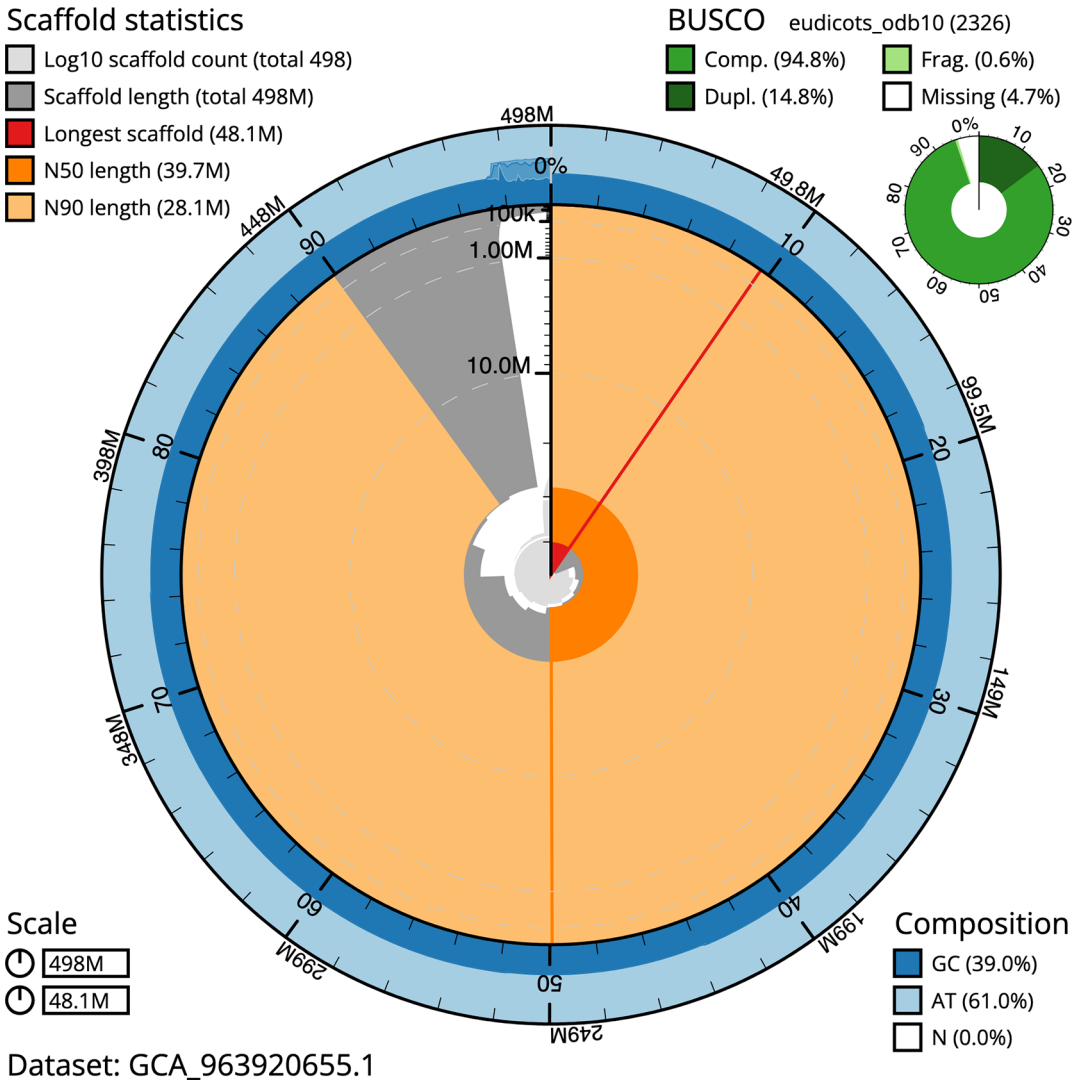

Figure 5. Assembly metrics for drGerRotu1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1 000 bins around the circumference. The outermost blue 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 eudicots_odb10 set is presented at the top right. An interactive version of this figure can be accessed on the BlobToolKit viewer.

