DATA NOTE

The genome sequence of the Mediterranean Skipper, *Gegenes nostrodamus* (Fabricius, 1793) (Lepidoptera: Hesperiidae)

[version 1; peer review: 2 approved]

Roger Vila[1], Eric Toro-Delgado[1], Charlotte J. Wright[2], Joana I. Meier[2], Mark L. Blaxter[2],
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, Project Psyche Community

[1]Institut de Biologia Evolutiva (CISC-UPF), Barcelona, Catalonia, Spain
[2]Tree of Life Programme, Wellcome Sanger Institute, Hinxton, England, UK

v1 **First published:** 16 Feb 2026, **11**:119
https://doi.org/10.12688/wellcomeopenres.25889.1
Latest published: 16 Feb 2026, **11**:119
https://doi.org/10.12688/wellcomeopenres.25889.1

Abstract

We present a genome assembly from a female specimen of *Gegenes nostrodamus* (Mediterranean Skipper; Arthropoda; Insecta; Lepidoptera; Hesperiidae). The assembly contains two haplotypes with total lengths of 499.73 megabases and 419.20 megabases. Most of haplotype 1 (99.52%) is scaffolded into 16 chromosomal pseudomolecules, including the W and Z sex chromosomes. Haplotype 2 was assembled to scaffold level. The mitochondrial genome has also been assembled, with a length of 18.43 kilobases. This work is part of Project Psyche, a collaborative programme generating genomes for European butterflies and moths.

Keywords

Gegenes nostrodamus; Mediterranean Skipper; genome sequence; chromosomal; Lepidoptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ✓ ✓

	1	2
version 1 16 Feb 2026	✓ view	✓ view

1. **Oliver Rupp**, Justus-Liebig-University, Giessengiessen, Germany

2. **Fedor Čiampor**, Plant Science and Biodiversity Centre, Slovak Academy of Sciences, Bratislava, Slovakia

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

Corresponding author: Charlotte J. Wright (cw22@sanger.ac.uk)

Author roles: Vila R: Investigation, Resources, Supervision, Writing – Review & Editing; **Toro-Delgado E**: Writing – Original Draft Preparation; **Wright CJ**: Funding Acquisition, Project Administration, Supervision; **Meier JI**: Funding Acquisition, Project Administration, Supervision; **Blaxter ML**: Funding Acquisition, Project Administration, Supervision;

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome through core funding to the Wellcome Sanger Institute (220540). E. T-D. is supported by an FPU grant from the Spanish Ministry of Science, Innovation and Universities (MCIU; grant number FPU22/02358). R.V. is supported by grant PID2022-139689NB-I00 (MICIU/ AEI/ 10.13039/501100011033 and ERDF, EU) and by grant 2021-SGR-00420 (Departament de Recerca i Universitats, Generalitat de Catalunya).
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: Vila R, Toro-Delgado E, Wright CJ *et al.* **The genome sequence of the Mediterranean Skipper, *Gegenes nostrodamus* (Fabricius, 1793) (Lepidoptera: Hesperiidae) [version 1; peer review: 2 approved]** Wellcome Open Research 2026, **11**:119 https://doi.org/10.12688/wellcomeopenres.25889.1

First published: 16 Feb 2026, **11**:119 https://doi.org/10.12688/wellcomeopenres.25889.1

Species taxonomy

Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Protostomia; Ecdysozoa; Panarthropoda; Arthropoda; Mandibulata; Pancrustacea; Hexapoda; Insecta; Dicondylia; Pterygota; Neoptera; Endopterygota; Amphiesmenoptera; Lepidoptera; Glossata; Neolepidoptera; Heteroneura; Ditrysia; Obtectomera; Hesperioidea; Hesperiidae; Hesperiinae; Baorini; *Gegenes*; *Gegenes nostrodamus* (Fabricius, 1793) (NCBI:txid1107782)

Background

The Mediterranean Skipper (*Gegenes nostrodamus*) is a species of butterfly from the family Hesperiidae, subfamily Hesperiinae. Both males and females have a general brown colouration. The dorsal side is cinnamon brown, while the ventral side is a uniform grey-brown (Kolev & Shtinkov, 2016; Vila *et al.*, 2018). It displays sexual dimorphism, with females showing white postdiscal spots on the forewing dorsal side. This species is very similar to the Pigmy Skipper (*G. pumilio*), with which it overlaps for a large part of its distribution. *G. nostrodamus* can be distinguished by its larger size, the lack of light flecks on the hindwing ventral side, and the presence of a tuft of hair-like scales on the hindwing, which are scarcer and shorter on *G. pumilio*, as well as by the genitalia of both sexes (de Jong & Coutsis, 2017).

G. nostrodamus is found throughout the Mediterranean basin, including most of the Iberian and Italian peninsulas, and extending through Anatolia and the Middle East up to India (de Jong & Coutsis, 2017). It is a thermophilous species, found mostly along coastal wetlands and river valleys at low altitudes. It is highly dispersive, with records of occasional migrants outside the main distribution range (García-Barros *et al.*, 2013). It is a multivoltine species, with adults flying in Aprilto May and September to October. Females lay eggs singly on the leaves of a wide range of Poaceae plants, including *Zea mays*, *Phragmites australis*, *Aeluropus* spp. and *Panicum* spp., and also on the highly invasive *Arundo donax* (Vila *et al.*, 2018). Like many Hesperiinae, larvae build a refugia by rolling a leaf using silk threads. It is classified as Least Concern (LC) by the European Red List of Butterflies (van Swaay *et al.*, 2025).

