DATA NOTE

The genome sequence of the Regal Piercer, *Pammene regiana* (Zeller, 1849) (Lepidoptera: Tortricidae)

[version 1; peer review: 2 approved, 1 approved with reservations]

Douglas Boyes[1+], Clare Boyes[2],
University of Oxford and Wytham Woods Genome Acquisition Lab,
Natural History Museum Genome Acquisition Lab,
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

[1]UK Centre for Ecology & Hydrology, Wallingford, England, UK
[2]Independent researcher, Welshpool, Wales, UK

[+] Deceased author

v1 **First published:** 21 Apr 2026, **11**:232
https://doi.org/10.12688/wellcomeopenres.26362.1
Latest published: 21 Apr 2026, **11**:232
https://doi.org/10.12688/wellcomeopenres.26362.1

Abstract

We present a genome assembly from an individual female *Pammene regiana* (Regal Piercer; Arthropoda; Insecta; Lepidoptera; Tortricidae). The genome sequence has a total length of 863.59 megabases. Most of the assembly (99.44%) is scaffolded into 29 chromosomal pseudomolecules, including the W and Z sex chromosomes. The mitochondrial genome has also been assembled, with a length of 16.5 kilobases. This assembly was generated as part of the Darwin Tree of Life project, which produces genome assemblies for eukaryotic species found in Britain and Ireland.

Keywords

Pammene regiana, Regal Piercer, genome sequence, chromosomal, Lepidoptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ✓ ✓ ?

	1	2	3
version 1 21 Apr 2026	✓ view	✓ view	? view

1. **Jurate De Prins**, Commonwealth Scientific and Industrial Research Organisation (Ringgold ID: 2221), Canberra, Australia
2. **Axel Künstner**, University of Lübeck, Lübeck, Germany
3. **Christophe Klopp**, INRAE, Castanet-Tolosan, France

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: Boyes D: Investigation, Resources; **Boyes C**: 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 (220540) and the Darwin Tree of Life Discretionary Award [218328, https://doi.org/10.35802/218328].
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Boyes D, Boyes C, University of Oxford and Wytham Woods Genome Acquisition Lab *et al.* **The genome sequence of the Regal Piercer, *Pammene regiana* (Zeller, 1849) (Lepidoptera: Tortricidae) [version 1; peer review: 2 approved, 1 approved with reservations]** Wellcome Open Research 2026, **11**:232 https://doi.org/10.12688/wellcomeopenres.26362.1

First published: 21 Apr 2026, **11**:232 https://doi.org/10.12688/wellcomeopenres.26362.1

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; Apoditrysia; Tortricoidea; Tortricidae; Olethreutinae; Grapholitini; *Pammene*; *Pammene regiana* (Zeller, 1849) (NCBI:txid1660671).

Background

Pammene regiana is a micromoth in the family Tortricidae. It is found throughout the British Isles and central Europe (GBIF Secretariat, 2025). It occurs in woodlands, gardens and hedges where Sycamore and other *Acer* species occur (Hancock *et al.*, 2015).

This common moth (forewing length 6–7.5 mm) is dark brown with a large pale orange blotch on the dorsum and several buffish costal streaks. It flies between May and August and occasionally comes to light (Sterling *et al.*, 2023). The eggs are laid on Acer species and the larvae feed in the seeds. In September, the larvae pupate on the trunk or a lower branch of the host (Hancock *et al.*, 2015).

We present a chromosome-level genome sequence for *Pammene regiana*, produced using the Tree of Life pipeline from a specimen collected from Wytham Woods, Oxfordshire, UK (Figure 1). This assembly was generated as part of the Darwin Tree of Life Project, which aims to generate high-quality reference genomes for all named eukaryotic species in Britain and Ireland to support research, conservation, and the sustainable use of biodiversity (Darwin Tree of Life Project Consortium, 2022).

Methods

Sample acquisition and DNA barcoding

The specimen used for genome sequencing was an adult female *Pammene regiana* (specimen ID Ox000656, ToLID ilPamRegi1; Figure 1), collected from Wytham Woods, Oxfordshire, UK (latitude 51.772, longitude −1.338) on 2020-07-20. The specimen was collected and identified by Douglas Boyes (University of Oxford). A second specimen was used for Hi-C and RNA sequencing (specimen ID NHMUK013697073, ToLID ilPamRegi2). It was collected from Lucas Road, High Wycombe, England, United Kingdom (latitude 51.63, longitude −0.74) on 2022-05-14. The specimen was collected and identified by David Lees.

The initial identification was verified by an additional DNA barcoding process according to the framework developed by Twyford *et al.* (2024). A small sample was dissected from the specimen and stored in ethanol, while the remaining parts were shipped on dry ice to the Wellcome Sanger Institute (WSI) (see the protocol). The tissue was lysed, the COI marker region was amplified by PCR, and amplicons were sequenced and compared to the BOLD database, confirming the species identification (Crowley *et al.*, 2023). Following whole genome sequence generation, the relevant DNA barcode region was also used alongside the initial barcoding data for sample tracking at the WSI (Twyford *et al.*, 2024). The standard operating procedures for Darwin Tree of Life barcoding are available on protocols.io.

