DATA NOTE

The genome sequence of the European mistletoe, *Viscum album* L. (Santalales: Viscaceae)

[version 1; peer review: 1 approved, 3 approved with reservations]

Lucia Campos-Dominguez[1,2], Maarten J. M. Christenhusz[3,4], David Bell[5], Michael F. Fay[4,6], Michelle Hart[5], Peter M. Hollingsworth[5], Ilia J. Leitch[4], Sahr Mian[4], Kanae Nishii[5], Royal Botanic Gardens Kew, Royal Botanic Garden Edinburgh, 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, Alex D. Twyford[2,5]

[1]Center for Research in Agricultural Genomics, CRAG (CSIC-IRTA-UAB-UB), Campus UAB, Cerdanyola del Vallès, 08193, Barcelona, Spain
[2]The University of Edinburgh, Edinburgh, Scotland, UK
[3]Curtin University, Perth, Western Australia, Australia
[4]Royal Botanic Gardens Kew, Richmond, England, UK
[5]Royal Botanic Garden Edinburgh, Edinburgh, Scotland, UK
[6]The University of Western Australia, Perth, Western Australia, Australia

v1 **First published:** 15 Sep 2025, **10**:517
https://doi.org/10.12688/wellcomeopenres.24920.1
Latest published: 15 Sep 2025, **10**:517
https://doi.org/10.12688/wellcomeopenres.24920.1

Abstract

We present a genome assembly from a female specimen of *Viscum album* (European mistletoe; Streptophyta; Magnoliopsida; Santalales; Viscaceae). The genome sequence has a total length of 94 261.04 megabases. Most of the assembly (98.67%) is scaffolded into 49 chromosomal pseudomolecules, including the B_1 chromosome. 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

Viscum album; European mistletoe; genome sequence; chromosomal; Santalales

Open Peer Review

Approval Status ? ✓ ? ?

	1	2	3	4
version 1 15 Sep 2025	? view	✓ view	? view	? view

1. **Juan Francisco Ornelas**, Red de Biología Evolutiva, Instituto de Ecología, A.C. (INECOL), Xalapa, Mexico
Enrique Ibarra-Laclette, Red de Estudios Moleculares Avanzados (REMAv), Instituto de Ecología, A.C. (INECOL, Veracruz, Mexico
Saddan Morales-Saldaña, Universidad Nacional Autonoma de Mexico Facultad de

This article is included in the Tree of Life gateway.

Ciencias (Ringgold ID: 98804), Mexico City, Mexico

2. **Xin Liu**, Beijing Genomics Institute, Beijing, China
3. **Rizky Dwi Satrio**, The Republic of Indonesia Defense University, Bogor, Indonesia
4. **Sunil Kumar Sahu**, Key Laboratory of Genomics, Ministry of Agriculture, BGI Research, Shenzhen, China

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: Campos-Dominguez L: Data Curation, Validation, Writing – Review & Editing; **Christenhusz MJM**: Investigation, Resources, Writing – Review & Editing; **Bell D**: Resources; **Fay MF**: Writing – Review & Editing; **Hart M**: Supervision; **Hollingsworth PM**: Supervision; **Leitch IJ**: Supervision, Writing – Review & Editing; **Mian S**: Investigation, Resources; **Nishii K**: Writing – Review & Editing; **Twyford AD**: Investigation, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome through core funding to the Wellcome Sanger Institute (220540) and the Darwin Tree of Life Discretionary Award (218328).
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: Campos-Dominguez L, Christenhusz MJM, Bell D *et al.* **The genome sequence of the European mistletoe, *Viscum album* L. (Santalales: Viscaceae) [version 1; peer review: 1 approved, 3 approved with reservations]** Wellcome Open Research 2025, **10**:517 https://doi.org/10.12688/wellcomeopenres.24920.1

First published: 15 Sep 2025, **10**:517 https://doi.org/10.12688/wellcomeopenres.24920.1

Species taxonomy

Eukaryota; Viridiplantae; Streptophyta; Streptophytina; Embryophyta; Tracheophyta; Euphyllophyta; Spermatophyta; Magnoliopsida; Mesangiospermae; eudicotyledons; Gunneridae; Pentapetalae; Santalales; Viscaceae; *Viscum*; *Viscum album* L. (NCBI:txid3972)

Background

The European mistletoe (*Viscum album*) (Figure 1) is an evergreen, obligate hemiparasitic shrub that connect via a haustorium to obtain water and inorganic compounds directly from host xylem. They can reach up to 1.50 m across (Wangerin, 1937). Plants have unisexual scented flowers with four tepals. Fruits are white, globose berries with a single seed (Thomas *et al.*, 2023).

In Great Britain, mistletoe grows on broadleaf trees, especially poplar (*Populus*), lime (*Tilia*), hawthorn (*Crataegus*), apple trees (*Malus*) and sometimes on other deciduous species (e.g., *Acer, Aesculus*, *Betula*, *Corylus*, *Juglans*, *Pyrus*, *Quercus*, *Robinia* and *Sorbus*). It is most common in the south and east of Britain, occurring in just a few sites in Scotland although it appears to be expanding there (Stroh *et al.*, 2023). *Viscum album* is native to most of Europe, except the far north where it is restricted by frost, and in northwest Africa, West Asia and the Himalayas, where its distribution is limited by drought (Thomas *et al.*, 2023). It has been introduced to Ireland, Japan and California (Thomas *et al.*, 2023). There are five subspecies, all with different host preferences (Thomas *et al.*, 2023). In Europe, *Viscum album* subsp. *album* is the only subspecies growing on broadleaf trees, whereas all others prefer gymnosperms. It can be particularly common on trees growing on alkaline soils. Although it does not kill its hosts, the European mistletoe is considered a pest because it reduces fruit yields by up to 50% (Preston, 1977). Mistletoe branches are often harvested from orchards for sale in Christmas markets.

Because it is evergreen, it is more visible in winter, which likely is the reason for it being a symbol of fertility and vitality in mythology and in folklore of protection, love, winter solstice and Christmas. It was revered by the early herbalists, reputed to cure epileptic trances and tumours, used for divining treasure, keeping witches at bay and protecting crops. It was also reputed to have powers as a fertility potion and aphrodisiac (Fay *et al.*, 2010) and references therein; (Christenhusz, 2025; Fay, 2017).

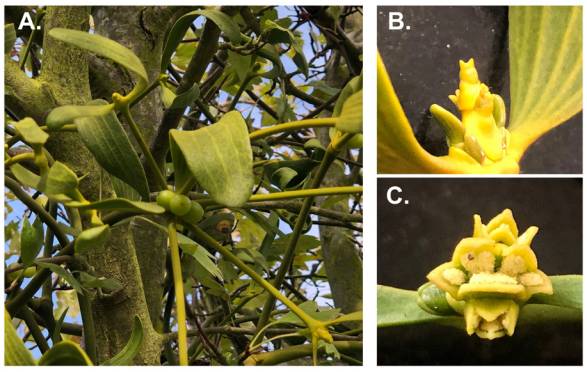

Figure 1. *Viscum album* on hawthorn at Home Park, Hampton Court. (**A**) The sequenced female individual on its hawthorn host, with developing fruits. (**B**–**C**) Close-ups of female and male flowers from other individuals in the same population (not the sequenced plant).