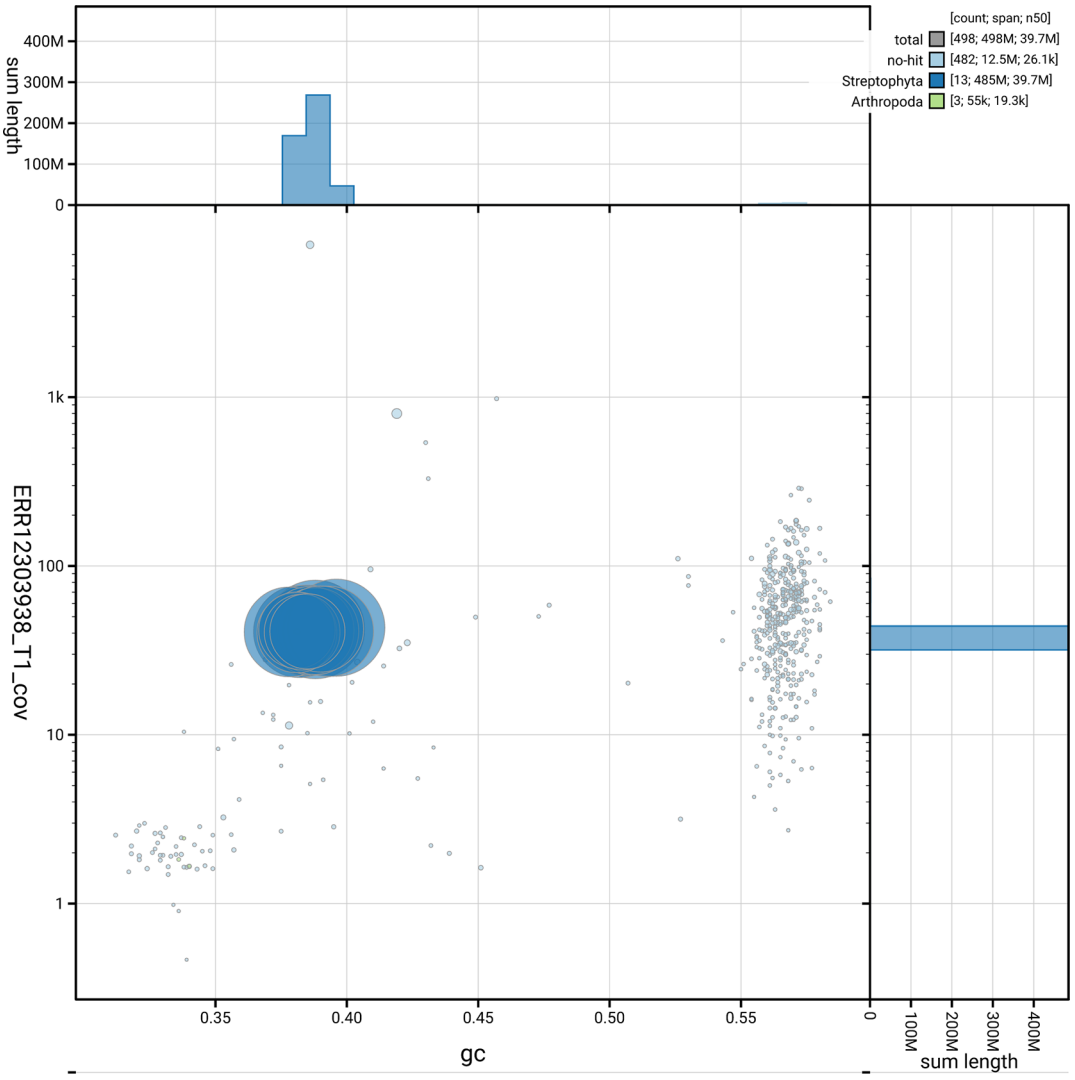

Figure 6. BlobToolKit blob plot for drGerRotu1.1. The plot shows base coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes display the total 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 *Geranium rotundifolium* assembly.

Measure	Value	Benchmark
EBP summary (primary)	6.C.Q54	6.C.Q40
Contig N50 length	1.10 Mb	≥ 1 Mb
Scaffold N50 length	39.70 Mb	= chromosome N50
Consensus quality (QV)	Primary: 54.0; alternate: 48.5; combined: 53.4	≥ 40
k-mer completeness	Primary: 99.20%; alternate: 5.12%; combined: 99.23%	≥ 95%
BUSCO	C:100.0% [S:60.0%, D:40.0%], F:0.0%, M:0.0%, n:255	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	97.57%	≥ 90%

Notes: The EBP summary uses log10(Contig N50); chromosome-level (C) or log10(Scaffold N50); Q (Merqury QV). BUSCO: C = complete; S = single-copy; D = duplicated; F = fragmented; M = missing; n = orthologues.

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.