Figure 1. Photograph of the *Pammene regiana* (ilPamRegi1) specimen used for genome sequencing.

Nucleic acid extraction

Detailed protocols for nucleic acid extraction developed at the Wellcome Sanger Institute (WSI) Tree of Life Core Laboratory are available on protocols.io (Howard *et al.*, 2025). The ilPamRegi1 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. High molecular weight (HMW) DNA was extracted 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 5.66 ng/μL and a yield of 2 207.40 ng.

RNA was extracted from abdomen tissue of ilPamRegi2 in the Tree of Life Laboratory at the WSI using the RNA Extraction: Automated MagMax™ *mir*Vana protocol. The RNA concentration was assessed using a Nanodrop spectrophotometer and a Qubit Fluorometer using the Qubit RNA Broad-Range Assay kit. Analysis of the integrity of the RNA was done using the Agilent RNA 6000 Pico Kit and Eukaryotic Total RNA assay.

PacBio HiFi library preparation and sequencing

Library preparation and sequencing were performed at the WSI Scientific Operations core. Libraries were prepared using the SMRTbell Prep Kit 3.0 (Pacific Biosciences, California, USA), following 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 using the Sequel IIe system (Pacific Biosciences, California, USA). The concentration of the library loaded onto the Sequel IIe was in the range 40–135 pM. The SMRT link software, a PacBio web-based end-to-end workflow manager, was used to set-up and monitor the run, and to perform primary and secondary analysis of the data upon completion.

Hi-C

Sample preparation and crosslinking

The Hi-C sample was prepared from 20–50 mg of frozen thorax tissue from the ilPamRegi2 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.

RNA library preparation and sequencing

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

Genome assembly

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

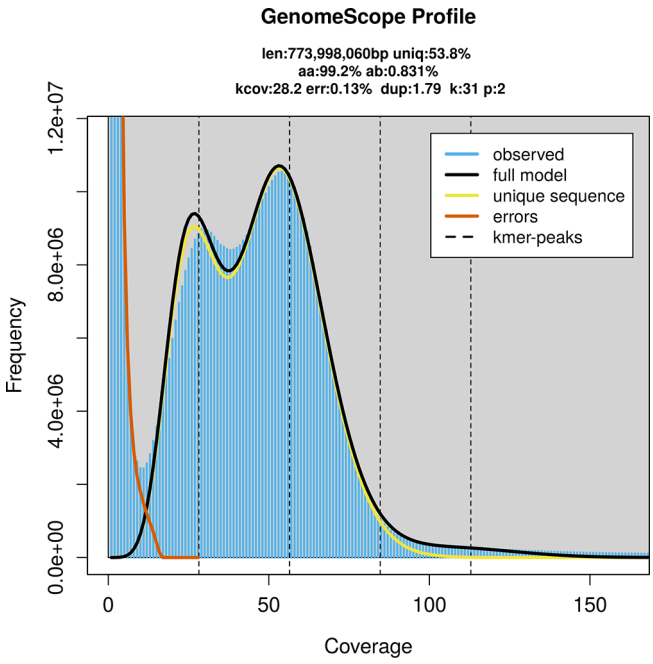

Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2. The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

Table 1. Specimen and sequencing data for BioProject PRJEB71190.

Platform	PacBio HiFi	Hi-C	RNA-seq
ToLID	ilPamRegi1	ilPamRegi2	ilPamRegi2
Specimen ID	Ox000656	NHMUK013697073	NHMUK013697073
BioSample (source individual)	SAMEA7701518	SAMEA115574683	SAMEA115574683
BioSample (tissue)	SAMEA7701700	SAMEA115599778	SAMEA115575071
Tissue	whole organism	thorax	abdomen
Instrument	Sequel IIe	Illumina NovaSeq X	Illumina NovaSeq X
Run accessions	ERR12370358; ERR12370359; ERR12370357	ERR13702753	ERR13999086
Read count total	6.93 million	341.71 million	37.84 million
Base count total	93.82 Gb	103.20 Gb	11.43 Gb

Table 2. **Genome assembly data for *Pammene regiana.***

Genome assembly	Primary assembly
Assembly name	ilPamRegi1.1
Assembly accession	GCA_965641965.1
Alternate haplotype accession	GCA_965642025.1
Assembly level	chromosome
Span (Mb)	863.59
Number of chromosomes	29
Number of contigs	238
Contig N50	9.25 Mb
Number of scaffolds	91
Scaffold N50	30.01 Mb
Sex chromosomes	W and Z
Organelles	Mitochondrion: 16.5 kb