Flowers are pollinated by insects attracted to its fruity scent, such as flies and ants in winter, and later in the spring by bees and bumblebees (Thomas *et al.*, 2023). Berries of the female plants are an important winter food for a number of bird species, including mistle thrush (*Turdus viscivorus*) and black caps (*Sylvia atricapilla*), which are immune to the toxic lectin viscumin in the berries (Olsnes *et al.*, 1982). These birds disperse the sticky seeds by rubbing them off their beaks or anuses on branches that are suitable for establishment (Beckman & Sullivan, 2023). The name mistletoe originates from Anglo-Saxon '*mistel*', dung, and '*tan*', twig. The sticky fruit pulp resulted in its Latin name, *viscum*, being adopted into English as the word 'viscous', sticky. In the past, the berries boiled down to a glue, which was used to catch songbirds for consumption. This resulted in mistletoe being referred to as 'birdlime'.

Mistletoes have some of the largest genomes among flowering plants: *Viscum* genomes range from 61 to 100 Gbp in size (Ulrich *et al.*, 1988; Zonneveld, 2010). The *Viscum album* subsp. *album* sampled here has a chromosome-level genome assembly ($2n$ = 20) of 94 Gbp (1C), making it the largest known genome of a plant species native to Britain and Ireland. This work not only represents an enormous technical challenge, but also provides an important genomic resource for the study of the biology of mistletoe, from its genome regulation machinery, its 3-D genome structure, and host-parasite interactions. Although unlikely to provide insights into its mythical powers, this high-quality reference genome is the first step towards the study of the full range of plant genome size and evolutionary consequences of large genomes.

Methods

Sample acquisition, flow cytometry and DNA barcoding

The specimen used for genome sequencing was a vegetative structure female *Viscum album* (specimen ID KDTOL10096, ToLID drVisAlbu1), collected from on planted trees along Hampton Court Rd, Home Park, Molesey, Kingston upon Thames, Surrey, United Kingdom (latitude 51.4061, longitude –0.3319) on 2020-09-07. The specimen was collected and identified by Maarten J. M. Christenhusz (Royal Botanic Gardens Kew).

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. For this species, was used for isolation of nuclei (Loureiro *et al.*, 2007), and the internal calibration standard was an *Allium cepa* L. standard as described in Doležel *et al.* (2007).

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 drVisAlbu1 sample was weighed and triaged to determine the appropriate extraction protocol.

Tissue from the leaf was homogenised by cryogenic disruption using the Covaris cryoPREP® Automated Dry Pulverizer.

HMW DNA was extracted using the Manual 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 by manual SPRI (solid-phase reversible immobilisation). 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.

RNA was extracted from flower tissue of drVisAlbu21 in the Tree of Life Laboratory at the WSI using the RNA Extraction: Automated MagMax™ mirVana protocol. The RNA concentration was assessed using a Nanodrop spectrophotometer and a Qubit Fluorometer using the Qubit RNA Broad-Range Assay kit. Analysis of the integrity of the RNA was done using the Agilent RNA 6000 Pico Kit and Eukaryotic Total RNA assay.

PacBio HiFi library preparation and sequencing

Library preparation and sequencing were performed at the WSI Scientific Operations core. Libraries were prepared using the PacBio Express Template Preparation Kit v2.0 (Pacific Biosciences, California, USA) according to the manufacturer's instructions. The kit includes the reagents required for removal of single-strand overhangs, DNA damage repair, end repair/A-tailing, adapter ligation, and nuclease treatment. Library preparation also included a library purification step using AMPure PB beads (Pacific Biosciences) and a size-selection step to remove templates <3 kb using AMPure PB modified SPRI. DNA concentration was quantified using the Qubit Fluorometer v2.0 (Thermo Fisher Scientific) and Qubit 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).

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

Hi-C

Sample preparation and crosslinking

Hi-C data were generated from frozen leaf tissue of the drVisAlbu1 using the Arima-HiC v2 kit (Arima Genomics). Tissue was finely ground using the Covaris cryoPREP Dry Pulverizer (Covaris), 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 and equimolar and/or weighted 2.8 nM pools. Pool concentrations were checked using the Agilent 4200 TapeStation (Agilent) with High Sensitivity D500 reagents before sequencing. Sequencing was performed using paired-end 150 bp reads on the Illumina NovaSeq 6000.

RNA library preparation and sequencing

Libraries were prepared using the NEBNext® Ultra™ II Directional RNA Library Prep Kit for Illumina (New England Biolabs), following the manufacturer's instructions. Poly(A) mRNA in the total RNA solution was isolated using oligo(dT) beads, converted to cDNA, and uniquely indexed; 14 PCR cycles were performed. Libraries were size-selected to produce fragments between 100–300 bp. Libraries were quantified, normalised, pooled to a final concentration of 2.8 nM, and diluted to 150 pM for loading. Sequencing was carried out on the Illumina NovaSeq X to generate 150-bp paired-end reads.

Genome assembly

Prior to assembly of the PacBio HiFi reads, a database of *k*-mer counts (k = 31) was generated from the filtered reads using FastK. GenomeScope2 (Ranallo-Benavidez *et al.*, 2020) was used to analyse the *k*-mer frequency distributions, providing estimates of genome size, heterozygosity, and repeat content.

The HiFi reads were assembled using Hifiasm (Cheng *et al.*, 2021) with the --primary option. Haplotypic duplications were identified and removed using purge_dups (Guan *et al.*, 2020). The Hi-C reads (Rao *et al.*, 2014) were mapped to the primary contigs using bwa-mem2 (Vasimuddin *et al.*, 2019), and the contigs were scaffolded in YaHS (Zhou *et al.*, 2023) with the --break option for handling potential misassemblies. The scaffolded assemblies were evaluated using Gfastats (Formenti *et al.*, 2022), BUSCO (Manni *et al.*, 2021) and MERQURY.FK (Rhie *et al.*, 2020).

Assembly curation

The assembly was decontaminated using the Assembly Screen for Cobionts and Contaminants (ASCC) pipeline. TreeVal was used to generate the flat files and maps for use in curation. Manual curation was conducted primarily in PretextView and HiGlass (Kerpedjiev *et al.*, 2018). Scaffolds were visually inspected and corrected as described by Howe *et al.* (2021). Manual corrections included 365 breaks, 353 joins, and removal of 500 haplotypic duplications. The curation process is documented at https://gitlab.com/wtsi-grit/rapid-curation. PretextSnapshot was used to generate a Hi-C contact map of the final assembly.

Assembly quality assessment

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

The genome was analysed using the BlobToolKit pipeline, a Nextflow implementation of the earlier Snakemake version (Challis *et al.*, 2020). The pipeline aligns PacBio reads using minimap2 (Li, 2018) and SAMtools (Danecek *et al.*, 2021) to generate coverage tracks. It runs BUSCO (Manni *et al.*, 2021) using lineages identified from NCBI Taxonomy (Schoch *et al.*, 2020). For the three domain-level lineages, BUSCO genes are aligned to the UniProt Reference Proteomes database (Bateman *et al.*, 2023) using DIAMOND blastp (Buchfink *et al.*, 2021). The genome is divided into chunks based on the density of BUSCO genes from the closest taxonomic lineage, and each chunk is aligned to the UniProt Reference Proteomes database with DIAMOND blastx. Sequences without hits are chunked using seqtk and aligned to the NT database with blastn (Altschul *et al.*, 1990). The BlobToolKit suite consolidates all outputs into a blobdir for visualisation. The BlobToolKit pipeline was developed using nf-core tooling (Ewels *et al.*, 2020) and MultiQC (Ewels *et al.*, 2016), with package management via Conda and Bioconda (Grüning *et al.*, 2018), and containerisation through Docker (Merkel, 2014) and Singularity (Kurtzer *et al.*, 2017).

