GENOME NOTE

The genome sequence of dewberry, *Rubus caesius* L. (Rosales: Rosaceae)

[version 1; peer review: awaiting peer review]

Maarten J. M. Christenhusz[1,2], José Ignacio Márquez-Corro[1], Isabel Larridon[3], Royal Botanic Gardens Kew Genome Acquisition Lab, Plant Genome Sizing Collective, Darwin Tree of Life Barcoding collective, Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team, Wellcome Sanger Institute Scientific Operations: Sequencing Operations, Wellcome Sanger Institute Tree of Life Core Informatics team, Tree of Life Core Informatics collective, Darwin Tree of Life Consortium

[1]Royal Botanic Gardens Kew, Richmond, Surrey, England, UK
[2]Botanical Society of Britain and Ireland, St Albans, Hertfordshire, England, UK
[3]Royal Botanic Gardens, Kew, Richmond, Surrey, England, UK

v1 **First published:** N/A, **N/A**: N/A N/A
Latest published: N/A, **N/A**: N/A N/A

Abstract

We present a genome assembly of the tetraploid *Rubus caesius* (dewberry; Streptophyta; Magnoliopsida; Rosales; Rosaceae). The two assemblies, designated haplotype 1 and haplotype 2, have total lengths of 628.60 megabases and 633.30 megabases. Most of the haplotype 1 assembly (97.18%) was scaffolded into 14 chromosomal pseudomolecules. Haplotype 2 was assembled to scaffold level. The mitochondrial genome assembly has a length of 423.35 kilobases and the plastid genome assembly has a length of 156.98 kilobases. 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

Rubus caesius, Dewberry, genome sequence, chromosomal, Rosales

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status *AWAITING PEER REVIEW*

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

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

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

Author roles: Christenhusz MJM: Investigation, Resources; **Márquez-Corro JI**: Investigation, Resources; **Larridon I**: Investigation, Resources;

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome through core funding to the Wellcome Sanger Institute [220540, https://doi.org/10.35802/220540] and the Darwin Tree of Life Discretionary Award [218328, https://doi.org/10.35802/218328].
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Christenhusz MJM, Márquez-Corro JI, Larridon I *et al.* **The genome sequence of dewberry, *Rubus caesius* L. (Rosales: Rosaceae) [version 1; peer review: awaiting peer review]** Wellcome Open Research , : https://doi.org/

First published: N/A, **N/A**: N/A N/A

Species taxonomy

Eukaryota; Viridiplantae; Streptophyta; Streptophytina; Embryophyta; Tracheophyta; Euphyllophyta; Spermatophyta; Magnoliopsida; Mesangiospermae; eudicotyledons; Gunneridae; Pentapetalae; rosids; fabids; Rosales; Rosaceae; Rosoideae; Rosoideae incertae sedis; *Rubus*; *Rubus caesius* L. (NCBI:txid75065)

Background

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

Dewberry (*Rubus caesius*) is a low-growing, deciduous bramble in the family Rosaceae. The species is tetraploid ($2n = 4x = 28$), with a basic chromosome number of $x = 7$ (Dowrick, 1966). Its slender stems trail along the ground or form low arches and bear short, weak, curved prickles. The leaves have three leaflets, of which the lateral two are stalkless (Figure 1). White flowers, 2–3 cm across, appear from June to August. The blue-black fruits consist of a few large drupelets covered by a distinctive whitish bloom (Harrap, 2025).

R. caesius is native to Britain and Ireland, and is locally common, particularly on boulder clay and other alkaline soils. It grows in hedgerows, roadside verges, ditches, woodland rides, scrub, rough grassland and sand dunes, principally in lowland areas (Harrap, 2025).

Figure 1. Photograph of *R. caesius* by David Perez.

Methods

Sample acquisition, flow cytometry and DNA barcoding

A specimen of *R. caesius* (specimen ID KDTOL10675, ToLID drRubCaes1) was used for genome sequencing. It was collected from Wicken Fen, Ely, Cambridgeshire, UK (latitude 52.3113, longitude 0.2899) on 2022-09-16. The specimen was collected by Maarten Christenhusz, Jose Ignacio Marquez-Corro and Isabel Larridon and identified by Maarten Christenhusz. The herbarium voucher for the sequenced plant (collector number MC9442) is deposited in the Herbarium, Royal Botanic Gardens, Kew (K), as specimen K001527036.

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

High molecular weight (HMW) DNA extraction protocols developed by the Tree of Life Core Laboratory at the Wellcome Sanger Institute (WSI) are available on protocols.io (Howard *et al.*, 2025). Tissue from the drRubCaes1 sample was weighed and triaged to select the appropriate extraction workflow. HMW DNA was prepared for one PacBio HiFi library: DTOL13456446, for which 67.0 mg of leaf tissue was homogenised by cryoPREP. HMW DNA was extracted using the Automated Plant MagAttract v2 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 following the Manual SPRI protocol.

The concentration of the sheared and purified DNA was assessed with a NanoDrop spectrophotometer and a Qubit Fluorometer using the Qubit dsDNA High Sensitivity Assay kit. Fragment size distribution was evaluated by running the sample on the Femto Pulse system. For this sample, final post-shearing DNA measurements included a Qubit concentration of 11.63 ng/μL, a yield of 523.35 ng, and an average fragment size of 10.6 kb. The 260/280 spectrophotometric ratio was 1.85, and the 260/230 ratio was 1.39. The Genomic Quality Number (GQN) was 4.4.

