DATA NOTE

The genome sequence of *Veronica scutellata* L., 1753 (Lamiales: Plantaginaceae)

[version 1; peer review: awaiting peer review]

Andrew Griffiths[1], José Ignacio Márquez-Corro[2],
Royal Botanic Garden Edinburgh Genome Acquisition Lab,
Plant Genome Sizing Collective, Darwin Tree of Life Barcoding collective,
Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team,
Wellcome Sanger Institute Scientific Operations: Sequencing Operations,
Wellcome Sanger Institute Tree of Life Core Informatics team,
Tree of Life Core Informatics collective, Darwin Tree of Life Consortium

[1]Royal Botanic Garden Edinburgh, Edinburgh, Scotland, UK
[2]Royal Botanic Gardens Kew, Richmond, England, UK

v1 **First published:** 19 Aug 2026, **11**:531
https://doi.org/10.12688/wellcomeopenres.27363.1
Latest published: 19 Aug 2026, **11**:531
https://doi.org/10.12688/wellcomeopenres.27363.1

Abstract

We present a genome assembly of *Veronica scutellata* (Marsh Speedwell; Streptophyta; Magnoliopsida; Lamiales; Plantaginaceae). The assembly consists of two haplotypes with total lengths of 767.32 megabases and 770.96 megabases. Most of haplotype 1 (97.74%) is scaffolded into 9 chromosomal pseudomolecules. Haplotype 2 was assembled to scaffold level. The mitochondrial sequences have lengths of 132.89 and 200.86 kilobases and the plastid genome assembly has a length of 150.99 kilobases. Gene annotation of this assembly on Ensembl identified 23 525 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

Veronica scutellata, Marsh Speedwell, genome sequence, chromosomal, Lamiales

Open Peer Review

Approval Status *AWAITING PEER REVIEW*

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

This article is included in the Tree of Life gateway.

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

Author roles: Griffiths A: Investigation, Resources, Writing – Original Draft Preparation; **Márquez-Corro JI**: Investigation, Resources;

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome through core funding to the Wellcome Sanger Institute (220540) 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: Griffiths A, Márquez-Corro JI, Royal Botanic Garden Edinburgh Genome Acquisition Lab *et al.* **The genome sequence of *Veronica scutellata* L., 1753 (Lamiales: Plantaginaceae) [version 1; peer review: awaiting peer review]** Wellcome Open Research 2026, **11**:531 https://doi.org/10.12688/wellcomeopenres.27363.1

First published: 19 Aug 2026, **11**:531 https://doi.org/10.12688/wellcomeopenres.27363.1

Species taxonomy

Eukaryota; Viridiplantae; Streptophyta; Streptophytina; Embryophyta; Tracheophyta; Euphyllophyta; Spermatophyta; Magnoliopsida; Mesangiospermae; eudicotyledons; Gunneridae; Pentapetalae; asterids; lamiids; Lamiales; Plantaginaceae; Veroniceae; *Veronica*; *Veronica subgen. Veronica*; *Veronica scutellata* L., 1753 (NCBI:txid195138).

Background

Marsh Speedwell, *Veronica scutellata* L., is a perennial herb found in wetland habitats such as pond and lake margins, marshes, fens, wet grassland, bogs and wet heath, and often associated with acidic soils (Stroh *et al.*, 2023). In Britain and Ireland, *Veronica scutellata* has been recorded from sea level up to an elevation of 780 m. It is a widespread species, although it is in decline due to hydrological degradation of its habitat. Decline is most notable in central and southern England, and central Ireland, and the species has become uncommon across much of the lowlands. Belonging to the Eurosiberian Boreo-temperate biogeographic element, *Veronica scutellata* occurs across much of Europe, into northern Asia, and also in North America. It has naturalised widely outside its native range, for example in New Zealand (Sykes, 1981).

We present a chromosome-level genome sequence for *Veronica scutellata.* The assembly was produced using the Tree of Life pipeline from a specimen collected in Balavil Fen, Insh Marshes, East Inverness-shire VC96, Scotland, UK (Figure 1). It was generated 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 (Darwin Tree of Life Project Consortium, 2022). *Veronica scutellata* has a reported chromosome count of $2n = 18$ (Peruzzi & Cesca, 2002; Šmarda *et al.*, 2019; Zonneveld, 2019), consistent with the nine chromosome-scale pseudomolecules in haplotype 1 of the assembly presented here.