Genome sequence report

Sequence data

The genome of a specimen of *Viscum album* was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating 6 011.18 Gb (gigabases) from 393.33 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 93 521.60 Mb, with a heterozygosity of 1.13% and repeat content of 30.90% (Figure 2). Using flow cytometry, the genome size (1C value)

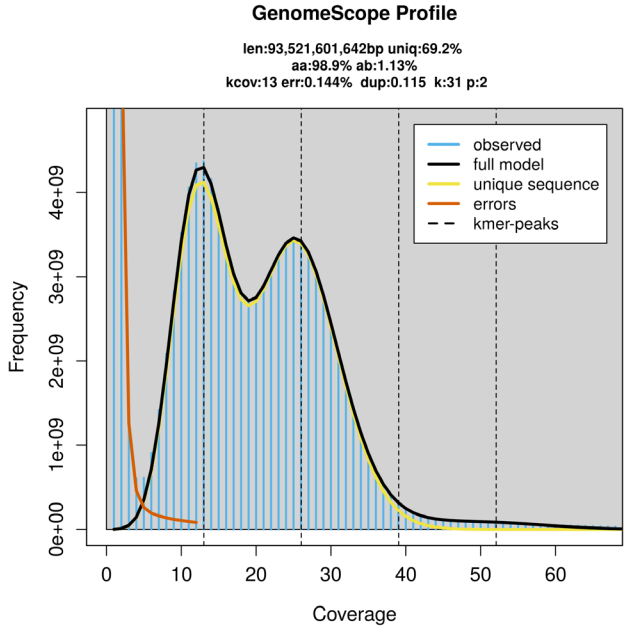

Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2. The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

of the sampled was estimated to be 93.78 pg, equivalent to 91 710.00 Mb. These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 26× coverage. Hi-C sequencing produced 1 942.64 Gb from 12 865.20 million reads, which were used to scaffold the assembly. RNA sequencing data were also generated and are available in public sequence repositories. Table 1 summarises the specimen and sequencing details.

Assembly statistics

The primary haplotype was assembled, and contigs corresponding to an alternate haplotype were also deposited in INSDC databases. The final assembly has a total length of 94 261.04 Mb in 5 096 scaffolds, with 6 368 gaps, and a scaffold N50 of 2 134.69 Mb (Table 2).

Most of the assembly sequence (98.67%) was assigned to 11 chromosomal-level scaffolds, representing 10 autosomes and the B_1 chromosome. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 3; Table 3). Because several chromosome sequences exceed the INSDC per-record length limit (2,147,483,647 bp), they were deposited in consecutive parts with separate accessions; in total, the 11 chromosomes correspond to 49 accessions.

Assembly quality metrics

The combined primary and alternate assemblies achieve an estimated QV of 67.3. The *k*-mer completeness is 76.68% for the primary assembly, 73.77% for the alternate haplotype, and 98.59% for the combined assemblies (Figure 4). BUSCO analysis using the eudicots_odb10 reference set (n = 2 326) identified 79.4% of the expected gene set (single = 73.7%, duplicated = 5.8%).

Table 1. Specimen and sequencing data for BioProject PRJEB47325.

Platform	PacBio HiFi	Hi-C	RNA-seq
ToLID	drVisAlbu1	drVisAlbu1	drVisAlbu21
Specimen ID	KDTOL10096	KDTOL10096	KDTOL11072
BioSample (source individual)	SAMEA7522516	SAMEA7522516	SAMEA114565187
BioSample (tissue)	SAMEA7522568	SAMEA7522562	SAMEA114565203
Tissue	leaf	leaf	flower
Sequencing platform and model	Sequel IIe	Illumina NovaSeq 6000	Illumina NovaSeq X
Run accessions	ERR6808049–ERR8755914	ERR6688777; ERR7220437; ERR7220436	ERR13148236; ERR13148237; ERR11641090; ERR13148235
Read count total	393.33 million	12 865.20 million	213.10 million
Base count total	6 011.18 Gb	1 942.64 Gb	32.18 Gb

Table 2. Genome assembly statistics.

Assembly name	drVisAlbu1.1
Assembly accession	GCA_963277665.1
Alternate haplotype accession	GCA_963082535.1
Assembly level	chromosome
Span (Mb)	94 261.04
Number of chromosomes	11
Number of contigs	11 464
Contig N50	34.83 Mb
Number of scaffolds	5 096
Scaffold N50	2134.69 Mb
Supernumerary chromosomes	B_1

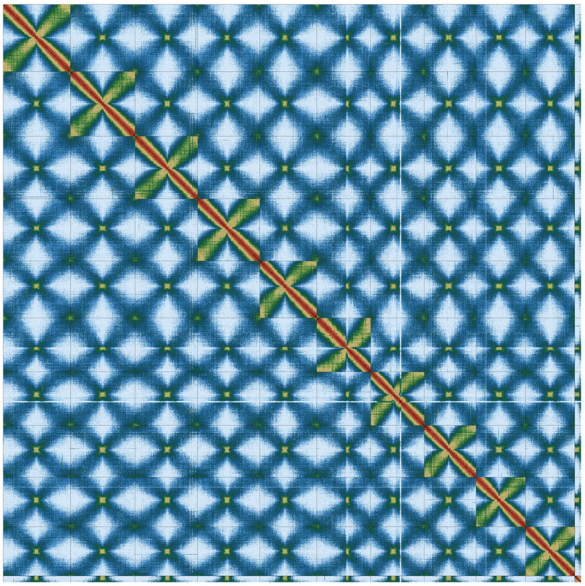

Figure 3. Hi-C contact map of the *Viscum album* genome assembly. Assembled chromosomes are shown in order of size. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in the primary genome assembly of *Viscum album* drVisAlbu1. The 11 chromosomes were submitted to the INSDC in multiple parts due to length constraints.

INSDC accession	Molecule	Length (Mb)	GC%
OY728119.1	1_1	2 143.53	31.50
OY728120.1	1_2	2 138.63	31
OY728121.1	1_3	2 132.99	31
OY728122.1	1_4	2 142.15	31
OY728123.1	1_5	2 142.78	31.50
OY728124.1	1_6	124.38	32.50
OY728125.1	2_1	2 112.40	31.50
OY728126.1	2_2	2 144.48	31
OY728127.1	2_3	2 133.12	31
OY728128.1	2_4	2 141.81	31
OY728129.1	2_5	1 870.27	31.50
OY728130.1	3_1	2 134.93	31.50
OY728131.1	3_2	2 108.66	31
OY728132.1	3_3	2 146.28	31
OY728133.1	3_4	2 117.02	31
OY728134.1	3_5	1 576.30	31.50
OY728135.1	4_1	2 067.10	31.50
OY728136.1	4_2	2 134.69	31
OY728137.1	4_3	2 136.66	31
OY728138.1	4_4	2 140.54	31
OY728139.1	4_5	1 531.58	31.50
OY728140.1	5_1	2 146.57	31.50
OY728141.1	5_2	2 138.19	31
OY728142.1	5_3	2 101.18	31
OY728143.1	5_4	2 146.23	31.50
OY728144.1	5_5	621.09	32
OY728145.1	6_1	2 138.61	31.50
OY728146.1	6_2	2 083.69	31
OY728147.1	6_3	2 144.31	31
OY728148.1	6_4	2 139.18	31
OY728149.1	6_5	172.72	32
OY728150.1	7_1	2 132.15	31
OY728151.1	7_2	2 133.92	31
OY728152.1	7_3	2 133.31	31
OY728153.1	7_4	2 100.93	31.50
OY728154.1	7_5	143.35	32
OY728155.1	8_1	2 134.14	31.50
OY728156.1	8_2	2 145.20	31
OY728157.1	8_3	2 137.73	31
OY728158.1	8_4	1 914.31	31.50
OY728159.1	9_1	2 145.48	31
OY728160.1	9_2	2 114.17	31
OY728161.1	9_3	2 146.42	31
OY728162.1	9_4	1 555.47	31.50
OY728163.1	10_1	2 141.25	31.50
OY728164.1	10_2	2 119.19	31
OY728165.1	10_3	2 142.18	31
OY728166.1	10_4	1 498.83	31.50
OY728167.1	B_1	1 020.19	32

The snail plot in Figure 5 summarises the scaffold length distribution and other assembly statistics for the primary assembly. The blob plot in Figure 6 shows the distribution of scaffolds by GC proportion and coverage.