Karyotyping studies determined the haploid chromosome number to be $n = 15$ (Larsen, 1982; Saitoh, 1979), based on specimens from Israel and Egypt (García-Barros *et al.*, 2013). Mitochondrial DNA analyses suggest no geographic structure and modest genetic diversity, with a single mitochondrial lineage spread throughout most of its European range, with some additional mitotypes (Dapporto *et al.*, 2022).

We present a chromosome-level genome sequence for *G. nostrodamus*, sequenced as part of Project Psyche (Wright *et al.*, 2025). The sequence data were derived from a female specimen (Figure 1) collected from Cardedeu, Vallès Oriental, Barcelona, Catalonia, Spain.

Methods

Sample acquisition

The specimen used for genome sequencing was an adult female *Gegenes nostrodamus* (specimen ID SAN28000302, ToLID ilGegNost1; Figure 1), collected from Cardedeu, Barcelona, Catalonia, Spain (latitude 41.6365, longitude 2.3442; elevation 186 m) on 2023-08-09. The specimen was collected and identified by Roger Vila (Institut de Biologia Evolutiva).

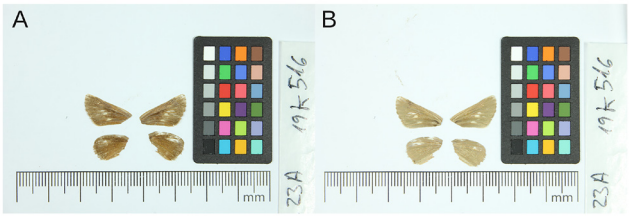

Figure 1. Voucher photographs of the *Gegenes nostrodamus* (ilGegNost1) specimen used for genome sequencing: **A**. Dorsal view (upperside), **B**. Ventral view (underside).

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 ilGegNost1 sample was weighed and triaged to determine the appropriate extraction protocol. Tissue from the whole organism was homogenised by powermashing using a PowerMasher II tissue disruptor.

HMW DNA was extracted in the WSI Scientific Operations core using the Automated 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 automated 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. For this sample, the final post-shearing DNA had a Qubit concentration of 15.5 ng/μL and a yield of 728.50 ng, with a fragment size of 13.7 kb.

PacBio HiFi library preparation and sequencing

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

The sample was sequenced on a Revio instrument (Pacific Biosciences). The prepared library was normalised to 2 nM, and 15 μL was used for making complexes. Primers were

annealed and polymerases bound to generate circularised complexes, following the manufacturer's instructions. Complexes were purified using 1.2X SMRTbell beads, then diluted to the Revio loading concentration (200–300 pM) and spiked with a Revio sequencing internal control. The sample was sequenced on a Revio 25M SMRT cell. The SMRT Link software (Pacific Biosciences), a web-based workflow manager, was used to configure and monitor the run and to carry out primary and secondary data analysis. Specimen details, sequencing platforms, and data yields are summarised in Table 1.

Hi-C

Sample preparation and crosslinking

The Hi-C sample was prepared from 20–50 mg of frozen tissue from the whole organism of the ilGegNost1 sample using the Arima-HiC v2 kit (Arima Genomics). Following the manufacturer's instructions, tissue was fixed and DNA crosslinked using TC buffer to a final formaldehyde concentration of 2%. The tissue was homogenised using the Diagnocine Power Masher-II. Crosslinked DNA was digested with a restriction enzyme master mix, biotinylated, and ligated. Clean-up was performed with SPRISelect beads before library preparation. DNA concentration was measured with the Qubit Fluorometer (Thermo Fisher Scientific) and Qubit HS Assay Kit. The biotinylation percentage was estimated using the Arima-HiC v2 QC beads.

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 to create equimolar and/or weighted 2.8 nM pools. Pool concentrations were checked using the Agilent 4200 TapeStation (Agilent) with High Sensitivity D500 reagents before sequencing. Sequencing was performed using paired-end 150 bp reads on the Illumina NovaSeq X. Specimen details, sequencing platforms, and data yields are summarised in Table 1.

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 in Hi-C phasing mode (Cheng *et al.*, 2021; Cheng *et al.*, 2022), producing two haplotypes. Hi-C reads (Rao *et al.*, 2014) were mapped to the primary contigs using bwa-mem2 (Vasimuddin *et al.*, 2019). Contigs were further scaffolded with Hi-C data in YaHS (Zhou *et al.*, 2023), using the --break option for handling potential misassemblies. The scaffolded assemblies were evaluated using Gfastats (Formenti *et al.*, 2022), BUSCO (Manni *et al.*, 2021) and MERQURY.FK (Rhie *et al.*, 2020).

The mitochondrial genome was assembled using MitoHiFi (Uliano-Silva *et al.*, 2023), which runs MitoFinder (Allio *et al.*, 2020) and uses these annotations to select the final mitochondrial contig and to ensure the general quality of the sequence.

Table 1. Specimen and sequencing data for BioProject PRJEB80921.

Platform	PacBio HiFi	Hi-C
ToLID	ilGegNost1	ilGegNost1
Specimen ID	SAN28000302	SAN28000302
BioSample (source individual)	SAMEA115768765	SAMEA115768765
BioSample (tissue)	SAMEA115768891	SAMEA115768891
Tissue	whole organism	whole organism
Instrument	Revio	Illumina NovaSeq X
Run accessions	ERR13800478	ERR13802622
Read count total	2.84 million	787.52 million
Base count total	35.13 Gb	118.92 Gb

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 one break and 173 joins. This reduced the scaffold count by 61.1%, increased scaffold N50 by 2.3%, and increased the total assembly length by 2.5%. The curation process is described 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), run in a Singularity container (Kurtzer *et al.*, 2017), was used to evaluate *k*-mer completeness and assembly quality for both haplotypes using the *k*-mer database (k = 31) computed prior to genome assembly. The analysis outputs included assembly QV scores and completeness statistics.

