DATA NOTE

The genome sequence of common vervain, *Verbena officinalis* L. (Verbenaceae)

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

Maarten J. M. Christenhusz[0], Ilia J. Leitch[0], José Ignacio Márquez-Corro[0],
Royal Botanic Gardens Kew Genome Acquisition Lab,
Plant Genome Sizing collective, Darwin Tree of Life Barcoding collective,
Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team,
Wellcome Sanger Institute Scientific Operations: Sequencing Operations,
Wellcome Sanger Institute Tree of Life Core Informatics team,
Tree of Life Core Informatics collective, Darwin Tree of Life Consortium

v1 **First published:** 18 Oct 2024, **9**:612
https://doi.org/10.12688/wellcomeopenres.23188.1
Latest published: 18 Oct 2024, **9**:612
https://doi.org/10.12688/wellcomeopenres.23188.1

Abstract

We present a genome assembly of a specimen of common vervain, *Verbena officinalis* (Streptophyta; Magnoliopsida; Lamiales; Verbenaceae). The genome sequence has a total length of 289.20 megabases. Most of the assembly is scaffolded into 7 chromosomal pseudomolecules. The mitochondrial and plastid genome assemblies have lengths of 495.81 kilobases and 153.46 kilobases, respectively. Gene annotation of this assembly on Ensembl identified 25,194 protein-coding genes.

Keywords

Verbena officinalis, common vervain, genome sequence, chromosomal, Lamiales

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ✓ ?

	1	2
version 1 18 Oct 2024	✓ view	? view

1. **Daniel G Peterson**, Mississippi State University, Mississippi, USA

2. **Martin Laforest**, Research and Development Centre, Agriculture and Agri-Food Canada, Saint-Jean-sur-Richelieu, Canada

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; **Leitch IJ**: Writing – Original Draft Preparation, Writing – Review & Editing; **Márquez-Corro JI**: Writing – Original Draft Preparation;

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome through core funding to the Wellcome Sanger Institute [206194, https://doi.org/10.35802/206194] and the Darwin Tree of Life Discretionary Award [218328, https://doi.org/10.35802/218328].
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Christenhusz MJM, Leitch IJ, Márquez-Corro JI *et al.* **The genome sequence of common vervain, *Verbena officinalis* L. (Verbenaceae) [version 1; peer review: 1 approved, 1 approved with reservations]** Wellcome Open Research 2024, **9**:612 https://doi.org/10.12688/wellcomeopenres.23188.1

First published: 18 Oct 2024, **9**:612 https://doi.org/10.12688/wellcomeopenres.23188.1

Species taxonomy

Eukaryota; Viridiplantae; Streptophyta; Streptophytina; Embryophyta; Tracheophyta; Euphyllophyta; Spermatophyta; Magnoliopsida; Mesangiospermae; eudicotyledons; Gunneridae; Pentapetalae; asterids; lamiids; Lamiales; Verbenaceae; Verbeneae; *Verbena*; *Verbena officinalis* L. (NCBI:txid79772).

Background

Common vervain, *Verbena officinalis* L., is a herbaceous perennial plant with opposite, deeply lobed leaves on square hollow stems that are topped with bifurcating spikes bearing pink to purple flowers. It flowers from June to October and provides ample nectar for a variety of insect pollinators. It prefers chalky soils and can be seen growing in grassland and on rough ground, mostly in the temperate and subtropical regions of Eurasia, Africa and Australia (POWO, 2024).

Verbena officinalis has been used in folk medicine for its analgesic and anti-inflammatory properties (Akerreta *et al.*, 2007; Calvo, 2006; Kubica *et al.*, 2020). It was widely cultivated as a medicinal herb in medieval gardens, but frequently returned to the wild, and is now found in many countries around the world, including Britain and Ireland (POWO, 2024).

The presence of *V. officinalis* in Britain has been dated back to the Neolithic, so it is one of the approximately 100 archaeophytes known from these islands (Preston *et al.*, 2004; Williamson *et al.*, 2008). It was probably introduced from southern Europe, where it is a component of the Mediterranean flora. In Britain and Ireland, the species usually inhabits rough grassland, coastal cliffs and roadside verges, field margins and waste grounds, but it is also frequently found in urban areas, along walls and pavements, because it thrives in hot, dry and disturbed habitats and can be weedy in some situations (Xu & Chang, 2017). Here we sampled a specimen (Figure 1) from the verge on Lower Ham Road along the River Thames in Canbury Gardens, Kingston upon Thames, Surrey. This verge frequently floods in winter, but is dry and hot in summer, and the population there consists of a few hundred individuals.

The genus name *Verbena* comes from classical Latin, referring to scented twigs used for ceremonial purposes. Sacred to the Romans, *V. officinalis* was presented with other scented herbs on altars dedicated to Jupiter. It was equally esteemed among Celtic tribes, who crowned their bards with this herb to receive divine inspiration. The association with divinity has remained in local folklore across Europe. There are still large Verbena celebrations in Spain in honour of local saints. The herb has also long been used as a magic charm for protection against the devil, witchcraft or even vampires. It was worn in shoes or around the ankles for endurance during long hikes in the Netherlands, Germany and Switzerland, and in England the roots were worn around the neck to cure lymph ailments and infections. Crusaders spread the rumour that vervain sprouted on Mount Calvary when Jesus Christ was crucified, hence it is sometimes referred to in England as 'herb-of-the-cross' or 'holy herb'. This has led to the belief that it has helped staunch the wounds of Jesus (Christenhusz *et al.*, 2017; De Cleene & Lejeune, 2000).