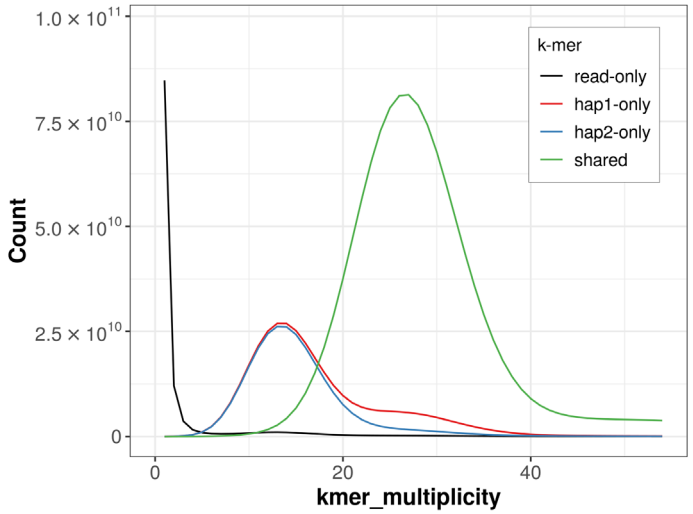

Figure 4. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers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

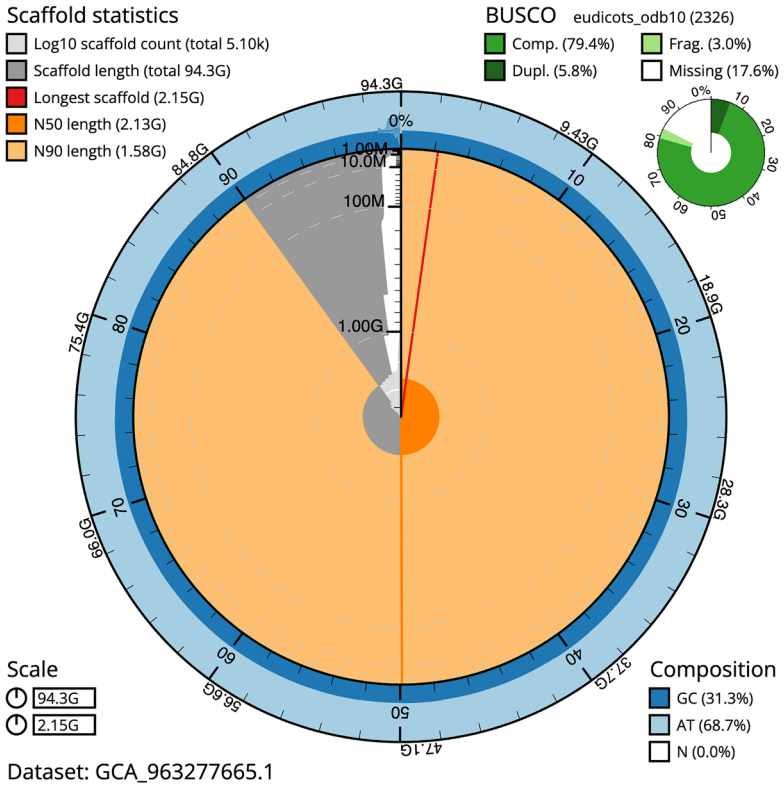

Figure 5. Assembly metrics for drVisAlbu1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the circumference. The outermost blue 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.

Table 4 lists the assembly metric benchmarks adapted from Rhie *et al.* (2021) the Earth BioGenome Project Report on Assembly Standards September 2024. The EBP metric calculated for the primary assembly is **7.C.Q67**, meeting the recommended reference standard.

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

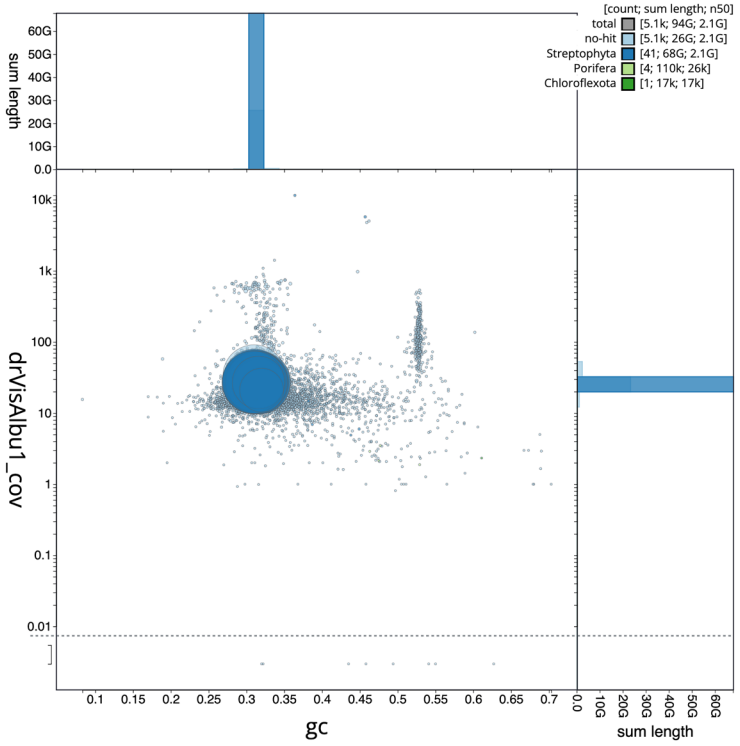

Figure 6. BlobToolKit GC-coverage plot for drVisAlbu1.1. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes display the total length of sequences distributed across different levels of coverage and GC content. An interactive version of this figure is available on the BlobToolKit viewer.

Table 4. Earth Biogenome Project summary metrics for the *Viscum album* assembly.

Measure	Value	Benchmark
EBP summary (primary)	7.C.Q67	6.C.Q40
Contig N50 length	34.83 Mb	≥ 1 Mb
Scaffold N50 length	2 134.69 Mb	= chromosome N50
Consensus quality (QV)	Primary: 67.3; alternate: 67.2; combined: 67.3	≥ 40
k-mer completeness	Primary: 76.68%; alternate: 73.77%; combined: 98.59%	≥ 95%
BUSCO	C:79.4%[S:73.7%;D:5.8%]; F:3.0%; M:17.6%; n:2 326	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	98.67%	≥ 90%

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: Viscum album (European mistletoe). Accession number PRJEB47325. The genome sequence is released openly for reuse. The *Viscum album* 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. 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.