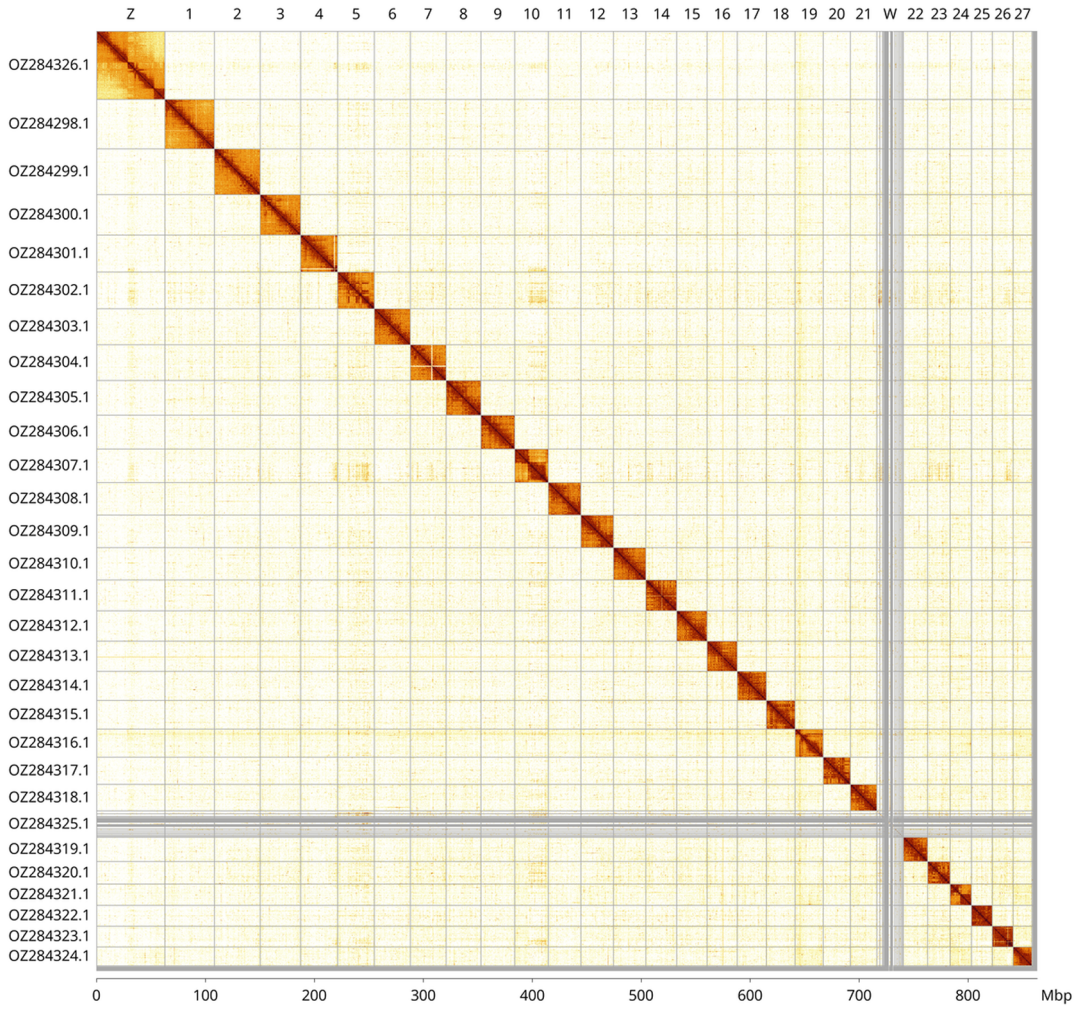

Figure 3. **Hi-C contact map of the *Pammene regiana* genome assembly.** Assembled chromosomes are shown in order of size and labelled along the axes, with a megabase scale shown below. The plot was generated using PretextSnapshot.

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

The mitochondrial genome was assembled using MitoHiFi (Uliano-Silva *et al.*, 2023).

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 48 breaks and 58 joins. This increased the scaffold count by 24.3% and increased the total assembly length by 0.8%. 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.

Table 3. Chromosomal pseudomolecules in the primary genome assembly of *Pammene regiana* ilPamRegi1.

INSDC accession	Molecule	Length (Mb)	GC%
OZ284298.1	1	45.35	37
OZ284299.1	2	42.20	37
OZ284300.1	3	37.01	37
OZ284301.1	4	34.07	37.50
OZ284302.1	5	33.73	37.50
OZ284303.1	6	32.95	37
OZ284304.1	7	32.86	37.50
OZ284305.1	8	31.90	37
OZ284306.1	9	31.11	37
OZ284307.1	10	30.77	37.50
OZ284308.1	11	30.01	37.50
OZ284309.1	12	29.93	37.50
OZ284310.1	13	29.69	37
OZ284311.1	14	28.22	37
OZ284312.1	15	27.93	37.50
OZ284313.1	16	27.60	37.50
OZ284314.1	17	26.76	37.50
OZ284315.1	18	26.27	37.50
OZ284316.1	19	25.93	38
OZ284317.1	20	25	37.50
OZ284318.1	21	24.33	37.50
OZ284319.1	22	22.12	37
OZ284320.1	23	20.87	37
OZ284321.1	24	19.52	39
OZ284322.1	25	19.10	38
OZ284323.1	26	19.07	38
OZ284324.1	27	17.31	38.50
OZ284325.1	W	24.28	37
OZ284326.1	Z	62.86	37

Assembly quality assessment

The MerquryFK tool (Rhie *et al.*, 2020) was run in a Singularity container (Kurtzer *et al.*, 2017) to evaluate *k*-mer completeness and assembly quality for the primary and alternate haplotypes using the *k*-mer 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 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).