Methods

Sample acquisition, flow cytometry and DNA barcoding

A specimen of *Veronica scutellata* (specimen ID EDTOL06286, ToLID daVerScut1; Figure 1) was used for genome sequencing. It was collected from Balavil Fen, Insh Marshes, East Inverness-Shire Vc96, Scotland, UK (latitude 57.0971, longitude −3.9793) on 2023-07-19. The specimen was collected by Andy Griffiths and José Ignacio Márquez-Corro and identified by Andy Griffiths (Royal Botanic Garden Edinburgh). The herbarium voucher associated with the sequenced plant has been deposited in the herbarium of RBG Edinburgh (E) https://data.rbge.org.uk/herb/E01358569.

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 Life Core Laboratory are available on protocols.io (Howard *et al.*, 2025). The daVerScut1 sample was weighed and triaged to determine the appropriate extraction protocol. For HMW DNA extraction, 64.0 mg of leaf tissue 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 67.4 ng/μL and a yield of 3 707.00 ng, with a fragment size of 13.1 kb. The 260/280 spectrophotometric ratio was 1.76, and the 260/230 ratio was 1.17. The Genomic Quality Number (GQN) was 7.

Figure 1. Photograph of the *Veronica scutellata* plant (daVerScut1) from which samples were taken for genome sequencing.

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, California, USA). Prepared libraries selected for multiplexing were pooled based on genome size, library molarity, and the intended plex level. Primers were annealed and polymerases were bound to generate circularised complexes, following the manufacturer's instructions. Complexes were purified using SMRTbell beads, diluted to the Revio loading concentration, and spiked with a Revio sequencing internal control. The pooled libraries were sequenced on a Revio 25 M Plex SMRT cell in a 2-plex run. 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 leaf tissue of daVerScut1 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

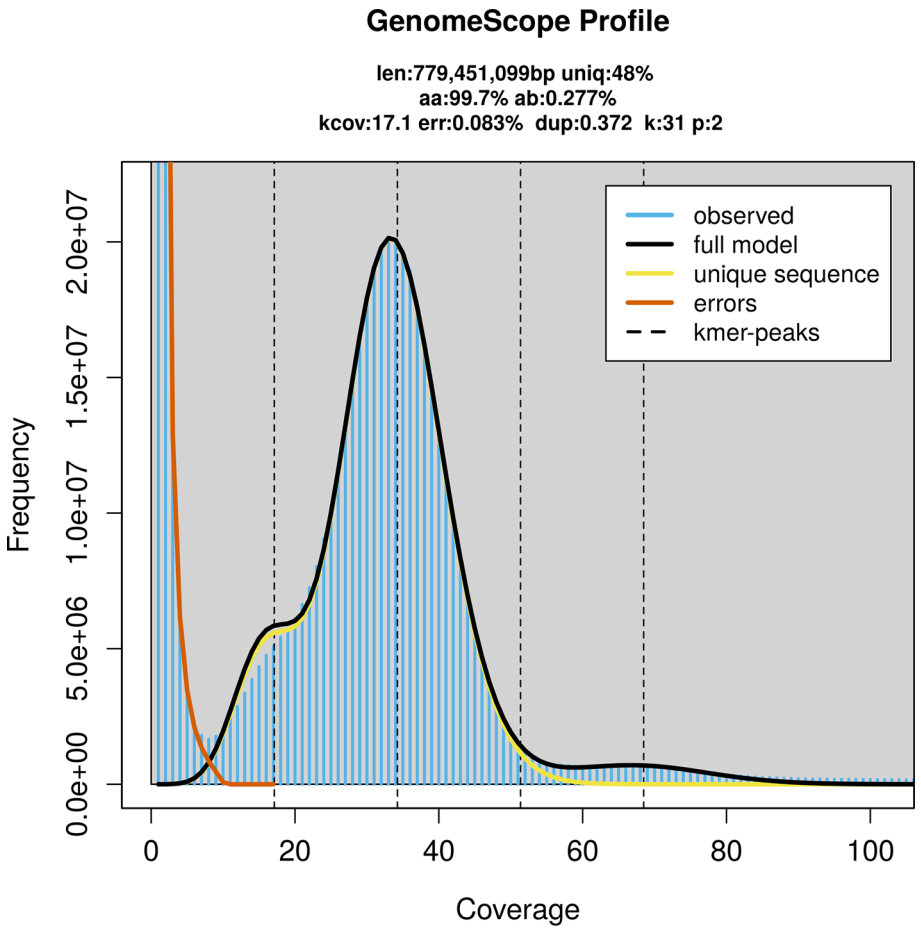

Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2. The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

