DATA NOTE

# The genome sequence of the Hook-streak Grass-veneer moth, *Crambus lathoniellus* (Zincken, 1817) [version 1; peer review: awaiting peer review]

Douglas Boyes[1+],
University of Oxford and Wytham Woods 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

[+] Deceased author

**v1** **First published:** 14 Aug 2024, **9**:473
https://doi.org/10.12688/wellcomeopenres.22864.1
**Latest published:** 14 Aug 2024, **9**:473
https://doi.org/10.12688/wellcomeopenres.22864.1

## Abstract

We present a genome assembly from an individual male *Crambus lathoniellus* (the Hook-streak Grass-veneer moth; Arthropoda; Insecta; Lepidoptera; Crambidae). The genome sequence spans 893.80 megabases. Most of the assembly is scaffolded into 30 chromosomal pseudomolecules, including the Z sex chromosome. The mitochondrial genome has also been assembled and is 16.49 kilobases in length. Gene annotation of this assembly on Ensembl identified 24,061 protein-coding genes.

## Keywords

Crambus lathoniellus, Hook-streak Grass-veneer moth, genome sequence, chromosomal, Lepidoptera

This article is included in the Tree of Life gateway.

**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.

**Corresponding author:** Darwin Tree of Life Consortium (mark.blaxter@sanger.ac.uk)

**Author roles: Boyes D**: Investigation, Resources;

**Competing interests:** No competing interests were disclosed.

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



**How to cite this article:** Boyes D, University of Oxford and Wytham Woods Genome Acquisition Lab, Darwin Tree of Life Barcoding collective *et al.* **The genome sequence of the Hook-streak Grass-veneer moth, *Crambus lathoniellus* (Zincken, 1817) [version 1; peer review: awaiting peer review]** Wellcome Open Research 2024, **9**:473 https://doi.org/10.12688/wellcomeopenres.22864.1

**First published:** 14 Aug 2024, **9**:473 https://doi.org/10.12688/wellcomeopenres.22864.1

## 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; Pyraloidea; Crambidae; Crambinae; *Crambus*; *Crambus lathoniellus* (NCBI:txid1100958).

## Background

The genome of the Hook-streak Grass-veneer, *Crambus lathoniellus*, was sequenced as part of the Darwin Tree of Life Project, a collaborative effort to sequence all named eukaryotic species in the Atlantic Archipelago of Britain and Ireland. Here we present the first chromosomally complete genome sequence for *Crambus lathoniellus*, based on one male specimen from Wytham Woods, Oxfordshire, UK.

## Genome sequence report

The genome of an adult male *Crambus lathoniellus* (Figure 1) was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating a total of 20.94 Gb (gigabases) from 1.81 million reads, providing approximately 22-fold coverage. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data, which produced 101.21 Gbp from 670.27 million reads, yielding an approximate coverage of 113-fold. Specimen and sequencing information is summarised in Table 1.

Manual assembly curation corrected 213 missing joins or mis-joins and 155 haplotypic duplications, reducing the assembly length by 5.49% and the scaffold number by 61.36%, and increasing the scaffold N50 by 1.87%. The final assembly has a total length of 893.80 Mb in 84 sequence scaffolds with a scaffold N50 of 31.7 Mb (Table 2). The snail plot in Figure 2 provides a summary of the assembly statistics, while the distribution of assembly scaffolds on GC proportion and coverage is shown in Figure 3. The cumulative assembly plot in Figure 4 shows curves for subsets of scaffolds assigned to different phyla. Most (99.74%) of the assembly sequence was assigned to 30 chromosomal-level scaffolds, representing 29 autosomes and the Z sex chromosome. Chromosome-scale scaffolds confirmed by the Hi-C data are named in order of size (Figure 5; Table 3). While not fully phased, the assembly deposited is of one haplotype. Contigs corresponding to the second haplotype have also been deposited. The mitochondrial genome was also assembled and can be found as a contig within the multifasta file of the genome submission.


