Check for updates

DATA NOTE

ERGA-BGE Reference Genome: *Corema album* ((L.) D. Don, 1830) - the Portuguese crowberry, endemic to the Iberian Peninsula.

[version 1; peer review: awaiting peer review]

João Pedro Marques[1-3*], Bruno Nevado[4,5*], Tom Brown[6,7*], Alessio Iannucci[8], Alice Mouton[8,9], Ann M Mc Cartney[10], Carla Alegria[4,11], Chiara Bortoluzzi[12], Chiara Natali[8], Christian de Guttry[12], Claudio Ciofi[8], Cristina Máguas[4,5], Diego De Panis[6,7], Giulio Formenti[13], Joana Jesus[4], João Jacinto[4,11], José Melo-Ferreira[1,2,14], Joseph Kirangwa[15,16], Judit Salces-Ortiz[17], M. Judite Alves[4,18], Marco Morselli[19], Maria Angela Diroma[8], Philipp H. Schiffer[16], Robert M. Waterhouse[12], Rosa Fernández[17], Sara Vicente[4,20], Sergio Chozas[4,5], Vitor C. Sousa[4,5], Helena Trindade[4,5*], Manuela Sim-Sim[4,5,18*]

[1]CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, InBIO Laboratório Associado, Campus de Vairão, Universidade do Porto, Vairão, 4485-661, Portugal
[2]BIOPOLIS Program in Genomics, Biodiversity and Land Planning, CIBIO, Campus de Vairão, Vairão, 4485-661, Portugal
[3]Institut des Sciences de l'Evolution de Montpellier (ISEM), Université de Montpellier, CNRS, IRD, Montpellier, 34095, France
[4]CE3C, Centre for Ecology, Evolution and Environmental Changes & CHANGE—Global Change and Sustainability Institute, Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal
[5]Departamento Biologia, Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal
[6]Leibniz Institute for Zoo and Wildlife Research, Alfred-Kowalke-Straße 17, Berlin, 10315, Germany
[7]Berlin Center for Genomics in Biodiversity Research (BeGenDiv), Koenigin-Luise-Str 6-8, Berlin, 14195, Germany
[8]Department of Biology, University of Florence, Sesto Fiorentino, 50019, Italy
[9]Laboratoire des Transitions, University of Liege, Place Delcour 17, Liège, 4020, Belgium
[10]Institute of Clinical and Translational Science, University of California, Irvine, CA, USA
[11]INIAV, Instituto Nacional de Investigação Agrária e Veterinária, I.P., Av. da República, Quinta do Marquês, Oeiras, 2780-157, Portugal
[12]SIB Swiss Institute of Bioinformatics, Environmental Bioinformatics group, Amphipôle, Quartier UNIL-Sorge, Lausanne, 1015, Switzerland
[13]The Vertebrate Genome Laboratory, The Rockefeller University, 1230 York Ave, New York City, USA
[14]Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, Porto, 4169-007, Portugal
[15]Department of Biochemistry, University of Oxford, Dorothy Crowfoot Hodgkin Building, South Parks Rd, Oxford, OX1 3QU, UK
[16]worm~lab/BioC2, Institute for Zoology, University of Cologne, Cologne, Germany
[17]Metazoa Phylogenomics and Genome Evolution lab, Institute of Evolutionary Biology (CSIC-Universitat Pompeu Fabra), Passeig Marítim de la Barceloneta 37-49, Barcelona, 08003, Spain
[18]MUHNAC, Museu Nacional de História Natural e da Ciência, Universidade de Lisboa, Rua da Escola Politécnica, 58, Lisboa, 1250-102, Portugal
[19]Department of Chemistry, Life Sciences and Environmental Sustainability, University of Parma, Parco Area delle Scienze 23/A, Parma, 43124, Italy
[20]ERISA—Escola Superior de Saúde Ribeiro Sanches, Instituto Politécnico da Lusofonia (IPLUSO), Rua do Telhal aos Olivais nº 8—8a, Lisboa, 1950-396, Portugal

* Equal contributors

v1 **First published:** 22 Sep 2026, **6**:366
https://doi.org/10.12688/openreseurope.23076.1
Latest published: 22 Sep 2026, **6**:366
https://doi.org/10.12688/openreseurope.23076.1

Abstract

Corema album is a perennial, dioecious shrub endemic to the Atlantic coast of the Iberian Peninsula and the Azores. It inhabits coastal dune ecosystems from Galicia to the Strait of Gibraltar, where it plays an important ecological role. The goal of this study was to assemble and annotate a chromosome-level reference genome for *C. album.* The entirety of the genome sequence was assembled into 13 contiguous chromosomal pseudomolecules, one mitochondrial, and one chloroplast genome. This chromosome-level assembly encompasses 548 Mb, composed of 96 contigs and 68 unplaced scaffolds, with contig and scaffold N50 values of 23.7 Mb and 40.9 Mb, respectively. Structural annotation revealed that 51% of the genome of *C. album* consists of repeats, identified 53,794 genes and encompassed 94.6% of the Viridiplantae BUSCO genes set, confirming a high gene representativeness in the assembly.

