DATA NOTE

The genome sequence of the sweet violet, *Viola odorata* L. (Malpighiales: Violaceae)

[version 1; peer review: 3 approved]

Maarten J. M. Christenhusz[1,2], Michael F. Fay[1,3], Ilia J. Leitch[1],
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]The University of Western Australia, Crawley, Western Australia, Australia

v1 **First published:** 18 Mar 2026, **11**:177
https://doi.org/10.12688/wellcomeopenres.26151.1
Latest published: 18 Mar 2026, **11**:177
https://doi.org/10.12688/wellcomeopenres.26151.1

Abstract

We present a genome assembly of *Viola odorata* (Sweet violet; Streptophyta; Magnoliopsida; Malpighiales; Violaceae). The genome sequence has a total length of 698.62 megabases. Most of the assembly (99.9%) is scaffolded into 10 chromosomal pseudomolecules. The mitochondrial sequence has a length of 482.11 kilobases and the plastid genome assembly has a length of 158.28 kilobases. Gene annotation of this assembly on Ensembl identified 34 902 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

Viola odorata; Sweet violet; genome sequence; chromosomal; Malpighiales

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ✓ ✓ ✓

	1	2	3
version 1 18 Mar 2026	✓ view	✓ view	✓ view

1. **Bruno Contreras-Moreira**, Estacion Experimental de Aula Dei-CSIC (Ringgold ID: 54439), Zaragoza, Spain
2. **Stalin Nithaniyal**, Western Regional Centre, Pune, India
3. **Zoé Postel**, Stockholm University, Stockholm, Sweden

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

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

Author roles: Christenhusz MJM: Investigation, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Fay MF**: Writing – Review & Editing; **Leitch IJ**: 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).
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, Fay MF, Leitch IJ *et al.* **The genome sequence of the sweet violet, *Viola odorata* L. (Malpighiales: Violaceae) [version 1; peer review: 3 approved]** Wellcome Open Research 2026, **11**:177 https://doi.org/10.12688/wellcomeopenres.26151.1

First published: 18 Mar 2026, **11**:177 https://doi.org/10.12688/wellcomeopenres.26151.1

Species taxonomy

Eukaryota; Viridiplantae; Streptophyta; Streptophytina; Embryophyta; Tracheophyta; Euphyllophyta; Spermatophyta; Magnoliopsida; Mesangiospermae; eudicotyledons; Gunneridae; Pentapetalae; rosids; fabids; Malpighiales; Violaceae; *Viola*; *Viola* subgen. *Viola*; *Viola* sect. *Viola*; *Viola* subsect. *Viola*; *Viola odorata* L. (NCBI:txid97441).

Background

Sweet violet, *V. odorata* (Violaceae) is a perennial, creeping herb with heart-shaped leaves, and in early spring, flowers with violet, purple or white petals. The flowers are spurred and highly scented, and pollinated by solitary bees. Cleistogamous, self-pollinating flowers are also often formed. Plants usually grow in shaded places, under hedges, in woodland margins or in clearings. It is widespread across Europe, western Asia and Northwest Africa, and is naturalised in the Nordic countries, North and South America, South and East Asia, Australia and New Zealand (POWO, 2026). It can be found in most of Britain and Ireland, but it is very local in Scotland, Ireland, West Wales and the Channel Islands. Due to its popularity, some populations may be the result of garden escapes (Stace, 2019).

Because of its early flowering and sweet and attractive scent, special powers have been attributed to sweet violets since ancient times. In Greek mythology, the flower is known as *ion porfuroun*, and it was the symbol of Persephone, the goddess of the underworld, because it flowered so early that Persephone would not yet have emerged from the nether realm. Its violet flowers were a symbol of purity and virginity, of faithfulness and durability, and a sign of perpetual love. In the latter half of the 19th century it became popular for men to wear small bouquets of violets. For instance, Oscar Wilde is depicted wearing a small posy on a string attached to his belt (De Cleene & Lejeune, 2000). As an emblem, *V. odorata* was also chosen for the imperial Napoleonic party (1769–1821), but during the Restoration this symbol was outlawed (De Cleene & Lejeune, 2000).

V. odorata is probably the most economically important species of violet (Wyse Jackson, 2014). It has been used to make perfume for over 2000 years, and it is grown commercially in southern France for the production of the essential oils for this purpose. It is also used as a skin cleanser and in soaps and creams, not only for its scent and colour, but also for its medicinal properties. Extracts of violets are used in syrups, liqueurs, confectionery and cakes, especially in England and France (De Cleene & Lejeune, 2000).

