



DATA NOTE

The genome sequence of petty spurge, *Euphorbia peplus* L. (Euphorbiaceae)

[version 1; peer review: 3 approved]

Maarten J. M. Christenhusz ^{1,2}, Michael F. Fay ¹, Ilia J. Leitch ¹,
Royal Botanic Gardens Kew Genome Acquisition Lab,
Darwin Tree of Life Barcoding collective, Plant Genome Sizing collective,
Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory
team,
Wellcome Sanger Institute Scientific Operations: Sequencing Operations,
Wellcome Sanger Institute Tree of Life Core Informatics team,
Tree of Life Core Informatics collective, Darwin Tree of Life Consortium

¹Royal Botanic Gardens Kew, Richmond, England, UK²Curtin University, Perth, Western Australia, Australia

V1 First published: 25 Apr 2025, 10:219
<https://doi.org/10.12688/wellcomeopenres.24030.1>

Latest published: 25 Apr 2025, 10:219
<https://doi.org/10.12688/wellcomeopenres.24030.1>

Abstract

We present a genome assembly from a specimen of *Euphorbia peplus* (petty spurge; Streptophyta; Magnoliopsida; Malpighiales; Euphorbiaceae). The genome sequence has a total length of 277.10 megabases. Most of the assembly is scaffolded into 8 chromosomal pseudomolecules. We also assembled six multipartite mitochondrial molecules and one plastid genome.

Keywords

Euphorbia peplus, petty spurge, genome sequence, chromosomal, Malpighiales



This article is included in the [Tree of Life gateway](#).

Open Peer Review

Approval Status

	1	2	3
version 1			
25 Apr 2025	view	view	view

1. **Thomas Brazier** , University of Rennes, Rennes, France
2. **Quentin Cronk** , University of British Columbia, Vancouver, Canada
Kaede Hirabayashi, The University of British Columbia, Vancouver, Canada
3. **Bimal K Chetri** , Indian Institute of Technology, Guwahati, India

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

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

Author roles: **Christenhusz MJM:** Investigation, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Fay MF:** Writing – Original Draft Preparation; **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, <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.

Copyright: © 2025 Christenhusz MJM *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Christenhusz MJM, Fay MF, Leitch IJ *et al.* **The genome sequence of petty spurge, *Euphorbia peplus* L. (Euphorbiaceae) [version 1; peer review: 3 approved]** Wellcome Open Research 2025, 10:219 <https://doi.org/10.12688/wellcomeopenres.24030.1>

First published: 25 Apr 2025, 10:219 <https://doi.org/10.12688/wellcomeopenres.24030.1>

Species taxonomy

Eukaryota; Viridiplantae; Streptophyta; Streptophytina; Embryophyta; Tracheophyta; Euphyllophyta; Spermatophyta; Magnoliopsida; Mesangiospermae; eudicotyledons; Gunneridae; Pentapetales; rosids; fabids; Malpighiales; Euphorbiaceae; Euphorbioideae; Euphorbieae; *Euphorbia*; *Euphorbia* subgen. *Esula*; *Euphorbia* sect. *Tithymalus*; *Euphorbia peplus* L. (NCBI: txid38846)

Background

Petty spurge (*Euphorbia peplus*, Euphorbiaceae) is an annual, weedy plant, commonly found in arable land, gardens, along paths, pavements and waste places. It prefers well-drained, nutrient-rich soils on sunny, warm locations. While it can grow up to 30 cm tall, it is usually much smaller, especially in exposed sites. The entire plant is pale green with lanceolate to ovate, pointed leaves and three-rayed inflorescences (Figure 1). Like other spurges, the flowers are minute and reduced, but several male flowers and a single female flower are surrounded by fused bracts called a cyathium, which are topped by glands. The glands of this species are kidney-shaped with two thin horns.

In the UK, this species is an archaeophyte (e.g. Stace, 2010), introduced from southern Europe in ancient times as a medicinal plant. It is now widely naturalised outside its native range and is one of the most common garden weeds across the temperate regions of the world, even reaching remote Easter Island (POWO, 2024; Zizka, 1991).

When damaged, the plant emits a milky latex, which is toxic to rapidly replicating cells. In humans this has traditionally been used for treating common skin lesions, including those caused by exposure to UV light from the sun, and basal cell carcinomas. The active ingredient, the diterpene ester ingenol mebutate, which is cytotoxic and is currently used medicinally for the treatment of squamous cell skin cancers and



Figure 1. Photograph of the *Euphorbia peplus* (ddEupPepu3) from which samples were taken for genome sequencing.

actinic keratoses (Braun *et al.*, 2014; Siller *et al.*, 2009). In preclinical trials, the data suggest that ingenol mebutate may also be effective in treating other types of cancers such as pancreatic, colorectal and epithelial cancers (reviewed in Shafombabi *et al.* 2025). However, there are safety concerns due to its cytotoxicity, so ongoing research is exploring how to minimize these risks, including the exploration of advanced delivery methods to enable its potential as an anticancer treatment to be expanded.

Despite a wide diversity of chromosome numbers reported for the genus *Euphorbia* ($2n = 12$ to ~ 200 ; Hans, 1973; Perry, 1943) arising from polyploidy, aneuploidy and structural rearrangements, *E. peplus* has so far only been reported to be diploid with $2n = 2x = 16$, based on a single chromosome count reported from the UK (Al-Bermani *et al.*, 1993; Perry, 1943) and several counts from mainland Europe (e.g. see Hans, 1973).

For the Darwin Tree of Life project, plants were collected from flower beds at the Royal Botanic Gardens, Kew. During the COVID-19 pandemic lockdowns, the beds had not been weeded, and petty spurge grew there in abundance. Here we present a high-quality genome of *E. peplus*. It complements the previously published chromosome level genome assembly for the same species, but sourced from a wild-grown individual plant in the Cornell Botanical Gardens, USA (Johnson *et al.*, 2023). Both resources, together with a chromosome level genome assembly of a related species, *E. lathyris* (Wang *et al.*, 2021), will be useful for enabling further studies into its medicinal properties and latex production.

Genome sequence report

Sequencing data

The genome of a specimen of *Euphorbia peplus* was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating 22.40 Gb (gigabases) from 1.57 million reads. GenomeScope analysis of the PacBio HiFi data estimated the haploid genome size at 355.40 Mb, with a heterozygosity of 0.06% and repeat content of 46.13%. These values provide an initial assessment of genome complexity and the challenges anticipated during assembly. Based on this estimated genome size, the sequencing data provided approximately 60.0x coverage of the genome. Using flow cytometry, the genome size (1C-value) was estimated to be 0.4 pg, equivalent to 390 Mb. Hi-C sequencing produced 98.03 Gb from 649.22 million reads. Table 1 summarises the specimen and sequencing information.