Genome sequence report

Sequence data

PacBio sequencing of the *Pammene regiana* specimen generated 93.82 Gb (gigabases) from 6.93 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 776.06 Mb, with a heterozygosity of 0.83% and repeat content of 46.22% (Figure 2). These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 60× coverage. Hi-C sequencing produced 103.20 Gb from 341.71 million reads, which were used to scaffold the assembly. RNA sequencing data were also generated and are available in public sequence repositories. Table 1 summarises the specimen and sequencing details.

Assembly statistics

The primary haplotype was assembled, and contigs corresponding to an alternate haplotype were also deposited in INSDC databases. The final assembly has a total length of 863.59 Mb in 91 scaffolds, with 147 gaps, and a scaffold N50 of 30.01 Mb (Table 2).

Most of the assembly sequence (99.44%) was assigned to 29 chromosomal-level scaffolds, representing 27 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). Chromosomes Z and W were assigned based on copy number in the diploid assembly and by Hi-C signal.

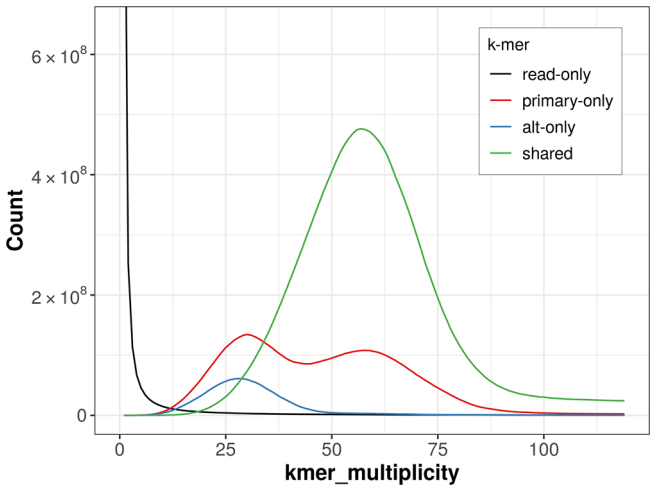

Figure 4. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers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

The mitochondrial genome was also assembled (length 16.5 kb, OZ284327.1). This sequence is included as a contig in the multifasta file of the genome submission and as a standalone record.

Assembly quality metrics

The combined primary and alternate assemblies achieve an estimated QV of 60.4. The *k*-mer completeness is 89.44% for the primary assembly, 69.74% for the alternate haplotype, and 98.15% for the combined assemblies (Figure 4).

BUSCO v.5.8.3 analysis using the lepidoptera_odb10 reference set (n = 5 286) identified 98.6% of the expected gene set (single = 97.8%, duplicated = 0.7%). The snail plot in Figure 5 summarises the scaffold length distribution and other assembly statistics for the primary assembly. The blob plot in Figure 6 shows the distribution of scaffolds by GC proportion and coverage.