Taxonomically, *Verbena officinalis* has two accepted varieties (POWO, 2024), of which the typical variety (*V. officinalis* var. *officinalis*) is the one found in Britain and Ireland and sequenced here. The other variety, *V. officinalis* var. *africana* (R.Fern. & Verdc.) Munir, has more divided and serrated lobes, as well as floral bracts half to two-thirds of the calyx length (Munir, 2002), and is found in Africa, East Asia and Australia.

Karyotypically, *Verbena officinalis* is considered to be a stable diploid species with $2n = 14$ chromosomes (Henniges *et al.*, 2022; Munir, 2002). Here, we present the first high-quality common vervain genome. It is the first for the tribe

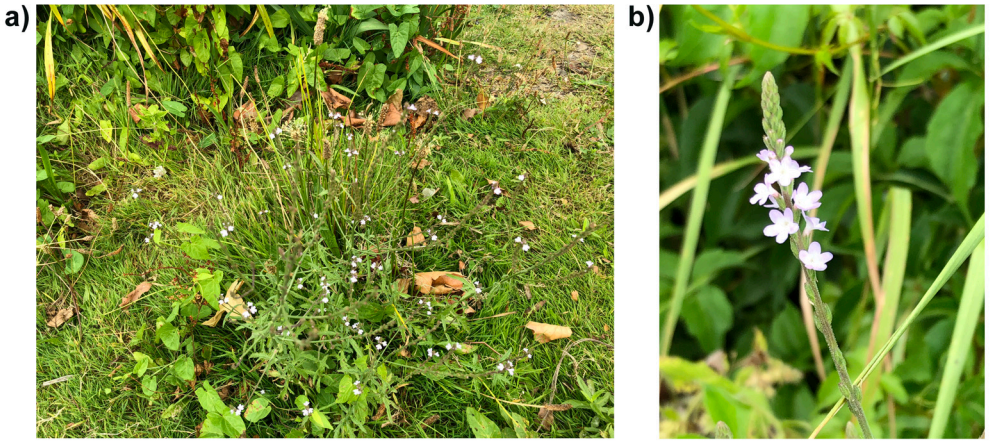

Figure 1. Photographs of the *Verbena officinalis* (daVerOffi1) plant that was sampled for genome sequencing. a) habit. **b**) inflorescence.

Verbeneae, and only the third species of Verbenaceae to have its entire genome sequenced. The other two species with whole genomic data are *Lantana* × *strigocamara* R.W.Sanders (as '*L. camara*'; Joshi *et al.*, 2022; Sanders, 2006) and *Petrea volubilis L.* (Hamilton *et al.*, 2023). The output of this work constitutes an important genome resource to improve investigations on the biosynthetic pathways for the many compounds that have been reported to be of medicinal value (Kubica *et al.*, 2020).

Genome sequence report

Using flow cytometry, the genome size (1C-value) of the *Verbena officinalis* specimen was estimated to be 0.36 pg, equivalent to 350 Mb. The genome was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating a total of 9.37 Gb (gigabases) from 0.72 million reads, providing approximately 27-fold coverage. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data, which produced 118.43 Gb from 784.31 million reads, yielding an approximate coverage of 410-fold. Specimen and sequencing details are provided in Table 1.

Manual assembly curation corrected two missing joins or mis-joins, reducing the assembly length by 0.22%, and increasing the scaffold N50 by 17.74%. The final assembly has a total length of 289.20 Mb in 12 sequence scaffolds with a scaffold N50 of 42.3 Mb (Table 2) with 7 gaps. The snail plot in Figure 2 provides a summary of the assembly statistics, while the distribution of assembly scaffolds based on GC proportion and coverage is shown in Figure 3. The cumulative assembly plot in Figure 4 shows curves for subsets of scaffolds assigned to different phyla. Most (99.76%) of the assembly sequence was assigned to 7 chromosomal-level scaffolds. Chromosome-scale scaffolds confirmed by the Hi-C data are named in order of size (Figure 5; Table 3). While not fully phased, the assembly deposited is of one haplotype. Contigs corresponding to the second haplotype have also been deposited. The mitochondrial and plastid genomes were also assembled and can be found as contigs within the multifasta file of the genome submission.

The estimated Quality Value (QV) of the final assembly is 63.1 with *k*-mer completeness of 100.0%, and the assembly has a BUSCO v5.4.3 completeness of 97.8% (single = 92.7%, duplicated = 5.1%), using the eudicots_odb10 reference set (n = 2,326).

Metadata for specimens, BOLD barcode results, spectra estimates, sequencing runs, contaminants and pre-curation assembly statistics are given at https://links.tol.sanger.ac.uk/species/79772.

Genome annotation report

The *Verbena officinalis* genome assembly (GCA_958496215.1) was annotated at the European Bioinformatics Institute (EBI) on Ensembl Rapid Release. The resulting annotation includes 43,020 transcribed mRNAs from 25,194 protein-coding and 5,008 non-coding genes (Table 2; https://rapid.ensembl.org/Verbena_officinalis_GCA_958496215.1/Info/Index). The average

Table 1. Specimen and sequencing data for *Verbena officinalis*.