Fothergill (1944) reported a diploid complement of $2n = 20$ across several named varieties of *V. odorata*, with a karyotype formula of 2(B, 9C) based on four size classes (A very long, B long, C medium, D small). He noted that the chromosomes are generally similar to those of *V. hirta*, but that *V. odorata* consistently includes a pair of long B-type chromosomes that are absent in *V. hirta*, whereas *V. hirta* has a pair of D-type chromosomes absent in *V. odorata.*

As part of the Darwin Tree of Life Project, a collaborative effort to sequence all named eukaryotic species in the Atlantic Archipelago of Britain and Ireland, we sequenced the genome of the sweet violet, *V. odorata* L. Here we present a chromosome-level genome sequence based on a wild specimen from Petersham Common, Surrey, UK (Figure 1).

Methods

Sample acquisition, flow cytometry and DNA barcoding

A specimen of *V. odorata* (specimen ID KDTOL10129, ToLID ddVioOdor1; Figure 1) was used for genome sequencing. It was collected from Richmond, Surrey, United Kingdom (latitude 51.4472, longitude −0.298) on 2021-03-09. The specimen was collected and identified by Maarten J. M. Christenhusz (Royal Botanic Gardens Kew). The herbarium voucher associated with the sequenced plant is K001400828 and is deposited in the herbarium of RBG Kew (K).

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.

Figure 1. Photograph of the *Viola odorata* (ddVioOdor1) specimen from which samples were taken for genome sequencing.

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 ddVioOdor1 sample was weighed and triaged to determine the appropriate extraction protocol. Leaf tissue was homogenised by cryogenic disruption using the Covaris cryoPREP® Automated Dry Pulverizer. Two different extractions of HMW were performed, one using the Automated Plant MagAttract v2 protocol, and the other using the 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 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 17.1 ng/μL and a yield of 2 223.00 ng.

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 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 ddVioOdor1 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.

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 MerquryFK (Rhie *et al.*, 2020).

The organelle genomes were assembled using MitoHiFi (Uliano-Silva *et al.*, 2023) and OATK (Zhou *et al.*, 2025).

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 17 breaks and 95 joins. This reduced the scaffold count by 73.8% and increased the scaffold N50 by 6.7%. The curation process is described at https://gitlab.com/wtsi-grit/rapid-curation. PretextSnapshot was used to generate a Hi-C contact map of the final assembly.

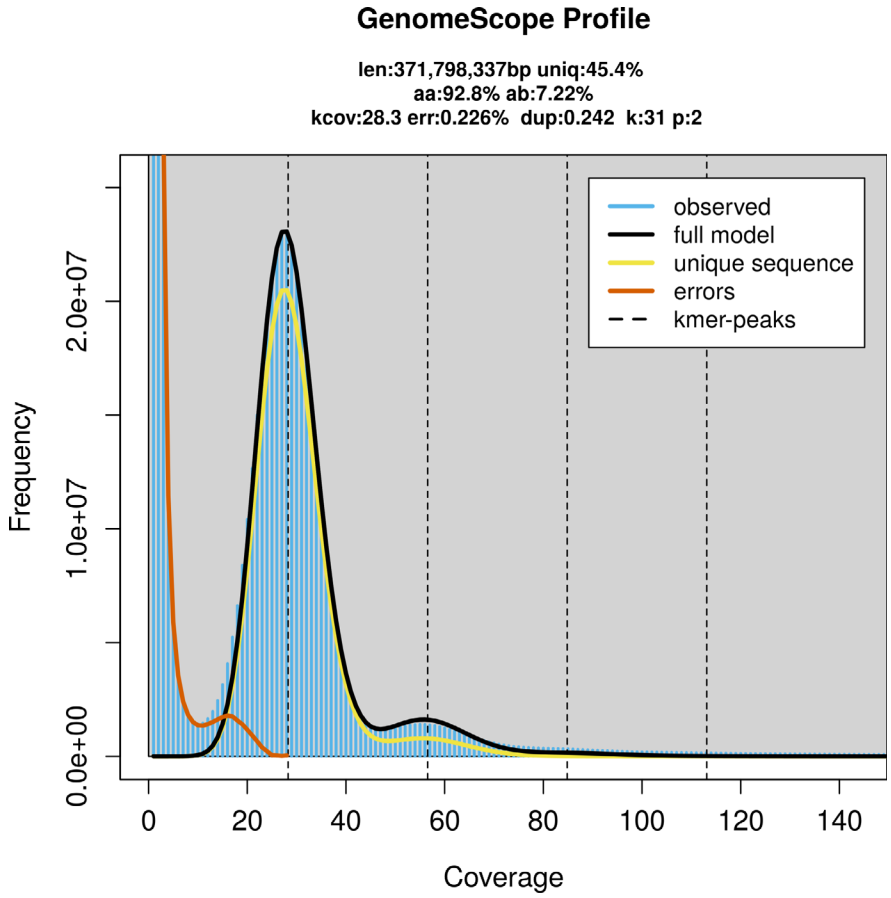

Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2. The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

Table 1. Specimen and sequencing data for BioProject PRJEB69496.