Assembly statistics

The primary haplotype was assembled, and contigs corresponding to an alternate haplotype were also deposited in INSDC databases. The assembly was improved by manual curation, which corrected 5 misjoins or missing joins. The interventions increased the scaffold count by 90.91% and decreased the scaffold N50 by 46.68%. The final assembly has a total length of 277.12 Mb in 14 scaffolds, with 5 gaps, and a scaffold N50 of 33.77 Mb (Table 2).

Table 1. Specimen and sequencing data for *Euphorbia peplus*.

Project information			
Study title	Euphorbia peplus		
Umbrella BioProject	PRJEB52211		
Species	<i>Euphorbia peplus</i>		
BioSample	SAMEA7521821		
NCBI taxonomy ID	38846		
Specimen information			
Technology	ToLID	BioSample accession	Organism part
PacBio long read sequencing	ddEupPepu3	SAMEA7521875	leaf
Hi-C sequencing	ddEupPepu3	SAMEA7521873	leaf
RNA sequencing	ddEupPepu3	SAMEA7521875	leaf
Sequencing information			
Platform	Run accession	Read count	Base count (Gb)
Hi-C Illumina NovaSeq 6000	ERR9580485	6.49e+08	98.03
PacBio Sequel IIe	ERR9630945	1.57e+06	22.4
RNA Illumina HiSeq 4000	ERR9580484	4.47e+07	6.74

The snail plot in [Figure 2](#) provides a summary of the assembly statistics, indicating the distribution of scaffold lengths and other assembly metrics. [Figure 3](#) shows the distribution of scaffolds by GC proportion and coverage. [Figure 4](#) presents a cumulative assembly plot, with separate curves representing different scaffold subsets assigned to various phyla, illustrating the completeness of the assembly.

Most of the assembly sequence (99.54%) was assigned to 8 chromosomal-level scaffolds. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size ([Figure 5](#); [Table 3](#)).

The mitochondrial and plastid genomes were also assembled. These sequences are included as contigs in the multifasta file of the genome submission and as standalone records in GenBank.

Assembly quality metrics

The estimated Quality Value (QV) and *k*-mer completeness metrics, along with BUSCO completeness scores, were calculated for each haplotype and the combined assembly. The QV reflects the base-level accuracy of the assembly, while *k*-mer completeness indicates the proportion of expected *k*-mers identified in the assembly. BUSCO scores provide a measure of completeness based on benchmarking universal single-copy orthologues.

The primary haplotype has a QV of 54.2, and the combined primary and alternate assemblies achieve an estimated QV of

53.9. The *k*-mer recovery for the primary haplotype is 97.97%, and for the alternate haplotype 1.18%; the combined primary and alternate assemblies have a *k*-mer recovery of 98.49%. BUSCO analysis using the eudicots_odb10 reference set (*n* = 2,326) identified 97.6% of the expected gene set (single = 94.4%, duplicated = 3.2%).

[Table 2](#) provides assembly metric benchmarks adapted from [Rhie et al. \(2021\)](#) and the Earth BioGenome Project (EBP) Report on Assembly Standards [September 2024](#). The primary assembly achieves the EBP reference standard of **7.C.Q54**.

Methods

Sample acquisition, DNA barcoding and genome size estimation

A sample of *Euphorbia peplus* (specimen ID KDTOL10012, ToLID ddEupPepu3) was collected from the Royal Botanic Gardens, Kew, Surrey, UK (latitude 51.47, longitude -0.30) on 2020-08-05. The specimen was collected and identified by Maarten J. M. Christenhusz (collection number MC9009) and preserved by freezing at -80 °C. The herbarium voucher associated with the sequenced plant (K001400680) is deposited in the herbarium of RBG Kew (K).

The initial species identification was verified by an additional DNA barcoding process following the framework developed by [Twyford et al. \(2024\)](#). Part of the plant specimen was preserved in silica gel desiccant ([Chase & Hills, 1991](#)). DNA was extracted from the dried specimen, then PCR was used

Table 2. Genome assembly data for *Euphorbia peplus*, ddEupPepu3.1

Genome assembly		
Assembly name	ddEupPepu3.1	
Assembly accession	GCA_964200815.1	
Alternate haplotype accession	GCA_964200835.1	
Assembly level for primary assembly	chromosome	
Span (Mb)	277.12	
Number of contigs	19	
Number of scaffolds	14	
Longest scaffold (Mb)	41.2	
Assembly metric	Measure	Benchmark
Contig N50 length	27.11 Mb	≥ 1 Mb
Scaffold N50 length	33.77 Mb	= chromosome N50
Consensus quality (QV)	Primary: 54.2; alternate: 47.4; combined: 53.9	≥ 40
<i>k</i> -mer completeness	Primary: 97.97%; alternate: 1.18%; combined: 98.49%	$\geq 95\%$
BUSCO*	C:97.6%[S:94.4%,D:3.2%], F:0.5%,M:1.9%,n:2,326	$S > 90\%$; $D < 5\%$
Percentage of assembly mapped to chromosomes	99.54%	$\geq 90\%$
Organelles	Mitochondrial genome: six separate molecules; Plastid genome: 159.6 kb	complete single alleles

* Assembly metric benchmarks are adapted from [Rhie et al. \(2021\)](#) and the Earth BioGenome Project Report on Assembly Standards [September 2024](#).

** BUSCO scores based on the eudicotyledons_odb10 BUSCO set using version 5.4.3. C = complete [S = single copy, D = duplicated], F = fragmented, M = missing, n = number of orthologues in comparison. A full set of BUSCO scores is available at https://blobtoolkit.genomehubs.org/view/Euphorbia_peplus/dataset/GCA_964200815.1/busco.

to amplify standard barcode regions. The resulting amplicons were sequenced and compared to public sequence databases including GenBank and the Barcode of Life Database (BOLD). The barcode sequences for this specimen are available on BOLD ([Ratnasingham & Hebert, 2007](#)). Following whole genome sequence generation, DNA barcodes were also used alongside the initial barcoding data for sample tracking through the genome production pipeline at the Wellcome Sanger Institute ([Twyford et al., 2024](#)). The standard operating procedures for the Darwin Tree of Life barcoding have been deposited on protocols.io ([Beasley et al., 2023](#)).