The genome was analysed using the BlobToolKit pipeline, a Nextflow (Di Tommaso *et al.*, 2017) 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 the 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 containerisation through Docker (Merkel, 2014) and Singularity (Kurtzer *et al.*, 2017).

We used lep_busco_painter to paint Merian elements along chromosomes (Wright *et al.*, 2024). Merian elements represent the 32 ancestral linkage groups in Lepidoptera. The painting process utilised BUSCO gene locations from the lepidoptera_odb10 set (Kriventseva *et al.*, 2019) and chromosome lengths from NCBI Datasets. Each complete BUSCO gene (both single-copy and duplicated) was assigned to a Merian element based on a reference database, then plotted along chromosomes drawn to scale.

Genome sequence report

Sequence data

PacBio sequencing of the *Gegenes nostrodamus* specimen generated 35.13 Gb (gigabases) from 2.84 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 458.34 Mb, with a heterozygosity of 0.88% and repeat content of 24.09% (Figure 2). These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 75× coverage. Hi-C sequencing produced 118.92 Gb from 787.52 million reads, which were used to scaffold the assembly. Table 1 summarises the specimen and sequencing details.

Assembly statistics

The genome was assembled into two haplotypes using Hi-C phasing. Haplotype 1 was curated to chromosome level, while haplotype 2 was assembled to scaffold level. The final assembly has a total length of 499.73 Mb in 60 scaffolds, with 203 gaps, and a scaffold N50 of 33.79 Mb (Table 2).

Most of the assembly sequence (99.52%) was assigned to 16 chromosomal-level scaffolds, representing 14 autosomes and the W and Z sex chromosomes. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 3; Table 3). Chromosome painting with Merian elements illustrates the distribution of orthologues along chromosomes and highlights patterns of chromosomal evolution relative to Lepidopteran ancestral linkage groups (Figure 4). Chromosomes Z and W were identified by Hi-C signal (single copy in a two-haplotype map).

The mitochondrial genome was also assembled (length 18.43 kb, OZ198218.1). This sequence is included as a contig in the multifasta file of the genome submission and as a standalone record.

Assembly quality metrics

For haplotype 1, the estimated QV is 64.6, and for haplotype 2, 64.7. When the two haplotypes are combined, the assembly achieves an estimated QV of 64.6. The *k*-mer completeness is 83.93% for haplotype 1, 77.56% for haplotype 2, and 99.75% for the combined haplotypes (Figure 5).

BUSCO analysis using the lepidoptera_odb10 reference set (n = 5 286) identified 98.4% of the expected gene set (single = 98.0%, duplicated = 0.4%) in haplotype 1. For haplotype 2, BUSCO analysis identified 93.3% of the expected gene set (single = 93.0%, duplicated = 0.3%).

The snail plot in Figure 6 summarises the scaffold length distribution and other assembly statistics for haplotype 1. The blob plot in Figure 7 shows the distribution of scaffolds by GC proportion and coverage for haplotype 1.

Table 4 lists the assembly metric benchmarks adapted from Rhie *et al.* (2021) and the Earth BioGenome Project Report on Assembly Standards September 2024. The EBP metric, calculated for the haplotype 1, is **7.C.Q64**, meeting the recommended reference standard.

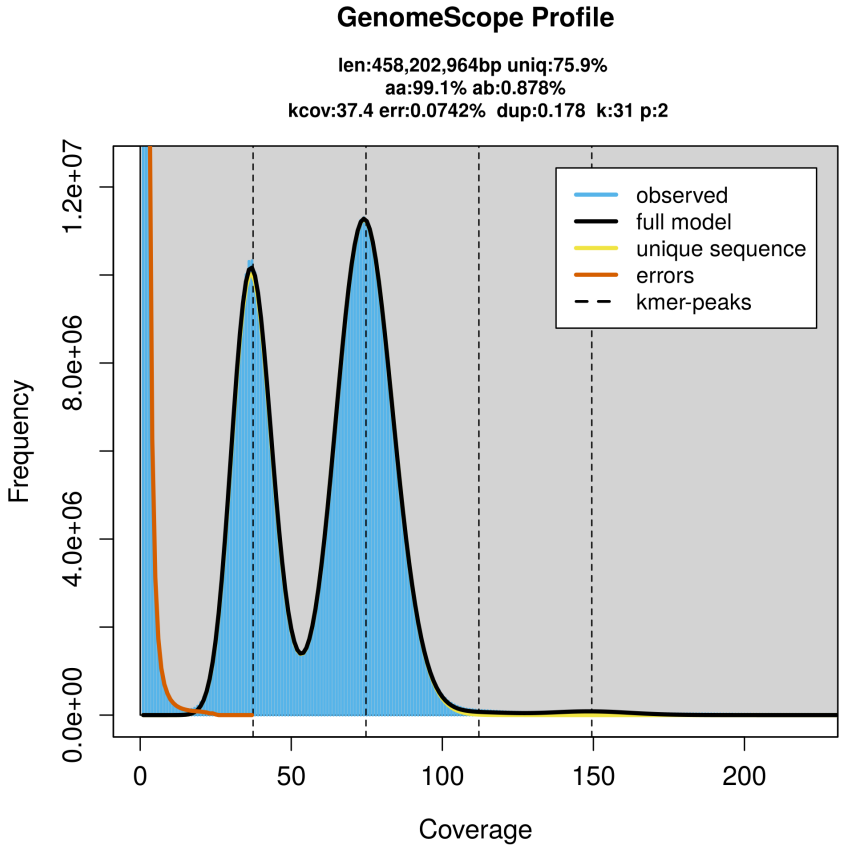

Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2. The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

Table 2. Genome assembly statistics.