PacBio HiFi library preparation and sequencing

Library preparation and sequencing took place in the WSI Scientific Operations core. Library DTOL13456446 was prepared with the SMRTbell Prep Kit 3.0 (Pacific Biosciences) according to the manufacturer's instructions. The kit includes reagents for end repair/A-tailing, adapter ligation, post-ligation SMRTbell bead clean-up, and nuclease treatment. Size selection and clean-up were performed using diluted AMPure PB beads (Pacific Biosciences). DNA concentration was quantified with a Qubit Fluorometer v4.0 (Thermo Fisher Scientific) and the Qubit 1X dsDNA HS assay kit. Final library fragment size was assessed with the Agilent Femto Pulse Automated Pulsed Field CE Instrument (Agilent Technologies) using the gDNA 55 kb BAC analysis kit.

The library was sequenced using the Sequel IIe system (Pacific Biosciences, California, USA). The concentration of the library loaded onto the Sequel system was in the range 40–135 pM. SMRT Link software, a PacBio web-based 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 drRubCaes1 leaf sample using the Arima-HiC v2 kit (Arima Genomics). Tissue was finely ground using the Covaris cryoPREP Dry Pulverizer (Covaris), and then subjected to nuclei isolation. Nuclei were isolated using a modified protocol based on the Qiagen QProteome Cell Compartment Kit (Qiagen), in which only the Lysis and CE2 buffers were used, with QIAshredder spin columns. After isolation, nuclei were fixed using formaldehyde to a final concentration of 2%

to crosslink the DNA. The crosslinked DNA was then digested and biotinylated according to the manufacturer's instructions. A clean-up step was performed with SPRIselect beads before library preparation. DNA concentration was quantified using the Qubit Fluorometer v4.0 (Thermo Fisher Scientific) and the Qubit HS Assay Kit, following the manufacturer's instructions.

Hi-C library preparation and sequencing

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

Prior to sequencing, libraries were normalised to 10 ng/μL. Normalised libraries were quantified again to create equimolar and/or weighted 2.8 nM pools. Pool concentrations were checked using the Agilent 4200 TapeStation (Agilent) with High Sensitivity D500 reagents before sequencing. Libraries were sequenced on the Illumina NovaSeq 6000, generating paired-end 150-bp reads.

Genome assembly

Prior to assembly of the PacBio HiFi reads, a database of *k*-mer counts ($k = 31$) was generated from the filtered reads using FastK. The *k*-mer frequency distribution was analysed with GenomeScope (version 2.1.0) (Ranallo-Benavidez *et al.*, 2020), using a tetraploid ($p = 4$) model to obtain fitted point estimates of the 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 assemblies designated haplotype 1 and haplotype 2. Hi-C reads (Rao *et al.*, 2014) were mapped to contigs using bwa-mem2 (Vasimuddin *et al.*, 2019). Contigs were further scaffolded with Hi-C data in YaHS (Zhou *et al.*, 2023), using the --break option for handling potential misassemblies. The scaffolded assemblies were evaluated using Gfastats (Formenti *et al.*, 2022), BUSCO (Manni *et al.*, 2021) and MerquryFK (Rhie *et al.*, 2020). The organelle genomes were assembled using OATK (Zhou *et al.*, 2025). The comparator for the mitochondrial assembly was the mitochondrial genome of *Rubus chingii* (NC_065238.1).

Assembly curation

The assembly was decontaminated using the Assembly Screen for Cobionts and Contaminants (ASCC) pipeline. The flat files and maps for curation were generated with TreeVal. Manual curation was conducted primarily in PretextView and HiGlass (Kerpedjiev *et al.*, 2018). Scaffolds were visually inspected and corrected as described by Howe *et al.* (2021). The two haplotype assemblies were combined for curation. Manual corrections included 26 breaks, 54 joins, and removal of three haplotypic duplications. This reduced the scaffold count by 6.3% and reduced the scaffold N50 by 44.1%. The curation process is described at sanger-tol/curation-resources. A Hi-C contact map of the haplotype 1 assembly was generated with 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 both haplotypes using the *k*-mer database ($k = 31$) computed prior to genome assembly. The analysis outputs included assembly QV scores and completeness statistics.

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

Genome sequence report

Sequence data

The genome of a specimen of *R. caesius* was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating 19.18 Gb (gigabases) from 1.98 million reads. We assembled the genome from these reads. GenomeScope analysis using a $p = 4$ model gave a *k*-mer-based genome size estimate of 337.59 Mb for one basic chromosome set, with fitted heterozygosity of 6.97% and repeat content of 53.79%; the full-model fit was 95.01% (Figure 2). Using flow cytometry, the haploid genome size (1C-value) of the sample was estimated to be 0.83 pg, equivalent to 810.00 Mb.

The haploid chromosome complement of this tetraploid species contains two basic chromosome sets, giving a *k*-mer-based haploid genome size estimate of 675.18 Mb, approximately 17% below the flow-cytometric estimate (Ranallo-Benavidez *et al.*, 2020).

The NCBI assembly record reports 27× genome coverage. Hi-C sequencing produced 74.69 Gb from 247.30 million read pairs, which were used to scaffold the haplotype 1 assembly. Table 1 summarises the specimen and sequencing details.