The genome size was estimated by flow cytometry using the fluorochrome propidium iodide and following the ‘one-step’ method as outlined in [Pellicer et al. \(2021\)](#). For this species, the General Purpose Buffer (GPB) supplemented with 3% PVP and 0.08% (v/v) beta-mercaptoethanol was used for isolation of nuclei ([Loureiro et al., 2007](#)), and the internal calibration

standard was *Solanum lycopersicum* ‘Stupiké polní rané’ with an assumed 1C-value of 968 Mb ([Doležel et al., 2007](#)).

Nucleic acid extraction

The workflow for high molecular weight (HMW) DNA extraction at the WSI Tree of Life Core Laboratory includes a sequence of procedures: sample preparation and homogenisation, DNA extraction, fragmentation and purification. Detailed protocols are available on protocols.io ([Denton et al., 2023](#)). In sample preparation, the ddEupPepu3 sample was weighed and dissected on dry ice ([Jay et al., 2023](#)) and leaf tissue was cryogenically disrupted using the Covaris cryoPREP® Automated Dry Pulverizer ([Narváez-Gómez et al., 2023](#)).

HMW DNA was extracted using the Automated Plant MagAttract v2 protocol ([Todorovic et al., 2023](#)). HMW DNA was sheared into an average fragment size of 12–20 kb in a Megaruptor 3 system ([Bates et al., 2023](#)). Sheared DNA was

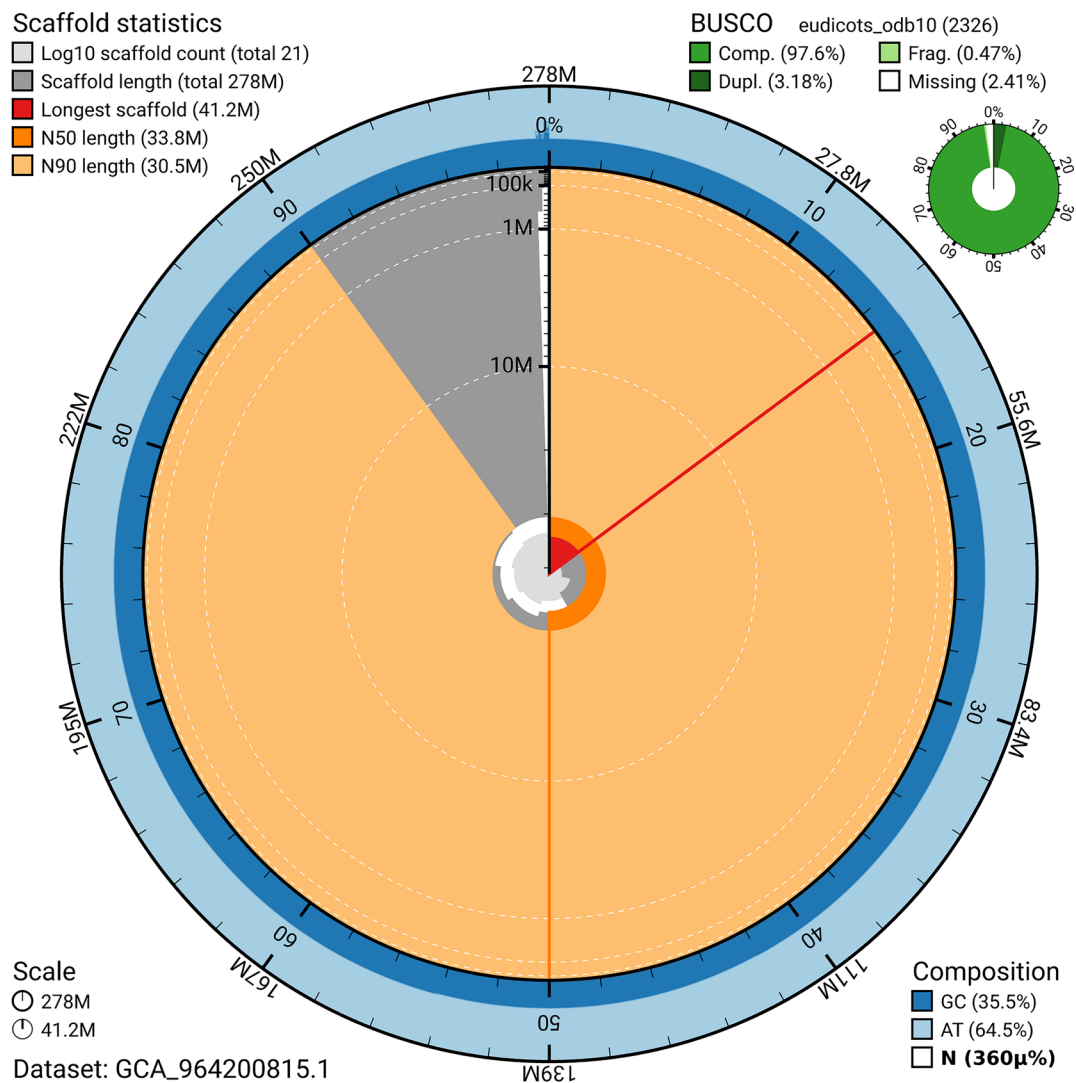


Figure 2. Snail plot summary of assembly statistics for assembly ddEupPepu3.1: metrics. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around 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 eudicotyledons_odb10 set is presented at the top right. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/GCA_964200815.1/dataset/GCA_964200815.1/snail.

purified by solid-phase reversible immobilisation, using AMPure PB beads to eliminate shorter fragments and concentrate the DNA (Oatley *et al.*, 2023). The concentration of the sheared and purified DNA was assessed using a Nanodrop spectrophotometer and Qubit Fluorometer and Qubit dsDNA High Sensitivity Assay kit. Fragment size distribution was evaluated by running the sample on the FemtoPulse system.