Project information			
Study title	*Verbena officinalis*		
Umbrella BioProject	PRJEB64064		
BioSample	SAMEA7521929		
NCBI taxonomy ID	79772		
Specimen information			
Technology	**ToLID**	**BioSample accession**	**Organism part**
PacBio long read sequencing	daVerOffi1	SAMEA7521946	flower
Hi-C sequencing	daVerOffi1	SAMEA7521949	leaf
RNA sequencing	daVerOffi1	SAMEA7521947	leaf
Sequencing information			
Platform	**Run accession**	**Read count**	**Base count (Gb)**
Illumina NovaSeq 6000 (Hi-C)	ERR11679374	7.84e+08	118.43
PacBio Sequel IIe (WGS)	ERR11673227	1.02e+05	1.5
PacBio Sequel IIe (WGS)	ERR11673228	6.14e+05	7.86
RNA Illumina HiSeq 4000 (RNA)	ERR11679375	9.13e+07	13.79

Table 2. Genome assembly data for *Verbena officinalis*, daVerOffi1.1.

Genome assembly		
Assembly name	daVerOffi1.1	
Assembly accession	GCA_958496215.1	
Accession of alternate haplotype	*GCA_958496265.1*	
Span (Mb)	289.20	
Number of contigs	21	
Contig N50 length (Mb)	35.9	
Number of scaffolds	12	
Scaffold N50 length (Mb)	42.3	
Longest scaffold (Mb)	51.29	
Assembly metrics*		***Benchmark***
Consensus quality (QV)	63.1	*≥ 50*
k-mer completeness	100.0%	*≥ 95%*
BUSCO**	C:97.8%[S:92.7%,D:5.1%], F:0.3%,M:1.8%,n:2,326	*C ≥ 95%*
Percentage of assembly mapped to chromosomes	99.76%	*≥ 95%*
Organelles	Mitochondrial genome: 495.81 kb; plastid genome: 153.46 kb	*complete single alleles*
Genome annotation at Ensembl		
Number of protein-coding genes	25,194	
Number of non-coding genes	5,008	
Number of gene transcripts	43,020	

* Assembly metric benchmarks are adapted from column VGP-2020 of "Table 1: Proposed standards and metrics for defining genome assembly quality" from Rhie *et al.* (2021).

** BUSCO scores based on the eudicots_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/daVerOffi1_1/dataset/daVerOffi1_1/busco.

transcript length is 3,176.10. There are 1.42 coding transcripts per gene and 5.54 exons per transcript.

Methods

Sample acquisition, DNA barcoding and genome size estimation

A specimen of *Verbena officinalis* (specimen ID KDTOL10020, ToLID daVerOffi1) was collected from Canbury Gardens, Kingston Upon Thames, Surrey, UK (latitude 51.42, longitude –0.31). The specimen was picked by hand from a riparian road verge by Maarten Christenhusz. The specimen was also formally identified by Maarten Christenhusz, and then preserved by freezing at –80°C. The herbarium specimen *Christenhusz 9017* of the sequenced plant is deposited at the Royal Botanic Gardens, Kew (K001400635).

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 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