Platform	PacBio HiFi	Hi-C
ToLID	ddVioOdor1	ddVioOdor1
Specimen ID	KDTOL10129	KDTOL10129
BioSample (source individual)	SAMEA9143020	SAMEA9143020
BioSample (tissue)	SAMEA9143630	SAMEA9143629
Tissue	leaf	leaf
Instrument	Sequel IIe	Illumina NovaSeq 6000
Run accessions	ERR12303932	ERR12318578
Read count total	1.84 million	693.43 million
Base count total	22.61 Gb	104.71 Gb

Table 2. **Genome assembly statistics.**

Assembly name	ddVioOdor1.1
Assembly accession	GCA_963691705.1
Alternate haplotype accession	GCA_963691845.1
Assembly level	chromosome
Span (Mb)	698.62
Number of chromosomes	10
Number of contigs	509
Contig N50	2.78 Mb
Number of scaffolds	25
Scaffold N50	70.51 Mb
Organelles	Mitochondrion: 482.11 kb; Plastid: 158.28 kb

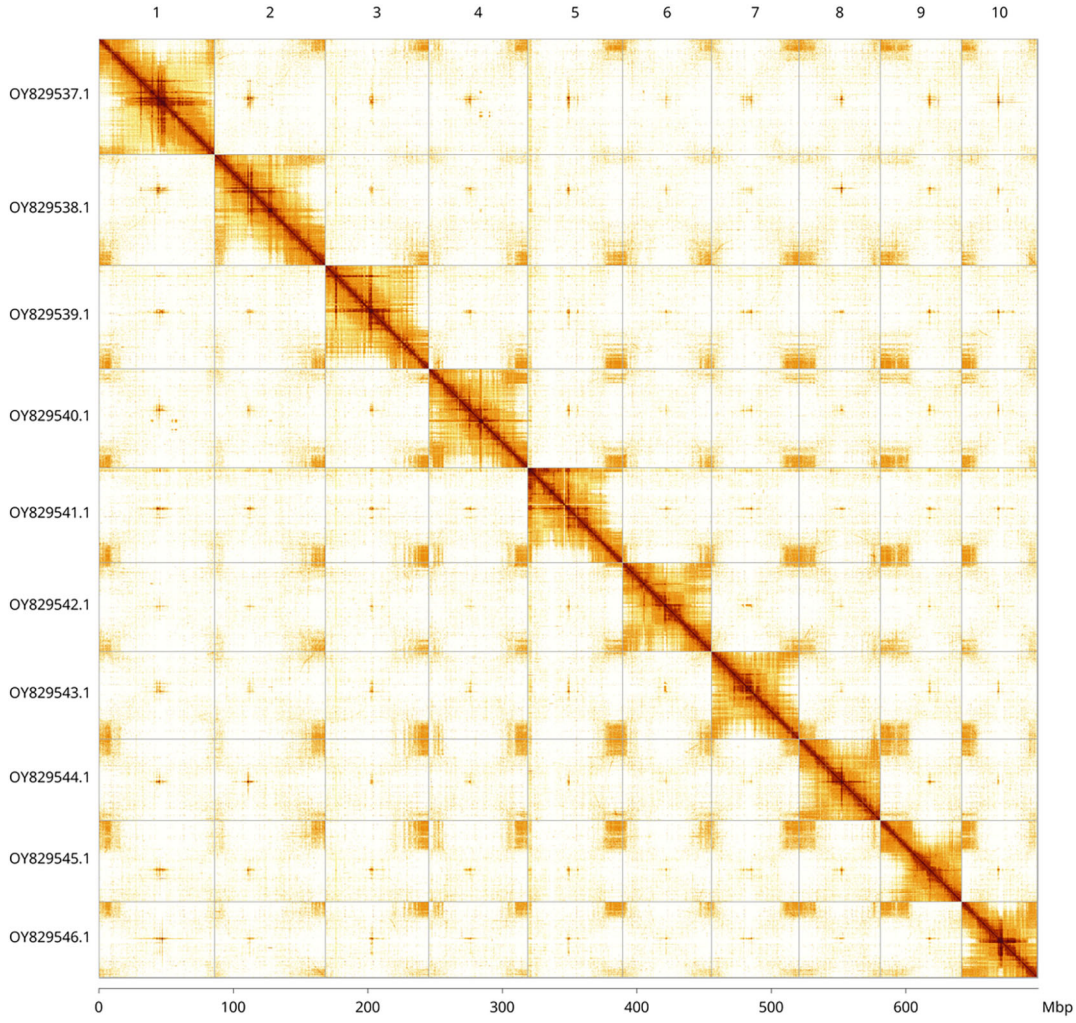

Figure 3. **Hi-C contact map of the *Viola odorata* genome assembly.** Assembled chromosomes are shown in order of size and labelled along the axes, with a megabase scale shown below. The plot was generated using PretextSnapshot.

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 the primary and alternate haplotypes using the *k*-mer database ($k = 31$) 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 NCBI datasets (O'Leary *et al.*, 2024). 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