Hi-C sample preparation

Hi-C data were generated from leaf tissue of the ddEupPepu3 sample at the WSI Scientific Operations core, using the

Arima-HiC v2 kit. Tissue was finely ground using cryoPREP, and then subjected to nuclei isolation using a modified protocol of the Qiagen QProteome Kit. After isolation, the nuclei were fixed, and the DNA crosslinked using a 37% formaldehyde solution (final concentration 2%). The crosslinked DNA was then digested using the restriction enzyme master mix. The 5'-overhangs were then filled in and labelled with biotinylated nucleotides and proximally ligated. An overnight incubation was carried out for enzymes to digest remaining proteins and for crosslinks to reverse. A clean up was performed with SPRIselect beads prior to library preparation. DNA concentration

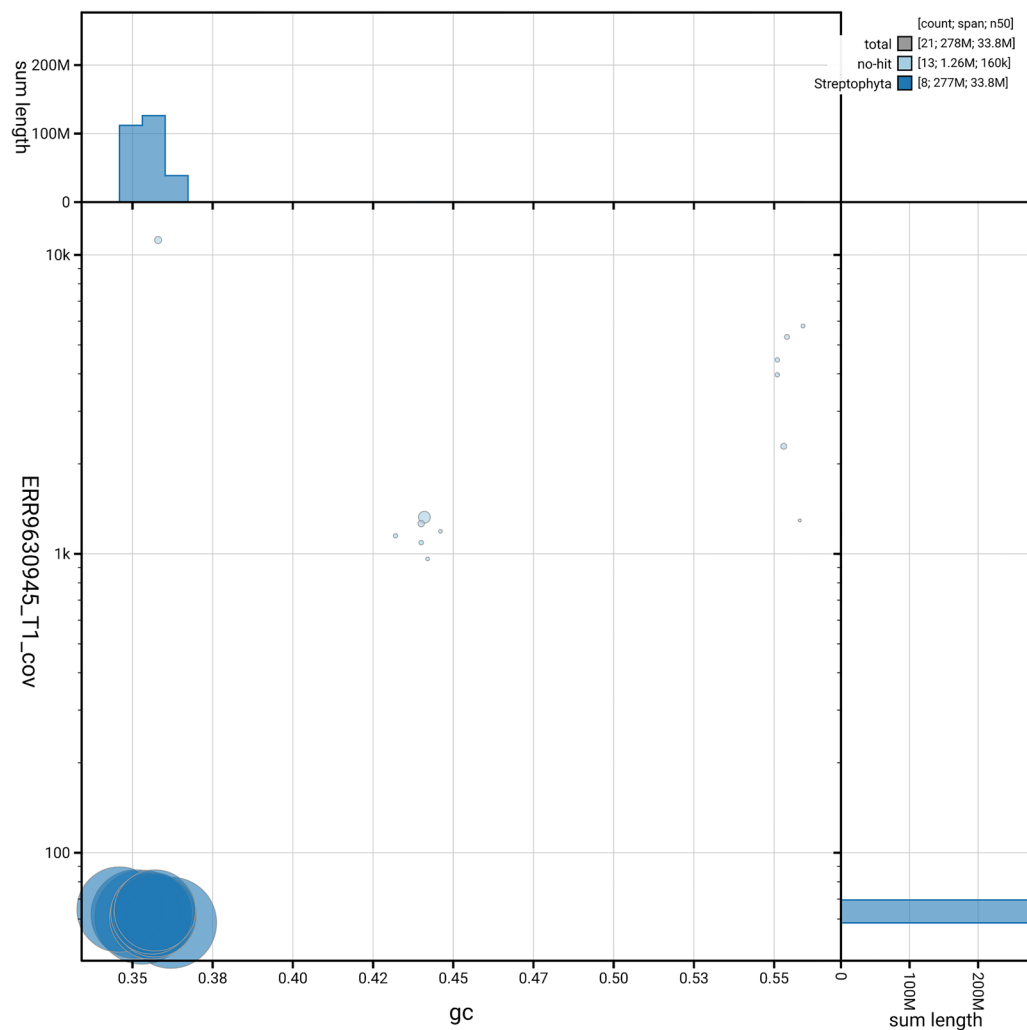


Figure 3. BlobToolKit blob plot for assembly ddEupPepu3.1: BlobToolKit GC-coverage plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes display the total length of sequences distributed across different levels of coverage and GC content. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/GCA_964200815.1/dataset/GCA_964200815.1/blob.

was quantified using the Qubit Fluorometer v2.0 and Qubit HS Assay Kit according to the manufacturer's instructions.

Library preparation and sequencing

Library preparation and sequencing were performed at the WSI Scientific Operations core.

PacBio HiFi

At a minimum, samples were required to have an average fragment size exceeding 8 kb and a total mass over 400 ng to proceed to the low input SMRTbell Prep Kit 3.0 protocol (Pacific Biosciences, California, USA). Libraries were prepared using the SMRTbell Prep Kit 3.0 (Pacific Biosciences, California, USA) as per the manufacturer's instructions. The kit includes the reagents required for end repair/A-tailing, adapter

ligation, post-ligation SMRTbell bead cleanup, and nuclease treatment. Following the manufacturer's instructions, size selection and clean up was carried out using diluted AMPure PB beads (Pacific Biosciences, California, USA). DNA concentration was quantified using the Qubit Fluorometer v4.0 (Thermo Fisher Scientific) with Qubit 1X dsDNA HS assay kit and the final library fragment size analysis was carried out using the Agilent Femto Pulse Automated Pulsed Field CE Instrument (Agilent Technologies) and gDNA 55kb BAC analysis kit.

Samples were sequenced on a Sequel IIe instrument (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

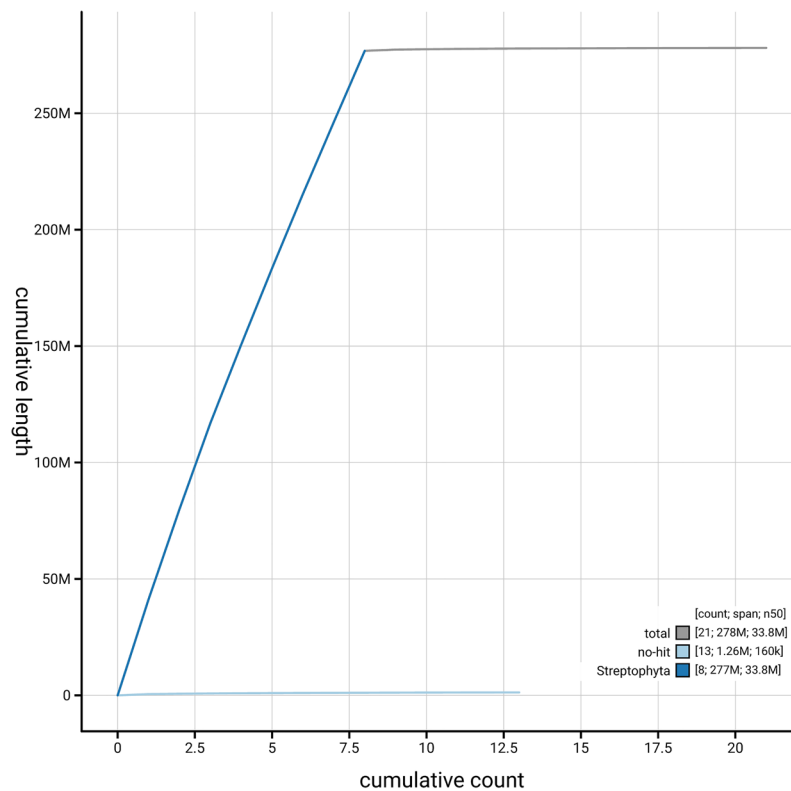


Figure 4. BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/GCA_964200815.1/dataset/GCA_964200815.1/cumulative.