Plain Language summary

Corema album is a small shrub that grows naturally along the Atlantic coasts of the Iberian Peninsula and the Azores. It is especially common in sandy coastal dunes, where it plays an important role in maintaining ecosystem stability and supporting local biodiversity. Despite its ecological importance, there is still limited genetic information available for this species.

In this study, we generated a high-quality reference genome for *C. album*, which can aid scientific discoveries surrounding how the plant grows and functions. Using advanced DNA sequencing technologies, we were able to reconstruct its genetic material and organise it into 13 chromosome-scale sequences. This provides a highly detailed and reliable representation of its genome.

We also annotated almost 54 thousands genes and repeats that constitute a large portion of the genome. The high completeness of this genome means it can serve as a valuable resource for future research.

Overall, this work provides an essential foundation for studying the biology, evolution, and conservation of *C. album*, particularly in the face of environmental changes affecting coastal ecosystems.

Keywords

Corema album, genome assembly, European Reference Genome Atlas, Biodiversity Genomics Europe, Earth Biogenome Project, Portuguese crowberry

Open Peer Review

Approval Status *AWAITING PEER REVIEW*

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

This article is included in the Genome Reports from the Biodiversity Genomics Europe Project collection.

Corresponding authors: João Pedro Marques (joao.marques@cibio.up.pt), Manuela Sim-Sim (mmsim-sim@ciencias.ulisboa.pt)

Author roles: Marques JP: Conceptualization, Formal Analysis, Investigation, Methodology, Resources, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Nevado B**: Conceptualization, Formal Analysis, Investigation, Methodology, Resources, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Brown T**: Data Curation, Formal Analysis, Writing – Review & Editing; **Iannucci A**: Investigation, Writing – Review & Editing; **Mouton A**: Funding Acquisition, Project Administration, Writing – Review & Editing; **M Mc Cartney A**: Funding Acquisition, Project Administration, Writing – Review & Editing; **Alegria C**: Investigation, Resources, Writing – Review & Editing; **Bortoluzzi C**: Project Administration, Writing – Review & Editing; **Natali C** : Formal Analysis, Methodology, Resources, Writing – Review & Editing; **de Guttry C**: Project Administration, Writing – Review & Editing; **Ciofi C**: Funding Acquisition, Methodology, Resources, Supervision, Writing – Review & Editing; **Máguas C**: Project Administration, Resources, Supervision, Writing – Review & Editing; **De Panis D**: Data Curation, Software, Writing – Review & Editing; **Formenti G**: Project Administration, Writing – Review & Editing; **Jesus J**: Investigation, Methodology, Writing – Review & Editing; **Jacinto J**: Investigation, Methodology, Resources, Writing – Review & Editing; **Melo-Ferreira J**: Project Administration, Writing – Review & Editing; **Kirangwa J**: Investigation, Methodology, Resources; **Salces-Ortiz J**: Investigation, Methodology, Resources, Writing – Review & Editing; **Alves MJ**: Resources, Writing – Review & Editing; **Morselli M**: Investigation, Methodology; **Diroma MA**: Data Curation, Writing – Review & Editing; **Schiffer PH**: Funding Acquisition, Project Administration, Resources; **Waterhouse RM**: Funding Acquisition, Supervision, Writing – Review & Editing; **Fernández R**: Funding Acquisition, Investigation, Methodology, Resources, Writing – Review & Editing; **Vicente S**: Investigation, Methodology, Writing – Review & Editing; **Chozas S**: Conceptualization, Investigation, Writing – Review & Editing; **Sousa VC** : Conceptualization, Methodology, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Trindade H**: Investigation, Methodology, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Sim-Sim M**: Conceptualization, Data Curation, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work is also part of the Biodiversity Genomics Europe (BGE) project (Grant No. 101059492), funded by Horizon Europe under the Biodiversity, Circular Economy and Environment call (REA.B.3), and co-funded by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract numbers 22.00173 and 24.00054, and by UK Research and Innovation (UKRI) through the Department for Business, Energy and Industrial Strategy's Horizon Europe Guarantee Scheme.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2026 Marques JP *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Marques JP, Nevado B, Brown T *et al.* **ERGA-BGE Reference Genome: *Corema album* ((L.) D. Don, 1830) - the Portuguese crowberry, endemic to the Iberian Peninsula. [version 1; peer review: awaiting peer review]** Open Research Europe 2026, **6**:366 https://doi.org/10.12688/openreseurope.23076.1

First published: 22 Sep 2026, **6**:366 https://doi.org/10.12688/openreseurope.23076.1

Introduction

The Portuguese crowberry, *Corema album* (L.) D. Don (Ericaceae), is a dioecious perennial woody shrub endemic to the Atlantic coast of the Iberian Peninsula (ssp. *album*), and to the Azores Islands (ssp. *azoricum*). In the Iberian Peninsula, it inhabits coastal regions that extend from Galicia to Gibraltar. This Iberian endemism is a crucial species in sand dune habitats, holding a priority conservation status under Council Directive 92/43/EEC of 21 May 1992 on the Conservation of Natural Habitats and of Wild Fauna and Flora. It is also listed as a species of concern in terms of conservation in PROGEN by ICNF (Institute for Nature Conservation and Forests). The influence that climate has on the species and, particularly, on the different responses of the sexes to aridity, has raised concerns about the vulnerability of *C. album* populations inhabiting the southern limit of the distribution range (Díaz-Barradas et al., 2018).