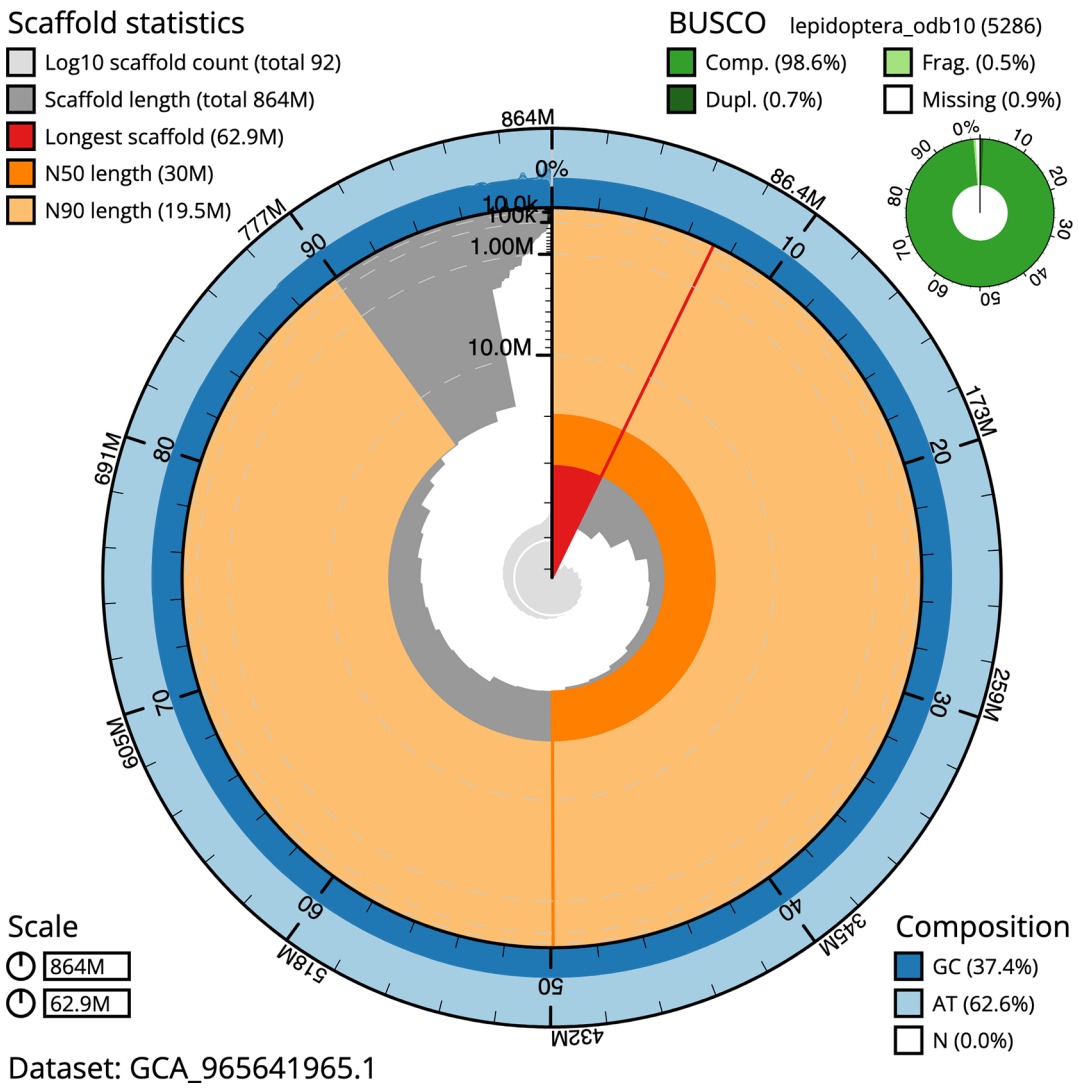

Figure 5. Assembly metrics for ilPamRegi1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1 000 bins around the circumference. The outermost blue 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 lepidoptera_odb10 set is presented at the top right. An interactive version of this figure can be accessed on the BlobToolKit viewer.