**Figure 1. Photograph of the *Crambus lathoniellus* (ilCraLath2) specimen used for genome sequencing.**

**Table 1. Specimen and sequencing data for *Crambus lathoniellus*.**

| **Project information** | | | |
|---|---|---|---|
| **Study title** | Crambus lathoniellus (hook-streak grass-veneer) | | |
| **Umbrella BioProject** | PRJEB57115 | | |
| **Species** | *Crambus lathoniellus* | | |
| **BioSample** | SAMEA10979181 | | |
| **NCBI taxonomy ID** | 1100958 | | |
| **Specimen information** | | | |
| **Technology** | **ToLID** | **BioSample accession** | **Organism part** |
| **PacBio long read sequencing** | ilCraLath2 | SAMEA10979617 | Whole organism |
| **Hi-C sequencing** | ilCraLath2 | SAMEA10979617 | Whole organism |
| **Sequencing information** | | | |
| **Platform** | **Run accession** | **Read count** | **Base count (Gb)** |
| **Hi-C Illumina NovaSeq 6000** | ERR10446392 | 6.70e+08 | 101.21 |
| **PacBio Sequel IIe** | ERR10439753 | 1.81e+06 | 20.94 |

**Table 2. Genome assembly data for *Crambus lathoniellus*, ilCraLath2.1.**

| **Genome assembly** | | |
|---|---|---|
| Assembly name | ilCraLath2.1 | |
| Assembly accession | GCA_949710035.1 | |
| *Accession of alternate haplotype* | *GCA_949710025.1* | |
| Span (Mb) | 893.80 | |
| Number of contigs | 524 | |
| Contig N50 length (Mb) | 3.5 | |
| Number of scaffolds | 84 | |
| Scaffold N50 length (Mb) | 31.7 | |
| Longest scaffold (Mb) | 67.47 | |
| **Assembly metrics*** | | ***Benchmark*** |
| Consensus quality (QV) | 61.9 | *≥ 50* |
| *k*-mer completeness | 100.0% | *≥ 95%* |
| BUSCO** | C:98.2%[S:97.6%,D:0.6%], F:0.5%,M:1.3%,n:5,286 | *C ≥ 95%* |
| Percentage of assembly mapped to chromosomes | 99.74% | *≥ 95%* |
| Sex chromosomes | Z | *localised homologous pairs* |
| Organelles | Mitochondrial genome: 16.49 kb | *complete single alleles* |
| **Genome annotation of assembly GCA_949710035.1 at Ensembl** | | |
| Number of protein-coding genes | 24,061 | |
| Number of gene transcripts | 24,374 | |

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

** BUSCO scores based on the lepidoptera_odb10 BUSCO set using version 5.3.2. C = complete [S = single copy, D = duplicated], F = fragmented, M = missing, n = number of orthologues in comparison. A full set of BUSCO scores is available at https://blobtoolkit.genomehubs.org/view/ilCraLath2_1/dataset/ilCraLath2_1/busco.

The estimated Quality Value (QV) of the final assembly is 61.9 with *k*-mer completeness of 100.0%, and the assembly has a BUSCO v5.3.2 completeness of 98.2% (single = 97.6%, duplicated = 0.6%), using the lepidoptera_odb10 reference set ($n$ = 5,286).

Metadata for specimens, BOLD barcode results, spectra estimates, sequencing runs, contaminants and pre-curation assembly statistics are given at https://links.tol.sanger.ac.uk/species/1100958.

## Genome annotation report

The *Crambus lathoniellus* genome assembly (GCA_949710035.1) was annotated at the European Bioinformatics Institute (EBI) on Ensembl Rapid Release. The resulting annotation includes 24,374 transcribed mRNAs from 24,061 protein-coding genes (Table 2; https://rapid.ensembl.org/Crambus_lathoniellus_GCA_949710035.1/Info/Index). The average transcript length is 7,917.62. There are 1.01 coding transcripts per gene and 4.45 exons per transcript.

## Methods

### Sample acquisition

An adult male *Crambus lathoniellus* (specimen ID Ox001918, ToLID ilCraLath2) was collected from Wytham Woods, Oxfordshire (biological vice-county Berkshire), UK (latitude 51.77, longitude –1.34) on 2021-06-16 using a light trap. The specimen was collected and identified by Douglas Boyes (University of Oxford) and preserved on dry ice.