From a conservation perspective, coastal dune systems are deemed natural habitats of high conservation value, particularly concerning vegetation. Considering this, *C. album* is an important contributor to ecosystem functionality, as its roots help in fixing the sand, preventing erosion and maintaining the structure of the dunes. However, dune systems are experiencing escalating disturbances due to various economic and recreational activities linked to the growth of the coastal population (Antunes et al., 2018).

Additionally, the Portuguese crowberry produces edible fruits that are highly nutritious and rich in antioxidants. However, the fruits and leaves contain interesting bioactive profiles with compounds that have yet to be fully characterised (Jacinto et al., 2021).

This presents a promising opportunity for the sustainable valorisation of this natural resource and may facilitate future research into the development of biotechnological applications.

Obtaining a reference genome for *C. album* will contribute to future conservation and evolutionary genomic studies and open the door to studying closely related species. This high-quality reference genome will allow the integration of further resequencing data, with the aim of characterising gene variation within populations and the identification of genes involved in adaptation and other ecologically relevant traits. It will provide a foundation for future studies exploring the genetic basis of secondary metabolites and other traits of ecological or potential economic interest. Taken together, a reference genome for the Portuguese crowberry will create new opportunities for research in conservation biology, evolutionary genomics and plant biology.

The generation of this reference resource was coordinated by the European Reference Genome Atlas (ERGA) Pilot Project (Mc Cartney et al., 2024) and the ERGA Biodiversity Genomics Europe (BGE) project, supporting ERGA's aims of promoting transnational cooperation to promote advances in the application of genomics technologies to protect and restore biodiversity (Mazzoni et al., 2023).

Materials & methods

Sample and sampling information

For genome assembly, a male adult plant was sampled during two field trips (24/11/2021 and 11/01/2022) to Azeitão, near Arrábida Natural Park (38.5325, −9.00694). Young expanding leaves were collected after 48 hours of dark treatment (plant covered with light-opaque paper sheets with a few holes that allowed air flow), according to permits from ICNF (21PTLX00657D). Samples from leaves for DNA extraction were flash frozen at −20 °C and − 80 °C, respectively.

For annotation of the genome assembly and in order to maximise the representativeness of expressed genes available for annotation, we obtained 16 samples representing different tissues, individuals and time points. Specifically, we sampled four different tissues (roots, stems, leaves and flowers) from the same male individual used for genome assembly and from eight unrelated females. Additionally, for each individual x tissue combination, sampling was repeated at two time points, once during early morning and again in the late afternoon. Samples were collected in the Azeitão region, near Arrábida Natural Park (38.5325, −9.00694), at dawn on 7 May 2024, at sunset on 9 May 2024, and at both dawn and sunset on 9 January 2025, and were immediately frozen in a dry shipper previously refrigerated with liquid nitrogen and stored at −80 °C upon arrival at the lab and until RNA extraction.

Vouchering information

Physical reference materials for the here sequenced specimen have been deposited in the Herbarium LISU from the Museu Nacional de História Natural e da Ciência, Universidade de Lisboa under the accession number LISU 270092.

Frozen reference tissue material of leaves, stems, and roots is available from the same individual at the Biobank [Corema] [http://biotropical.ihmt.unl.pt] under the voucher IDs BTRA1075, BTRA1076 and BTRA1077.

An electronic voucher image of the sequenced individual is available from the BioImageArchive dataset https://www.ebi.ac.uk/biostudies/studies/S-BSST2685, under accession ID S-BSST2685 and doi:10.6019/S-BSST2685.

Data availability

Corema album and the related genomic study were assigned to Tree of Life ID (ToLID) 'ddCorAlbu1', and all sample, sequence, and assembly information are available under the umbrella BioProject PRJEB76683. The sample information is available at the following BioSample accession: SAMEA12997959. The genome assembly is accessible from ENA under accession number GCA_964187925.1. Raw DNA sequencing data have been deposited in the ENA under sample accession SAMEA12997959. Raw RNA sequencing data have been deposited in the ENA under sample accessions SAMEA118800024 to SAMEA118800033. Documentation related to the genome assembly and curation can be found in the ERGA Assembly Report (EAR) document available at https://github.com/ERGA-consortium/EARs/blob/main/Assembly_Reports/Corema_album/ddCorAlbu1/ddCorAlbu1_EAR.pdf. Further details and data about the project are hosted on the ERGA portal at https://portal.erga-biodiversity.eu/data_portal/199168. The genome annotation files are described and available at https://doi.org/10.5281/zenodo.20555339.

Genetic information

The estimated genome size, based on ancestral taxa, is 1.4 Gb. This corresponds to a diploid genome with a haploid number of 13 chromosomes (2n = 26). All information for this species was retrieved from Genomes on a Tree (Challis R et al., 2023).