Table 1. **Specimen and sequencing data for *Veronica scutellata* (BioProject PRJEB79447).**

Platform	PacBio HiFi	Hi-C
ToLID	daVerScut1	daVerScut1
Specimen ID	EDTOL06286	EDTOL06286
BioSample (source individual)	SAMEA114782739	SAMEA114782739
BioSample (tissue)	SAMEA114788155	SAMEA114788155
Tissue	leaf	leaf
Instrument	Revio	Illumina NovaSeq X
Run accessions	ERR13615511	ERR13621442
Read count total	2.93 million reads	319.80 million read pairs
Base count total	27.47 Gb	96.58 Gb

(Agilent) with High Sensitivity D500 reagents before sequencing. Sequencing was performed using paired-end 150 bp reads on the Illumina NovaSeq X.

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, 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 MerquryFK (Rhie *et al.*, 2020). The organelle genomes were assembled using 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.*

Table 2. **Genome assembly statistics for *Veronica scutellata*.**

Genome assembly	Haplotype 1	Haplotype 2
Assembly name	daVerScut1.hap1.1	daVerScut1.hap2.1
Assembly accession	GCA_965198585.1	GCA_965198465.1
Assembly level	chromosome	scaffold
Span (Mb)	767.32	770.96
Number of chromosomes	9	-
Number of contigs	548	457
Contig N50	4.95 Mb	4.24 Mb
Number of scaffolds	288	196
Scaffold N50	86.66 Mb	83.34 Mb
Longest scaffold length (Mb)	91.99	91.81
Organelles	Mitochondrial genomes: 132.89 and 200.86 kb; Plastid genome: 150.99 kb	-