The initial species identification was verified by an additional DNA barcoding process according to the framework

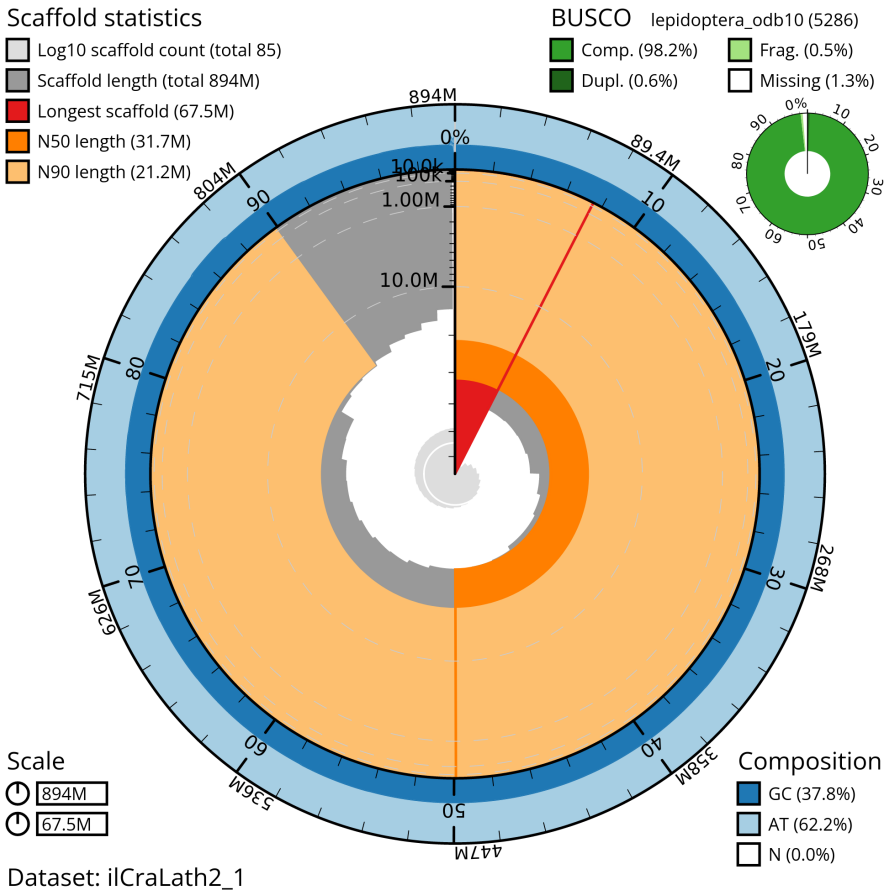


**Figure 2. Genome assembly of *Crambus lathoniellus*, ilCraLath2.1: metrics.** The BlobToolKit snail plot shows N50 metrics and BUSCO gene completeness. The main plot is divided into 1,000 size-ordered bins around the circumference with each bin representing 0.1% of the 893,782,834 bp assembly. The distribution of scaffold lengths is shown in dark grey with the plot radius scaled to the longest scaffold present in the assembly (67,472,359 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 scaffold lengths (31,726,729 and 21,161,173 bp), respectively. The pale grey spiral shows the cumulative scaffold count on a log scale with white scale lines showing successive orders of magnitude. The blue and pale-blue area around the outside of the plot shows the distribution of GC, AT and N percentages in the same bins as the inner plot. A summary of complete, fragmented, duplicated and missing BUSCO genes in the lepidoptera_odb10 set is shown in the top right. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/ilCraLath2_1/dataset/ilCraLath2_1/snail.

developed by Twyford *et al.* (2024). A small sample was dissected from the specimen and stored in ethanol, while the remaining parts of the specimen were shipped on dry ice to the Wellcome Sanger Institute (WSI). 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 have been deposited on protocols.io (Beasley *et al.*, 2023).

## Nucleic acid extraction

The workflow for high molecular weight (HMW) DNA extraction at the Wellcome Sanger Institute (WSI) Tree of Life Core Laboratory includes a sequence of core procedures: sample preparation; sample homogenisation, DNA extraction, fragmentation, and clean-up. In sample preparation, the ilCraLath2 sample was weighed and dissected on dry ice (Jay *et al.*, 2023). Tissue from whole organism was homogenised using a PowerMasher II tissue disruptor (Denton *et al.*, 2023a).

HMW DNA was extracted at the WSI Scientific Operations core using the Automated MagAttract v2 protocol (Oatley *et al.*, 2023). The DNA was sheared into an average fragment size of 12–20 kb in a Megaruptor 3 system (Bates *et al.*, 2023). Sheared DNA was purified by solid-phase reversible immobilisation (Strickland *et al.*, 2023): in brief, the method employs AMPure PB beads to eliminate shorter fragments and concentrate the DNA. The concentration of the sheared and purified DNA was assessed using a Nanodrop