DNA/RNA processing

Cell lysis was performed by first grinding approximately 2 g of leaves in a mortar with liquid nitrogen and then incubating the resulting powder for 20 min at 74 °C in a Carlson lysis buffer (100 mM Tris Cl, pH 9.5; 2% CTAB; 1.4 M NaCl; 1% PEG 6000 or 8000; 20 mM EDTA) with β-mercaptoethanol and RNase A. The lysate was stabilized at pH 7 with HCl and high molecular weight DNA was extracted using the GenomicTip G500 kit (Qiagen) following the manufacturer's protocol. Nucleic acids yield was quantified in a Qubit fluorimeter using a Qubit 1X dsDNA HS Kit (Life Technologies), and DNA purity was assessed by comparing absorbance values at 260, 280 and 230 nm in a Tecan Infinite M200 Pro spectrophotometer using a NanoQuant plate for quantification of small volumes of nucleic acids (Tecan). Integrity of DNA was then assessed by pulsed-field gel electrophoresis. DNA was then preserved at −80 °C.

RNA was extracted using a lithium chloride protocol based on Azevedo et al. (2003). RNA was extracted from leaves, stems and roots from male and female plants. RNA quality was evaluated on an electrophoresis gel, and approximate concentration was measured on a Nanodrop. RNA was kept at −80 °C until use.

Library preparation and sequencing

Whole DNA was sheared in a Megaruptor 2 DNA shearing system (Diagenode) using large fragment hydropores with a target mean fragment length of 25 kb. Fragment profile was assessed by capillary electrophoresis on a 5200 Fragment Analyzer system using an HS large fragment 50 kb kit (Agilent Technologies). Repair and A-tailing of DNA fragments and SMRTbell adapters ligation were performed according to the SMRTbell prep kit 3.0 protocol (Pacific Biosciences). A BluePippin automated pulsed-field gel electrophoresis system (Sage Science) was employed to remove fragments shorter than 6 kb. Primer annealing, polymerase binding and preparation of internal DNA control were performed using the Pacific Biosciences Sequel II binding kit and DNA internal control complex 3.2 according to the manufacturer's protocol. The library was sequenced in HiFi mode in a Pacific Biosciences Sequel IIe platform using a Sequel II Sequencing plate 2.0 and two 8 M ZMW SMRT cells with a 30-hour movie time and 2 hours of pre-extension time for a target 50X genome coverage. The Hi-C library was performed from young leaves using the Arima Hi-C High Coverage kit (Arima Genomics) through a collaboration, in the Metazoa Phylogenomics and Genome Evolution Lab, Institute of Evolutionary Biology (CSIC-University Pompeu Fabra). The Arima HiC library was loaded for paired-end sequencing on an Illumina NovaSeq 6000 system. The library was designed in order to attain a 50X mean genome coverage and was run in single index running mode of 6:151:151:0 bp cycles using a NovaSeq 6000 SP Reagent Kit v1.5 (300 cycles). In total, a coverage of 54x PacBio HiFi and 49x Hi-C data was obtained to generate the assembly.

For RNA sequencing, messenger RNA was purified from total RNA using poly-T oligo-attached magnetic beads. After fragmentation, the first strand cDNA was synthesised using random hexamer primers. Then the second strand cDNA was synthesised using dUTP, instead of dTTP. The directional library was ready after end repair, A-tailing, adapter ligation, size selection, USER enzyme digestion, amplification, and purification. Samples were multiplexed, and sequencing was

performed on a NovaSeq X Plus series, with 150 bp Paired-End reads, and a target of at least 50 M pairs of reads per sample. Library preparation and sequencing were outsourced to Novogene.

Genome assembly methods

The genome was assembled using the ERGA HiFi & HiC assembly pipeline (De Panis 2024a,b,c). Briefly, Illumina reads were preprocessed for quality and length using fastp v0.23.4, and PacBio HiFi reads were preprocessed for adapter sequences using cutadapt v4.9 and initial contigs were assembled using HiFiasm v0.19.9, followed by removal of retained haplotigs using purge-dups v1.2.6 and scaffolding with YaHS v1.2a.2. Contaminant sequences were then identified and removed using fcs-gx v0.5.0. Finally, assembled scaffolds were curated via manual inspection using Pretext v0.2.5 with the Rapid Curation Toolkit (https://gitlab.com/wtsi-grit/rapid-curation) to remove any false joins and incorporate any sequences not automatically scaffolded into their respective locations in the chromosomal pseudomolecules (or super-scaffolds). The mitochondrial and chloroplast genomes were assembled as single circular contigs of 587,474 bp and 172,206 bp, respectively, using oatk v1.0 and included in the released assembly (GCA_964187925.1). Summary analysis of the released assembly was performed using the ERGA-BGE Genome Report ASM Galaxy workflow (De Panis & ERGA, 2024).