Assembly name	ilGegNost1.hap1.1	ilGegNost1.hap2.1
Assembly accession	GCA_964291895.1	GCA_964291775.1
Assembly level	chromosome	scaffold
Span (Mb)	499.73	419.20
Number of chromosomes	16	scaffold-level
Number of contigs	263	112
Contig N50	10.42 Mb	11.58 Mb
Number of scaffolds	60	38
Scaffold N50	33.79 Mb	33.36 Mb
Longest scaffold length (Mb)	43.17	-
Sex chromosomes	W and Z	-
Organelles	Mitochondrion: 18.43 kb	-

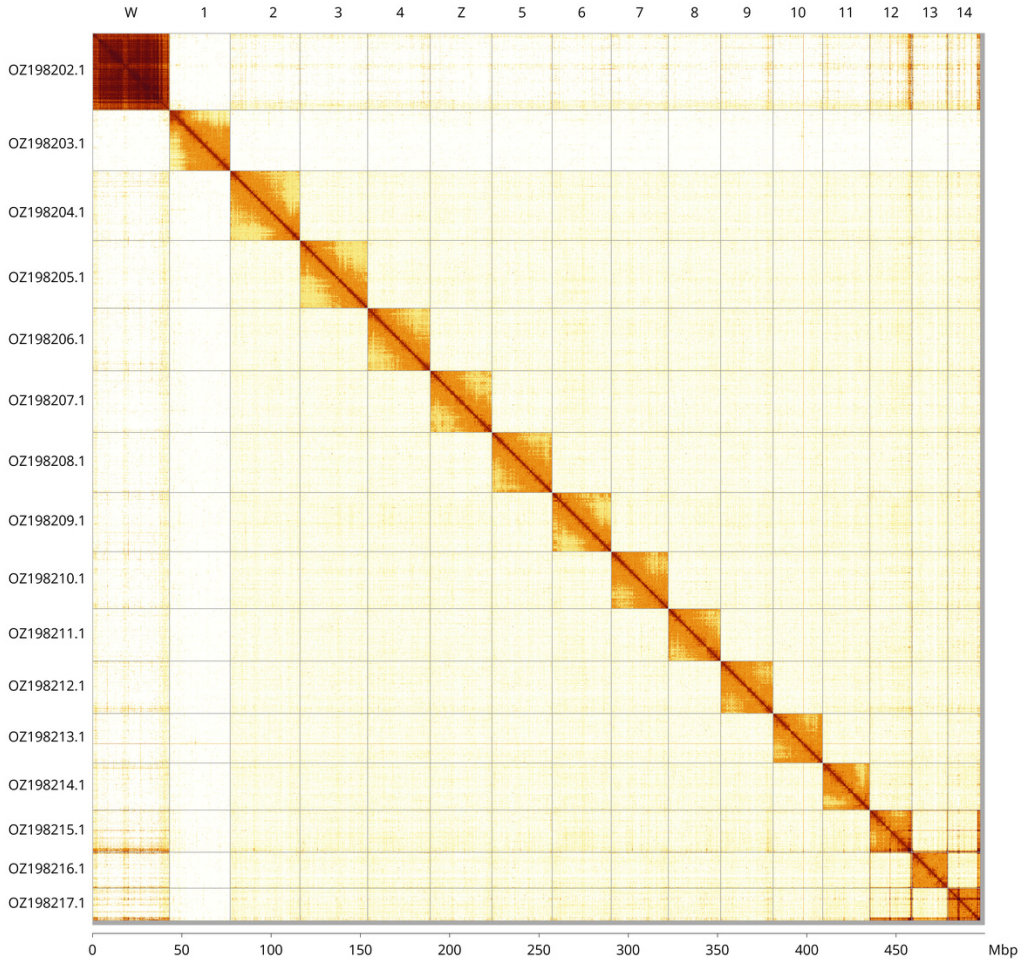

Figure 3. Hi-C contact map of the *Gegenes nostrodamus* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes, with a megabase scale shown below. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in the haplotype 1 genome assembly of *Gegenes nostrodamus* ilGegNost1.

INSDC accession	Molecule	Length (Mb)	GC%	Assigned Merian elements
OZ198203.1	1	39.08	35	M26;M28;M8
OZ198204.1	2	37.94	35	M14;M2;M3
OZ198205.1	3	34.97	35	M1;M6
OZ198206.1	4	34.52	35.50	M11;M15;M25
OZ198208.1	5	33.79	35	M16;M9
OZ198209.1	6	33.04	35	M17;M20;M27;M29
OZ198210.1	7	31.93	35.50	M22;M5
OZ198211.1	8	29.40	35	M10;M21
OZ198212.1	9	29.37	35	M13;M7
OZ198213.1	10	27.76	35	M12;M24
OZ198214.1	11	26.34	35.50	M19;M4
OZ198215.1	12	23.64	37	M14;M30
OZ198216.1	13	20.02	36	M18
OZ198217.1	14	18.34	36	M23
OZ198202.1	W	43.17	37	-
OZ198207.1	Z	34.02	35.50	M31;MZ