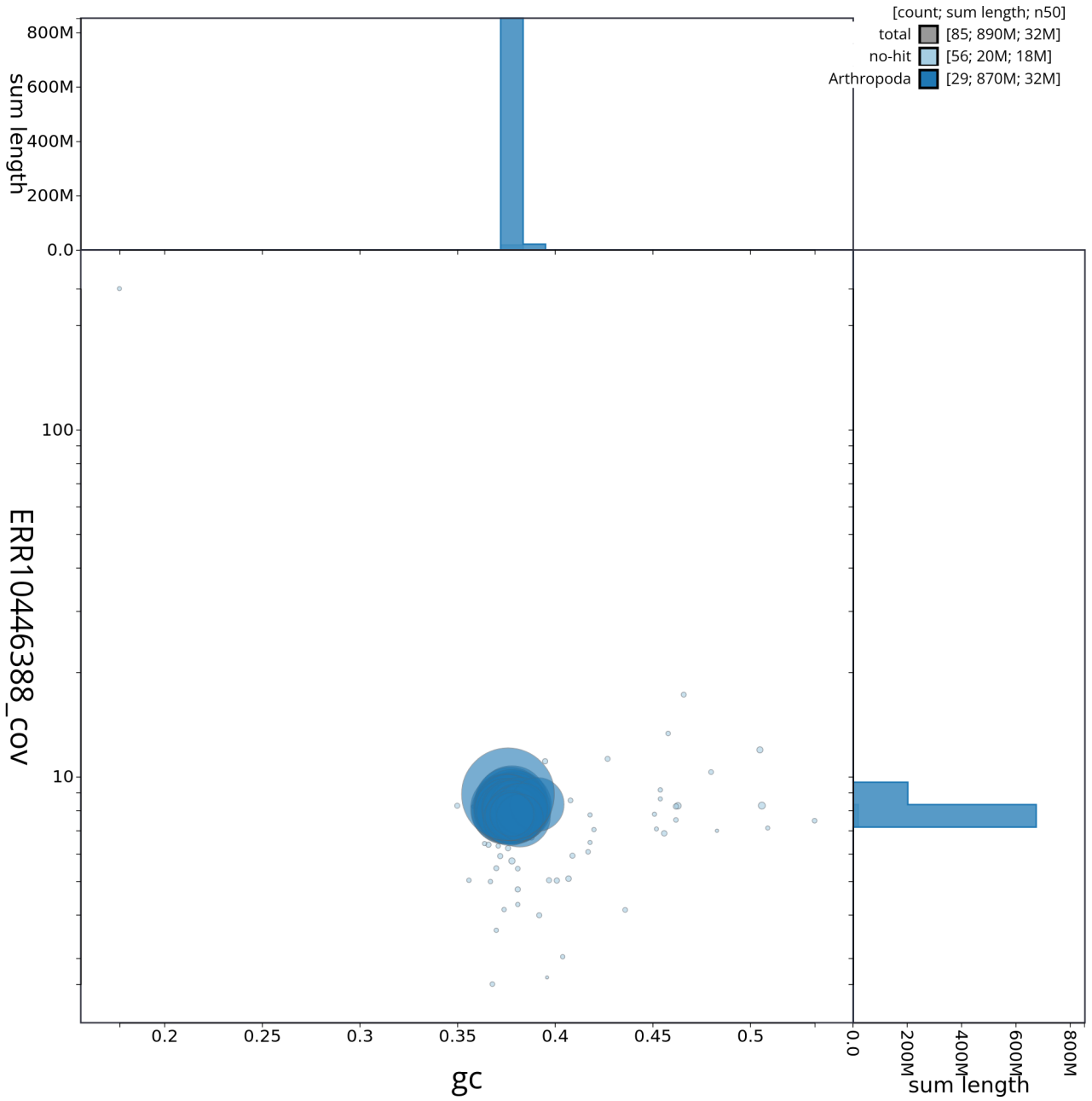


**Figure 3. Genome assembly of *Crambus lathoniellus*, ilCraLath2.1: BlobToolKit GC-coverage plot.** Sequences are coloured by phylum. Circles are sized in proportion to sequence length. Histograms show the distribution of sequence length sum along each axis. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/ilCraLath2_1/dataset/ilCraLath2_1/blob.

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.

Protocols developed by the WSI Tree of Life laboratory are publicly available on protocols.io (Denton *et al.*, 2023b).