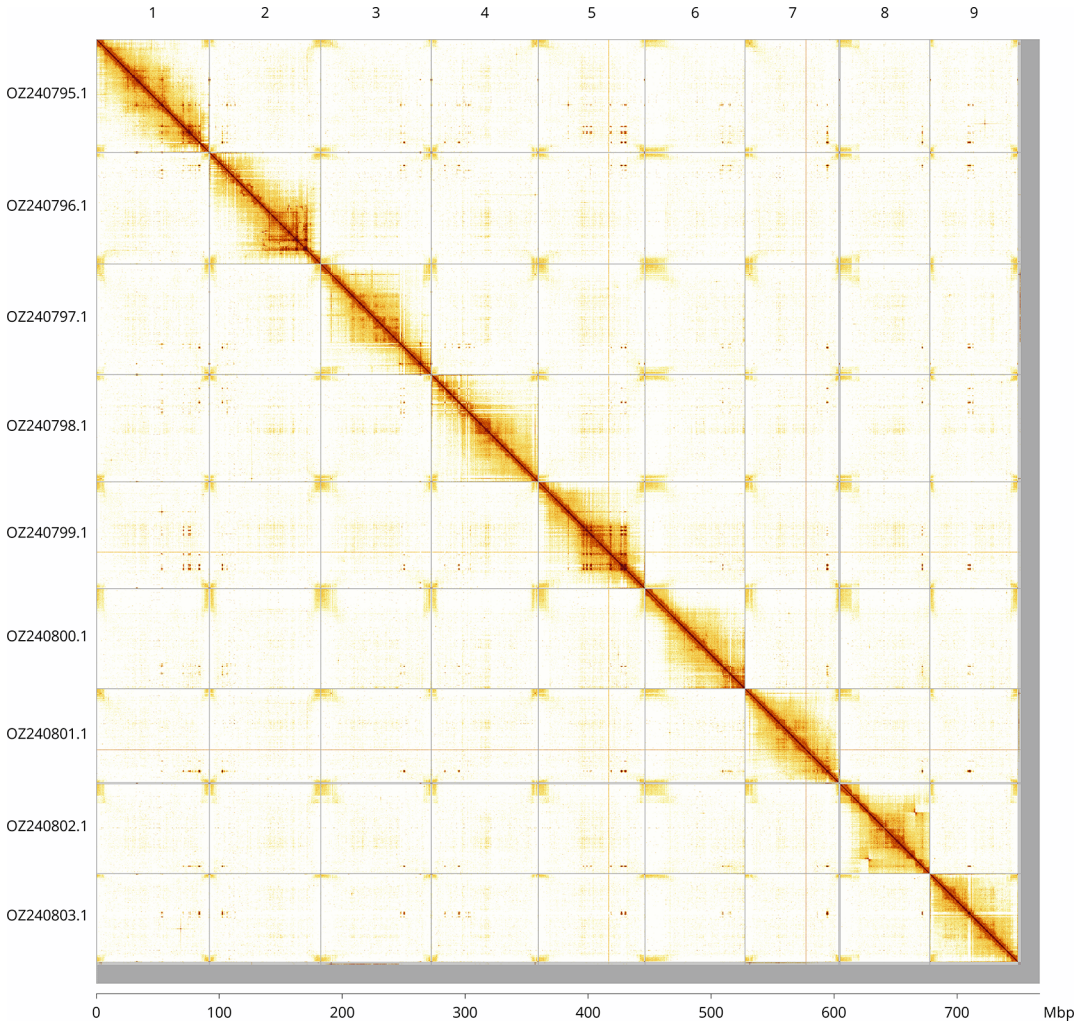

Figure 3. Hi-C contact map of the *Veronica scutellata* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes, with a megabase scale shown below. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in the haplotype 1 genome assembly of *Veronica scutellata* daVerScut1.

INSDC accession	Molecule	Length (Mb)	GC%
OZ240795.1	1	91.99	39
OZ240796.1	2	90.76	40
OZ240797.1	3	89.81	39.50
OZ240798.1	4	87.13	39
OZ240799.1	5	86.66	38.50
OZ240800.1	6	81.41	39
OZ240801.1	7	77.15	39.50
OZ240802.1	8	73.07	39.50
OZ240803.1	9	71.96	39

(2021). Manual corrections included 12 breaks and 31 joins. This reduced the scaffold count by 7.3%, increased the scaffold N50 by 10.4%, and reduced the total assembly length by 1.1%. The curation process is described at sanger-tol/curation-resources. PretextSnapshot was used to generate a Hi-C contact map of the final assembly.

Assembly quality assessment

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

The genome was analysed using the BlobToolKit pipeline, a Nextflow implementation of the earlier Snakemake version (Challis *et al.*, 2020). PacBio reads were aligned to the assembly using minimap2 (Li, 2018) and processed with SAMtools (Danecek *et al.*, 2021) to generate coverage tracks. The pipeline runs BUSCO (Manni *et al.*, 2021) using lineages identified from the NCBI Taxonomy (Schoch *et al.*, 2020). For the three basal BUSCO lineages, eukaryota, bacteria and archaea, BUSCO genes are aligned to the UniProt Reference Proteomes database (Bateman *et al.*, 2023) using DIAMOND BLASTp (Buchfink *et al.*, 2021). The genome assembly 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 using DIAMOND BLASTx. Sequences without DIAMOND BLASTx hits are extracted using seqtk and aligned to the NT database using BLASTn (Altschul *et al.*, 1990). The outputs are consolidated into a BlobDir for visualisation in BlobToolKit. 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 *Veronica scutellata* was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating 27.47 Gb (gigabases) from 2.93 million reads, which were used to assemble the genome. GenomeScope2.0 analysis using a $p = 2$ model gave a *k*-mer-based genome size estimate of 780.45 Mb, with a heterozygosity of 0.28% and repeat content of 52.02% (Figure 2). Using flow cytometry, the genome size (1C-value) of the sample was estimated to be 0.92 pg, equivalent to 900.00 Mb. These estimates guided expectations for the assembly. Based on the