Table 3. Chromosomal pseudomolecules in the primary genome assembly of *Viola odorata* ddVioOdor1.

INSDC accession	Molecule	Length (Mb)	GC%
OY829537.1	1	86.17	40.50
OY829538.1	2	82.27	40
OY829539.1	3	76.95	40
OY829540.1	4	73.59	40
OY829541.1	5	70.51	40.50
OY829542.1	6	66.10	39.50
OY829543.1	7	65.05	40
OY829544.1	8	60.60	39.50
OY829545.1	9	60.34	40
OY829546.1	10	56.31	39.50

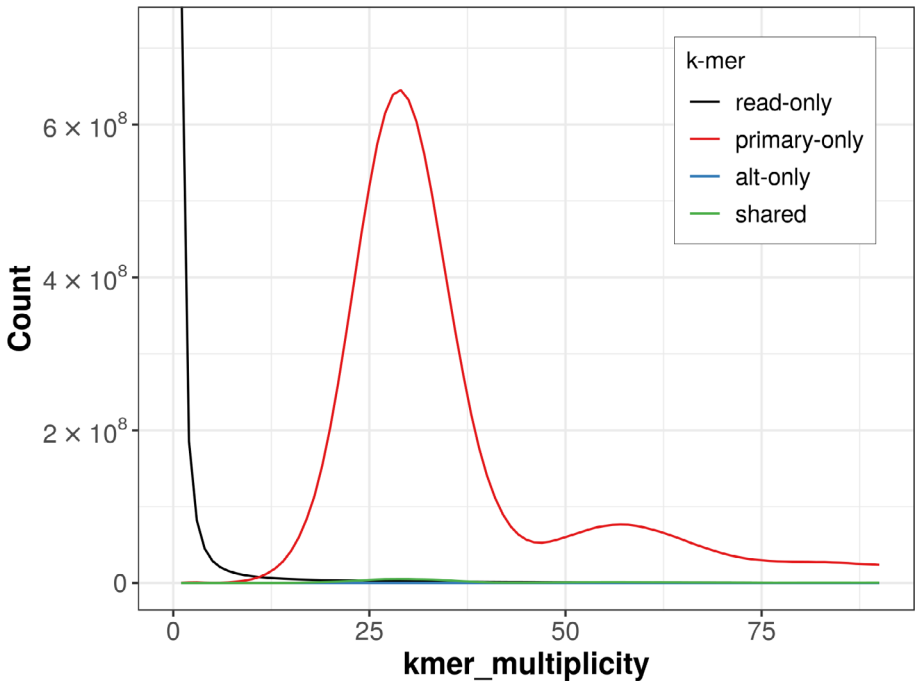

Figure 4. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers 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.

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 containerisation through Docker (Merkel, 2014) and Singularity (Kurtzer *et al.*, 2017).

Genome sequence report

Sequence data

The genome of a specimen of *V. odorata* was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating 22.61 Gb (gigabases) from 1.84 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 371.80 Mb, with a heterozygosity of 7.22% and repeat content of 54.83% (Figure 2). Using flow cytometry, the genome size (1C-value) of the sample was estimated to be 1.05 pg, equivalent to 1 030.00 Mb. These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing

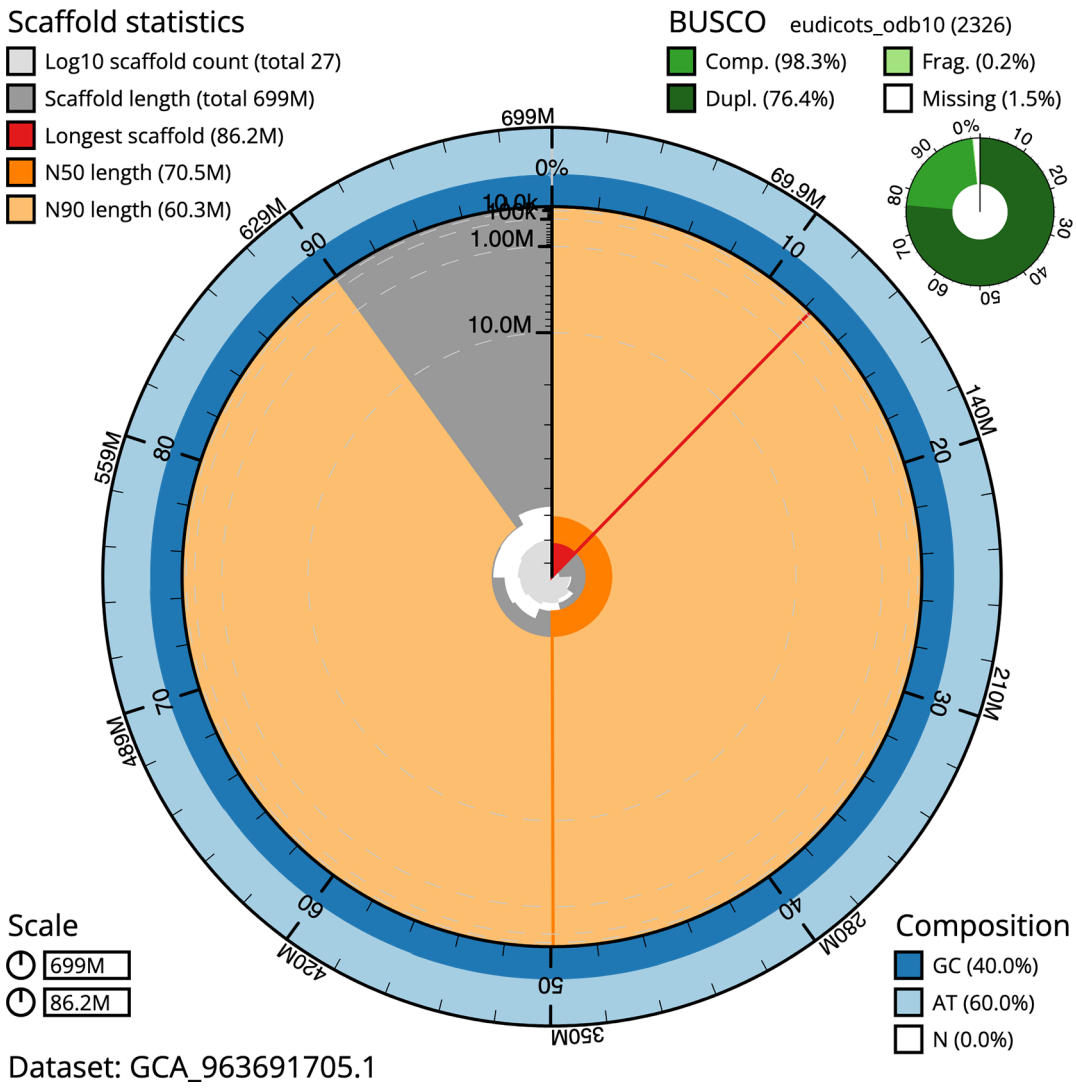

Figure 5. Assembly metrics for ddVioOdor1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the circumference. The outermost blue 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 eukaryota_odb10 set is presented at the top right. An interactive version of this figure can be accessed on the BlobToolKit viewer.

data provided approximately 57× coverage. Hi-C sequencing produced 104.71 Gb from 693.43 million reads, which were used to scaffold the assembly. Table 1 summarises the specimen and sequencing details.

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 698.62 Mb in 25 scaffolds, with 484 gaps, and a scaffold N50 of 70.51 Mb (Table 2).

Most of the assembly sequence (99.9%) was assigned to 10 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 482.11 kb, OY829547.1) and plastid genome (length 158.28 kb, OY829548.1) were also assembled. These sequences are included as contigs in the multifasta file of the genome submission and as standalone records.