## Sequencing

Pacific Biosciences HiFi circular consensus DNA sequencing libraries were constructed according to the manufacturers' instructions. DNA sequencing was performed by the Scientific Operations core at the WSI on a Pacific Biosciences Sequel IIe instrument. Hi-C data were also generated from whole organism tissue of ilCraLath2 using the Arima-HiC v2

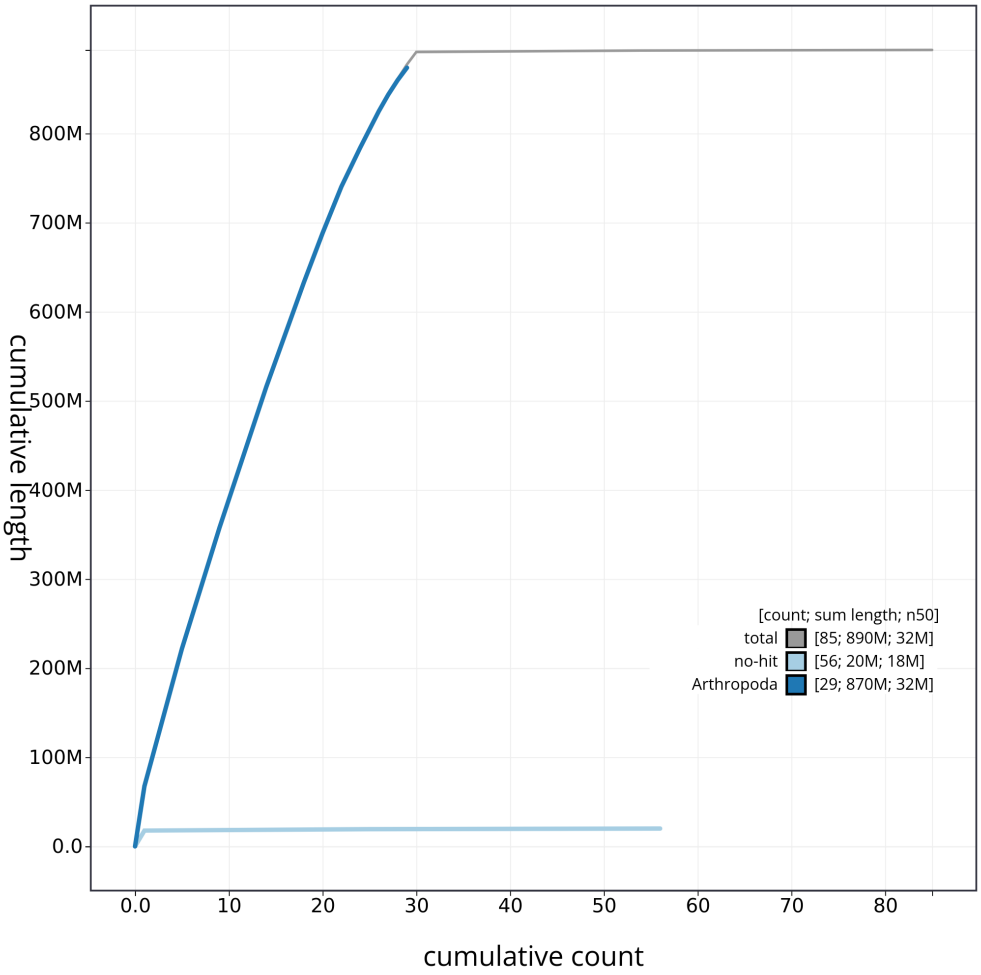


**Figure 4. Genome assembly of *Crambus lathoniellus* ilCraLath2.1: BlobToolKit cumulative sequence plot.** The grey line shows cumulative length for all sequences. Coloured lines show cumulative lengths of sequences assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/ilCraLath2_1/dataset/ilCraLath2_1/cumulative.

kit. The Hi-C sequencing was performed using paired-end sequencing with a read length of 150 bp on the Illumina NovaSeq 6000 instrument.

## Genome assembly, curation and evaluation

### *Assembly*

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

The 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.

### *Assembly curation*

The assembly was decontaminated using the Assembly Screen for Cobionts and Contaminants (ASCC) pipeline (article in preparation). Manual curation was primarily conducted using PretextView (Harry, 2022), with additional insights provided by JBrowse2 (Diesh *et al.*, 2023) and HiGlass (Kerpedjiev *et al.*, 2018). Scaffolds were visually inspected and corrected as described by Howe *et al.* (2021). Any identified contamination, missed joins, and mis-joins were corrected, and duplicate sequences were tagged and removed. The entire process