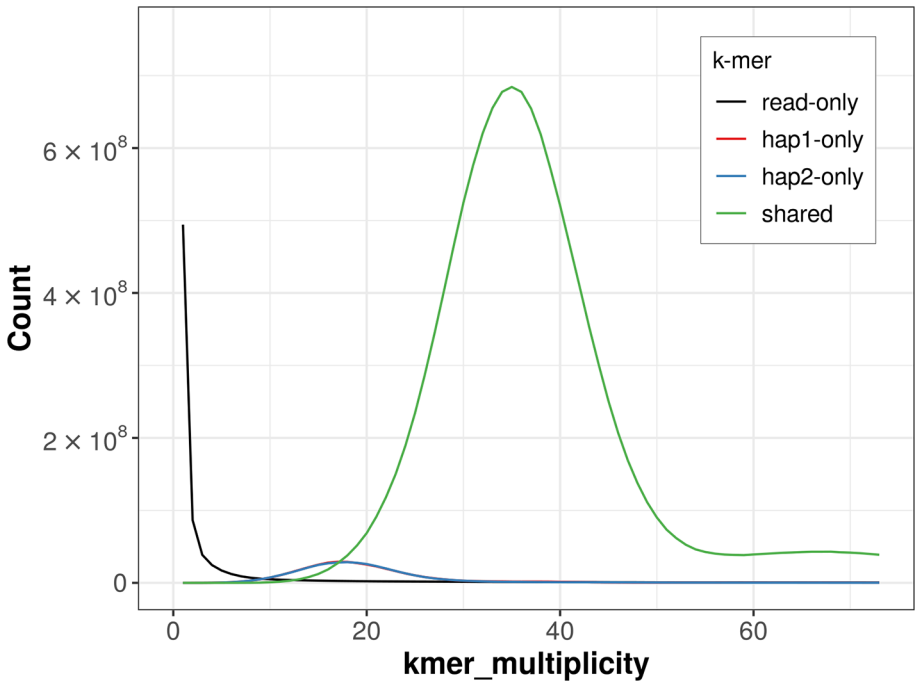

Figure 4. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers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