Author information

- Members of the Royal Botanic Gardens Kew Genome Acquisition Lab
- Members of the Royal Botanic Garden Edinburgh Genome Acquisition Lab
- Members of the Plant Genome Sizing 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

Table 5. Software versions and sources.

Software	Version	Source
BEDTools	2.30.0	https://github.com/arq5x/bedtools2
BLAST	2.14.0	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/
BlobToolKit	4.4.4	https://github.com/blobtoolkit/blobtoolkit
bwa-mem2	2.2.1	https://github.com/bwa-mem2/bwa-mem2
Cooler	0.8.11	https://github.com/open2c/cooler
DIAMOND	2.1.8	https://github.com/bbuchfink/diamond
fasta_windows	0.2.4	https://github.com/tolkit/fasta_windows
FastK	1.1	https://github.com/thegenemyers/FASTK
GenomeScope2.0	2.0.1	https://github.com/tbenavi1/genomescope2.0
Gfastats	1.3.6	https://github.com/vgl-hub/gfastats
GoaT CLI	0.2.5	https://github.com/genomehubs/goat-cli
Hifiasm	0.16.1-r375	https://github.com/chhylp123/hifiasm
HiGlass	1.13.4	https://github.com/higlass/higlass
MerquryFK	1.1.2	https://github.com/thegenemyers/MERQURY.FK
Oatk	pre-release	https://github.com/c-zhou/oatk
MultiQC	1.14; 1.17 and 1.18	https://github.com/MultiQC/MultiQC

Software	Version	Source
Nextflow	24.10.4	https://github.com/nextflow-io/nextflow
PretextSnapshot	N/A	https://github.com/sanger-tol/PretextSnapshot
PretextView	0.2.5	https://github.com/sanger-tol/PretextView
purge_dups	1.2.3	https://github.com/dfguan/purge_dups
samtools	1.21	https://github.com/samtools/samtools
sanger-tol/ascc	0.1.0	https://github.com/sanger-tol/ascc
sanger-tol/blobtoolkit	v0.7.0	https://github.com/sanger-tol/blobtoolkit
Seqtk	1.3	https://github.com/lh3/seqtk
Singularity	3.9.0	https://github.com/sylabs/singularity
TreeVal	1.2.0	https://github.com/sanger-tol/treeval
YaHS	yahs-1.1.91eebc2	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 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

Beckman NG, Sullivan LL: **The causes and consequences of seed dispersal.** *Annu Rev Ecol Evol Syst.* 2023; **54**(1): 403–27.
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, 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): 215–20.
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

Christenhusz M: **Liefde in de plantenwereld, een botanische kijk op de evolutie van seks**. Sterck & De Vreese, 2025.
Reference Source

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

Doležel J, Greilhuber J, Suda J: **Estimation of nuclear DNA content in plants using flow cytometry.** *Nat Protoc.* 2007; **2**(9): 2233–44.
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

Fay MF: **Is mistletoe more than just an excuse for a kiss?** 2017.
Reference Source

Fay MF, Bennett JR, Dixon KW, *et al.*: **Parasites, their relationships and the disintegration of scrophulariaceae *Sensu lato*.** *Curtis's Botanical Magazine.* 2010; **26**(4): 286–313.
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

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): 875–88.
PubMed Abstract | Publisher Full Text | Free Full Text

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

Olsnes S, Stirpe F, Sandvig K, *et al.*: **Isolation and characterization of viscumin, a toxic lectin from *Viscum album* L. (mistletoe).** *J Biol Chem.* 1982; **257**(22): 13263–70.
PubMed Abstract | 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: P. Besse (ed), *Methods Mol Biol.* New York, NY, 2021; **2222**: 325–61.
PubMed Abstract | **Publisher Full Text**

Preston AP: **Effects of mistletoe (*Viscum album*) on young apple trees.** *Hortic Res.* 1977; **17**: 33–38.

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): 355–64.
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 (Oxford).* 2020; **2020**: baaa062.
PubMed Abstract | **Publisher Full Text** | **Free 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

Thomas PA, Dering M, Giertych MJ, *et al.*: **Biological flora of Britain and Ireland: *Viscum album*.** *J Ecol.* 2023; **111**(3): 701–39.
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**

Ulrich I, Fritz B, Ulrich W: **Application of DNA fluorochromes for flow cytometric DNA analysis of plant protoplasts.** *Plant Sci.* 1988; **55**(2): 151–58.
Publisher 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

Wangerin W: **Loranthaceae.** In: O. v. Kirchner, E. Loew, and C. Schroeter (eds), *Lebensgeschichte der Blütenpflanzen Mitteleuropas.* Stuttgart: Ulmer, 1937; **II/1**: 953–1146.
Reference Source

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: **New record holders for maximum genome size in eudicots and monocots.** *J Bot.* 2010; **2010**(1): 527357.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ? ✔ ? ?

Version 1

Reviewer Report 23 April 2026

https://doi.org/10.21956/wellcomeopenres.27450.r151072

?

Sunil Kumar Sahu
Key Laboratory of Genomics, Ministry of Agriculture, BGI Research, Shenzhen, China

This Data Note provides clear value by presenting a chromosome-scale nuclear genome for *Viscum album*, a species notable for its unusually large genome and biological significance as an obligate hemiparasite. The assembly is technically impressive and fills an important gap in mistletoe genomics, particularly for studies of parasitism, genome expansion, and chromosome-scale organization.
That said, the manuscript would benefit from more explicitly positioning this advance in relation to prior *V. album* genomic resources, including previously published gene-space and organelle-focused studies referenced in the review materials.

1. The sequencing and assembly strategy is broadly appropriate for a large and complex plant genome, incorporating PacBio HiFi, Hi-C, RNA-seq, Hifiasm, YaHS, and manual curation. However, several methodological details are not described with sufficient clarity to ensure reproducibility. In particular, the flow cytometry protocol, the DNA barcoding loci used, and the provenance of the RNA-seq samples relative to the genome specimen should be more explicitly documented.
2. The manuscript should also clarify the curation workflow. Specifically, it would be helpful to distinguish which steps were performed using PretextView versus HiGlass, rather than listing both tools without context. As written, the methods are generally sound but not fully reproducible.
3. The assembly statistics are strong, with a chromosome-scale genome, high scaffold N50, and a high QV score. However, there is a notable internal inconsistency in the reported HiFi sequencing coverage. The stated 6,011.18 Gb of HiFi data and an approximately 91.7 Gb genome size are not consistent with the reported 26-fold coverage. This discrepancy should be corrected or clearly explained, including whether the coverage estimate reflects filtered reads only and, if so, the criteria used for filtering.
4. In Figure 3, the Hi-C contact map lacks appropriate labeling. Please update the figure to improve interpretability.
5. In Figure 4, the proportions appear slightly inconsistent: if the unique fraction is 69.2%, the repeat fraction should be 30.8%, not 30.9% as stated in the text. This should be corrected.
6. The phrase “49 chromosomal pseudomolecules” may be misinterpreted as indicating 49

chromosomes. It would be clearer to specify that the assembly represents 11 chromosomes, with sequences split across multiple INSDC accessions due to sequence-length limitations.

7. While the manuscript is generally readable, several presentation issues reduce overall clarity and confidence. These include minor grammatical errors, inconsistent unit formatting, and awkward phrasing (e.g., “collected from on planted trees”), all of which should be corrected through careful language editing.

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: Plant genomics and evolution

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, however I have significant reservations, as outlined above.