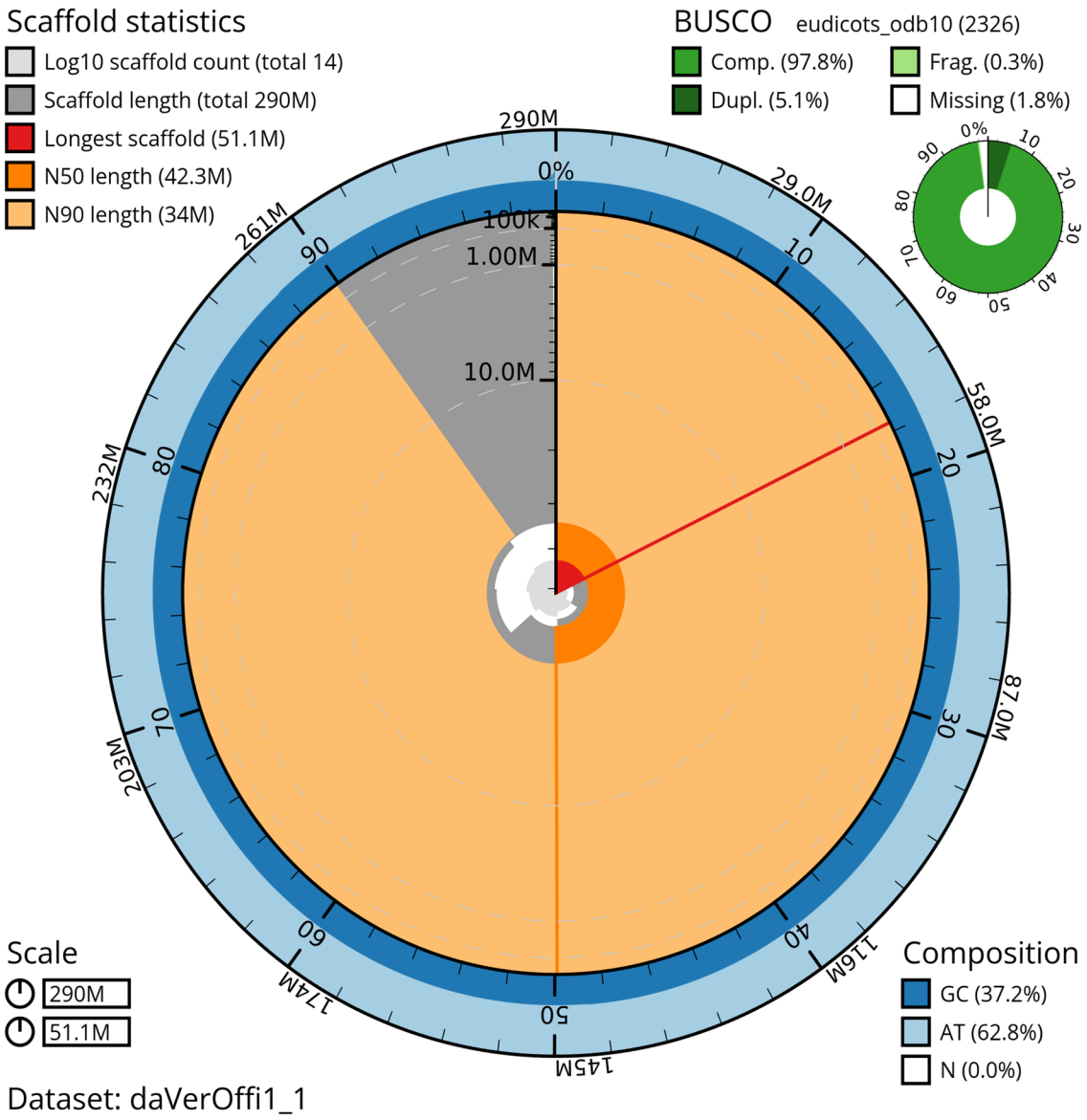

Figure 2. Genome assembly of *Verbena officinalis*, daVerOffi1.1: metrics. The BlobToolKit snail plot shows N50 metrics and BUSCO gene completeness. The main plot is divided into 1,000 size-ordered bins around the circumference with each bin representing 0.1% of the 289,850,359 bp assembly. The distribution of scaffold lengths is shown in dark grey with the plot radius scaled to the longest scaffold present in the assembly (51,139,607 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 scaffold lengths (42,305,780 and 33,961,201 bp), respectively. The pale grey spiral shows the cumulative scaffold count on a log scale with white scale lines showing successive orders of magnitude. The blue and pale-blue area around the outside of the plot shows the distribution of GC, AT and N percentages in the same bins as the inner plot. A summary of complete, fragmented, duplicated and missing BUSCO genes in the eudicots_odb10 set is shown in the top right. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/daVerOffi1_1/dataset/daVerOffi1_1/snail.

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 plant's 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