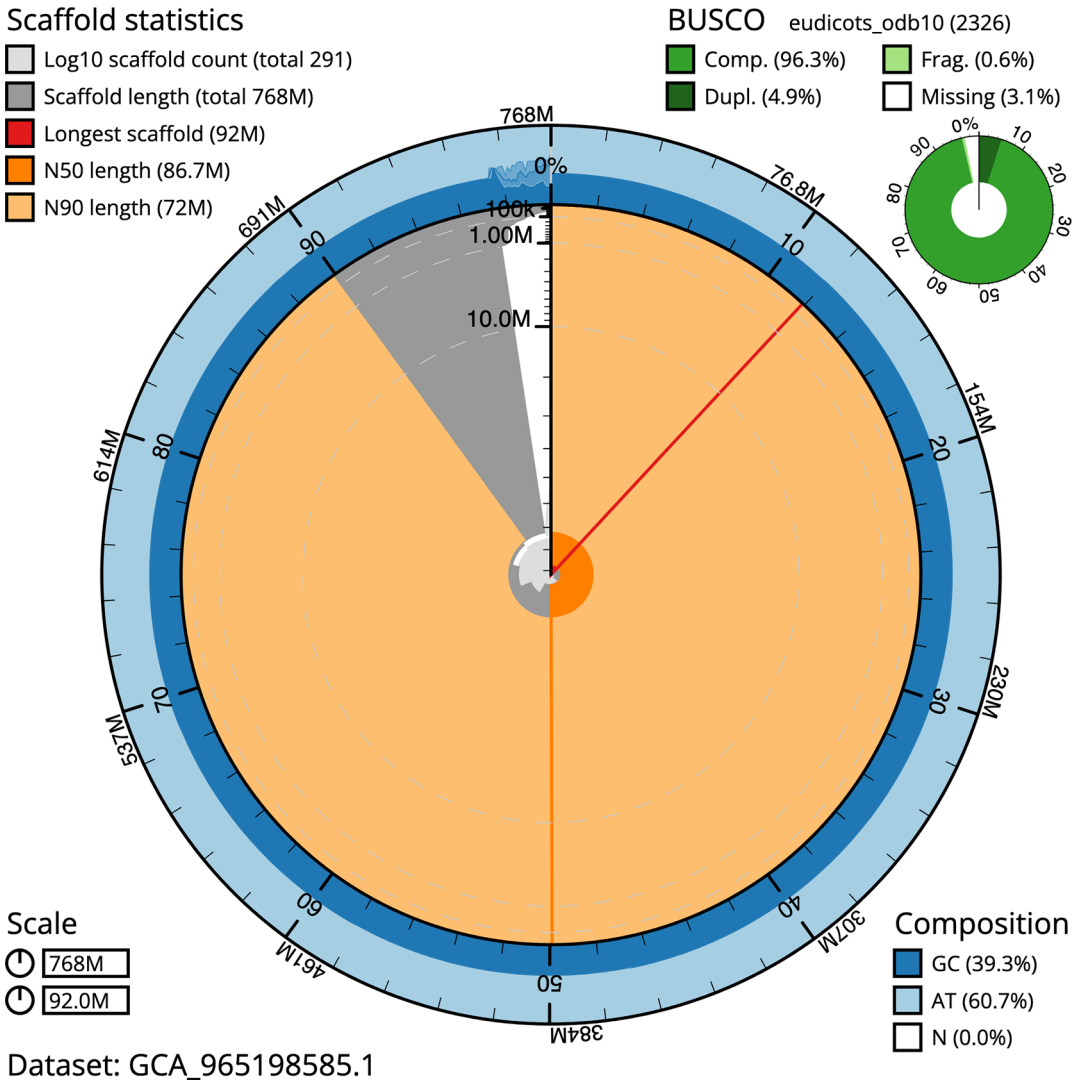

Figure 5. Assembly metrics for daVerScut1.hap1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1 000 bins around the circumference. The outermost 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 eudicots_odb10 set is presented at the top right. An interactive version of this figure can be accessed on the BlobToolKit viewer.

estimated genome size, the sequencing data provided approximately 34× coverage. Hi-C sequencing produced 96.58 Gb from 319.80 million read pairs, which were 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 767.32 Mb in 288 scaffolds, with 260 gaps, and a scaffold N50 of 86.66 Mb (Table 2).

Most of the assembly sequence (97.74%) was assigned to 9 chromosomal-level scaffolds. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 3; Table 3).