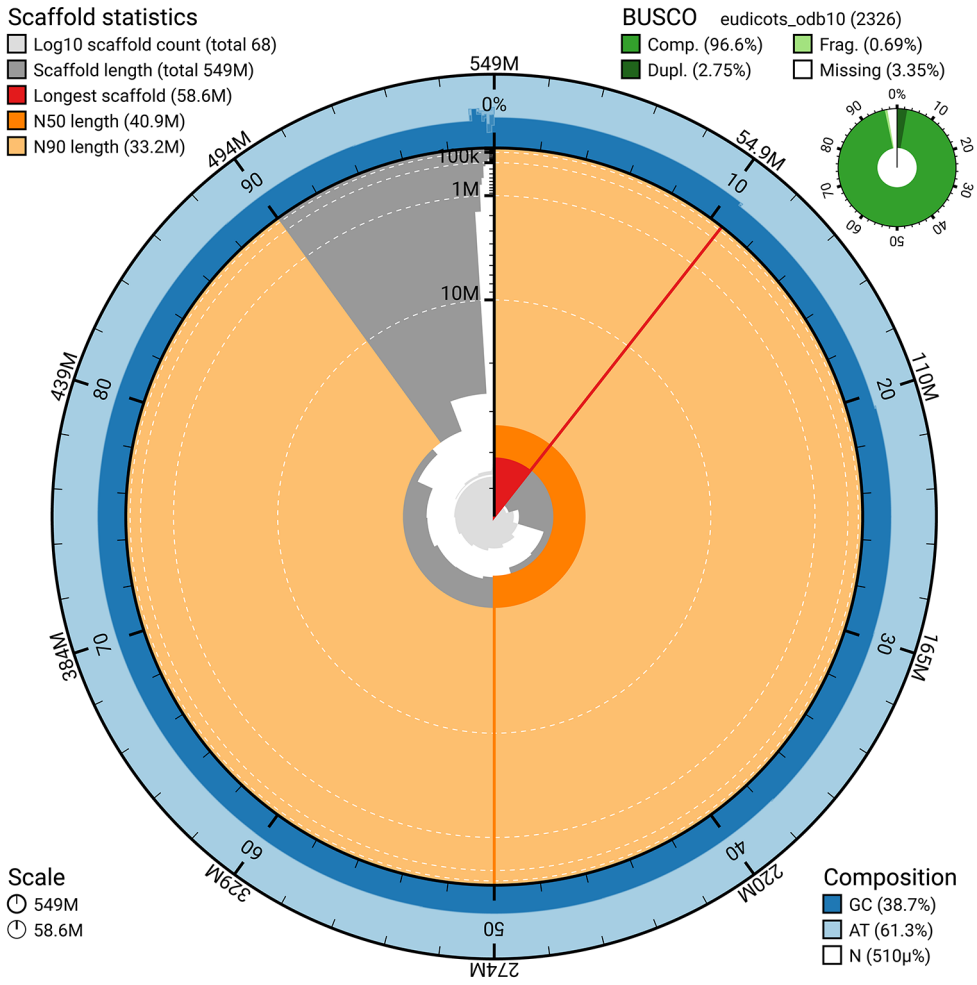


Figure 1. Snail plot summary of assembly statistics. The main plot is divided into 1,000 size-ordered bins around the circumference, with each bin representing 0.1% of the 548,992,474 bp assembly, including the mitochondrial and chloroplast genomes. The distribution of sequence lengths is shown in dark grey, with the plot radius scaled to the longest sequence present in the assembly (58,605,797 bp, shown in red). Orange and pale-orange arcs show the scaffold N50 and N90 sequence lengths (40,943,406 and 33,182,486 bp), respectively. The pale grey spiral shows the cumulative sequence 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 found in the assembled genome from the Eudicots database (odb10) is shown in the top right.

Genome annotation methods

The structural annotation pipeline focused on the 13 longest scaffolds of the assembly, which represent the n = 13 chromosomes of *C. album*, and identified both repeat elements and genes using a combination of de novo and homology-based information.

For repeat annotation, we used RepeatModeler v 4.1.8 (Flynn et al. 2020) and RepeatMasker v2.0.6 (Smit et al. 2013-2015) to identify repeats in the genome of *C. album.* Repeats identified were then classified with TEsorter v1.5.0 (Zhang et al. 2022) and the results plotted in R using the repeatR package (available from github.com/dwinter/repeatR).

For gene model annotation, we used both evidence-based and ab initio prediction tools, and merged both sources of evidence using the maker v 3.01.04 pipeline (Holt and Yandell 2011). For evidence-based annotation using the RNAseq data from multiple individuals, tissues and time points of *C. album*, we trimmed the RNAseq reads using trim_galore v0.6.10 (github.com/FelixKrueger/TrimGalore), mapped them to the genome assembly using star v2.7.11 (Dobin and Gingeras 2015), and used stringtie v3.0.1 (Pertea et al. 2015) and transdecoder v5.7.1 (Haas et al. 2013) to identify gene transcripts. For evidence-based annotation, we additionally used the closely related species *Rhododendron vialii* (accession number: GCA_030253575.1) with GeMoMa v1.9 (Keilwagen et al. 2019) to produce gene models based on protein homology. For *ab initio* gene prediction, we used both SNAP v20060728 (Korf 2004) and Augustus v3.5.0 (Stanke et al. 2008). These sources of information were then combined using the MAKER v 3.01.04 pipeline (Holt and Yandell 2011). The genome annotation pipeline is described on GitHub and available at https://doi.org/10.5281/zenodo.20554699.