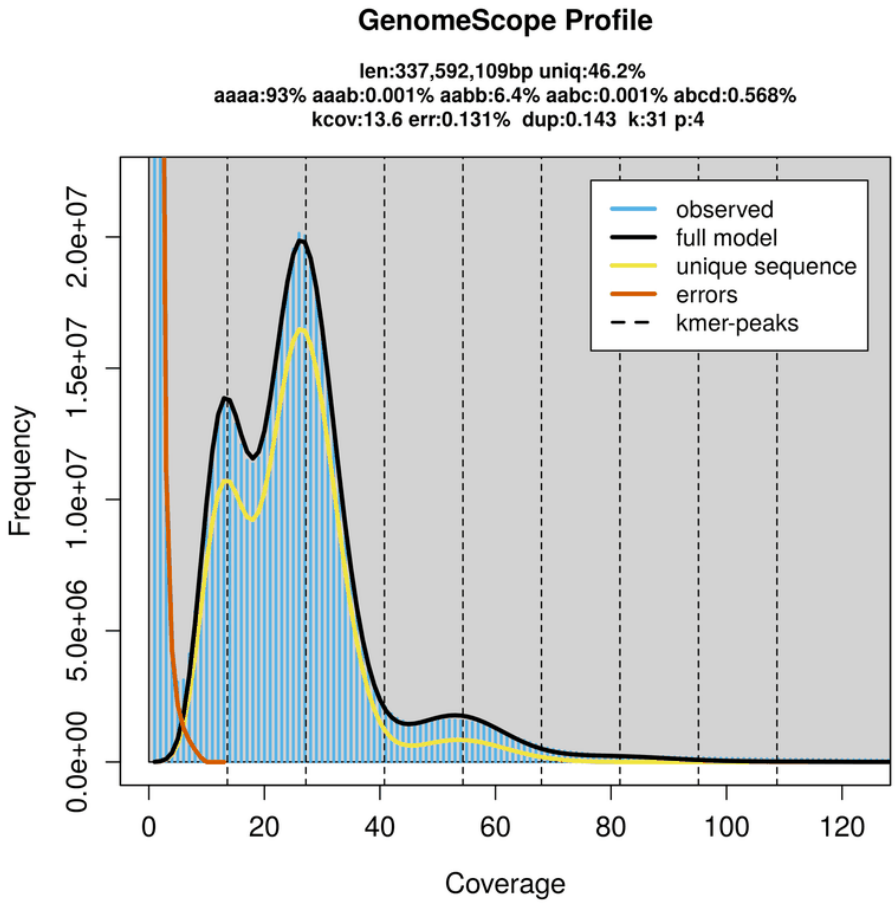

Figure 2. GenomeScope model fitted to the *k*-mer frequency distribution. The plot shows the observed and modelled *k*-mer spectra for the tetraploid ($p = 4$) model. The displayed length (337,592,109 bp) is the estimate for one basic chromosome set, corresponding to a haploid estimate (1C) of approximately 675.18 Mb. Heterozygosity and repeat content were also estimated from the unassembled sequencing reads.

Table 1. Specimen and sequencing data for *R. caesius* (BioProject PRJEB74720)

Platform	PacBio HiFi	Hi-C
ToLID	drRubCaes1	drRubCaes1

Platform	PacBio HiFi	Hi-C
Specimen ID	KDTOL10675	KDTOL10675
BioSample (source individual)	SAMEA112287521	SAMEA112287521
BioSample (tissue)	SAMEA112287602	SAMEA112287602
Tissue	leaf	leaf
Instrument	Sequel IIe (1 run)	Illumina NovaSeq 6000
Library ID & run accession(s)	DTOL13456446: ERR12875186	ERR12893042
Read count total	1.98 million reads	247.30 million read pairs
Base count total	19.18 Gb	74.69 Gb

Assembly statistics

Hi-C phasing produced two assemblies, designated haplotype 1 and haplotype 2 in the genome submissions. The haplotype 1 assembly was curated to chromosome level, while haplotype 2 was assembled to scaffold level. The haplotype 1 assembly, which is proposed as the reference assembly for this species, has a total length of 628.60 Mb in 428 scaffolds (excluding organelle sequences), with 413 gaps, and a scaffold N50 of 42.94 Mb (Table 2). The haplotype 1 assembly has a scaffold auN of 43.4 Mb.

The assembly spans of 628.60 and 633.30 Mb are approximately 6–7% shorter than the ploidy-adjusted *k*-mer 1C estimate of 675.18 Mb and approximately 22% shorter than the flow-cytometric 1C estimate of 810.00 Mb.

Table 2. Genome assembly statistics for *R. caesius*

Genome assembly	Haplotype 1	Haplotype 2
Assembly name	drRubCaes1.hap1.1	drRubCaes1.hap2.1
Assembly accession	GCA_964235055.1	GCA_964235075.1
Assembly level	chromosome	scaffold
Span (Mb)	628.60	633.30
Number of chromosomes	14	-
Number of contigs	841	665
Contig N50 (Mb)	2.3	2.22
Number of scaffolds (excluding organelle sequences)	428	295
Scaffold N50 (Mb)	42.94	76.24
Scaffold auN (Mb)	43.4	66.9
Longest scaffold length (Mb)	59.5	94.85
Organelles	Mitochondrial genome: 423.35 kb; Plastid genome: 156.98 kb	-

Most of the haplotype 1 assembly sequence (97.18%) was assigned to 14 chromosome-level scaffolds, consistent with the haploid chromosome complement ($n = 2x = 14$) of this tetraploid species. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 3; Table 3). Region 15 - 18.5 Mbp in Chromosome 8 is highly repetitive and the order and orientation may be incertain..