Reviewer Report 07 April 2026

https://doi.org/10.21956/wellcomeopenres.27450.r147744

?

Rizky Dwi Satrio
The Republic of Indonesia Defense University, Bogor, Indonesia

This manuscript presents a chromosome-level genome assembly of *Viscum album*, an obligate hemiparasitic plant with an exceptionally large genome (~94 Gb). The study represents a significant technical achievement, and the generation and analysis of genomic data at this scale is clearly a non-trivial effort that deserves strong appreciation, given the substantial computational and methodological challenges involved. The integration of PacBio HiFi sequencing, Hi-C scaffolding, and comprehensive assembly curation is appropriate and aligns with current best practices.

However, as a Data Note, the manuscript is primarily descriptive and lacks deeper biological interpretation, functional insights, and critical discussion of assembly limitations. While the dataset is valuable, several methodological clarifications, inconsistencies in reporting, and issues related to genome quality metrics and completeness should be addressed to improve clarity, reproducibility, and scientific rigor.

1. Novelty and Biological Significance
The manuscript strongly emphasizes the technical achievement but provides limited biological insight into *Viscum album*. The authors should include a brief discussion on genome evolution in parasitic plants, implications of extremely large genome size (e.g., transposable element expansion and gene family evolution), and comparisons with other large plant genomes.

2. BUSCO Completeness is Relatively Low
The reported BUSCO completeness of 79.4% is below the recommended benchmark (>90%) and raises concerns about incomplete gene space representation. The authors should provide a clear explanation for this result (e.g., genome size, repeat content, or lineage divergence) and consider including additional completeness metrics or lineage-specific BUSCO analyses.

3. K-mer Completeness Discrepancy
The primary assembly shows relatively low k-mer completeness (76.68%) compared to the combined assembly (98.59%), suggesting missing or partitioned genomic information. The authors should clarify whether the primary assembly is intended as the reference and discuss the implications of this discrepancy for downstream analyses.

4. Limited Discussion of Repeat Content
Although repeat content (~30.9%) is reported, it is not further analyzed or discussed in the manuscript. Given the extremely large genome size, the authors should provide a repeat annotation summary (e.g., TE classes and proportions) and compare these findings with other large plant genomes.

5. Lack of Gene Annotation Results
No gene annotation results are presented, and this is only mentioned as future work, which limits the utility of the dataset. The authors should include at least a preliminary gene prediction or clearly justify its omission as a scope limitation of this Data Note.

6. Coverage is Relatively Low for Genome Size
The sequencing depth (~26×) may be relatively low considering the extremely large genome size, potentially affecting assembly completeness and accuracy. The authors should justify the adequacy of this coverage for HiFi-based assembly and discuss possible biases or missing genomic regions.

7. RNA-seq Sample Inconsistency
RNA-seq data were generated from a different individual (drVisAlbu21) than the genome assembly sample (drVisAlbu1), which may introduce inconsistencies for annotation. The authors should justify the use of a different individual and flower tissue, clarify its relevance, and discuss implications for downstream annotation, while ideally providing genome annotation and making it publicly accessible even within the Data Note framework.

8. Chromosome Number vs Assembly Representation
The species is reported as 2n = 20, yet the assembly is described as 10 autosomes plus a B chromosome, which may cause confusion. The authors should clearly explain the haploid versus diploid representation and explicitly describe the chromosome naming and counting scheme.

9. Overly Descriptive Background Section
The background section includes folklore and cultural descriptions that are not directly relevant to the genomic focus of the study. The authors should streamline this section and focus more on genomic, evolutionary, and biological context.

10. Limited Critical Evaluation of Assembly Quality
Although the assembly meets EBP standards (7.C.Q67), the manuscript lacks critical discussion of its limitations. The authors should include a concise limitations paragraph addressing BUSCO incompleteness, potential collapsed repeats, and unresolved heterozygosity.

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: My expertise lies in plant genomics and bioinformatics, with a strong focus on genome assembly, annotation, analysis, and interpretation of complex plant genomes.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 27 March 2026

https://doi.org/10.21956/wellcomeopenres.27450.r151067

Xin Liu

Beijing Genomics Institute, Beijing, China

This Data Note describes the sequencing and assembly of the European mistletoe genome. For a non-UK reviewer like myself, the authors' engaging background section provided valuable insight into the cultural (and even medicinal) significance of this species. As a genomics researcher, I was already aware of the unusually large genome size in Viscum species, and it is an honor to review this article, which adds meaningful knowledge to plant genomics.
The Darwin Tree of Life project has supported numerous genome studies across plants and animals. Publishing a Data Note is an effective way to rapidly benefit the wider research community. I also recognize that a Data Note differs from a research article: the goal is to clearly describe methods, present the data, and summarize the results. Within this framework, I find this Data Note already very strong. I have no major concerns, but I would like to share a few comments and suggestions—these are not scientific criticisms, but rather ideas for clarifying the interpretation of the genomic data:

1. **Genome assembly quality** for a genome exceeding 90 Gb, the assembly achieved here is impressive and beyond my expectations. It is clear that the authors invested substantial effort and computational resources in the analysis. The methods and workflows described will serve as a valuable reference for future large-genome projects. This is a major contribution, and I thank the authors for advancing plant genomics.
2. **Assembly validation through sequencing depth** Sequence alignment already provides strong evidence of assembly quality. Since the authors have performed alignments, I suggest adding a simple figure showing the distribution of sequencing depth across bases. Such a plot would clearly illustrate accuracy (e.g., whether it resembles a Poisson distribution), coverage, and completeness (mapping rate). I do not usually request this in reviews, but for a genome of this scale, such a straightforward visualization would provide highly informative quality metrics.
3. **Scaffold statistics beyond chromosomes** Although the assembly contains over 5,000 scaffolds, the main chromosomes are relatively complete, as noted in the manuscript. For the smaller scaffolds, I recommend providing basic statistics—such as their length distribution after excluding the chromosome-level sequences. While deeper biological interpretation may be reserved for a future research article, even preliminary statistics in the Data Note would help readers understand the assembly outcome and infer characteristics of the Viscum genome.
4. **Repeat content from k-mer analysis** The k-mer analysis suggests that approximately 30% of the genome consists of repetitive sequences. This is intriguing, and although I expect the authors will explore this further in subsequent work, it would be useful to briefly highlight possible reasons here.

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: genomics

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

Reviewer Report 10 March 2026

https://doi.org/10.21956/wellcomeopenres.27450.r147747

?

Juan Francisco Ornelas
Red de Biología Evolutiva, Instituto de Ecología, A.C. (INECOL), Xalapa, Veracruz, Mexico
Enrique Ibarra-Laclette
Red de Estudios Moleculares Avanzados (REMAv), Instituto de Ecología, A.C. (INECOL, Veracruz, Mexico
Saddan Morales-Saldaña
Herbario, Universidad Nacional Autonoma de Mexico Facultad de Ciencias (Ringgold ID: 98804), Mexico City, Mexico

Review of Wellcome Open Research WOR-v1-2026-0517

Lucia Campos-Dominguez, Maarten J. M. Christenhusz, David Bell, Michael F. Fay, Michelle Hart, Peter M. Hollingsworth, Ilia J. Leitch, Sahr Mian, Kanae Nishii, Royal Botanic Gardens Kew, Royal Botanic Garden Edinburgh, 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, and Alex D. Twyford

“The genome sequence of the European mistletoe, *Viscum album* L. (Santalales: Viscaceae)”

In this Data Note the authors present a genome assembly from a female specimen of *Viscum album* (European mistletoe; Streptophyta; Magnoliopsida; Santalales; Viscaceae), a scaffolded genome sequence assembly of 94 261.04 megabases mostly (98.67%) into 49 chromosomal pseudomolecules, including the B1 chromosome. Collectively, the findings of the study, the nuclear genome of *Viscum album*, represents a significant advancement because contributes to our understanding on the evolution of parasitic lifestyles in plants. Perhaps a sentence to be added on how the genome of *Viscum album* would improve our understanding of mistletoes biology. However, to strengthen the manuscript, the analysis would benefit from being placed within a broader genomic context by integrating existing organellar genomes and previously

published genomic data for this species to enhance the annotation support. Some opportunities for the authors were identified to improve the manuscript in a revision before the manuscript is considered suitable for resubmission.