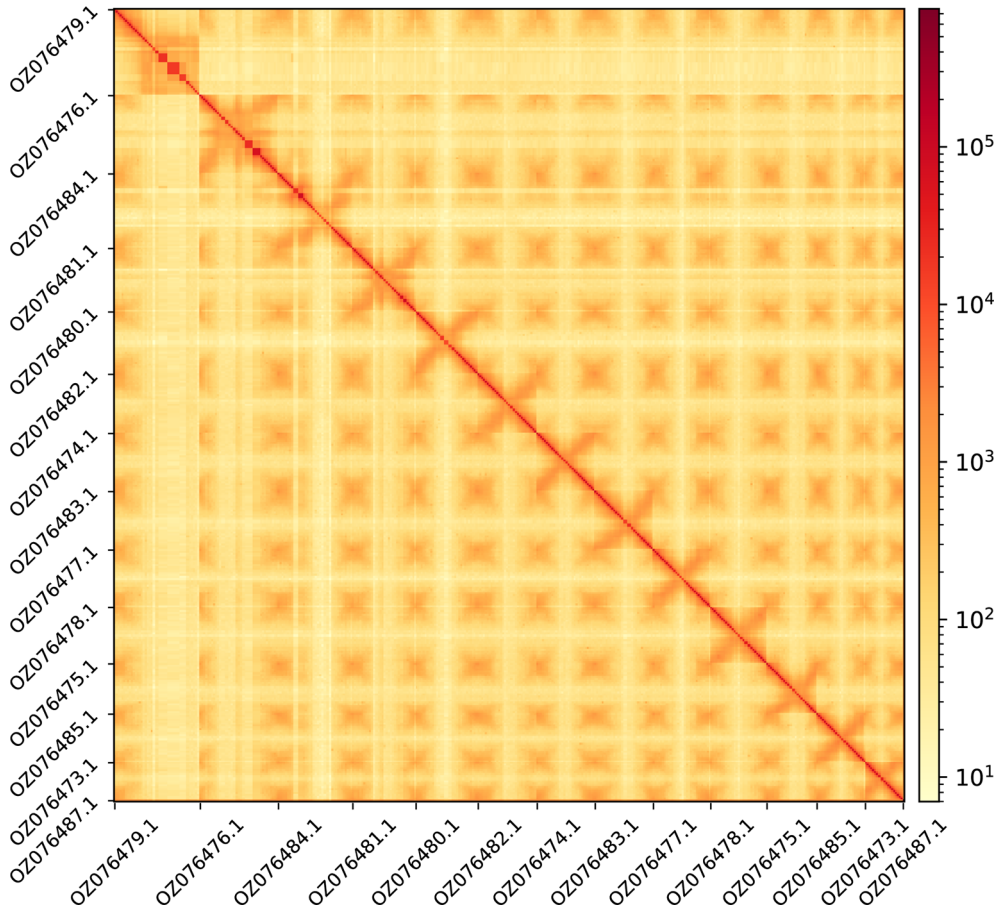


Figure 2. Hi-C contact map showing spatial interactions between regions of the genome. The diagonal corresponds to intra-chromosomal contacts, depicting chromosome boundaries. The frequency of contacts is shown on a logarithmic heatmap scale. Hi-C matrix bins were merged into a 25 kb bin size for plotting.

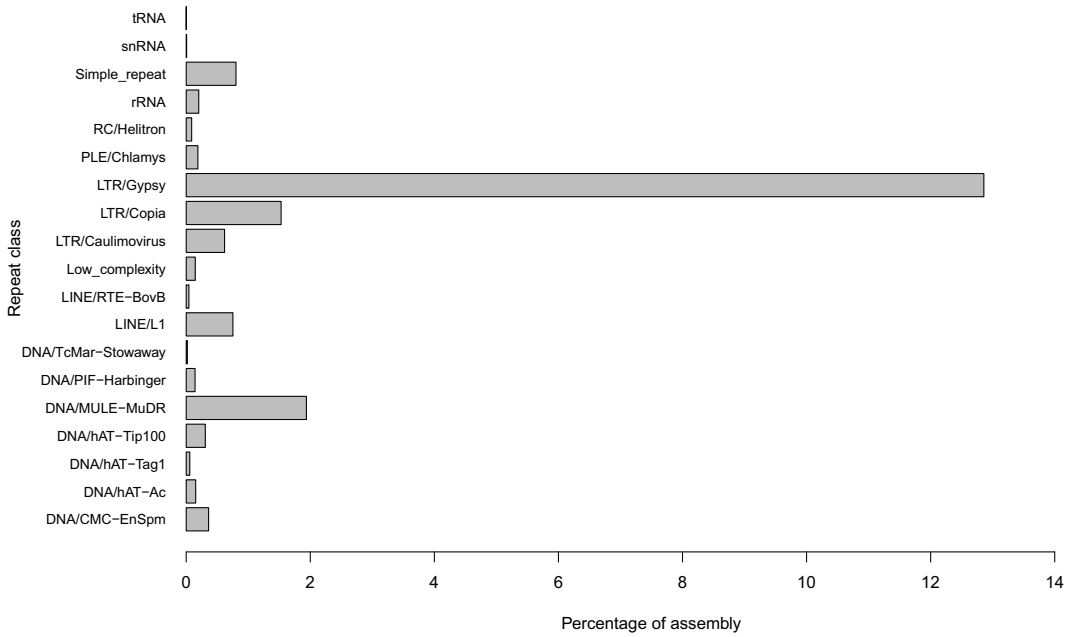


Figure 3. Relative contribution of different repeat classes to the Corema album genome assembly, expressed as the proportion of genomic bases assigned to each repeat category. LTR/Gypsy elements were the most abundant classified repeats, accounting for approximately 13% of the genome.