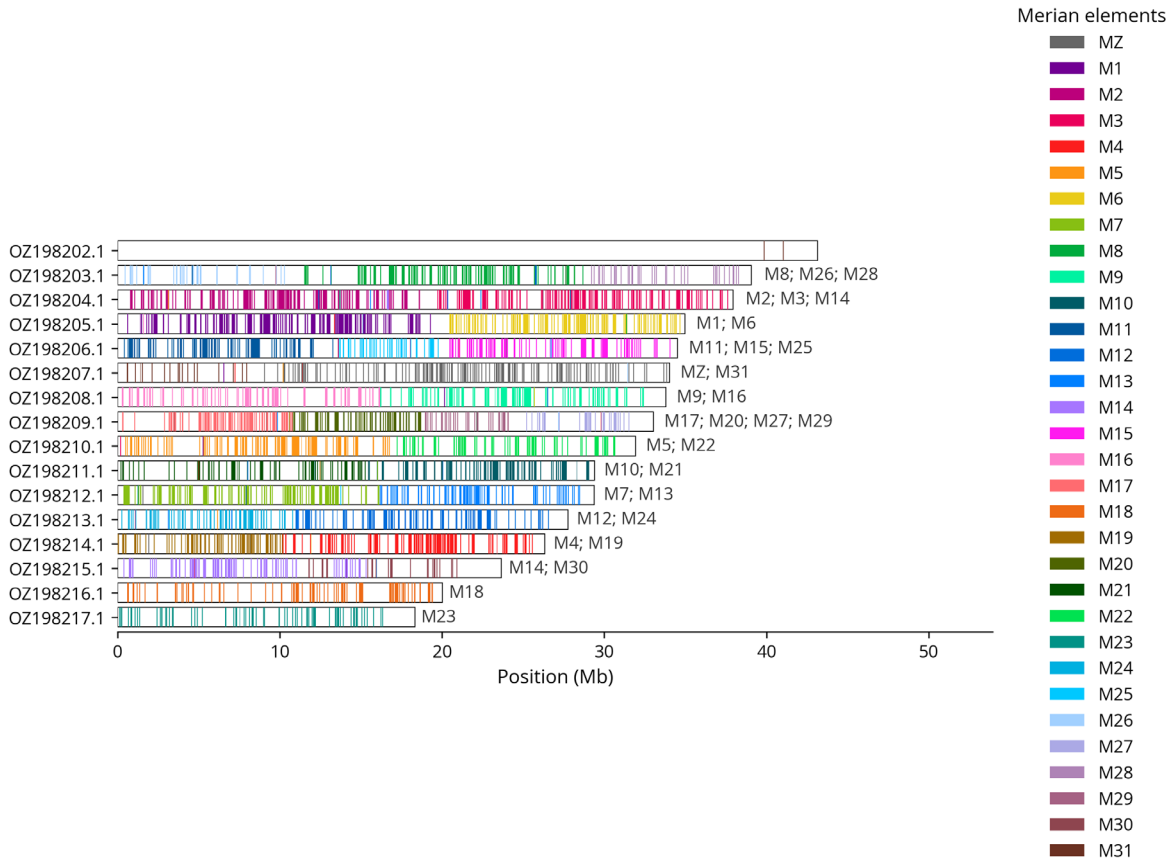

Figure 4. Merian elements painted across chromosomes in the ilGegNost1.hap1.1 assembly of *Gegenes nostrodamus*. Chromosomes are drawn to scale, with the positions of orthologues shown as coloured bars. Each orthologue is coloured by the Merian element that it belongs to. All orthologues which could be assigned to Merian elements are shown.

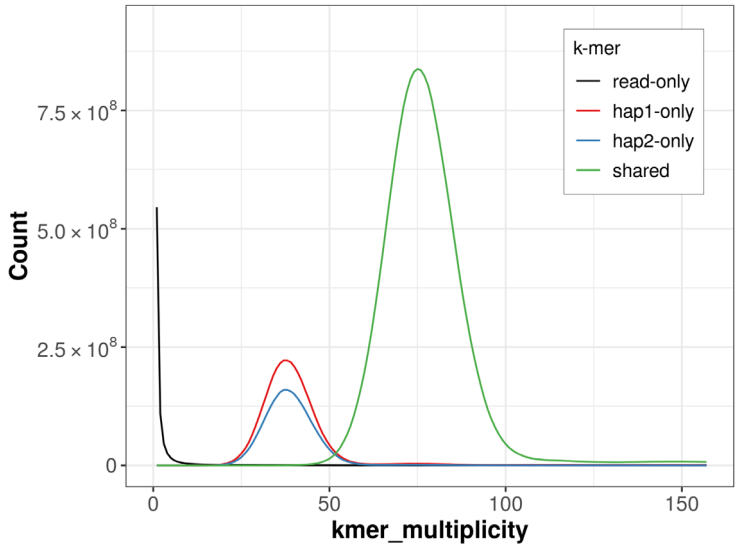

Figure 5. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers 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.