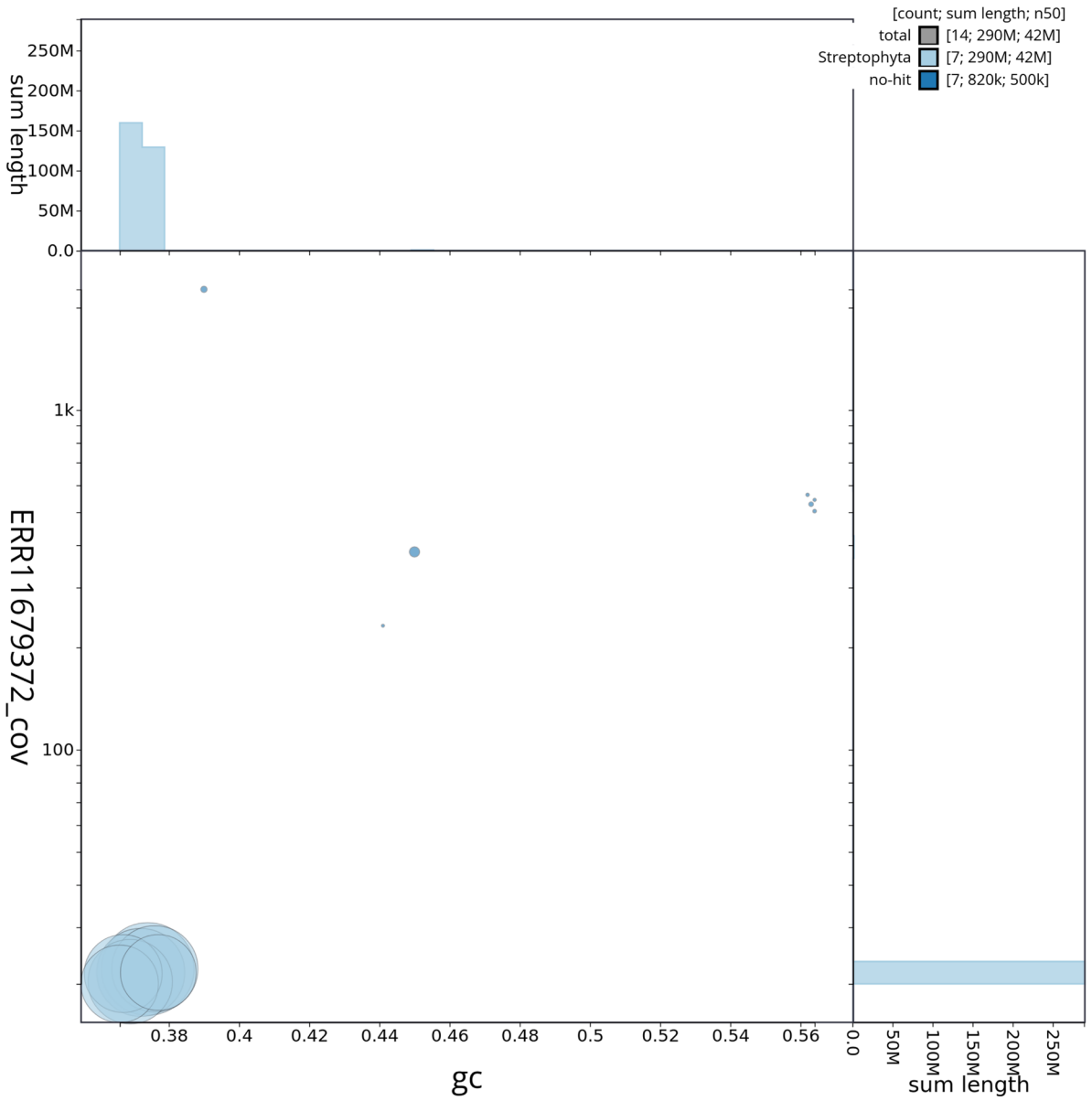

Figure 3. Genome assembly of *Verbena officinalis*, daVerOffi1.1: Blob plot of base coverage against GC proportion for sequences in assembly daVerOffi1.1 Scaffolds are coloured by phylum. Circles are sized in proportion to scaffold length. Histograms show the distribution of scaffold length sum along each axis. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/daVerOffi1_1/dataset/daVerOffi1_1/blob.

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 Wellcome Sanger Institute (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). The daVerOffi1 sample was weighed and dissected on dry ice (Jay *et al.*, 2023) and flower tissue was cryogenically disrupted using the Covaris cryoPREP® Automated Dry Pulverizer (Narváez-Gómez *et al.*, 2023).

HMW DNA was extracted using the Plant Organic Extraction protocol (Jackson & Howard, 2023b). HMW DNA was