Results

Genome assembly

The genome assembly has a total length of 548,992,474 bp in 68 scaffolds, including the mitogenome and chloroplast genome (Figure 1), with a GC content of 38.7%. Hi-C data supported the reconstruction of chromosome-scale scaffolds and revealed strong intra-chromosomal interaction patterns consistent with the final assembly structure (Figure 2). The assembly has a contig N50 of 23,686,000 bp and L50 of 9, and a scaffold N50 of 40,943,406 bp and L50 of 6. The assembly has a total of 28 gaps. The single-copy gene content analysis using the Eudicots database (odb10) with BUSCO v5.8.0 (Manni et al. 2021) resulted in 96.6% completeness (93.9% single and 2.8% duplicated). 88.4% of reads k-mers were present in the assembly, and the assembly has a base accuracy Quality Value (QV) of 68.3 as estimated by Merqury (Rhie et al., 2020).

Genome annotation

Annotation of repeats in the *C. album* genome revealed that ca. 51% of the genome consists of repeats, of which ca. 60% could not be classified, suggesting a large set of repeats in *C. album* are not currently represented in existing databases. Among the classified repeats, LTR/Gypsy retrotransposons were the most abundant, accounting for approximately 13% of the genome (Figure 3).

Gene model annotation identified 53,794 genes and 229,093 exons (average of 4.3 exons per gene), a mean gene length of 3,694 bp, and a mean exon length of 349 bp. Assessment with BUSCO against the Viridiplantae v10 database revealed that 94.6% of the BUSCO genes were found and complete in the resulting annotation, with 92% corresponding to single-copy genes and 2.6% duplicated in *C. album.*

Conflict of interest

The authors declare no conflict of interest related to this study. The funding sources had no involvement in the study design, collection, analysis, or interpretation of data; in the writing of the manuscript; or in the decision to submit the article for publication. All authors have participated sufficiently in the work to take public responsibility for the content and agree to the submission of this manuscript.

Funder information

This work is also part of the Biodiversity Genomics Europe (BGE) project (Grant No. 101059492), funded by Horizon Europe under the Biodiversity, Circular Economy and Environment call (REA.B.3), and co-funded by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract numbers 22.00173 and 24.00054, and by UK

Research and Innovation (UKRI) through the Department for Business, Energy and Industrial Strategy's Horizon Europe Guarantee Scheme.

Author contributions

MS-S, HT, VCS and JPM coordinated the project; MS-S, CM, SC, MJA and JJ collected the species; AM, AMMC and GF led the pilot and secured resources; JM-F, VCS, CG, CN and RW provided sampling and metadata support and management; HT, SV, JJ and JJ extracted DNA; JS-O, JK, AI, CB, MM and MAD prepared libraries, and performed sequencing under the supervision of PS, RF and CC; JPM, DP and TB performed genome assembly and curation; BN performed genome annotation; TB and BN generated the analysis and report. All authors contributed to the writing, review, and editing of this genome note and read and approved the final version.

Acknowledgements

We acknowledge the Biobank (Biotropical Resources, BIOTROP, Global Health and Tropical Medicine, Institute of Hygiene and Tropical Medicine, NOVA University Lisbon, GHTM IHMT NOVA) for its collaboration in depositing the samples. We also thank the assembly reviewer, Diego De Panis (Leibniz Institute for Zoo and Wildlife Research), for his contribution to the assembly review process.

This work received additional support from CE3C through FCT Unit funding UID/00329/2025 (https://doi.org/10.54499/UID/00329/2025). The authors also acknowledge the use of the Deucalion HPC server, supported by FCT funding (https://doi.org/10.54499/2023.14346.CPCA.A3).

We acknowledge the essential contribution of the ERGA Pilot Project coordinators, Ann Mc Cartney, Giulio Formenti and Alice Mouton, whose efforts were instrumental in establishing the infrastructure required for this work. This included the development of sample metadata collection procedures, the ERGA manifest GitHub repository, the provision of in-kind reagents from commercial partners, and the continuous coordination activities carried out throughout the ERGA Pilot Project.

We acknowledge access to the ERGA storage resources hosted at the Barcelona Supercomputing Center, partially funded by the European Union H2020-INFRAEOSC-2018-2020 programme through the DICE project (Grant Agreement No. 101017207).

We thank Alisha Ahamed, Josephine Burgin, Joana Paupério, Jeena Rajan and Guy Cochrane from the European Nucleotide Archive (ENA) for their support with data coordination and submission. We also thank Felix Shaw, Aaliyah Providence, Debby Ku, Robert P. Davey, Seanna McTaggart and Alice Minotto for their support with metadata coordination and submission.

RF acknowledges support from the Ramón y Cajal fellowship (grant agreement no. RYC2017-22492 funded by MCIN/AEI/10.13039/501100011033 and ESF "Investing in your Future"), the Agencia Estatal de Investigación (project PID2019-108824GA-I00 funded by MCIN/AEI/10.13039/501100011033), the European Research Council (ERC; grant agreement no. 948281 under the European Union's Horizon 2020 research and innovation programme), and the Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya (AGAUR 2021-SGR00420).