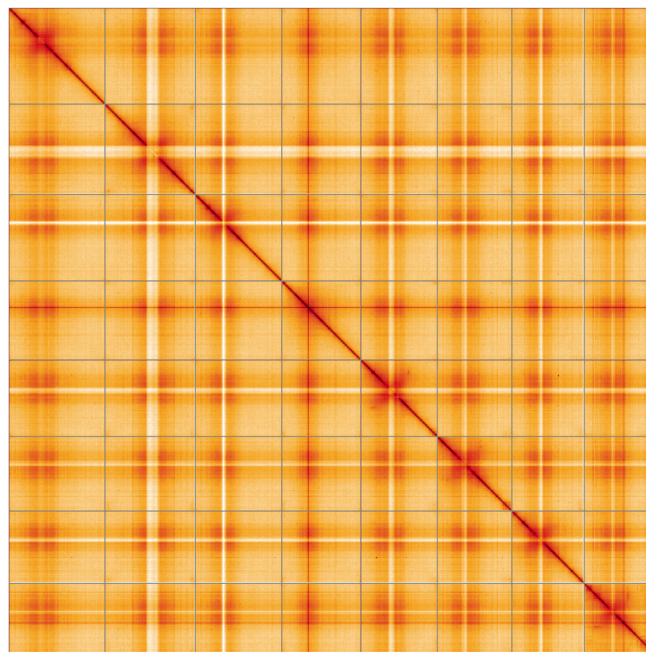


Figure 5. Genome assembly of *Euphorbia peplus*, ddEupPepu3.1: Hi-C contact map of the ddEupPepu3.1 assembly, visualised using HiGlass. Chromosomes are shown in order of size from left to right and top to bottom. Darker shades indicate more frequent physical contacts between regions, while lighter areas represent fewer contacts. An interactive version of this figure may be viewed at https://genome-note-higlass.tol.sanger.ac.uk/l/?d=NjzabIgvS9aw_nejOLH5XQ.

Table 3. Chromosomal pseudomolecules in the genome assembly of *Euphorbia peplus*, ddEupPepu3.

INSDC accession	Name	Length (Mb)	GC%
OZ123146.1	1	41.2	35.5
OZ123147.1	2	38.69	36.0
OZ123148.1	3	36.98	35.0
OZ123149.1	4	33.77	34.5
OZ123150.1	5	32.81	35.5
OZ123151.1	6	31.94	35.5
OZ123152.1	7	30.89	35.5
OZ123153.1	8	30.53	35.5
OZ123160.1	Pltd	0.16	36.0
OZ123154.1	MT1	0.52	44.0
OZ123155.1	MT2	0.03	44.5
OZ123156.1	MT3	0.03	44.5
OZ123157.1	MT4	0.13	44.0
OZ123158.1	MT5	0.04	43.0
OZ123159.1	MT6	0.05	44.0

workflow manager, was used to set-up and monitor the run, as well as perform primary and secondary analysis of the data upon completion.

Hi-C

For Hi-C library preparation, DNA was fragmented to a size of 400 to 600 bp using a Covaris E220 sonicator. The DNA was then enriched, barcoded, and amplified using the NEBNext Ultra II DNA Library Prep Kit (New England Biolabs) following manufacturer's instructions. Hi-C sequencing was performed using paired-end sequencing with a read length of 150 bp on an Illumina NovaSeq 6000 instrument.

Genome assembly, curation and evaluation

Assembly

The HiFi reads were first assembled using Hifiasm (Cheng *et al.*, 2021) with the --primary option. Haplotypic duplications were identified and removed using purge_dups (Guan *et al.*, 2020). The Hi-C reads were mapped to the primary contigs using bwa-mem2 (Vasimuddin *et al.*, 2019). The contigs were further scaffolded using the provided Hi-C data (Rao *et al.*, 2014) in YaHS (Zhou *et al.*, 2023) using the --break option. 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.*, 2024).

Curation

The assembly was decontaminated using the Assembly Screen for Cobionts and Contaminants (ASCC) pipeline (article in

preparation). Flat files and maps used in curation were generated in TreeVal (Pointon *et al.*, 2023). Manual curation was primarily conducted using PretextView (Harry, 2022), with additional insights provided by JBrowse2 (Diesh *et al.*, 2023) and HiGlass (Kerpedjiev *et al.*, 2018). Scaffolds were visually inspected and corrected as described by Howe *et al.* (2021). Any identified contamination, missed joins, and mis-joins were corrected, and duplicate sequences were tagged and removed. The process is documented at <https://gitlab.com/wtsi-grit/rapid-curation> (article in preparation).

Evaluation of the final assembly

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

A Hi-C contact map was generated for the final version of the assembly. The Hi-C reads were aligned using bwa-mem2 (Vasimuddin *et al.*, 2019) and the alignment files were combined using SAMtools (Danecek *et al.*, 2021). The Hi-C alignments were converted into a contact map using BEDTools (Quinlan & Hall, 2010) and the Cooler tool suite (Abdennur & Mirny, 2020). The contact map was visualised in HiGlass (Kerpedjiev *et al.*, 2018).

The blobtoolkit pipeline is a Nextflow port of the previous Snakemake Blobtoolkit pipeline (Challis *et al.*, 2020). It aligns the PacBio reads in SAMtools and minimap2 (Li, 2018) and generates coverage tracks for regions of fixed size. In parallel, it queries the GoAT database (Challis *et al.*, 2023) to identify all matching BUSCO lineages to run BUSCO (Manni *et al.*, 2021). For the three domain-level BUSCO lineages, the pipeline aligns the BUSCO genes to the UniProt Reference Proteomes database (Bateman *et al.*, 2023) with DIAMOND (Buchfink *et al.*, 2021) blastp. The genome is also split into chunks according to the density of the BUSCO genes from the closest taxonomic lineage, and each chunk is aligned to the UniProt Reference Proteomes database with DIAMOND blastx. Genome sequences with no hits are chunked with seqtk and aligned to the NT database with blastn (Altschul *et al.*, 1990). The blobtools suite combines all these 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), relying on the Conda package manager, the Bioconda initiative (Grünig *et al.*, 2018), the Biocontainers infrastructure (da Veiga Leprevost *et al.*, 2017), as well as the Docker (Merkel, 2014) and Singularity (Kurtzer *et al.*, 2017) containerisation solutions.