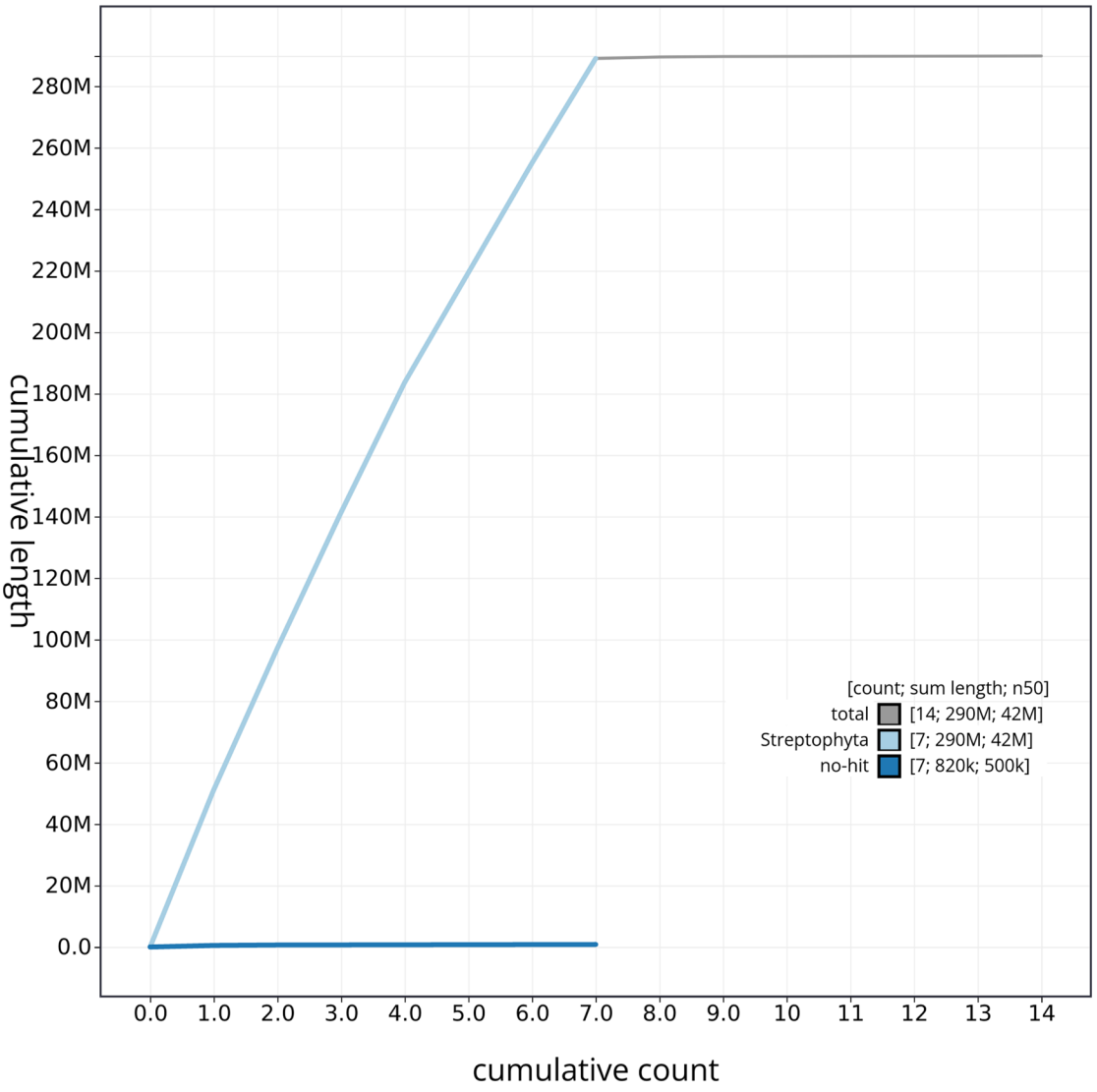

Figure 4. Genome assembly of *Verbena officinalis*, daVerOffi1.1: BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/daVerOffi1_1/dataset/daVerOffi1_1/cumulative.

extracted using the Automated Plant MagAttract v4 protocol (Jackson & Howard, 2023a). 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 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.

RNA was extracted from leaf tissue of daVerOffi1 in the Tree of Life Laboratory at the WSI using the RNA Extraction: Automated MagMax™ *mir*Vana protocol (do Amaral *et al.*, 2023). 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.

Hi-C preparation

Leaf tissue of the daVerOffi1 sample was processed at the WSI Scientific Operations core, using the Arima-HiC v2 kit. In brief, frozen tissue (stored at –80 °C) was fixed, and the DNA crosslinked using a TC buffer with 22% formaldehyde. After crosslinking, the tissue was homogenised using the Diagnocine Power Masher-II and BioMasher-II tubes and pestles.

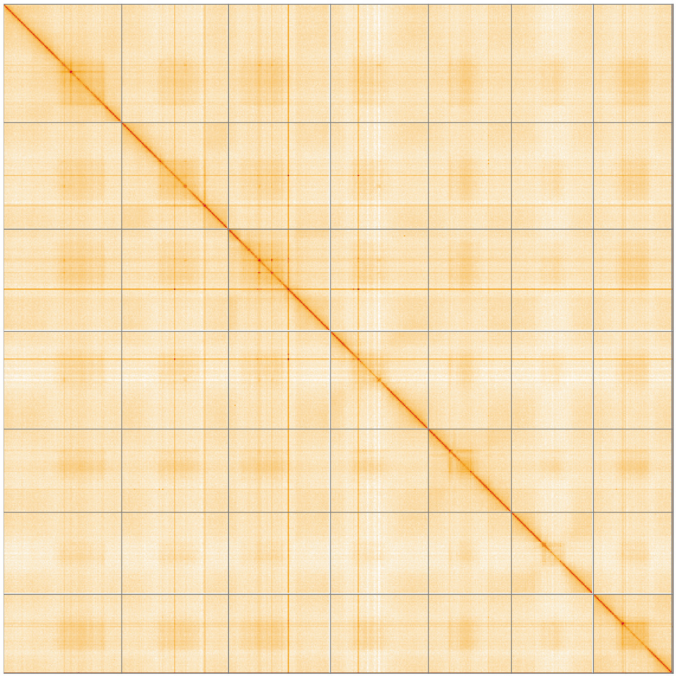

Figure 5. Genome assembly of *Verbena officinalis*, daVerOffi1.1: Hi-C contact map of the daVerOffi1.1 assembly, visualised using HiGlass. Chromosomes are shown in order of size from left to right and top to bottom. An interactive version of this figure may be viewed at https://genome-note-higlass.tol.sanger.ac.uk/l/?d=TwrNDtIuRxewJTJmfbSgiQ.

Table 3. Chromosomal pseudomolecules in the genome assembly of *Verbena officinalis*, daVerOffi1.

INSDC accession	Name	Length (Mb)	GC%
OY292385.1	1	51.14	37.5
OY292386.1	2	46.09	37.0
OY292387.1	3	44.14	37.5
OY292388.1	4	42.31	37.0
OY292389.1	5	35.93	36.5
OY292390.1	6	35.45	36.5
OY292391.1	7	33.96	37.5
OY292392.1	MT	0.5	45.0
OY292393.1	Pltd	0.15	39.0

Following the kit manufacturer's instructions, crosslinked DNA was digested using a 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.

Library preparation and sequencing

Library preparation and sequencing were performed at the WSI Scientific Operations core. Pacific Biosciences HiFi circular consensus DNA sequencing libraries were prepared using the PacBio Express Template Preparation Kit v2.0 (Pacific Biosciences, California, USA) as per 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, California, USA) and size selection step to remove templates <3kb using AMPure PB modified SPRI. DNA concentration was quantified using the Qubit Fluorometer v2.0 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 and gDNA 165kb gDNA and 55kb BAC analysis kit. Samples were sequenced using the Sequel IIe system (Pacific Biosciences, California, USA). The concentration of the library loaded onto the Sequel IIe was between 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, as well as perform primary and secondary analysis of the data upon completion.

RNA-Seq libraries were constructed using the NEB Ultra II RNA Library Prep kit, following the manufacturer's instructions. RNA sequencing was performed on the Illumina HiSeq 4000 instrument.