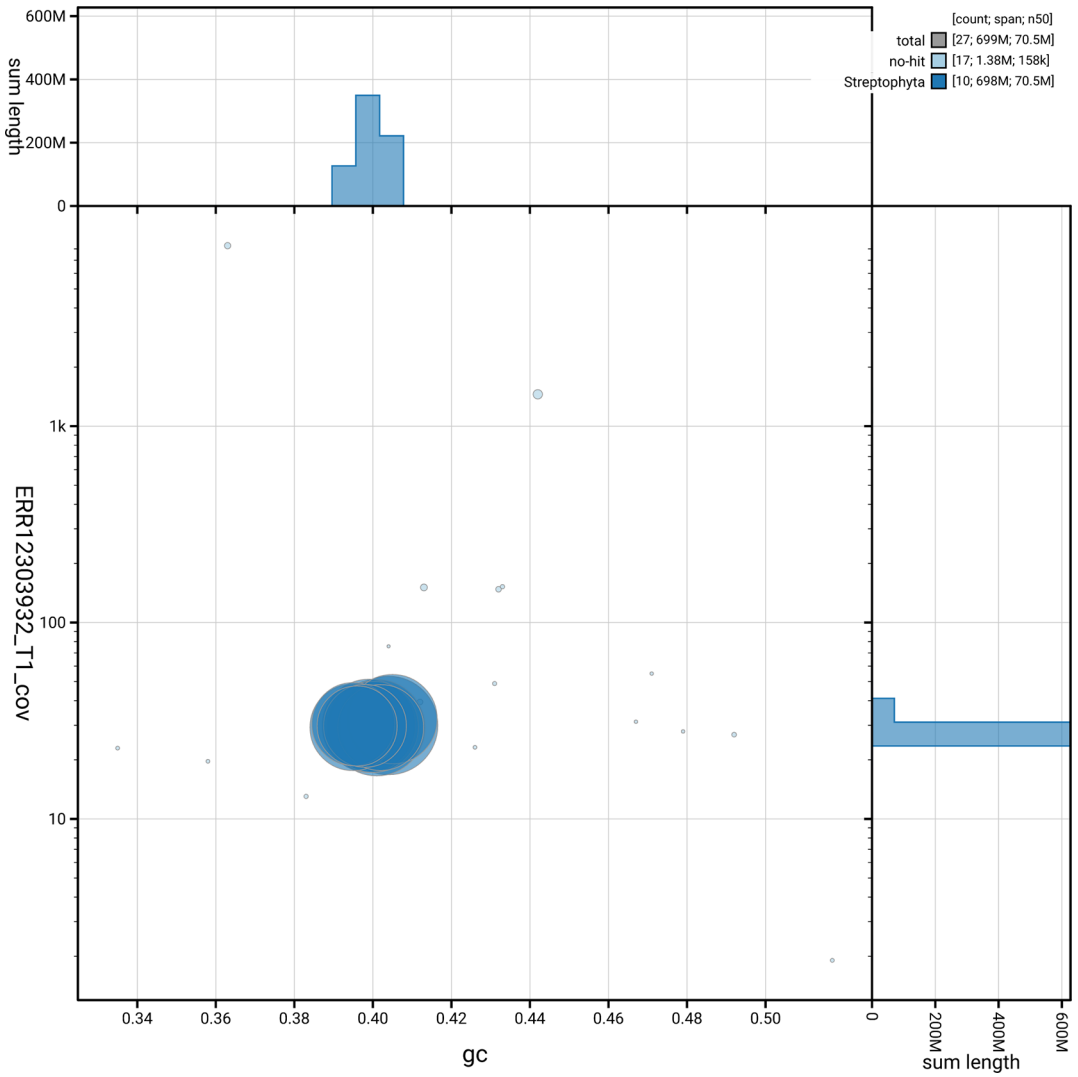

Figure 6. BlobToolKit blob plot for ddVioOdor1.1. The plot shows base coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes display the total 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 *Viola odorata* assembly.

Measure	Value	Benchmark
EBP summary (primary)	6.C.Q59	6.C.Q40
Contig N50 length	2.78 Mb	≥ 1 Mb
Scaffold N50 length	70.51 Mb	= chromosome N50
Consensus quality (QV)	Primary: 59.3; alternate: 51.8; combined: 59.0	≥ 40
k-mer completeness	Primary: 98.43%; alternate: 1.30%; combined: 98.44%	≥ 95%
BUSCO	C:99.2% [S:13.3%, D:85.9%], F:0.0%, M:0.8%, n:255	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	99.90%	≥ 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.

Assembly quality metrics

The combined primary and alternate assemblies achieve an estimated QV of 59.0. The *k*-mer completeness is 98.43% for the primary assembly, 1.30% for the alternate haplotype, and 98.44% for the combined assemblies (Figure 4).