Main comments

1. This Data Note reports a chromosome-scale nuclear genome assembly for *Viscum album* generated using PacBio HiFi, 10X assembled reads, Hi-C, and RNA-seq, followed by substantial curation. The dataset is likely to be useful. However, several statements in the Background require factual/citation correction, as explained in detail below, and key methodological/interpretive points (coverage calculations; chromosome/pseudomolecule wording; benchmarking language) need clarification to prevent misinterpretation. We also noted that the PDF of the paper to be reviewed includes public Reader Comments (pp. 13–14) that raise substantive issues; where appropriate, those issues are reiterated here (see below) because they are visible to readers and should be addressed explicitly in a revised version.

1. Organelle Genomes. The current manuscript focuses exclusively on the nuclear genome. It is strongly recommended that the authors clarify whether the organelle genomes were also sequenced and assembled as part of their project. If not, a brief justification would be helpful for the reader. If those organelle genomes (chloroplast and mitochondrial) were assembled, their inclusion in the analysis would greatly enrich our understanding of V. album biology. The authors could strengthen their study considerably if these data are acknowledged and integrated, which would, in turn, add robustness to their genome annotation. Moreover, a valuable foundation of prior genomic work on V. album already exists (see Schröder, L., Hohnjec, N., Senkler, M., Senkler, J., Küster, H. and Braun, H.-P. (2022), The gene space of European mistletoe (*Viscum album*). Plant J, 109: 278-294.(Refer to reference 1) and Schröder L, Rupp O, Senkler M, Rugen N, Hohnjec N, Goesmann A, Küster H and Braun H-P (2023) The *Viscum album* Gene Space database. Front. Plant Sci. 14:1193122. (Refer to reference 1) .

1. Clarify the Curation Workflow. The text mentions that data curation was conducted “primarily” in PretextView and HiGlass. To be more precise, it would be helpful to specify if there was a distinct workflow or division of labor between the two tools. For example, you could state: “PretextView was used for the initial contact map scrutiny and scaffolding, while HiGlass was employed for fine-scale border validation and subtle adjustments.” This distinction would give readers a clearer understanding of how each tool contributed to the final assembly.

1. Coverage statement appears internally inconsistent. Please verify the coverage calculations. There is a discrepancy that must be checked, the text reports 6,011.18 Gb of HiFi data. Given the estimated genome size of ~91.71 Gb, this would yield approximately 65× coverage, not the stated 26×. The 26× figure would correspond to only ~2.384 Gb of data. This discrepancy suggests either a miscalculation or confusion. Location: Sequence data / Results text stating “~26× coverage” (PDF pp. 5–6). Please reconcile the stated PacBio HiFi coverage (~26×) with the reported sequencing yield and the estimated haploid genome size. If “26×” refers to post-filtered usable bases, define the filtering criteria and report the post-filter yield actually used for assembly. As written, the numbers suggest a substantially higher raw

yield/coverage.

1. The abstract wording (“49 chromosomal pseudomolecules”) can be read as 49 chromosomes. Please clarify explicitly (in Abstract and Results) that the assembly represents 11 chromosomes (10 A + 1 B), and that chromosomes were submitted as 49 INSDC accessions due to per-record length constraints (if that is the reason). This is important because public assembly summaries may display “49 chromosomes,” which will confuse readers. Please Clarify “49 chromosomal pseudomolecules” vs 11 chromosomes (risk of misinterpretation). Location: Abstract (PDF p. 1) and Results/Assembly description (PDF pp. 6–7; Table 3).

1. The manuscript reports an excellent contiguity/QV profile, but the primary assembly shows lower BUSCO/*k*-mer completeness than the stated benchmark thresholds (while the combined haplotypes are high). Please (*i*) define precisely what you mean by “meets the recommended reference standard,” and (*ii*) explain why BUSCO/*k*-mer behave differently for the primary assembly (e.g., primary vs combined; biology; limitations of the chosen assessment). Revise wording to avoid over-claiming. Benchmarking language (“meets recommended reference standard”) should be qualified. Location: Table 4 and adjacent text (PDF p. 9).

1. Scope clarification: organelle genomes and prior *V. album* genomic resources (align with Reader Comment). Please state explicitly whether plastid and mitochondrial genomes were recovered/assembled and deposited. If not, say so and explain why (even briefly). Also, please cite and contextualize the most relevant prior *V. album* resources that are directly useful for annotation/comparative work (e.g., organelle genome work and the “Gene Space” resources). Location: Background and/or Data availability (PDF pp. 3 and 10).

1. Background: correct/qualify statements already flagged by Reader Comments. The following statements require verification and alignment between claim and citation: (a) Distribution/trends in Britain (ensure the cited source supports the precise statement; revise if needed). (b) Yield impact (“up to 50%”): either narrow to what the cited study supports (context-specific) or add broader/recent synthesis. (c) Bird “immunity” to viscumin: “immune” is too strong unless directly supported; re-check the cited paper and rephrase conservatively (tolerance/limited sensitivity) with appropriate citation(s). Location: Background (PDF p. 3). In general, the background information on the natural history of *Viscum album* needs to be updated and improved according to comments and available information.

1. Methods: flow cytometry text is incomplete / not reproducible as written. The sentence describing nuclei isolation appears grammatically incomplete and does not clearly specify the buffer/protocol used. Please correct the sentence and provide the minimum details needed for reproducibility (buffer name/composition or an unambiguous protocol identifier, and any modifications). Location: Flow cytometry section (PDF p. 3).

1. Methods: DNA barcoding is underspecified. “Standard barcode markers” is too generic. Please specify which loci were amplified/sequenced (e.g., rbcL/matK/ITS, as applicable), and provide primer/protocol identifiers and matching/threshold criteria (GenBank/BOLD), to support traceability of specimen identity. Location: DNA barcoding section (PDF pp. 3–4).

1. RNA-seq provenance relative to the genome specimen. Please clarify whether RNA-seq derives from the same individual as the genome specimen. If RNA-seq is from a different specimen/tissue, state this explicitly and justify briefly, noting any implications for annotation. Location: Nucleic acid extraction / Table 1 (PDF pp. 4 and 6).

1. Methods. Sample acquisition. Why it is important to indicate that the sampled individual plant was a female plant? Would you expect a different result if the sampled individual was a male plant?

Minor concerns

1. Terminology: “10 autosomes and the B chromosome”. Consider revising to “10 A chromosomes plus one B chromosome” and add a short note that B chromosomes are supernumerary and can vary among individuals/populations, so users interpret this assembly correctly. Location: Assembly description (PDF p. 6).

1. Reference list / software list formatting. The “Software Version Source” block appears misplaced/duplicated relative to the bibliographic references. Please separate software reporting from the reference list and format both consistently. Location: Start of References (PDF p. 11).