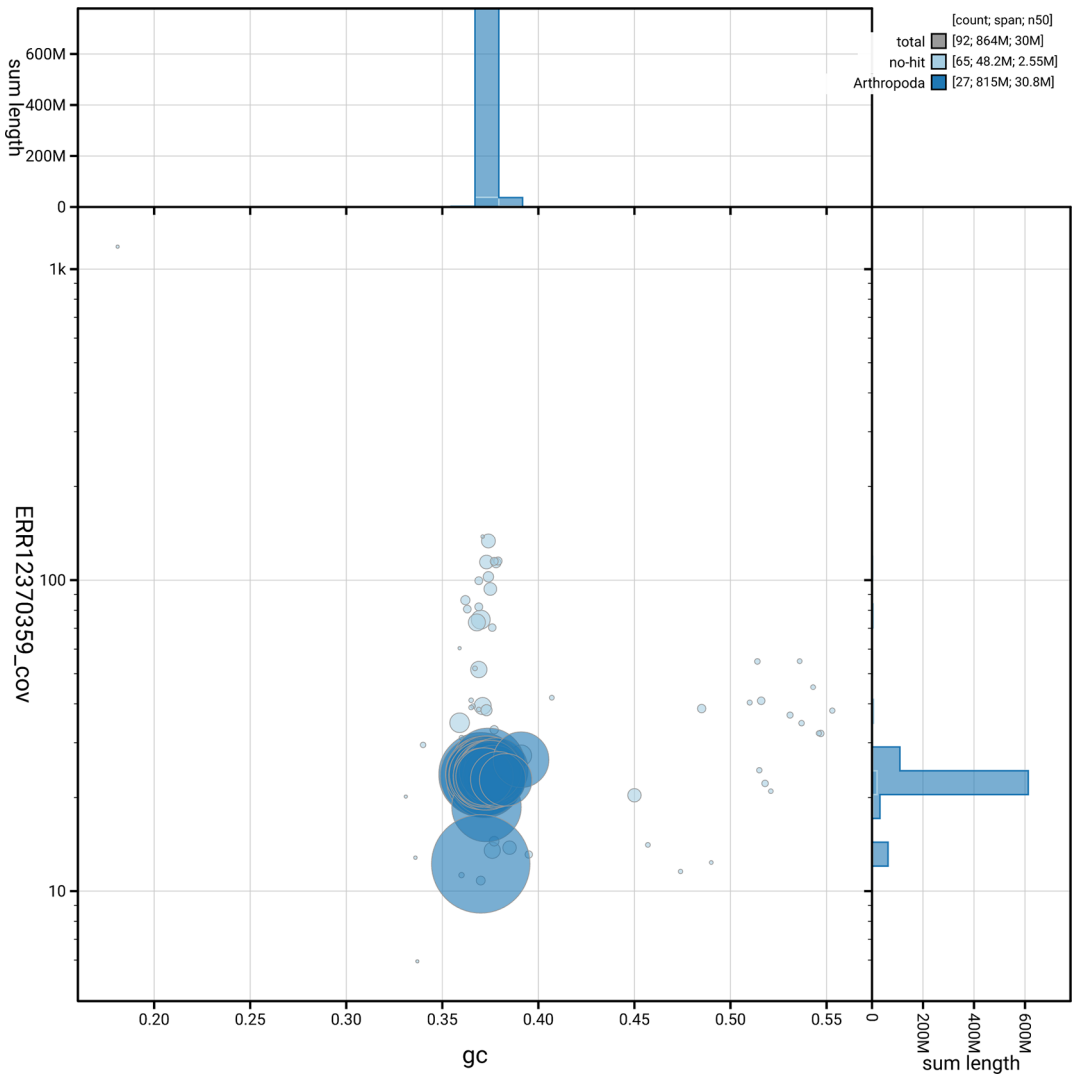

Figure 6. BlobToolKit blob plot for ilPamRegi1.1. The plot shows base coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes display the total 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 *Pammene regiana* assembly.

Measure	Value	Benchmark
EBP summary (primary)	6.C.Q63	6.C.Q40
Contig N50 length	9.25 Mb	≥ 1 Mb
Scaffold N50 length	30.01 Mb	= chromosome N50
Consensus quality (QV)	Primary: 63.0; alternate: 59.4; combined: 60.4	≥ 40
k-mer completeness	Primary: 89.44%; alternate: 69.74%; combined: 98.15%	≥ 95%

Table 4. *Continued*

Measure	Value	Benchmark
BUSCO	C:98.6% [S:97.8%, D:0.7%], F:0.5%, M:0.9%, n:5 286	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	99.44%	≥ 90%

Notes: The 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.

Table 4 lists the assembly metric benchmarks adapted from Rhie *et al.* (2021) and the Earth BioGenome Project Report on Assembly Standards January 2026. The EBP metric, calculated for the primary assembly, is **6.C.Q63**, meeting the recommended reference standard.

Author information

Contributors are listed at the following links:

- Members of the University of Oxford and Wytham Woods Genome Acquisition Lab
- Members of the Natural History Museum Genome Acquisition Lab
- Members of the Darwin Tree of Life Barcoding collective
- Members of the Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team
- Members of Wellcome Sanger Institute Scientific Operations – Sequencing Operations
- Members of the Wellcome Sanger Institute Tree of Life Core Informatics team
- Members of the Tree of Life Core Informatics collective
- Members of the Darwin Tree of Life Consortium

Wellcome Sanger Institute – Legal and Governance

The materials that have contributed to this genome note have been supplied by a Darwin Tree of Life Partner. The submission of materials by a Darwin Tree of Life Partner is subject to the **'Darwin Tree of Life Project Sampling Code of Practice'**, which can be found in full on the Darwin Tree of Life website. By agreeing with and signing up to the Sampling Code of Practice, the Darwin Tree of Life Partner agrees they will meet the legal and ethical requirements and standards set out within this document in respect of all samples acquired for, and supplied to, the Darwin Tree of Life Project. Further, the Wellcome Sanger Institute employs a process whereby due diligence is carried out proportionate to the nature of the materials themselves, and the circumstances under which they have been/are 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: Pammene regiana (regal piercer). Accession number PRJEB71190. The genome sequence is released openly for reuse. The *Pammene regiana* genome sequencing initiative is part of the Darwin Tree of Life Project (PRJEB40665), 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 the Ensembl pipeline at the European Bioinformatics Institute. Raw data and assembly accession identifiers are reported in Tables 1 and 2.

Production code used in genome assembly at the WSI Tree of Life is available at https://github.com/sanger-tol. Table 5 lists software versions used in this study.

Table 5. Software versions and sources.

Software	Version	Source
BLAST	2.14.0	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/
BlobToolKit	4.4.6	https://github.com/blobtoolkit/blobtoolkit
BUSCO	5.8.3	https://gitlab.com/ezlab/busco
bwa-mem2	2.2.1	https://github.com/bwa-mem2/bwa-mem2
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
MerquryFK	1.1.2	https://github.com/thegenemyers/MERQURY.FK
Minimap2	2.28-r1209	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	24.10.4	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	1.21	https://github.com/samtools/samtools
sanger-tol/ascc	0.1.0	https://github.com/sanger-tol/ascc
sanger-tol/blobtoolkit	v0.8.0	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

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 M-J, 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 H-G: **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

Crowley L, Allen H, Barnes I, *et al.*: **A sampling strategy for genome sequencing the British terrestrial Arthropod fauna.** *Wellcome Open Res.*

2023; **8**: 123.
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).
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Darwin Tree of Life Project Consortium: **Sequence locally, think globally: The Darwin Tree of Life Project.** *Proc. Natl. Acad. Sci. USA.* 2022; **119**(4): e2115642118.
PubMed Abstract | **Publisher Full Text** | **Free 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**

GBIF Secretariat: ***Pammene regiana* (Zeller, 1849).** 2025.
Reference Source

Hancock EF, Bland KP, Razowski J: *The Moths and Butterflies of Great Britain and Ireland: Tortricidae Olethreutinae.* Leiden: Brill; 2015.