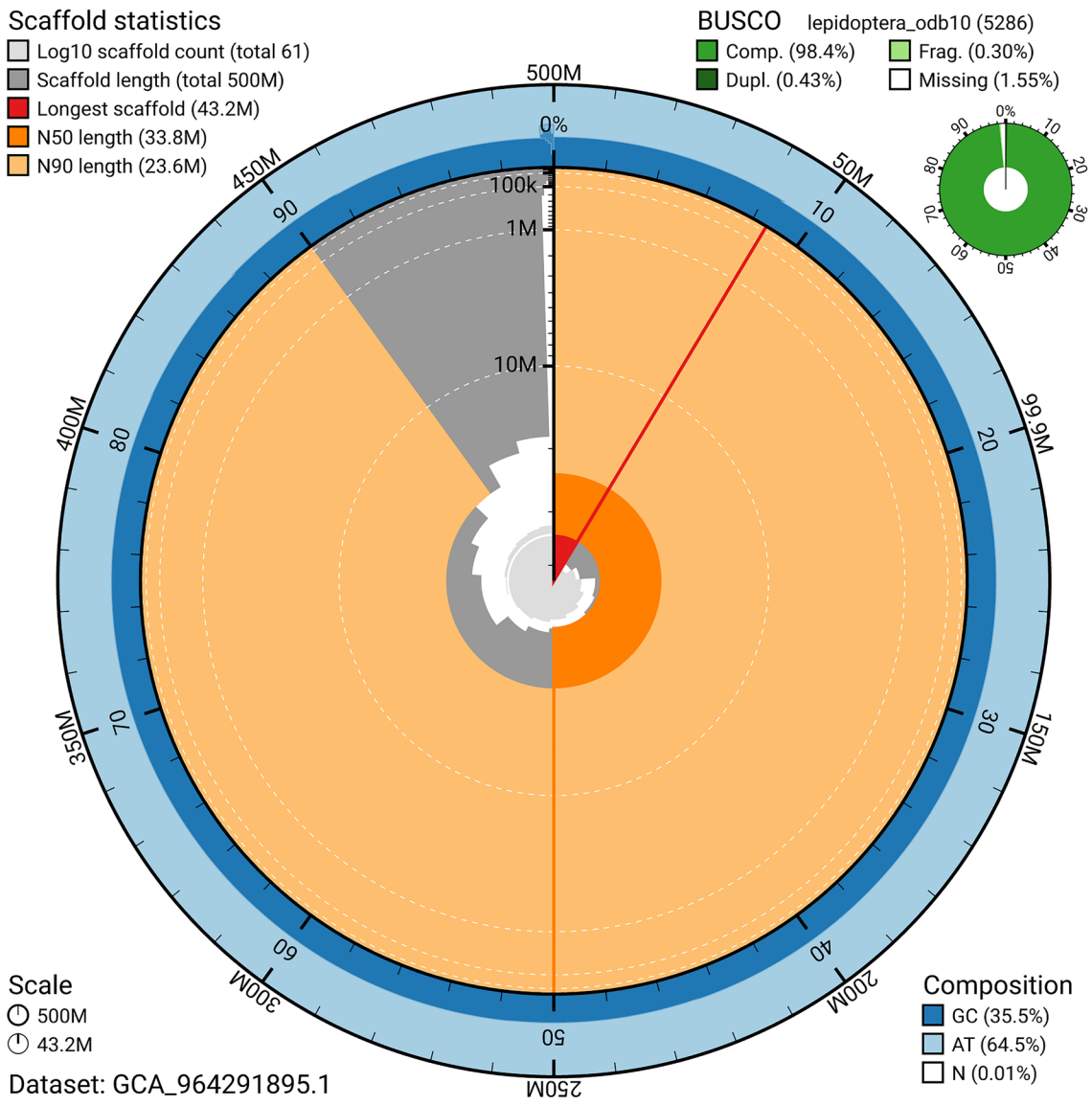

Figure 6. Assembly metrics for ilGegNost1.hap1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the circumference. The outermost 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 set is presented at the top right. An interactive version of this figure can be accessed on the BlobToolKit viewer.

Wellcome Sanger Institute – Legal and Governance

The materials that have contributed to this genome note have been supplied by a Tree of Life collaborator. 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 undertaken according to a Research Collaboration Agreement or Material Transfer Agreement entered into by the Tree of Life collaborator, Genome Research