The mitochondrial genome (length 423.35 kb, OZ174144.1) and plastid genome (length 156.98 kb, OZ174145.1) were also assembled. These sequences are included as contigs in the multifasta file of the haplotype 1 genome submission and as standalone records.

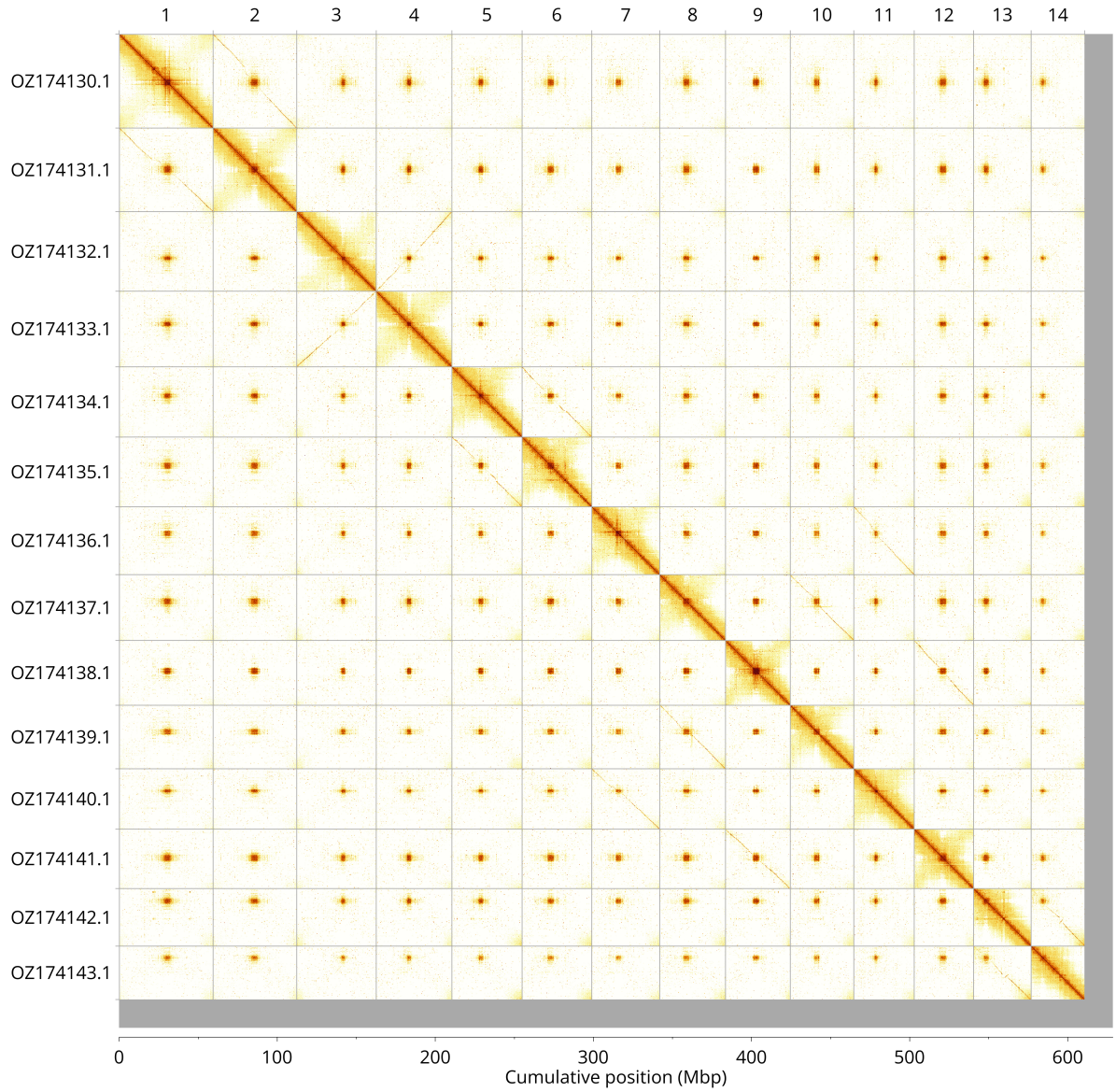

Figure 3. Hi-C contact map of the *R. caesius* haplotype 1 assembly. Assembled chromosomes are shown in order of size and labelled along the axes, with a megabase scale shown below. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in the haplotype 1 genome assembly of *R. caesius* drRubCaes1

INSDC accession	Molecule	Length (Mb)	GC%
OZ174130.1	1	59.50	37.50
OZ174131.1	2	52.87	37.50
OZ174132.1	3	50.34	37
OZ174133.1	4	47.88	37.50
OZ174134.1	5	44.28	37.50
OZ174135.1	6	44.17	37.50
OZ174136.1	7	42.94	37.50

INSDC accession	Molecule	Length (Mb)	GC%
OZ174137.1	8	41.68	37.50
OZ174138.1	9	40.91	37.50
OZ174139.1	10	40.20	37
OZ174140.1	11	38.20	37.50
OZ174141.1	12	37.58	37.50
OZ174142.1	13	36.39	37.50
OZ174143.1	14	33.90	38

Assembly quality metrics

For the haplotype 1 assembly, the estimated QV is 61.5, and for the haplotype 2 assembly, 61.7. When the two haplotypes are combined, the assembly achieves an estimated QV of 61.6. The *k*-mer completeness is 85.01% for the haplotype 1 assembly, 85.37% for the haplotype 2 assembly, and 98.63% for the combined haplotypes (Figure 4).