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).
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Kerpedjiev P, Abdennur N, Lekschas F, *et al.*: **HiGlass: Web-based visual exploration and analysis of genome interaction maps.** *Genome Biol.* 2018; **19**(1): 125.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Kurtzer GM, Sochat V, Bauer MW: **Singularity: Scientific containers for mobility of compute.** *PLoS One.* 2017; **12**(5): e0177459.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Li H: **Minimap2: Pairwise alignment for nucleotide sequences.** *Bioinformatics.* 2018; **34**(18): 3094–3100.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

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).
Publisher Full Text

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).
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Schoch CL, Ciufo S, Domrachev M, *et al.*: **NCBI taxonomy: A comprehensive update on curation, resources and tools.** *Database.* 2020; **2020**: baaa062.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Sterling P, Parsons M, Lewington R: *Field Guide to the Micro Moths of Great Britain and Ireland.* Dorset: British Wildlife Publishing; 2023.

Twyford AD, Beasley J, Barnes I, *et al.*: **A DNA barcoding framework for taxonomic verification in the Darwin Tree of Life Project.** *Wellcome Open Res.* 2024; **9**: 339.
PubMed Abstract | **Publisher Full Text** | **Free 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.** *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS).* IEEE; 2019; pp. 314–324.
Publisher Full Text

Zhou C, McCarthy SA, Durbin R: **YaHS: Yet another Hi-C scaffolding tool.** *Bioinformatics.* 2023; **39**(1).
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Open Peer Review

Current Peer Review Status: ✔ ✔ ?

Version 1

Reviewer Report 22 June 2026

https://doi.org/10.21956/wellcomeopenres.29033.r157326

?

Christophe Klopp
INRAE, Castanet-Tolosan, France

The authors present the first chromosomal assembly of the Pammene regiana genome. The methods are usual and the article is standard.

The authors should indicate if hifiasm was run with Hi-C or not.
The authors should have provided hap1 and hap2 which are easier to exploit in further analyses than primary and alternate which are now outdated.

The authors should give some insight on the difference between kmer genome size estimation (776.06 Mb) and genome assembly size (863.59) with is close to 12%.
Regarding: "Based on the estimated genome size, the sequencing data provided approximately 60× coverage." If you divide 93.82 Gb by 0.776 Gb you obtain 120x and not 60. This should be checked.
Looking at the Hi-C map, it is unclear how the W chromosome has been scaffolded. This should be explained.
The "primary only" kmer peak located at the homozygous position (60x) is not awaited. It usually indicates that your primary assembly contains duplicated haplotypic contigs. But the Busco scores do not show this duplication. This is really strange and should be checked.
The authors should use the same diagram type (meaning Y axis) for genomescope2 and multiplicity plot to ease comparison.
When you dot plot primary and alternate assemblies some of the chromosomes show configuration with very low probability of occurrence = over 10 rearrangements covering the complete chromosome. This is very unusual and should be validated by the Hi-C. The authors should provide Hi-C maps for both haplotypic chromosomes in such cases.

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?
No

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

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: genome assembly and annotation

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 15 June 2026

https://doi.org/10.21956/wellcomeopenres.29033.r157322

Axel Künstner
University of Lübeck, Lübeck, Germany

This is a standard Darwin Tree of Life genome note for *Pammene regiana*, following the well-established DToL template. The assembly metrics are strong: contig N50 of 9.25 Mb, scaffold N50 of 30.01 Mb, 99.44% of the sequence assigned to chromosomes, and BUSCO completeness of 98.6% (S:97.8%, D:0.7%). The EBP grade of 6.C.Q63 meets the reference standard.

The assembly resolves 27 autosomes plus the W and Z. Counted conventionally (autosomes plus the Z), this is n = 28 - the modal number documented for Olethreutinae and specifically for the tribe Grapholitini, to which *P. regiana* belongs (Šíchová et al. 2013, PLoS ONE 8:e64520; cf. *Cydia pomonella*, n = 28, WZ/ZZ). The 29 pseudomolecules are this n = 28 complement plus the female-limited W. I would suggest the authors add a sentence placing the chromosome count in this clade-specific context: it is fully consistent with expectation for Olethreutinae and is not a reduction relative to the relevant baseline (n = 28 in this subfamily, versus n = 30 in Tortricinae and n = 31 ancestrally in Lepidoptera).This genome data note reports a chromosome-level reference assembly for Pammene regiana (Regal Piercer; Lepidoptera: Tortricidae: Olethreutinae), produced as part of the Darwin Tree of Life (DToL) project. The assembly was generated from a wild-caught female specimen from Wytham Woods, UK, using PacBio HiFi long reads for contig assembly (Hifiasm), Illumina Hi-C for scaffolding (YaHS), and supplementary RNA-seq data. The article follows the standard DToL data note format and contributes a high-quality reference genome for a species of interest as a representative of the speciose family Tortricidae and a herbivore associated with Acer.