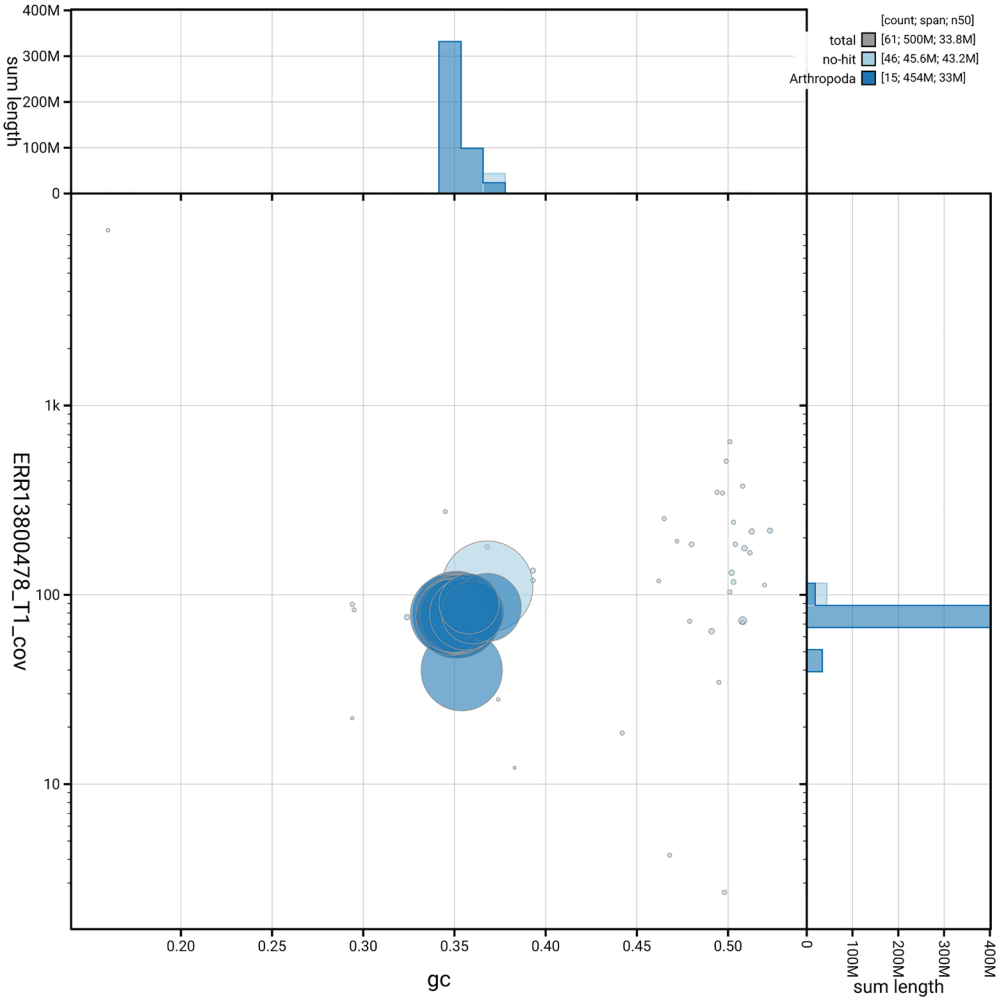

Figure 7. BlobToolKit GC-coverage plot for ilGegNost1.hap1.1. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes 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 *Gegenes nostrodamus* assembly.

Measure	Value	Benchmark
EBP summary (haplotype 1)	7.C.Q64	6.C.Q40
Contig N50 length	10.42 Mb	≥ 1 Mb
Scaffold N50 length	33.79 Mb	= chromosome N50
Consensus quality (QV)	Haplotype 1: 64.6; haplotype 2: 64.7; combined: 64.6	≥ 40
k-mer completeness	Haplotype 1: 83.93%; Haplotype 2: 77.56%; combined: 99.75%	≥ 95%
BUSCO	C:98.4% [S:98.0%; D:0.4%]; F:0.3%; M:1.2%; n:5 286	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	99.52%	≥ 90%

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

Limited (operating as the Wellcome Sanger Institute), and in some circumstances, other Tree of Life collaborators.

Data availability

European Nucleotide Archive: Gegenes nostrodamus. Accession number PRJEB80921. The genome sequence is released openly for reuse. The *Gegenes nostrodamus* genome sequencing initiative is part of the Sanger Institute Tree of Life Programme (PRJEB43745) and Project Psyche (PRJEB71705). 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 Ensembl 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

Contributors are listed at the following links:

- 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 Project Psyche Community

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	4.3.9	https://github.com/blobtoolkit/blobtoolkit
BUSCO	5.5.0	https://gitlab.com/ezlab/busco
bwa-mem2	2.2.1	https://github.com/bwa-mem2/bwa-mem2
Cooler	0.8.11	https://github.com/open2c/cooler
DIAMOND	2.1.8	https://github.com/bbuchfink/diamond
fasta_windows	0.2.4	https://github.com/tolkit/fasta_windows
FastK	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
Hifiasm	0.19.8-r603	https://github.com/chhylp123/hifiasm
HiGlass	1.13.4	https://github.com/higlass/higlass
lep_busco_painter	1.0.0	https://github.com/charlottewright/lep_busco_painter
MerquryFK	1.1.2	https://github.com/thegenemyers/MERQURY.FK
Minimap2	2.24-r1122	https://github.com/lh3/minimap2
MitoHiFi	3	https://github.com/marcelauliano/MitoHiFi
MultiQC	1.14; 1.17 and 1.18	https://github.com/MultiQC/MultiQC
Nextflow	23.10.0	https://github.com/nextflow-io/nextflow
PretextSnapshot	0.0.5	https://github.com/sanger-tol/PretextSnapshot
PretextView	1.0.3	https://github.com/sanger-tol/PretextView
samtools	1.19.2	https://github.com/samtools/samtools

Software	Version	Source
sanger-tol/ascc	0.1.0	https://github.com/sanger-tol/ascc
sanger-tol/blobtoolkit	0.6.0	https://github.com/sanger-tol/blobtoolkit
sanger-tol/curationpretext	1.4.2	https://github.com/sanger-tol/curationpretext
Seqtk	1.3	https://github.com/lh3/seqtk
Singularity	3.9.0	https://github.com/sylabs/singularity
TreeVal	1.4.0	https://github.com/sanger-tol/treeval
YaHS	1.2a.2	https://github.com/c-zhou/yahs

References

Allio R, Schomaker-Bastos A, Romiguier J, *et al.*: **MitoFinder: efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics.** *Mol Ecol Resour.* 2020; **20**(4): 892–905.
PubMed Abstract | Publisher Full Text | Free Full Text

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

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

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

Cheng H, Jarvis ED, Fedrigo O, *et al.*: **Haplotype-resolved assembly of diploid genomes without parental data.** *Nat Biotechnol.* 2022; **40**(9): 1332–1335.
PubMed Abstract | Publisher Full Text | Free Full Text

Danecek P, Bonfield JK, Liddle J, *et al.*: **Twelve years of SAMtools and BCFtools.** *GigaScience.* 2021; **10**(2): giab008.
PubMed Abstract | Publisher Full Text | Free Full Text