BUSCO analysis using the eudicots_odb10 reference set (n = 2,326) identified 98.6% of the expected gene set (single = 8.1%, duplicated = 90.5%) in the haplotype 1 assembly. For the haplotype 2 assembly, BUSCO v.5.5.0 analysis identified 98.9% of the expected gene set (single = 9.2%, duplicated = 89.7%).

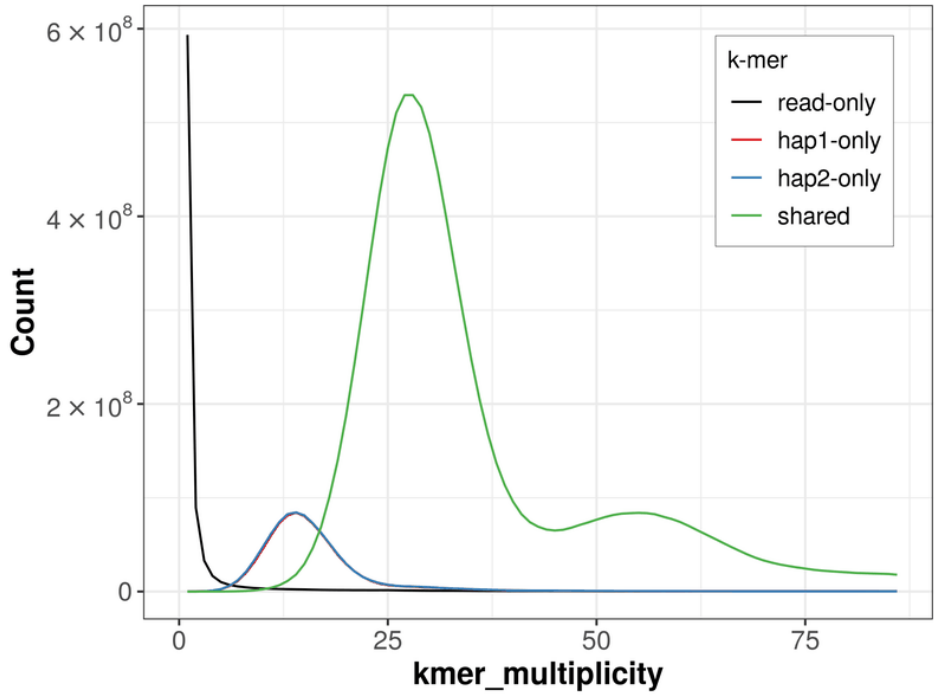

Figure 4. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the haplotype 1 assembly, the haplotype 2 assembly and the combined haplotypes. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows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