BUSCO v.5.5.0 analysis using the eukaryota_odb10 reference set (n = 255) identified 99.2% of the expected gene set (single = 13.3%, duplicated = 85.9%). 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.

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.Q59**, meeting the recommended reference standard.

Genome annotation report

The *V. odorata* genome assembly (GCA_963691705.1) was annotated by Ensembl at the European Bioinformatics Institute (EBI). This annotation includes 59 356 transcribed mRNAs from 34 902 protein-coding and 14 061 non-coding genes. The average transcript length is 2 175.49 bp, with an average of 1.21 coding transcripts per gene and 4.51 exons per transcript. For further information about the annotation, please refer to the annotation page on Ensembl.

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: *V. odorata* (sweet violet). Accession number PRJEB69496. The genome sequence is released openly for reuse. The *V. odorata* 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. 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
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
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
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
MitoHiFi	3	https://github.com/marcelauliano/MitoHiFi
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	1.0.3	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
sanger-tol/curationpretext	1.4.2	https://github.com/sanger-tol/curationpretext
Seqtk	1.3	https://github.com/lh3/seqtk

Table 5. *Continued*

Software	Version	Source
Singularity	3.9.0	https://github.com/sylabs/singularity
TreeVal	1.4.0	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.
Reference Source . | 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): D523–31.
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

De Cleene M, Lejeune MC: *Compendium van Rituele Planten in Europa.* Gent: Uitgeverij Stichting Mens en Kultuur; 2000.

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

Fothergill PG: **Studies in Viola. IV. The somatic cytology and taxonomy of our British species of the genus *Viola.*** *New Phytol.* 1944; **43**(1): 23–35.
Publisher 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

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): 875–88.
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

O'Leary NA, Cox E, Holmes JB, *et al.*: **Exploring and retrieving sequence and metadata for species across the tree of life with NCBI Datasets.** *Sci. Data.* 2024; **11**(1): 732.
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.
Publisher Full Text

Pellicer J, Powell RF, Leitch IJ: **The application of flow cytometry for estimating genome size, ploidy level endopolyploidy, and reproductive modes in plants.**Besse P, editor. *Methods in Molecular Biology.* New York, NY: 2021; **2222**: 325–61.
PubMed Abstract

POWO: **Plants of the World online.** 2026.
Reference Source

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): 1665–80.
PubMed Abstract | Publisher Full Text | Free Full Text

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

Stace C: *New Flora of the British Isles.* Middlewood Green: C&M Floristics;4th ed 2019.

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

Uliano-Silva M, Ferreira JGRN, Krasheninnikova K, *et al.*: **MitoHiFi: A Python pipeline for mitochondrial genome assembly from PacBio high fidelity reads.** *BMC Bioinformatics.* 2023; **24**(1): 288.
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

Wyse Jackson P: *Ireland's Generous Nature. The Past and Present Use of Wild Plants in Ireland.* St Louis: Missouri Botanical Garden Press; 2014.

Zhou C, Brown M, Blaxter M, *et al.*: **Oatk: A de novo assembly tool for complex plant organelle genomes.** *Genome Biol.* 2025; **26**: 235.
Publisher Full Text

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

Open Peer Review

Current Peer Review Status: ✔ ✔ ✔

Version 1

Reviewer Report 02 June 2026

https://doi.org/10.21956/wellcomeopenres.28798.r151785

✔ **Zoé Postel**
Department of Ecology, Environment and Plant sciences, Stockholm University, Stockholm, Sweden

This report by Christenhusz et al. describes the genome assembly of *Viola odorata* (Malpighiales) based on an individual sampled in Surrey, UK. The motivation, methodology, and assessment of assembly quality are clearly presented and easy to follow. Overall, the manuscript is well written, and the level of detail provided should allow future users to readily assess the quality and usability of the genome resource. The assembly itself appears to be of high structural quality, with excellent BUSCO completeness and a very high proportion of the assembly anchored to chromosome-level scaffolds. This represents a valuable resource for future research in plant biology and genomics!
Additional details on the annotation process would be beneficial, although the authors appropriately refer to established protocols.
Overall, this work provides a useful and well-documented genomic resource.
Some more specific comments can be find below:

- p. 5 - In the Methods (genome assembly section), k-mer counts were performed on filtered reads. It would be helpful to clarify how these reads were filtered and which criteria were applied.
- p. 5 - For mitochondrial and chloroplast genome assembly, the authors used two different tools (MitoHiFi and OATK). Could the authors clarify the rationale for this choice, given that OATK can be applied to both organellar genomes?
- Figure 2 would benefit from additional explanation to help a broader audience interpret the different peaks.
- Figure 6 : a bit more help would be appreciated for interpretation from the reader
- p. 11 - The high proportion of duplicated BUSCOs would merit further discussion. Given the reported level of heterozygosity (7.22%) and very low k-mer completeness for the alternate assembly (1.3%), this may reflect incomplete collapsing of heterozygous regions in the primary assembly.
- p. 11 - Among the 34,902 protein-coding genes, it would be informative to specify how many are of mitochondrial or chloroplast origin.