Table 4 contains a list of relevant software tool versions and sources.

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

Table 4. Software tools: versions and sources.

Software tool	Version	source
BEDTools	2.30.0	https://github.com/arq5x/bedtools2
BLAST	2.14.0	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/
BlobToolKit	4.3.9	https://github.com/blobtoolkit/blobtoolkit
BUSCO	5.5.0	https://gitlab.com/ezlab/busco
bwa-mem2	2.2.1	https://github.com/bwa-mem2/bwa-mem2
Cooler	0.8.11	https://github.com/open2c/cooler
DIAMOND	2.1.8	https://github.com/bbuchfink/diamond
fasta_windows	0.2.4	https://github.com/tolkit/fasta_windows
FastK	427104ea91c78c3b8b8b49f1a7d6bbeaa869ba1c	https://github.com/thegenemyers/FASTK
Gfastats	1.3.6	https://github.com/vgl-hub/gfastats
Goat CLI	0.2.5	https://github.com/genomehubs/goat-cli
Hifiasm	0.16.1-r375	https://github.com/chhyllp123/hifiasm
HiGlass	44086069ee7d4d3f6f3f0012569789ec138f42b84aa44357826c0b6753eb28de	https://github.com/higlass/higlass
Mercury.FK	d00d98157618f4e8d1a9190026b19b471055b22e	https://github.com/thegenemyers/MERQURY.FK
MultiQC	1.14, 1.17, and 1.18	https://github.com/MultiQC/MultiQC
Nextflow	23.04.0-5857	https://github.com/nextflow-io/nextflow
OATK	1	https://github.com/c-zhou/oatk
PretextView	0.2.5	https://github.com/sanger-tol/PretextView
purge_dups	1.2.3	https://github.com/dfguan/purge_dups
samtools	1.16.1, 1.17, and 1.18	https://github.com/samtools/samtools
sanger-tol/ ascc	-	https://github.com/sanger-tol/ascc
sanger-tol/ blobtoolkit	0.5.1	https://github.com/sanger-tol/blobtoolkit
Seqtk	1.3	https://github.com/lh3/seqtk
Singularity	3.9.0	https://github.com/sylabs/singularity
TreeVal	1.2.0	https://github.com/sanger-tol/treeval
YaHS	yahs-1.1.91eebc2	https://github.com/c-zhou/yahs

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 [here](#). 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: *Euphorbia peplus*. Accession number PRJEB52211; <https://identifiers.org/ena.embl/PRJEB52211>. The genome sequence is released openly for reuse. The *Euphorbia peplus* genome sequencing initiative is part of the Darwin Tree of Life (DTOL) project. All raw sequence data and the assembly have been deposited in INSDC databases. The genome will be annotated using available RNA-Seq data and presented through the [Ensembl](#) pipeline at the European

Bioinformatics Institute. Raw data and assembly accession identifiers are reported in [Table 1](#).

Author information

Members of the Royal Botanic Gardens Kew Genome Acquisition Lab are listed here: <https://doi.org/10.5281/zenodo.12625079>.

Members of the Plant Genome Sizing collective are listed here: <https://doi.org/10.5281/zenodo.7994306>.

Members of the Darwin Tree of Life Barcoding collective are listed here: <https://doi.org/10.5281/zenodo.12158331>

Members of the Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team are listed here: <https://doi.org/10.5281/zenodo.12162482>.

Members of Wellcome Sanger Institute Scientific Operations: Sequencing Operations are listed here: <https://doi.org/10.5281/zenodo.12165051>.

Members of the Wellcome Sanger Institute Tree of Life Core Informatics team are listed here: <https://doi.org/10.5281/zenodo.12160324>.

Members of the Tree of Life Core Informatics collective are listed here: <https://doi.org/10.5281/zenodo.12205391>.

Members of the Darwin Tree of Life Consortium are listed here: <https://doi.org/10.5281/zenodo.4783558>

References

- Abdennur N, Mirny LA: **Cooler: scalable storage for Hi-C data and other genomically labeled arrays**. *Bioinformatics*. 2020; **36**(1): 311–316. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Al-Bermani AK, Al-Shammari KIA, Gornall RJ, *et al.*: **Contribution to a cytological catalogue of the British and Irish flora, 3**. *Watsonia*. 1993; **19**: 169–171. [Reference Source](#)
- Altschul SF, Gish W, Miller W, *et al.*: **Basic Local Alignment Search Tool**. *J Mol Biol*. 1990; **215**(3): 403–410. [PubMed Abstract](#) | [Publisher Full Text](#)
- Bateman A, Martin MJ, Orchard S, *et al.*: **UniProt: the universal protein knowledgebase in 2023**. *Nucleic Acids Res*. 2023; **51**(D1): D523–D531. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Bates A, Clayton-Lucey I, Howard C: **Sanger Tree of Life HMW DNA fragmentation: diagenode Megaruptor[®]3 for LI PacBio**. *protocols.io*. 2023. [Publisher Full Text](#)
- Beasley J, Uhl R, Forrest LL, *et al.*: **DNA barcoding SOPs for the Darwin Tree of Life project**. *protocols.io*. 2023; [Accessed 25 June 2024]. [Publisher Full Text](#)
- Braun SA, Homey B, Gerber PA: **Erfolgreiche Behandlung eines Morbus Bowen mit Ingenolmebutat**. *Der Hautarzt*. 2014; **65**(10): 848–850. [Publisher Full Text](#)
- Buchfink B, Reuter K, Drost HG: **Sensitive protein alignments at Tree-of-Life scale using DIAMOND**. *Nat Methods*. 2021; **18**(4): 366–368. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Challis R, Kumar S, Sotero-Caio C, *et al.*: **Genomes on a Tree (GoAT): a versatile, scalable search engine for genomic and sequencing project metadata across the eukaryotic Tree of Life [version 1; peer review: 2 approved]**. *Wellcome Open Res*. 2023; **8**: 24. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Challis R, Richards E, Rajan J, *et al.*: **BlobToolKit – interactive quality assessment of genome assemblies**. *G3 (Bethesda)*. 2020; **10**(4): 1361–1374. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Chase MW, Hills HH: **Silica gel: an ideal material for field preservation of leaf samples for DNA studies**. *Taxon*. 1991; **40**(2): 215–220. [Publisher Full Text](#)
- Cheng H, Concepcion GT, Feng X, *et al.*: **Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm**. *Nat Methods*. 2021; **18**(2): 170–175. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- da Veiga Leprevost F, Grüning BA, Alves Aflitos S, *et al.*: **BioContainers: an open-source and community-driven framework for software standardization**. *Bioinformatics*. 2017; **33**(16): 2580–2582. [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](#)
- Denton A, Yatsenko H, Jay J, *et al.*: **Sanger Tree of Life wet laboratory protocol collection V.1**. *protocols.io*. 2023. [Publisher Full Text](#)
- Diesh C, Stevens GJ, Xie P, *et al.*: **JBrowse 2: a modular genome browser with views of syntenic and structural variation**. *Genome Biol*. 2023; **24**(1): 74. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Doležel J, Greilhuber J, Suda J: **Estimation of nuclear DNA content in plants**