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

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

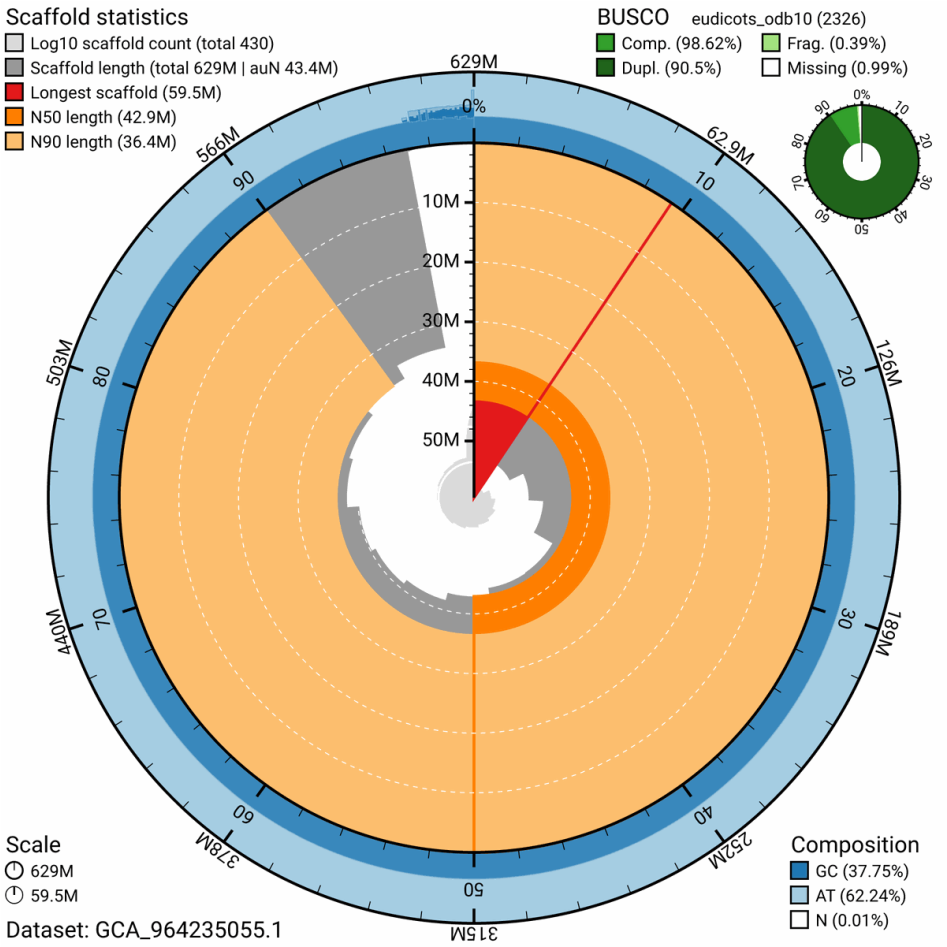

Figure 5. Assembly metrics for drRubCaes1.hap1.1, the haplotype 1 assembly. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the haplotype 1 assembly, and the main plot is divided into 1,000 bins around the circumference. The outermost blue tracks display the distribution of GC, AT, and N percentages across the bins. Scaffolds are arranged clockwise from longest to shortest and are depicted in dark grey. Scaffold length is shown on a linear radial scale, scaled to the longest scaffold. Red shading indicates the longest scaffold, and the deeper orange and pale orange shading represent the N50 and N90 lengths. The summary statistics include scaffold auN, the area under the scaffold Nx curve. The light grey spiral at the centre shows the cumulative scaffold count on a logarithmic scale. A summary of complete, fragmented, duplicated, and missing BUSCO genes in the eudicots_odb10 set is presented at the top right. The assembly can be explored interactively in the BlobToolKit viewer.

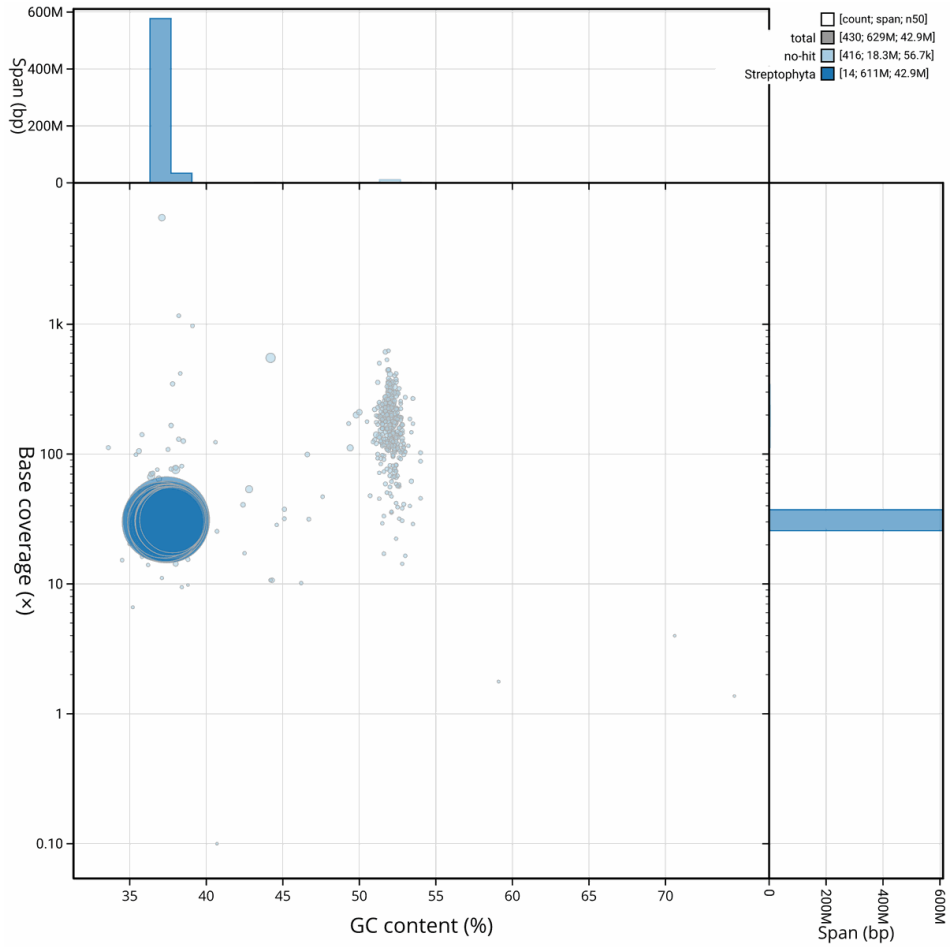

Figure 6. BlobToolKit blob plot for drRubCaes1.hap1.1, the haplotype 1 assembly. The plot shows base coverage (vertical axis) and GC content (horizontal axis). Circles represent scaffolds; marker area is proportional to scaffold length, and colour indicates taxonomic assignment. The histograms along the axes display the total length of sequences distributed across different levels of coverage and GC content. The assembly can be explored interactively in the BlobToolKit viewer.