Data availability

European Nucleotide Archive: *G. rotundifolium* (round-leaved crane's-bill). Accession number PRJEB69504. The genome sequence is released openly for reuse. The *G. rotundifolium* 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 Tables 1 and 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

Table 5. *Continued*

Software	Version	Source
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	0.19.5-r587	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	23.04.1	https://github.com/nextflow-io/nextflow
PretextSnapshot	0.0.5	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	0.4.0	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

References

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

Bateman A, Martin M-J, 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 H-G: **Sensitive protein alignments at tree-of-life scale using DIAMOND.** *Nat. Methods.* 2021; **18**(4): 366–368.
Publisher Full Text

Challis R, Richards E, Rajan J, *et al.*: **BlobToolKit – interactive quality assessment of genome assemblies.** *G3 Genes|Genomes|Genetics.* 2020; **10**(4): 1361–1374.
Publisher 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.
Publisher Full Text

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

Henniges MC, Powell RF, Mian S, *et al.*: **A taxonomic, genetic and ecological data resource for the vascular plants of Britain and Ireland.** *Sci. Data.* 2022; **9**: 1–8.
PubMed Abstract | Publisher Full Text | Free Full Text

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).
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.
Publisher Full Text

Kurtzer GM, Sochat V, Bauer MW: **Singularity: Scientific containers for mobility of compute.** *PLoS One.* 2017; **12**(5): e0177459.
Publisher 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.** *Wellcome Open Res.* 2022; **7**: 187.
Publisher Full Text

Li H: **Minimap2: Pairwise alignment for nucleotide sequences.** *Bioinformatics.* 2018; **34**(18): 3094–3100.
Publisher Full Text

Loureiro J, Rodriguez E, Doležel 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.
Publisher Full Text

Merkel D: **Docker: Lightweight Linux containers for consistent development and deployment.** *Linux J.* 2014; **2014**(239).
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.**Besse P, editor. *Methods in Molecular Biology.* New York, NY: 2021; Vol. **2222**: 325–61.
PubMed Abstract

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.
Publisher 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.
Publisher Full Text

Ratnasingham S, Hebert PDN: **BOLD: The Barcode of Life Data System.** *Mol. Ecol. Notes.* 2007; **7**(3): 355–364.
PubMed Abstract | **Publisher Full Text** | **Free Full Text** | **Reference Source**

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.
Publisher 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).
Publisher Full Text

Royal Botanic Gardens, Kew: **POWO: Plants of the World Online.** 2025.
Reference Source

Schoch CL, Ciufo S, Domrachev M, *et al.*: **NCBI taxonomy: A comprehensive update on curation, resources and tools.** *Database.* 2020; **2020**: baa062.
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.

Twyford AD, Beasley J, Barnes I, *et al.*: **A DNA barcoding framework for taxonomic verification in the Darwin Tree of Life Project.** *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.** *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS).* IEEE; 2019; 314–24.
Publisher Full Text

Zhou C, McCarthy SA, Durbin R: **YaHS: Yet another Hi-C scaffolding tool.** *Bioinformatics.* 2023; **39**(1).
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 1

Reviewer Report 20 April 2026

https://doi.org/10.21956/wellcomeopenres.28759.r152433

✔ **Margaret E Staton**
The University of Tennessee, Knoxville, Tennessee, USA

The authors present a chromosome scale haploid reference genome for Geranium rotundifolium. The genome used HiFi and HiC to yield a high quality final product.

The authors might comment on why they opted to use hifiasm with the -primary flag instead of haplotype resolved assemblies, which are usually possible when hi-C data is included.

Its unclear if the gene annotation utilized the RNASeq data or if the genes were functionally annotated. There's also no information on the repetitive content. The genome was annotated by the Ensembl pipeline, but some results on these points would be a welcome addition.

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: Computational 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 06 April 2026

https://doi.org/10.21956/wellcomeopenres.28759.r152432

Nils Stein
Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Seeland, Germany

The authors present a de novo nuclear and plastid genome assembly of G. rotundifolium. The species was selected as being endemic to UK and thus became a selected species within the DToL initiative. The study follows transparently the established standardized procedures and all assembly metrics and stats are presented and benchmarked against the Earth Biogenome Project standards.

I am reviewing the first time for this format but I am absolutely thrilled - this is, to my opinion the way genome sequences should be presented and made public rapidly in order to generate impact in the science community comments:

1) HiC interaction plot: I understand that the effort is limited that can be spent in a high throughput genome sequencing effort for individual curation steps. Individual chromosomes still have obvious traces of orientation/assembly artifacts. It also looks as if chromosomes are presented in random orientation. There is a convention that chromosomes shall be presented from short to long arm. It looks as if this species has a mix of acrocentric, telocentric and metacentric chromosomes. If there are any cytological data the chromosome order could be referred to this would add value.

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 Genetics, Plant Genomics, Plant Genetic Resources

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.