using flow cytometry. *Nat Protoc.* 2007; **2**(9): 2233–2244.

[PubMed Abstract](#) | [Publisher 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](#)

Grüning B, Dale R, Sjödin A, *et al.*: **Bioconda: sustainable and comprehensive software distribution for the life sciences.** *Nat Methods.* 2018; **15**(7): 475–476.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Guan D, McCarthy SA, Wood J, *et al.*: **Identifying and removing haplotypic duplication in primary genome assemblies.** *Bioinformatics.* 2020; **36**(9): 2896–2898.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Hans AS: **Chromosomal conspectus of the Euphorbiaceae.** *Taxon.* 1973; **22**(5–6): 591–636.

[Publisher Full Text](#)

Harry E: **PretextView (Paired REad TEXTure Viewer): a desktop application for viewing pretext contact maps.** 2022.

[Reference Source](#)

Howe K, Chow W, Collins J, *et al.*: **Significantly improving the quality of genome assemblies through curation.** *GigaScience.* 2021; **10**(1): gaa153.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Jay J, Yatsenko H, Narváez-Gómez JP, *et al.*: **Sanger Tree of Life sample preparation: triage and dissection.** *protocols.io.* 2023.

[Publisher Full Text](#)

Johnson AR, Yue Y, Carey SB, *et al.*: **Chromosome-level genome assembly of *Euphorbia peplus*, a model system for plant latex, reveals that relative lack of Ty3 transposons contributed to its small genome size.** *Genome Biol Evol.* 2023; **15**(3): evad018.

[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, Dolezel J, *et al.*: **Two new nuclear isolation buffers for plant DNA flow cytometry: a test with 37 species.** *Ann Bot.* 2007; **100**(4): 875–888.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Manni M, Berkeley MR, Seppely M, *et al.*: **BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes.** *Mol Biol Evol.* 2021; **38**(10): 4647–4654.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Merkel D: **Docker: lightweight Linux containers for consistent development and deployment.** *Linux J.* 2014; **2014**(239): 2. [Accessed 2 April 2024].

[Reference Source](#)

Narváez-Gómez JP, Mbye H, Oatley G, *et al.*: **Sanger Tree of Life sample homogenisation: Covaris cryoPREP® Automated Dry Pulverizer V.1.** *protocols.io.* 2023.

[Publisher Full Text](#)

Oatley G, Sampaio F, Howard C: **Sanger Tree of Life fragmented DNA clean up: automated SPRI.** *protocols.io.* 2023.

[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.** In: Besse, P. (ed.) *Methods Mol Biol.* New York, NY: Humana, 2021; **2022**: 325–361.

[PubMed Abstract](#) | [Publisher Full Text](#)

Perry BA: **Chromosome number and phylogenetic relationships in the Euphorbiaceae.** *Am J Bot.* 1943; **30**(7): 527–543.

[Publisher Full Text](#)

Pointon DL, Eagles W, Sims Y, *et al.*: **sanger-tol/treeval v1.0.0 – Ancient Atlantis.** 2023.

[Publisher Full Text](#)

POWO: **Plants of the World Online.** Royal Botanic Gardens, Kew, 2024.

[Reference Source](#)

Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics.* 2010; **26**(6): 841–842.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Rao SSP, Huntley MH, Durand NC, *et al.*: **A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping.** *Cell.* 2014; **159**(7): 1665–1680.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Ratnasingham S, Hebert PD: **bold: The Barcode of Life Data System (<http://www.barcodinglife.org>).** *Mol Ecol Notes.* 2007; **7**(3): 355–364.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Rhie A, McCarthy SA, Fedrigo O, *et al.*: **Towards complete and error-free genome assemblies of all vertebrate species.** *Nature.* 2021; **592**(7856): 737–746.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Rhie A, Walenz BP, Koren S, *et al.*: **Merquy: reference-free quality, completeness, and phasing assessment for genome assemblies.** *Genome Biol.* 2020; **21**(1): 245.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Shafombabi NF, Knott M, Kapewangolo P, *et al.*: **Ingenol Mebutate in cancer therapy: mechanisms, clinical applications and future directions.** *Med Oncol.* 2025; **42**(3): 69.

[PubMed Abstract](#) | [Publisher Full Text](#)

Siller G, Gebauer K, Welburn P, *et al.*: **PEP005 (Ingenol Mebutate) gel, a novel agent for the treatment of actinic keratosis: results of a randomized, double-blind, vehicle-controlled, multicentre, phase IIa study.** *Australas J Dermatol.* 2009; **50**(1): 16–22.

[PubMed Abstract](#) | [Publisher Full Text](#)

Stace C: **New flora of the British Isles.** Cambridge: Cambridge University Press, 2010.

[Reference Source](#)

Todorovic M, Sampaio F, Howard C: **Sanger Tree of Life HMW DNA Extraction: Automated Plant MagAttract v.2.** *protocols.io.* 2023.

[Publisher Full Text](#)

Twyford AD, Beasley J, Barnes I, *et al.*: **A DNA barcoding framework for taxonomic verification in the Darwin Tree of Life project [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2024; **9**: 339.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Vasimuddin M, Misra S, Li H, *et al.*: **Efficient architecture-aware acceleration of BWA-MEM for multicore systems.** In: *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS).* IEEE, 2019; 314–324.