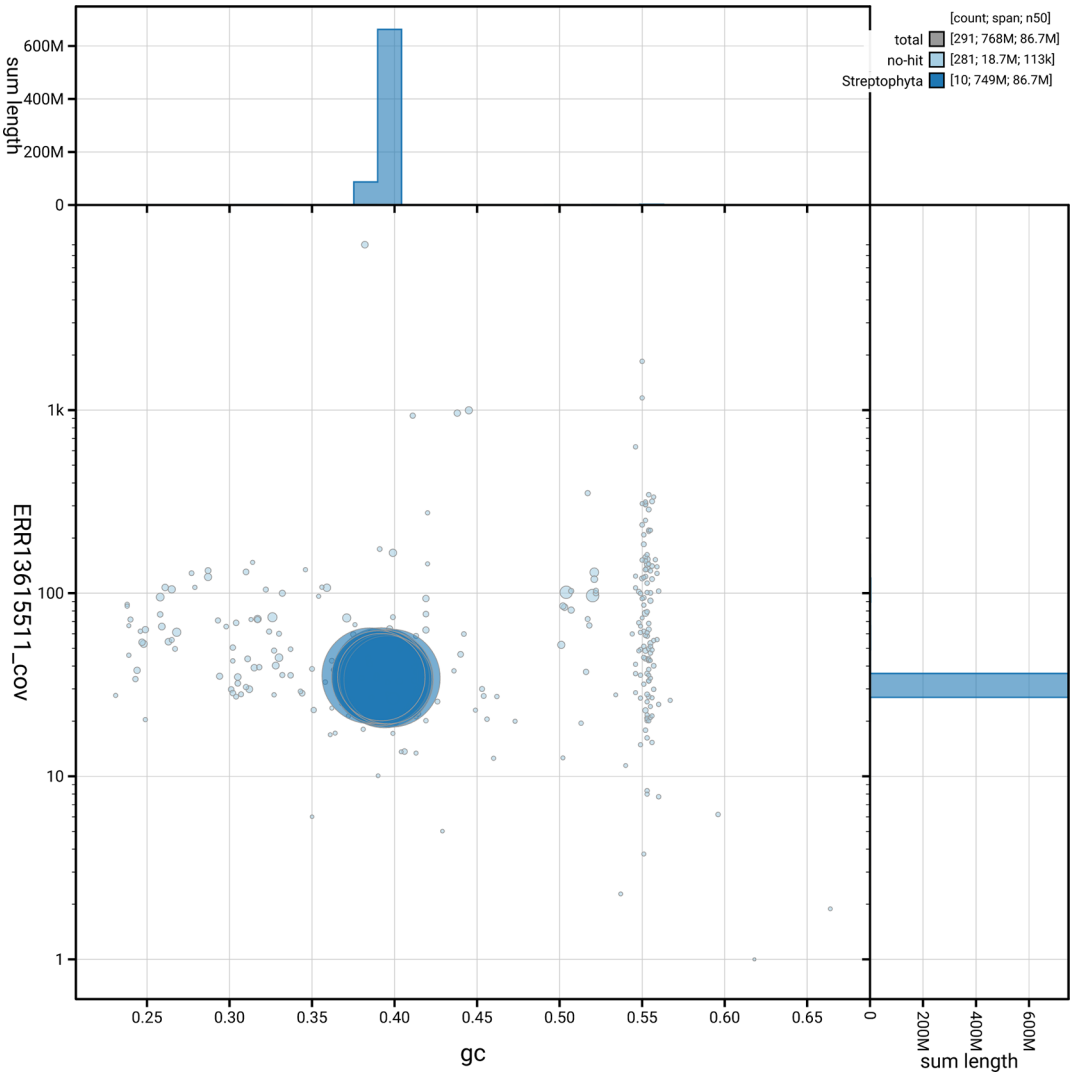

Figure 6. BlobToolKit blob plot for daVerScut1.hap1.1. The plot shows base coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes display the 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.

Two mitochondrial genomes (lengths 132.89 kb [OZ240804.1] and 200.86 kb [OZ240805.1]) and the plastid genome (length 150.99 kb, OZ240806.1) were also assembled. These sequences are included as contigs in the multifasta file of the genome submission and as standalone records.

Assembly quality metrics

For haplotype 1, the estimated QV is 63.7, and for haplotype 2, 63.2. When the two haplotypes are combined, the assembly achieves an estimated QV of 63.4. The *k*-mer completeness is 94.05% for haplotype 1, 94.01% for haplotype 2, and 98.84% for the combined haplotypes (Figure 4).

BUSCO analysis using the eudicots_odb10 reference set (n = 2 326) identified 96.3% of the expected gene set (single = 91.4%, duplicated = 4.9%) 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. Earth Biogenome Project summary metrics for the *Veronica scutellata* assembly.

Measure	Value	Benchmark
EBP summary (haplotype 1)	6.C.Q63	6.C.Q40
Contig N50 length	4.95 Mb	≥ 1 Mb
Scaffold N50 length	86.66 Mb	= chromosome N50
Consensus quality (QV)	Haplotype 1: 63.7; haplotype 2: 63.2; combined: 63.4	≥ 40
k-mer completeness	Haplotype 1: 94.05%; Haplotype 2: 94.01%; combined: 98.84%	≥ 90%
BUSCO	Haplotype 1: C:96.3% [S:91.4%, D:4.9%], F:0.6%, M:3.1%, n:2 326; Haplotype 2: C:96.0% [S:91.3%, D:4.7%], F:0.7%, M:3.3%, n:2 326	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	97.74%	≥ 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

Table 4 lists the assembly metric benchmarks adapted from Rhie *et al.* (2021) and the Earth BioGenome Project Report on Assembly Standards January 2026. The EBP metric calculated for haplotype 1 is **6.C.Q63**, meeting the recommended 6.C.Q40 reference standard.