Dapporto L, Menchetti M, Vodă R, *et al.*: **The atlas of mitochondrial genetic diversity for Western Palaearctic butterflies.** *Glob Ecol Biogeogr.* 2022; **31**(11): 2184–2190.
Publisher Full Text

Di Tommaso P, Chatzou M, Floden EW, *et al.*: **Nextflow enables reproducible computational workflows.** *Nat Biotechnol.* 2017; **35**(4): 316–319.
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

Garćia-Barros E, Munguira ML, Stefanescu C, *et al.*: **Fauna Ibérica, Vol. 37. Lepidoptera. Papilionoidea.** Madrid: Museo Nacional de Ciencias Naturales, 2013.
Reference Source

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

de Jong R, Coutsis JG: **A re-appraisal of *Gegenes* Hübner, 1819 (Lepidoptera, Hesperiidae) based on male and female genitalia, with the description of a new genus, *Afrogegenes*.** *Tijdschr Entomol.* 2017; **160**(1): 41–60.
Publisher 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

Kolev Z, Shtinkov N: **The Pygmy Skipper *Gegenes pumilio*: a new species to Bulgaria, and a confirmation of its occurrence in the eastern Balkan Peninsula (lepidoptera: hesperiidae).** *Phegea.* 2016; **44**(1): 17–22.
Reference Source

Kriventseva EV, Kuznetsov D, Tegenfeldt F, *et al.*: **OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs.** *Nucleic Acids Res.* 2019; **47**(D1): D807–D811.
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

Larsen TB: ***Gegenes pumilio* hoffmannsegg, 1804, a review with cytological evidence that two species are involved (hesperiidae).** *Nota Lepidopterol.* 1982; **5**(2/3): 103–10.
Reference Source

Li H: **Minimap2: pairwise alignment for nucleotide sequences.** *Bioinformatics.* 2018; **34**(18): 3094–3100.
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

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

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**

Saitoh K: **A note on the haploid karyotype of the Mediterranean skipper *Gegenes nostrodamus* (lepidoptera, hesperiidae) from Israel.** *Chromosome Inf Serv.* 1979; **27**(1): 8–9.
Reference Source

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**

van Swaay C, Warren M, Ellis S, *et al.*: **Measuring the pulse of European biodiversity.** *European Red List of Butterflies.* 2025.
Publisher Full Text

Uliano-Silva M, Ferreira JGRN, Krasheninnikova K, *et al.*: **MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio high fidelity reads.** *BMC Bioinformatics.* 2023; **24**(1): 288.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

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

Vila R, Stefanescu C, Sesma JM: **Guia de Les Papallones Diürnes de Catalunya.** Barcelona: Lynx Edicions, 2018.
Reference Source

Wright CJ, Stevens L, Mackintosh A, *et al.*: **Comparative genomics reveals the dynamics of chromosome evolution in Lepidoptera.** *Nat Ecol Evol.* 2024; **8**(4): 777–790.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Wright CJ, Wahlberg N, Vila R, *et al.*: **Project psyche: reference genomes for all Lepidoptera in Europe.** *Trends Ecol Evol.* 2025; **40**(12): 1234–1250.
PubMed Abstract | **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 25 March 2026

https://doi.org/10.21956/wellcomeopenres.28496.r150015

✔ **Fedor Čiampor**
Department of Biodiversity and Ecology, Plant Science and Biodiversity Centre, Slovak Academy of Sciences, Bratislava, Slovakia

The manuscript presents a high-quality chromosome-level genome assembly for the butterfly *Gegenes nostrodamus*, produced within the Tree of Life / Project Psyche initiative. The work is technically sound, clearly documented, and valuable as a genomic resource for Lepidoptera research. The paper fits well as a *data note*. The methodology and results are described concisely and clearly. Following are few comments:

* Is it necessary to repeat most of words from the title in the Keywords?
* Background paragraph 2: "flying in Aprilto May" - include space between "April" and "to"
* "Mitochondrial DNA analyses suggest no geographic structure and modest genetic diversity, with a single mitochondrial lineage spread throughout most of its European range, with some additional mitotypes" - To make that claim based on 45 people is a bit bold; I'd soften it a bit
* If, according to the background information, the first generation flies in April–May and the second in September–October, which generation did the individual used for analysis belong to, having been caught in August (2023-08-09)? This is quite important, as there have already been cases where it was discovered that two generations of a lepidopteran species are actually two cryptic species. It is necessary to specify which generation the analyzed individual belongs to.
* The published data indicate n=15, but in this study it appears to be n=16. It would be helpful to add a note explaining this discrepancy.

Methods are generally identical with previous manuscripts, I think they are well described
Despite comments above, this is a good genome resource paper providing a high-quality reference assembly with excellent technical standards and full reproducibility.

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: arthropod DNA metabarcoding, beetle systematics and phylogeny

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

Reviewer Report 25 March 2026

https://doi.org/10.21956/wellcomeopenres.28496.r150019

Oliver Rupp
Justus-Liebig-University, Giessengiessen, Germany

This article presents the sequencing and haplotype phased assembly of the butterfly Gegenes nostrodamus genome.
The data creation and data analyses follow the state-of-the art methods for eukaryotic genome assemblies, including PacBio Hifi and Hi-C sequencing. The assembly with hifiasm was followed by Hi-C scaffolding and manual curation.
The assembly quality assessment using k-mer and orthology based methods like Merqury.FK or BUSCO showed a high quality sequence for both haplotypes.
The methods are clearly presented and the quality of the data is shown.
The pipeline and data are available at public databases and a gene annotation will be performed through Ensembl.
Overall no major issues are to report.

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: eukaryotic genome assembly

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.