[Publisher Full Text](#)

Wang M, Gu Z, Fu Z, *et al.*: **High-quality genome assembly of an important biodiesel plant, *Euphorbia lathyris* L.** *DNA Res.* 2021; **28**(6): dsab022.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

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

[Publisher Full Text](#)

Zhou C, McCarthy SA, Durbin R: **YahS: yet another Hi-C scaffolding tool.** *Bioinformatics.* 2023; **39**(1): btac808.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Zizka G: **Flowering plants of Easter Island.** *Palmarum Hortus Francofurtensis.* 1991; **3**: 5–108.

[Reference Source](#)

Open Peer Review

Current Peer Review Status:   

Version 1

Reviewer Report 03 June 2025

<https://doi.org/10.21956/wellcomeopenres.26512.r122868>

© 2025 Chetri B. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Bimal K Chetri 

Indian Institute of Technology, Guwahati, Assam, India

This is a strong paper from the Darwin Tree of Life Consortium, presenting an impressive chromosome-level genome assembly for *Euphorbia peplus*. The authors clearly explain why sequencing this species is important, given its medicinal uses and its widespread weedy nature. The methods are well-described, covering everything from sample collection to the detailed assembly and evaluation steps.

The quality metrics (like the scaffold N50 of 33.77 Mb, 99.54% mapping to 8 chromosomes, QV of 54.2, and 97.6% BUSCO completeness) are excellent and show this is a high-quality resource that meets EBP standards. Making the data and interactive figures readily available is also a big plus for the research community, as is the clear information on the organelle genomes. This work is a valuable contribution to the Darwin Tree of Life project and will be a key resource for future studies on Euphorbiaceae, latex, and important plant chemicals.

I have a few points for clarification regarding the manual curation:

1. The paper says manual curation "improved" the assembly. However, it then states this "improvement" made the assembly have many more scaffold pieces (up by 90.91%) and a much lower scaffold N50 (down by 46.68%). This is a bit confusing because usually, an "improvement" means you get fewer, longer scaffolds (a higher N50). Could you please explain if the "improvement" here mainly means you fixed errors to make the assembly more accurate, even if that meant breaking up some of the longer, but incorrect, scaffolds?
2. You also mention correcting "missing joins." It would be helpful if you could briefly explain how fixing these "missing joins" also made the scaffold count go up. Normally, fixing a missing join would connect pieces, which should make the scaffold count go down.
3. While the final assembly quality shown in Table 2 looks very good (with great contig N50, BUSCO scores, etc.), the part about manual curation could use a bit more explanation. Specifically, helping us understand the impact on how connected the scaffolds are would make it clearer why this step was a good trade-off, even if it made the assembly look more fragmented in those specific stats.

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: Organellar genome and molecular phylogeny

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

Reviewer Report 30 May 2025

<https://doi.org/10.21956/wellcomeopenres.26512.r123308>

© 2025 Cronk Q et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Quentin Cronk 

University of British Columbia, Vancouver, Canada

Kaede Hirabayashi

The University of British Columbia, Vancouver, British Columbia, Canada

(1) This is an informative and well-written paper. We like the interactive Hi-C map, which was useful in reviewing the assembly quality (HiGlass). The authors present a contiguous assembly with very few gaps, which is a good achievement. We appreciate the detailed explanation in the Methods section of what Blobtools is doing, as it is difficult to determine just by looking at Figures 2, 3, and 4.

(2) We have not checked the voucher herbarium specimen, but the identification is obviously correct, based on the photographic voucher (Figure 1).

(3) There is some discrepancy between the 277 Mb genome size reported, and the Genomescope estimate from the PacBio reads of 355 Mb. This might be worth commenting on (some discrepancy is to be expected, but this is rather large).

(4) The presence of a repetitive-looking sequence at the gap inside chromosome/scaffold 4 (if you zoom into the Hi-C map) is concerning. It might be best to consider removing the repetitive region from the scaffold, as it is located at the gap. At a minimum, we suggest including a comment in the text. Compared to the repetitive section in scaffold 8, which has a clear interaction with only that one scaffold, this repeat appears not to indicate a very strong interaction with scaffold 4

specifically, but rather it's present throughout the genome.

(5) Purge_dups was used for purging duplicates, etc., in this paper. Please state whether any manual checks were involved to set the thresholds or whether only automatic thresholds were used. It isn't mentioned in the Methods. In our experience, this tool doesn't perform well when run with the automatic thresholds. Setting the values manually helps prevent over-purging sequences.

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: botany and genomics

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

Reviewer Report 28 May 2025

<https://doi.org/10.21956/wellcomeopenres.26512.r123307>

© 2025 Brazier T. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Thomas Brazier 

University of Rennes, Rennes, France

Christenhusz et al. report on the assembly of a reference genome for *Euphorbia peplus*, the petty spurge. They provide a brief background on the life history, distribution, and medical importance of the species, as the plant produces a cytotoxic milky latex used in cancer treatment. They also describe with details the sampling, DNA extraction and library preparation, as well as the bioinformatics protocols for assembling a chromosome-level genome assembly for the species from long-read PacBio HiFi data and short-read Hi-C data.

The study seems to follow modern and scientifically sound methods for collecting tissues, conducting lab work, and conducting de novo assembly of a reference genome, following the standards of the Darwin Tree of Life project. The summary statistics of the final assembly are above the minimum standards.

I answered 'partly' to the question "Are sufficient details of methods and materials provided to allow replication by others?". Globally, I think someone could easily replicate the work done here from their descriptions. However, at some point manual curation was done and not well described. In the 'Assembly statistics subsection', the authors say "The assembly was improved by manual curation, which corrected 5 misjoins or missing joins." The authors should give more details: why and how?

Moreover, the Assembly Quality metrics QV, k-mer recovery and EBP reference standards should be briefly described (how they are computed? how to interpret them?) and compared with the minimum standard expected for a good quality assembly.

In addition, I don't think "run in a Singularity container (Kurtzer et al., 2017)," and "The blobtoolkit pipeline was developed using nf-core tooling (Ewels et al., 2020) and MultiQC (Ewels et al., 2016), relying on the Conda package manager, the Bioconda initiative (Grüning et al., 2018), the Biocontainers infrastructure (da Veiga Leprevost et al., 2017), as well as the Docker (Merkel, 2014) and Singularity (Kurtzer et al., 2017) containerisation solutions." are necessary to understand and reproduce the study. It is good to cite tools, but these are the internal architecture of the pipeline used and are already included in the blobtoolkit citation.

Overall it is a good genome note with sufficient details to reproduce the experiment and evaluate the quality of the assembly.

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, population 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.