Genome annotation report

The *Veronica scutellata* genome assembly (GCA_965198585.1) was annotated by Ensembl at the European Bioinformatics Institute (EBI). This annotation includes 50 332 transcribed mRNAs from 23 525 protein-coding and 17 114 non-coding genes. The average transcript length is 2 052.05 bp, with an average of 1.24 coding transcripts per gene and 4.31 exons per transcript. For further information about the annotation, please refer to the annotation page on Ensembl.

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

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: Veronica scutellata (marsh speedwell). Accession number PRJEB79447; https://identifiers.org/ena.embl/PRJEB79447. The genome sequence is released openly for reuse. The *Veronica scutellata* genome sequencing initiative is part of the Darwin Tree of Life Project (PRJEB40665) and the Sanger Institute Tree of Life Programme (PRJEB43745). All raw sequence data and the 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 used for *Veronica scutellata*.

Software	Version	Source
BLAST	2.14.0	https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/
BlobToolKit	4.4.5	https://github.com/blobtoolkit/blobtoolkit
BUSCO	5.7.1	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	0.19.8-r603	https://github.com/chhylp123/hifiasm
HiGlass	1.13.4	https://github.com/higlass/higlass
MerquryFK	1.1.0-c1	https://github.com/thegenemyers/MERQURY.FK
Minimap2	2.28-r1209	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	24.10.4	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	1.21	https://github.com/samtools/samtools
sanger-tol/ascc	0.1.0	https://github.com/sanger-tol/ascc
sanger-tol/blobtoolkit	v0.7.1	https://github.com/sanger-tol/blobtoolkit
sanger-tol/curationpretext	1.4.2	https://github.com/sanger-tol/curationpretext
Seqtk	1.3	https://github.com/lh3/seqtk
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.
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.
PubMed Abstract | Publisher Full Text | Free 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.
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

Darwin Tree of Life Project Consortium: **Sequence locally, think globally: The Darwin Tree of Life Project.** *Proceedings of the National Academy of Sciences of the United States of America.* 2022; **119**(4): e2115642118.
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

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

Li H: **Minimap2: Pairwise alignment for nucleotide sequences.** *Bioinformatics.* 2018; **34**(18): 3094–3100.
PubMed Abstract | Publisher Full Text | Free 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.
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.
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; **2222**: 325–61.
PubMed Abstract | Publisher Full Text

Peruzzi L, Cesca G: **Chromosome numbers of flowering plants from Calabria, S Italy.** *Willdenowia.* 2002; **32**(1): 33–44.
Publisher Full Text

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

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

Schoch CL, Ciufo S, Domrachev M, *et al.*: **NCBI taxonomy: A comprehensive update on curation, resources and tools.** *Database.* 2020; **2020**: baaa062.
PubMed Abstract | Publisher Full Text | Free Full Text

Šmarda P, Knápek O, Březinová A, *et al.*: **Genome sizes and genomic guanine+cytosine (GC) contents of the Czech vascular flora with new estimates for 1700 species.** *Preslia.* 2019; **91**(2): 117–142.
Publisher Full Text

Stroh PA, Walker KJ, Humphrey TA, *et al.*: *Plant Atlas 2020: Mapping changes in the distribution of the British and Irish flora.* Princeton University Press; 2023.
Publisher Full Text

Sykes WR: **Checklist of dicotyledons naturalised in New Zealand 7. Scrophulariales.** *N. Z. J. Bot.* 1981; **19**(1): 53–57.
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.** *Wellcome Open Research.* 2024; **9**: 339.
PubMed Abstract | Publisher Full Text | Free Full Text

Vasimuddin M, Misra S, Li H, *et al.*: **Efficient architecture-aware acceleration of BWA-MEM for multicore systems.** *IEEE International Parallel and Distributed Processing Symposium (IPDPS).* IEEE; 2019; **2019**: 314–24.
Publisher Full Text

Zhou C, Brown M, Blaxter M, *et al.*: **Oatk: A de novo assembly tool for complex plant organelle genomes.** *Genome Biol.* 2025; **26**(1): 235.
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

Zonneveld BJM: **The DNA weights per nucleus (genome size) of more than 2350 species of the Flora of The Netherlands, of which 1370 are new to science, including the pattern of their DNA peaks.** *Forum Geobotanicum.* 2019; **8**: 24–78.
Publisher Full Text