PacBio HiFi sequencing and Illumina RNA sequencing were supported by the sequencing facility of the Department of Biology, University of Florence, through the Departments of Excellence programme funded by the Italian Ministry for University and Research.

We gratefully acknowledge all commercial partners that donated kits and reagents to the ERGA Pilot library preparation hubs, enabling the generation of high-quality genomes and annotations for species lacking dedicated funding. This support has been instrumental in fostering diversity, equity, inclusion and justice within the Pilot Project. In particular, we thank Dovetail Genomics (part of Cantata Bio LLC), especially Mark Daly, Thomas Swale and Lily Shuie; Arima Genomics; PacBio; Integrated DNA Technologies (IDT); MagBio Genomics Europe GmbH; Zymo Research; Agilent Technologies; Fisher Scientific Spain; and Illumina Inc.

Literature Cited

Antunes C, Díaz-Barradas MC, Zunzunegui M, *et al.*: **Water source partitioning among plant functional types in a semi-arid dune ecosystem.** *Journal of Vegetation Science.* 2018; **29**(4), 671–683.
Publisher Full Text

Azevedo H, Lino-Neto T, Tavares R: **An Improved Method for High-Quality RNA Isolation From Needles of Adult Maritime Pine Trees.** *Plant Mol. Biol. Report.* 2003; **21**: 333–338.
Publisher Full Text

Challis R, Kumar S, Sotero-Caio C: **Genomes on a Tree (GoaT): A versatile, scalable search engine for genomic and sequencing project metadata across the eukaryotic tree of life [version 1; peer review: 2 approved].** 2023; **8**(24).
Publisher Full Text

De Panis DERGA: **ERGA HiFi-only Assembly+QC Hifiasm.** 2024c, September 24; **v2409**(WF2).
Reference Source

De Panis DERGA: **ERGA HiFi PriAlt Purge+QC.** 2024b, September 24; **v2409**(WF3).
Reference Source

De Panis DERGA: **ERGA HiC Pri Scaffolding+QC YaHS.** 2024a, September 24; **v2409**(WF4).
Reference Source

De Panis DERGA: **ERGA-BGE Genome Report ASM analyses (one-asm HiFi + HiC).** 2024, August 20.
Reference Source

Díaz-Barradas MC, Zunzunegui M, Correia O, *et al.*: **Gender dimorphism in Corema album across its biogeographical area and implications under a scenario of extreme drought events.** *Environmental and Experimental Botany.* 2018; **155**: 609–618.
Publisher Full Text

Dobin A, Gingeras TR: **Mapping RNA-seq Reads with STAR.** *Curr. Protoc. Bioinformatics.* 2015 Sep 3; **51**: 11.14.1–11.14.19.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Flynn, Hubley R, Goubert C, *et al.*: **RepeatModeler2 for automated genomic discovery of transposable element families.** *Proc. Natl. Acad. Sci. USA.* 2020; **117**(17): 9451–9457.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Haas BJ, *et al.*: **De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference genomes.** *Nat. Protoc.* 2013; **8**(1): 1–13.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Holt C, Yandell M: **MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects.** *BMC Bioinformatics.* 2011; **12**: 491.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Jacinto J, Giovanetti M, Oliveira P, *et al.*: **Quality attributes of cultivated white crowberries (Corema album (L.) D. Don) from a multi-origin clonal field.** *Euphytica.* 2021; **217**(3).
Publisher Full Text

Keilwagen J, Hartung F, Grau J: **GeMoMa: Homology-Based Gene Prediction Utilizing Intron Position Conservation and RNA-seq Data.** *Methods in molecular biology (Clifton, N.J.), 1962.* 2019; 161–177.
Publisher Full Text

Korf I: **Gene finding in novel Genomes.** *BMC Bioinformatics.* 2004; **5**: 59.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Mc Cartney AM, Formenti G, Mouton A, *et al.*: **The European Reference Genome Atlas: piloting a decentralised approach to equitable biodiversity genomics.** *npj biodivers.* 2024; **3**: 28.
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**

Mazzoni C, Ciofi C, Waterhouse R: **Biodiversity: An atlas of European reference genomes.** *Nature.* 2023; **619**: 252–252.
Publisher Full Text

Pertea M, Pertea G, Antonescu C, *et al.*: **StringTie enables improved reconstruction of a transcriptome from RNA-seq reads.** *Nat. Biotechnol.* 2015; **33**: 290–295.
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**

Smit AFA, Hubley R, Green P: **RepeatMasker Open-4.0.** 2013-2015.
Reference Source

Stanke, Diekhans M, Baertsch R, *et al.*: **Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding.** *Bioinformatics.* March 2008; **24**(5): 637–644.
Publisher Full Text

The UniProt Consortium: **UniProt: A worldwide hub of protein knowledge.** *Nucleic Acids Res.* 2019; **47**(D1): D506–D515.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Zhang RG, Li GY, Wang XL, *et al.*: **TEsorter: An accurate and fast method to classify LTR-retrotransposons in plant genomes.** *Hortic. Res.* 2022; **9**: uhac017.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**