1. Grammar: “that connect via a haustorium” > “that connects via a haustorium.”

1. Wording/typo: “collected from on planted trees” > “collected from planted trees” (or equivalent).

1. Units/style: standardize Mb/Gb (and avoid ambiguous “megabases” phrasing); state the total length once clearly, then use one unit consistently.

1. Table 3 readability: clarify the chromosome-part naming convention (e.g., “Chr5_part1/part2”) so the split across accessions is immediately interpretable.

Suggested references (if not already included)

Briggs (2021), British & Irish Botany (status/trends in Britain & Ireland). DOI: 10.33928/bib.2021.03.419.
Petersen et al. (2015), Scientific Reports (mitochondrial genome features/gene loss). DOI: 10.1038/srep17588.
Schröder et al. (2022), The Plant Journal (gene space). DOI: 10.1111/tpj.15558.
Schröder et al. (2023), Frontiers in Plant Science (Gene Space database). DOI: 10.3389/fpls.2023.1193122.

(Optional, for nuanced resource acquisition language) Thomas et al. (2023) indicates discussion that mistletoe extracts at least some carbon from the host.

(Optional, for host performance under stress) González de Andrés et al. (2024), Tree Physiology,

on mistletoe impacts on tree carbon/water/nutrients under environmental stress.

Recommendation
Approved with Reservations. The dataset is likely sound and useful, but accuracy, clarity, and reproducibility require the revisions above, particularly (*i*) chromosome/pseudomolecule wording, (*ii*) coverage calculation transparency, (*iii*) qualified benchmarking language, and (*iv*) corrected Background claims with appropriate citations.

References
1. Schröder L, Hohnjec N, Senkler M, Senkler J, et al.: The gene space of European mistletoe (Viscum album). *The Plant Journal*. 2022; **109** (1): 278-294 Publisher Full Text
2. Schröder L, Rupp O, Senkler M, Rugen N, et al.: The Viscum album Gene Space database. *Frontiers in Plant Science*. 2023; **14**. Publisher Full Text
3. Briggs J: Mistletoe, Viscum album (Santalaceae), in Britain and Ireland; a discussion and review of current status and trends. *British & Irish Botany*. 2021; **3** (4). Publisher Full Text
4. Petersen G, Cuenca A, Møller I, Seberg O: Massive gene loss in mistletoe (Viscum, Viscaceae) mitochondria. *Scientific Reports*. 2015; **5** (1). Publisher Full Text

Is the rationale for creating the dataset(s) clearly described?
Partly

Are the protocols appropriate and is the work technically sound?
Partly

Are sufficient details of methods and materials provided to allow replication by others?
No

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

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Phylogeography, plant-hummingbird interactions, mistletoes, 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, however we have significant reservations, as outlined above.

Comments on this article

Version 1

Reader Comment 04 Jan 2026

Hans-Peter Braun, Plant Genetics, Leibniz University Hannover, Hanover, Germany

Plant cells have genomes in the cell nucleus, mitochondria and plastids. The authors only comment on the nuclear genome of V. album. Were the organelle genomes also sequenced? Previous genome and mRNA sequencing efforts for V. album should be mentioned and are likely useful for data annotation of the Viscum album genome sequence.

Publications:

Petersen G, Cuenca A, Møller IM, Seberg O. Massive gene loss in mistletoe (Viscum, Viscaceae) mitochondria. Sci Rep. 2015 Dec 2;5:17588. doi: 10.1038/srep17588. PMID: 26625950; PMCID: PMC4667250.

Petersen G, Cuenca A, Seberg O. Plastome Evolution in Hemiparasitic Mistletoes. Genome Biol Evol. 2015 Aug 29;7(9):2520-32. doi: 10.1093/gbe/evv165. PMID: 26319577; PMCID: PMC4607522.

Schröder L, Hohnjec N, Senkler M, Senkler J, Küster H, Braun HP. The gene space of European mistletoe (Viscum album). Plant J. 2022 Jan;109(1):278-294. doi: 10.1111/tpj.15558. Epub 2021 Nov 12. PMID: 34713513.

Schröder L, Rupp O, Senkler M, Rugen N, Hohnjec N, Goesmann A, Küster H, Braun HP. The *Viscum album* Gene Space database. Front Plant Sci. 2023 Jun 26;14:1193122. doi: 10.3389/fpls.2023.1193122. PMID: 37484460; PMCID: PMC10359728.

See also the "*Viscum album* Gene Space Database (2023)" at https://viscumalbum.pflanzenproteomik.de/ Hans-Peter Braun (coauthor of some of the above cited publications)

Competing Interests: No competing interests were disclosed.

Reader Comment 28 Sep 2025

Jonathan Briggs, Not Applicable, UK

My previous comment on the distribution sentence in this draft seems to have been deleted. But basically it was to state that the sentence "It is most common in the south and east of Britain, occurring in just a few sites in Scotland although it appears to be expanding there (Stroh *et al.*, 2023)." is incorrect. Stroh et al do not say this. They state that it is more common in the east than it used to be, not that is most common there. They also do not mention Scotland or that it is expanding there. It would be better to cite the sourec at least resaonably accurately! Or, better still, to cite a source that actually discusses the distribution trend. In this case, as Thomas et al 2023 is already being cited elsewhere in the paper, use that, which does discuss distribution trends and makes it clear that the species is most common in the south and west (defiintely not east!) but expanding generally. In Scotland it is very rare and there is no sign of any significant natural expansion. Jonathan Briggs (a co-author of Thomas et al)

Competing Interests: No competing interests were disclosed.

Reader Comment 21 Sep 2025

Jonathan Briggs, Not Applicable, UK

Another couple of comments on some factual issues in the introductory text: The sentence "Although it does not kill its hosts, the European mistletoe is considered a pest because it reduces fruit yields by up to 50% (Preston, 1977)" is oddly inadequate, referring only to a paper on fruit trees from 1977 and ignoring all much more recent work describing increasing mistletoe problems on fruit and many other trees, in Britain and across Europe. These are reviewed in the already cited Thomas et al 2023. The text regarding Viscum berry-feeding birds, saying that they "are immune to the toxic lectin viscumin in the berries (Olsnes et al., 1982)" is questionable and misquotes the source. The source quoted does not describe toxins in the berries (it used leaf tissue) nor does it reference immunity in any bird species. The toxicity it did assess was concentrated Viscocumin administered by injection, not unconcentrated leaf (or berry) tissue consumed. It did not suggest that consumption of mistletoe is problematic for herbivores. There is a review of the toxic elements of mistletoe and of mistletoe herbivory (widespread in insects, mammals and birds) in both Thomas et al (2023) and Briggs (2021) (refs below). It cannot be suggested that particular birds have immunity to toxins without any evidence that the toxins are a problem to any herbivory. Viscum tissues contain toxins that are dangerous when concentrated and injected, but it doesn't follow that Viscum is toxic to consume. Thomas, P. A., Dering, M., Giertych, M. J., Iszkuło, G., Tomaszewski, D., & Briggs, J. (2023). Biological Flora of Britain and Ireland: Viscum album. Journal of Ecology, 111, 701–739. https://doi.org/10.1111/1365-2745.14036 Briggs, Jonathan (2021). Mistletoe, Viscum album (Santalaceae), in Britain and Ireland: a discussion and review of current status and trends British & Irish Botany Dec 2021 3(4): 419-454 https://doi.org/10.33928/bib.2021.03.419

Competing Interests: No competing interests were disclosed.