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

Table 4. Earth BioGenome Project summary metrics for the *R. caesius* assembly

Measure	Value	*Benchmark*
EBP summary (haplotype 1)	6.C.Q61	*6.C.Q40*
Contig N50 length	2.30 Mb	*> 1 Mb*
Scaffold N50 length	42.94 Mb	*= chromosome N50*
Consensus quality (QV)	Haplotype 1: 61.5; haplotype 2: 61.7; combined: 61.6	*# 40*
k-mer completeness	Haplotype 1: 85.01%; Haplotype 2: 85.37%; combined: 98.63%	*> 90%*
BUSCO	Haplotype 1: C:98.6% [S:8.1%, D:90.5%], F:0.4%, M:1.0%, n:2,326;	*S > 90%; D < 5%*

Measure	Value	*Benchmark*
	Haplotype 2: C:98.9% [S:9.2%, D:89.7%], F:0.3%, M:0.8%, n:2,326	
Percentage of assembly assigned to chromosomes	97.18%	*# 90%*

Notes: The EBP summary uses log10(Contig N50); chromosome-level (C) or log10(Scaffold N50); Q (Merqury QV). BUSCO v5.5.0 using the eudicots_odb10 lineage dataset: C=complete; S=single-copy; D=duplicated; F=fragmented; M=missing; n=orthologues.

Data availability

European Nucleotide Archive: Rubus caesius. Accession number PRJEB74720; https://identifiers.org/ena.embl/PRJEB74720. The genome sequence is released openly for reuse. The *R. caesius* genome sequencing initiative is part of the Darwin Tree of Life Project (PRJEB40665) and the Sanger Institute Tree of Life Programme (PRJEB43745). All raw sequence data and the haplotype assemblies have been deposited in INSDC databases. The genome will be annotated using available RNA-Seq data and presented through the Ensembl pipeline at the European Bioinformatics Institute. Raw data and assembly accession identifiers are reported in Table 1 and Table 2.

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

Table 5. Software versions and sources

Software	Version	Source
BLAST	2.14.0	https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/
BlobToolKit	4.3.9	https://github.com/genomehubs/blobtoolkit
BlobTk (hosted BlobDir)	0.5.1	https://github.com/genomehubs/blobtk
BlobTk (snail and blob plot renderer)	0.8.3	https://github.com/genomehubs/blobtk
BUSCO	5.5.0; 6.1.0	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	1.2	https://github.com/thegenemyers/FASTK
GenomeScope	2.1.0	https://github.com/tbenavi1/genomescope2.0
Gfastats	1.3.6	https://github.com/vgl-hub/gfastats
Hifiasm	0.19.8-r603	https://github.com/chhylp123/hifiasm
HiGlass	1.13.4	https://github.com/higlass/higlass
MerquryFK	1.1.0-c1	https://github.com/thegenemyers/MERQURY.FK
Minimap2	2.24-r1122	https://github.com/lh3/minimap2
Oatk	0.9	https://github.com/c-zhou/oatk
MultiQC	1.14; 1.17 and 1.18	https://github.com/MultiQC/MultiQC
Nextflow	23.10.0	https://github.com/nextflow-io/nextflow
PretextSnapshot	0.0.6	https://github.com/sanger-tol/PretextSnapshot

Software	Version	Source
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	0.6.0	https://github.com/sanger-tol/blobtoolkit
sanger-tol/curationpretext	1.4.2	https://github.com/sanger-tol/curationpretext
Seqtk	1.4-r122	https://github.com/lh3/seqtk
Singularity	3.9.0	https://github.com/sylabs/singularity
TreeVal	1.0.2; 1.1.0	https://github.com/sanger-tol/treeval
YaHS	1.2a.2	https://github.com/c-zhou/yahs

Author information

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

Wellcome Sanger Institute – Legal and Governance

The materials that have contributed to this genome note have been supplied by a Darwin Tree of Life Partner. The submission of materials by a Darwin Tree of Life Partner is subject to the **'Darwin Tree of Life Project Sampling Code of Practice'**, which can be found in full on the Darwin Tree of Life website. By agreeing with and signing up to the Sampling Code of Practice, the Darwin Tree of Life Partner agrees they will meet the legal and ethical requirements and standards set out within this document in respect of all samples acquired for, and supplied to, the Darwin Tree of Life Project. Further, the Wellcome Sanger Institute employs a process whereby due diligence is carried out proportionate to the nature of the materials themselves, and the circumstances under which they have been/are to be collected and provided for use. The purpose of this is to address and mitigate any potential legal and/or ethical implications of receipt and use of the materials as part of the research project, and to ensure that in doing so we align with best practice wherever possible. The overarching areas of consideration are:

- Ethical review of provenance and sourcing of the material
- Legality of collection, transfer and use (national and international)

Each transfer of samples is further undertaken according to a Research Collaboration Agreement or Material Transfer Agreement entered into by the Darwin Tree of Life Partner, Genome Research Limited (operating as the Wellcome Sanger Institute), and in some circumstances, other Darwin Tree of Life collaborators.