Summary of findings

The primary assembly spans 863.59 Mb in 91 scaffolds, with 99.44% of sequence assigned to 29 chromosomal pseudomolecules representing 27 autosomes and the W and Z sex chromosomes. Key assembly metrics include a contig N50 of 9.25 Mb, scaffold N50 of 30.01 Mb, BUSCO completeness of 98.6% (single-copy 97.8%, duplicated 0.7%) against the lepidoptera_odb10 set, and an EBP summary grade of 6.C.Q63, exceeding the recommended 6.C.Q40 reference standard. The mitochondrial genome (16.5 kb) was also assembled. All raw data and assemblies are deposited under BioProject PRJEB71190.

Major comments

1. Chromosome count and phylogenetic context. The assembly resolves 29 pseudomolecules, comprising 27 autosomes, the Z, and the female-limited W. Counted in the conventional haploid notation (autosomes + Z), this corresponds to n = 28. The manuscript does not situate this number within the karyotypic context of the relevant clade, and`lepidopteran karyotype respectively, not of this subfamily. The chromosome complement of P. regiana is therefore fully consistent with expectation for its clade and does not represent a reduction. I recommend the authors add one or two sentences placing the observed n = 28 in this subfamily-level context, explicitly distinguishing it from the Tortricinae baseline that appears to motivate the reviewer's comment.

Minor comments

1. The k-mer completeness of the primary assembly alone (89.44%) falls below the EBP threshold of ≥95%, though the combined primary + alternate value of 98.15% is satisfactory. It would be helpful to add a brief sentence acknowledging this and noting that the combined assemblies meet the benchmark, as readers unfamiliar with haplotype-resolved assembly conventions may otherwise interpret the primary figure as a quality concern.

2. Table 3 reports GC content rounded to 0.5% intervals for all but a few chromosomes. It is unclear whether this reflects the true precision of the estimates or a reporting convention. If the latter, a note to this effect would aid readers wishing to use these values.

3. Manual curation involved 48 breaks and 58 joins, increasing scaffold count by 24.3% and assembly length by 0.8%. These figures are reported without further commentary. A brief statement on what types of misassemblies or joins were identified (e.g. chimeric contigs, proximity ligation artefacts) would add context, even if brief.

Conclusion

This is a technically sound and well-documented genome data note. The assembly quality is high by current standards and meets or exceeds EBP reference benchmarks. The primary concern — the interpretation of the chromosome count — is one of contextualisation rather than error, and can be addressed straightforwardly in a revised author response or in a subsequent version. I recommend approval subject to the addition of clade-appropriate karyotypic context.

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: Genetics, Microbiome, Cancer

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 05 May 2026

https://doi.org/10.21956/wellcomeopenres.29033.r153964

Jurate De Prins
National Collections and Marine Infrastructure, Commonwealth Scientific and Industrial Research Organisation (Ringgold ID: 2221), Canberra, Australian Capital Territory, Australia

It is very appreciated that the whole genome sequence is presented for this species. The authors found 29 chromosomal pseudomolecules, including W and Z chromosomes. This is one pair less than the usual set of chromosomes for microlepidoptera, which contains 30 pairs of chromosomes. Maybe the molecular data could provide a short explanation for this reduction?

My second comment is actually related to all the articles of the Tree of Life Microlepidoptera section. It is very important to have 100% correct species ID of the specimen that was used for the whole mitogenome sequencing. A small photo of an adult in a rest position is certainly not enough. In this particular case, the ID was verified against DNA barcode database. It is *Pammene regiana,* otherwise I would doubt that it could also be *Pammene populana* (please see https://projects.biodiversity.be/lepidoptera/species/4349/)

In the UK, and also in Flanders, every year about 20 species of Microlepidoptera are recorded that come to Europe and the UK with the plant and food trade from all over the world. Many of these species are not DNA barcoded, or some of them are misidentified or have the same DNA barcodes. Therefore, three independent character sets are used for correct species ID of the specimen, which is processed for whole mitogenome sequencing: 1) external characters; 2) male

and/or female genitalia characters; 3) molecular characters. All three character sets should lead to the same species ID. Otherwise, the danger exists that an error or misidentification can occur in the ID of the whole mitogenomic set. Here, in this article is not the case, but I would advise including micromorphology traits for Microlepidoptera in the protocols. This is the most reliable method to recognise the exotic species arriving via human and trade travels, and also a double check for native European species of Microlepidoptera.

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: Taxonomy, Microlepidoptera

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.