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 geneticist and population genomics, nuclear and organellar genomes

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 07 April 2026

https://doi.org/10.21956/wellcomeopenres.28798.r151135

Stalin Nithaniyal
Botanical Survey of India, Western Regional Centre, Pune, Maharashtra, India

Christenhusz and co-authors investigated genome sequence of sweet violet (*Viola odorata*) using Illumina NovaSeq 6000 as a part of Darwin Tree of Life project. This timely study focused on the notable plant *V. odorata*, sampled from Richmond, Surrey, United Kingdom, which is highly regarded for its cultural, horticultural, and medicinal value. The research was well-executed, utilizing a robust methodology covering DNA extraction, library preparation, genome assembly, and curation. The genome assembly represents a high-quality reference data, indicated by its high contiguity (99.9% chromosomal scaffolding) and high accuracy (QV 59.0). Furthermore, the EBP 6.C.Q59 metric demonstrates that it meets the highest international standards for reference-quality genomes. This assembly provides a strong foundation for evolutionary studies, boasting 99.2% gene-space completeness and a comprehensive Ensembl annotation of 34,902 protein-coding genes from mitochondrial, and plastid genomes. While the high gene duplication rate (85.9%) in the BUSCO analysis suggests a need for careful interpretation to distinguish between biological paralogs and potential assembly redundancies, the 99.2% overall BUSCO score confirms that the assembly captures nearly the entire expected gene set. Overall, the work is well-presented, and the genomic data will serve as a valuable resource for botanists and plant geneticists by facilitating future comparative studies.

Major outcomes of the study are,
1. Achieving 99.9% of the 698.62 Mb assembly and scaffolded into 10 chromosomal pseudomolecules is a significant result and provide near-complete physical map of the genome.
2. High QV of 59.0 demonstrates quality base-call accuracy required for the reference genomes
3. Successful assembly of mitochondrial (482.11 kb) and plastid (158.28 kb) genomes
4. Ensembl Annotation provides well-curated set of 34,902 protein-coding genes which act as valuable resource to the research community

Some minor comments includes,
1. Clarify the reason for high duplication rate in BUSCO analysis (85.9%) and selection of eukaryota_odb10 reference set instead of plant reference
2. A ratio of 1.2 transcripts per gene is considered low in the context of plant transcriptomics
3. Breakdown of 14 061 non-coding genes may be provided.

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: Molecular Biology, Plant Systematics, Ethnobotany

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 03 April 2026

https://doi.org/10.21956/wellcomeopenres.28798.r151782

Bruno Contreras-Moreira
Estacion Experimental de Aula Dei-CSIC (Ringgold ID: 54439), Zaragoza, Aragon, Spain

The manuscript by Christenhusz et al describes their work on the assembly and annotation of the genome of dicot Viola odorata out of DNA extracted from a plant in Surrey. I find that the methods

describe with sufficient precision all the steps from plant identification in a public park to the preparation of DNA libraries, assembly and QC. Probably the part that is explained with least detail is the gene annotation protocol, despite the info available externally. All software version/releases are also listed on Table 5.
Regarding the results, the paper reports efficiently what most readers will want to know about an assembly, its quality, and its annotation, linking back to the data sources and INSDC. In addition, the genome is made available at the Ensembl browser so that anyone can use it.

Here are some comments in case the authors want to address them:

Figure 2 legend: Please help readers understand the figure; as it stands some people won't be able to follow what the peaks are.

p8: Please provide details of the singularity container for reproducibility.

p11 "Assembly quality metrics":
a) I see the primary hifiasm assembly is reported; did the authors ever try a hap1/hap2 assembly or that pathway is deemed too complidated?
b) why is the BUSCO completeness reported here relate to lineage eukaryota_odb10 instead of eudicots_odb1o (Figure 5)? Please explain to avoid confusion.

p11 "Genome annotation report": did the gene annotation of this genome used any RNAseq data from related species or was instead based on de novo or protein-based gene calling?

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?
Partly

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

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics, 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.