References

Altschul S. F., Gish W., Miller W.: **Basic Local Alignment Search Tool**. *Journal of Molecular Biology*; 1990; **215**(3): 403–410. PubMed Abstract | Publisher Full Text |

Buchfink B., Reuter K., Drost H.-G.: **Sensitive protein alignments at tree-of-life scale using DIAMOND**. *Nature Methods*; 2021; **18**(4): 366–368. PubMed Abstract | Publisher Full Text |

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

Challis Richard, Blaxter Mark: **Snail plots are badges of genome assembly quality**. *G3 Genes|Genomes|Genetics*; 2026; **16**(6): jkag074. PubMed Abstract | Publisher Full Text |

Chase M. W., Hills H. H.: **Silica gel: An ideal material for field preservation of leaf samples for DNA studies**. *Taxon*; 1991; **40**(2): 215–220. Publisher Full Text |

Cheng H., Jarvis E. D., Fedrigo O.: **Haplotype-resolved assembly of diploid genomes without parental data**. *Nature Biotechnology*; 2022; **40**(9): 1332–1335. PubMed Abstract | Publisher Full Text |

Cheng H., Concepcion G. T., Feng X.: **Haplotype-resolved *de novo* assembly using phased assembly graphs with Hifiasm**. *Nature Methods*; 2021; **18**(2): 170–175. PubMed Abstract | Publisher Full Text |

Dowrick G. J.: **Breeding systems in tetraploid Rubus species**. *Genetics Research*; 1966; **7**(2): 245–253. Publisher Full Text |

Formenti G., Abueg L., Brajuka A.: **Gfastats: Conversion, evaluation and manipulation of genome sequences using assembly graphs**. *Bioinformatics*; 2022; **38**(17): 4214–4216. PubMed Abstract | Publisher Full Text |

Harrap Simon: *Harrap's wild flowers: A field guide to the wild flowers of Britain & Ireland*. Bloomsbury Publishing Plc; 2025. 2 ed.

Howard Caroline, Denton Amy, Jackson Benjamin W.: **On the path to reference genomes for all biodiversity: Laboratory protocols and lessons learned from processing over 2,000 species in the Sanger Tree of Life**. *GigaScience*; 2025; **14** giaf119. Reference Source Publisher Full Text |

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

Kerpedjiev P., Abdennur N., Lekschas F.: **HiGlass: Web-based visual exploration and analysis of genome interaction maps**. *Genome Biology*; 2018; **19**(1): 125. PubMed Abstract | Publisher Full Text |

Kurtzer G. M., Sochat V., Bauer M. W.: **Singularity: Scientific containers for mobility of compute**. *PLOS ONE*; 2017; **12**(5): e0177459. PubMed Abstract | Publisher Full Text |

Loureiro J., Rodriguez E., Doležel J.: **Two new nuclear isolation buffers for plant DNA flow cytometry: A test with 37 species**. *Annals of Botany*; 2007; **100**(4): 875–888. PubMed Abstract | Publisher Full Text |

Manni Mosè, Berkeley Matthew R, Seppey Mathieu: **BUSCO update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes**. *Molecular Biology and Evolution*; 2021; **38**(10): 4647–4654. PubMed Abstract | Publisher Full Text |

Obermayer R., Leitch I. J., Hanson L.: **Nuclear DNA C-values in 30 species double the familial representation in pteridophytes**. *Annals of Botany*; 2002; **90**(2): 209–217. PubMed Abstract | Publisher Full Text |

Pellicer J., Powell R. F., Leitch I. J.: **The application of flow cytometry for estimating genome size, ploidy level endopolyploidy, and reproductive modes in plants**. *Methods in Molecular Biology* Besse P.: New York, NY. 2021. **2222**: 325–361. PubMed Abstract |

Ranallo-Benavidez Timothy R., Jaron Kai S., Schatz Michael C.: **GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes**. *Nature Communications*; 2020; **11**(1): 1432. PubMed Abstract | Publisher Full Text |

Rao S. S. P., Huntley M. H., Durand N. C.: **A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping**. *Cell*; 2014; **159**(7): 1665–1680. PubMed Abstract | Publisher Full Text |

Ratnasingham S., Hebert P. D. N.: **BOLD: The Barcode of Life Data System (http://www.barcodinglife.org)**. *Molecular Ecology Notes*; 2007; **7**(3): 355–364. PubMed Abstract | Publisher Full Text |

Rhie A., Walenz B. P., Koren S.: **Merqury: Reference-free quality, completeness, and phasing assessment for genome assemblies**. *Genome Biology*; 2020; **21**(1): 245. PubMed Abstract | Publisher Full Text |

Rhie A., McCarthy S. A., Fedrigo O.: **Towards complete and error-free genome assemblies of all vertebrate species**. *Nature*; 2021; **592**(7856): 737–746. PubMed Abstract | Publisher Full Text |

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

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

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

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

Open Peer Review

Current Peer Review Status:

Version 1

Reviewer Report 01 January 2000

https://doi.org/

Automated Benchmarking
Wellcome Sanger Institute, Hinxton, UK

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

For Rubus caesius (drRubCaes1), all four checks passed.

Data available: pass. ENA study PRJEB74720, raw-read accessions ERR12875186 and ERR12893042, and assembly accessions GCA_964235055.1 and GCA_964235075.1 are public.

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

k-mer completeness: pass. The combined haplotype assemblies have 98.63% k-mer completeness, exceeding the Earth BioGenome Project target of 90%.

BUSCO: pass. Haplotype 1, which is proposed as the reference assembly for this species, has 98.6% complete BUSCOs (8.1% single-copy and 90.5% duplicated). BUSCO software version: 5.5.0; lineage dataset: eudicots_odb10. It meets the plant-specific target of greater than 90% complete BUSCOs.

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

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

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.