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 following manufacturers' instructions. The 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 original assembly of HiFi reads was performed using the Hifiasm (Cheng *et al.*, 2021) with the --primary option. Haplotypic duplications were identified and removed with purge_dups (Guan *et al.*, 2020). Hi-C reads were further mapped with bwa-mem2 (Vasimuddin *et al.*, 2019) to the primary contigs, which were further scaffolded using the provided Hi-C data (Rao *et al.*, 2014) in YaHS (Zhou *et al.*, 2023) using the --break option. Scaffolded assemblies were evaluated using Gfastats (Formenti *et al.*, 2022), BUSCO (Manni *et al.*, 2021) and MERQURY.FK (Rhie *et al.*, 2020). The organelle genomes were assembled using OATK (Zhou, 2023).

Curation

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

A Hi-C map for the final assembly was produced using bwa-mem2 (Vasimuddin *et al.*, 2019) in the Cooler file format (Abdennur & Mirny, 2020). To assess the assembly metrics, the *k*-mer completeness and QV consensus quality values were calculated in Merqury (Rhie *et al.*, 2020). This work was done using the "sanger-tol/readmapping" (Surana *et al.*, 2023a) and "sanger-tol/genomenote" (Surana *et al.*, 2023b) pipelines. The genome evaluation pipelines were 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.

The genome was also analysed within the BlobToolKit environment (Challis *et al.*, 2020) and BUSCO scores (Manni *et al.*, 2021) were calculated. 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 submission of materials by a Darwin Tree of Life Partner is subject to the **'Darwin Tree of Life Project Sampling Code of**

Table 4. Software tools: versions and sources.

Software tool	Version	Source
BlobToolKit	4.2.1	https://github.com/blobtoolkit/blobtoolkit
BUSCO	5.3.2	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
Gfastats	1.3.6	https://github.com/vgl-hub/gfastats
Hifiasm	0.19.5-r587	https://github.com/chhylp123/hifiasm
HiGlass	44086069ee7d4d3f6f3f00 12569789ec138f42b84aa4 4357826c0b6753eb28de	https://github.com/higlass/higlass
Merqury.FK	d00d98157618f4e8d1a91 90026b19b471055b22e	https://github.com/thegenemyers/MERQURY.FK
OATK	0.9	https://github.com/c-zhou/oatk
PretextView	0.2	https://github.com/wtsi-hpag/PretextView
purge_dups	1.2.5	https://github.com/dfguan/purge_dups
sanger-tol/genomenote	v1.0	https://github.com/sanger-tol/genomenote
sanger-tol/readmapping	1.1.0	https://github.com/sanger-tol/readmapping/tree/1.1.0
YaHS	1.2a.2	https://github.com/c-zhou/yahs

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: *Verbena officinalis*. Accession number PRJEB64064; https://identifiers.org/ena.embl/PRJEB64064 (Wellcome Sanger Institute, 2023). The genome sequence is released openly for reuse. The *Verbena officinalis* 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. 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

Akerreta S, Cavero RY, Calvo MI: **First comprehensive contribution to medical ethnobotany of Western Pyrenees.** *J Ethnobiol Ethnomed.* 2007; **3**(1): 26.
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

Calvo MI: **Anti-inflammatory and analgesic activity of the topical preparation of *Verbena officinalis* L.** *J Ethnopharmacol.* 2006; **107**(3): 380–382.
PubMed Abstract | Publisher 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

Christenhusz MJM, Fay MF, Chase MW: **Plants of the World: an illustrated encyclopedia of vascular plants.** The University of Chicago Press, 2017.
Publisher 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

De Cleene M, Lejeune MC: **Compendium van rituele planten in Europa.** Ghent: Uitgeverij Stichting Mens en Kultuur, 2000.
Reference Source

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 synteny and structural variation.** *Genome Biol.* 2023; **24**(1): 74.
PubMed Abstract | Publisher Full Text | Free Full Text

do Amaral RJV, Denton A, Yatsenko H, *et al.*: **Sanger Tree of Life RNA**

extraction: automated MagMax™ mirVana. *protocols.io.* 2023.
Publisher 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

Hamilton JP, Vaillancourt B, Wood JC, *et al.*: **Chromosome-scale assembly of the Verbenaceae species Queen's Wreath (*Petrea volubilis* L.).** *BMC Genom Data.* 2023; **24**(1): 14.
PubMed Abstract | Publisher Full Text | Free Full Text

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

Henniges MC, Powell RF, Mian S, *et al.*: **A taxonomic, genetic and ecological data resource for the vascular plants of Britain and Ireland.** *Sci Data.* 2022; **9**(1): 1.
PubMed Abstract | Publisher Full Text | Free 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

Jackson B, Howard C: **Sanger Tree of Life HMW DNA extraction: automated plant MagAttract v.4.** *protocols.io.* 2023a.
Publisher Full Text

Jackson B, Howard C: **Sanger Tree of Life HMW DNA extraction: Plant Organic HMW gDNA Extraction (POE).** *protocols.io.* 2023b.
Publisher 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

Joshi AG, Praveen P, Ramakrishnan U, *et al.*: **Draft genome sequence of an invasive plant *Lantana camara* L.** *Bioinformation.* 2022; **18**(9): 739–741.
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

Kubica P, Szopa A, Dominiak J, *et al.*: ***Verbena officinalis* (Common Vervain) – a review on the investigations of this medicinally important plant species.** *Planta Med.* 2020; **86**(17): 1241–1257.
PubMed Abstract | Publisher 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

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, 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. [Accessed 2 April 2024].
Reference Source

Munir AA: **A taxonomic revision of the genus *Verbena* L. (Verbenaceae) in Australia.** *J Adelaide Bot Garden.* 2002; **20**: 21–103.

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 (Clifton, N.J.).* New York, NY: Humana, 2021; **2222**: 325–361.
PubMed Abstract | Publisher Full Text

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

Preston CD, Pearman DA, Hall AR: **Archaeophytes in Britain.** *Botan J Linnean Soc.* 2004; **145**(3): 257–294.
Publisher Full Text

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

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

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

Rhie A, Walenz BP, Koren S, *et al.*: **Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies.** *Genome Biol.* 2020; **21**(1): 245.
PubMed Abstract | Publisher Full Text | Free Full Text

Sanders RW: **Taxonomy of *Lantana* sect. *Lantana* (Verbenaceae) I: correct application of *Lantana camara* and associated names.** *Sida.* 2006; **22**: 381–421.
Reference Source

Surana P, Muffato M, Qi G: **sanger-tol/readmapping: sanger-tol/readmapping v1.1.0 - Hebridean Black (1.1.0).** *Zenodo.* 2023a.
Publisher Full Text

Surana P, Muffato M, Sadasivan Baby C: **sanger-tol/genomenote (v1.0.dev).** *Zenodo.* 2023b.
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: 1 approved].** *Wellcome Open Res.* 2024; **9**: 339.
Publisher Full Text

Vasimuddin Md, 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

Williamson M, Stout JC, Dehnen-Schmutz K, *et al.*: **A provisional list of Irish archaeophytes.** *Irish Nat J.* 2008; **29**(1): 30–35.
Reference Source

Wellcome Sanger Institute: **The genome sequence of common vervain, *Verbena officinalis* L. (Verbenaceae).** European Nucleotide Archive. [dataset], accession number PRJEB64064, 2023.

Xu Z, Chang L: **Identification and control of common weeds: volume 3.** Singapore: Springer Singapore, 2017.
Publisher Full Text

Zhou C: **c-zhou/oatk: Oatk-0.1**. 2023.
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

Open Peer Review

Current Peer Review Status: ✔ ?

Version 1

Reviewer Report 15 November 2024

https://doi.org/10.21956/wellcomeopenres.25538.r107667

?

Martin Laforest
Research and Development Centre, Agriculture and Agri-Food Canada, Saint-Jean-sur-Richelieu, Quebec, Canada

Maarten J. M. Christenhusz, Ilia J. Leitch, José Ignacio Márquez-Corro are presenting a paper describing sequencing of the common vervain genome. The manuscript is titled “The genome sequence of common vervain, *Verbena officinalis* L. (Verbenaceae)”. Briefly the authors have harvested an individual, which was appropriately identified as belonging the Verbena officinalis species. They have used PacBio and Illumina technologies to sequence the genome, determine scaffold structures using Hi-C and sequenced the transcriptome to aid gene annotation. The results of their assembly is presented with many statistics to highlight the quality of the results. The article is well written and I only have a few points to bring forward. First, I am not sure the paragraph that discuss the role of vervain in different religions is in its place in a genome paper. Second, it is written that this report is the first genome for the tribe Verbeneae, and only the third species of Verbenaceae. I don't understand tribe here. The first of the tribe but third species. To my understanding, tribe is above genus, thus also species, in the taxonomic tree. Maybe I get it wrong, please check.
Finally, in the methods, the NEB Ultra II RNA Library Prep kit is used. If I recall correctly, another NEB kit for rRNA depletion (or mRNA purification) needs to be used. It is not specified which one was used.

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: Weed science and 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, however I have significant reservations, as outlined above.

Reviewer Report 13 November 2024

https://doi.org/10.21956/wellcomeopenres.25538.r106546

Daniel G Peterson
Mississippi State University, Mississippi, USA

Via a Data Note, Christenhusz et al. present a genome assembly for the common verain, *Verbena officinalis* L.

The Background section of their manuscript does an excellent job of introducing readers to *Verbena officinalis* while illuminating the species' historic and potential future value to human beings.

The genome sequencing and assembly techniques are scientifically sound and well documented.

I have a few questions and comments. Most of my comments simply state what I hope the authors (or others) will determine from the data.

[1] Is the herbarium specimen Christenhusz 9017 stored at -80C or is it stored as a dried pressing or in some other way? This is unclear.

[2] I am pleased that the authors did a genome size estimate using flow cytometry. Co-author Leitch is one of the world's foremost experts on plant genome size variation and genome size estimation. I hope to eventually see a version of this manuscript or a follow-up manuscript that discusses the potential variation between the genome assembly and the flow cytometric value.

[3] Are the centromeres absent from the assembly? What makes up the repetitive part of the genome? These are more results that I hope to learn more about when the author's publish their results.

[4] Are their signs of ancient genome duplications? Another result I look forward to hearing about.

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; genome sequencing; plant biology; bioinformatics; cytogenetics

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.