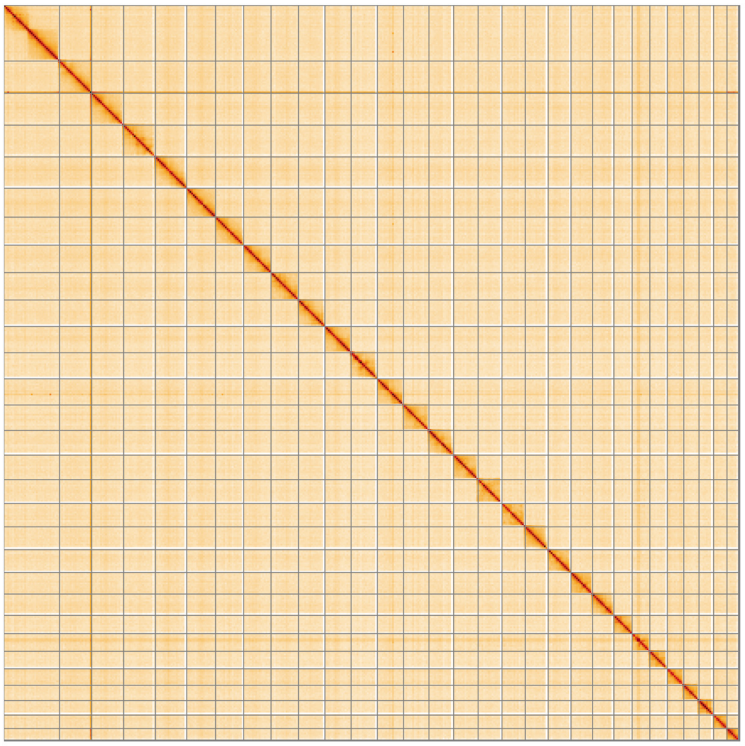

**Figure 5. Genome assembly of *Crambus lathoniellus*, ilCraLath2.1: Hi-C contact map of the ilCraLath2.1 assembly, visualised using HiGlass.** Chromosomes are shown in order of size from left to right and top to bottom. An interactive version of this figure may be viewed at https://genome-note-higlass.tol.sanger.ac.uk/l/?d=QtKd1UIBS0qEkIWdIsVXNQ.

**Table 3. Chromosomal pseudomolecules in the genome assembly of *Crambus lathoniellus*, ilCraLath2.**

| INSDC accession | Name | Length (Mb) | GC% |
|---|---|---|---|
| OX453297.1 | 1 | 39.0 | 38.0 |
| OX453298.1 | 2 | 38.77 | 38.0 |
| OX453299.1 | 3 | 38.59 | 37.5 |
| OX453300.1 | 4 | 38.05 | 37.5 |
| OX453301.1 | 5 | 35.04 | 38.0 |
| OX453302.1 | 6 | 33.93 | 38.0 |
| OX453303.1 | 7 | 33.3 | 38.0 |
| OX453304.1 | 8 | 33.29 | 37.5 |
| OX453305.1 | 9 | 32.07 | 37.5 |
| OX453306.1 | 10 | 31.9 | 37.5 |
| OX453307.1 | 11 | 31.73 | 37.5 |
| OX453308.1 | 12 | 31.73 | 37.5 |
| OX453309.1 | 13 | 30.93 | 38.0 |
| OX453310.1 | 14 | 29.96 | 37.5 |
| OX453311.1 | 15 | 29.7 | 37.5 |
| OX453312.1 | 16 | 28.96 | 38.0 |
| OX453313.1 | 17 | 28.32 | 38.0 |
| OX453314.1 | 18 | 27.79 | 38.0 |
| OX453315.1 | 19 | 27.62 | 38.0 |
| OX453316.1 | 20 | 26.12 | 38.0 |
| OX453317.1 | 21 | 25.96 | 38.0 |
| OX453318.1 | 22 | 22.13 | 38.0 |
| OX453319.1 | 23 | 21.47 | 39.0 |
| OX453320.1 | 24 | 21.16 | 37.5 |
| OX453321.1 | 25 | 20.08 | 38.5 |
| OX453322.1 | 26 | 18.45 | 37.5 |
| OX453323.1 | 27 | 17.7 | 37.5 |
| OX453324.1 | 28 | 16.26 | 38.0 |
| OX453325.1 | 29 | 14.03 | 38.0 |
| OX453296.1 | Z | 67.47 | 37.5 |
| OX453326.1 | MT | 0.02 | 18.0 |

is documented at https://gitlab.com/wtsi-grit/rapid-curation (article in preparation).

***Evaluation of the final assembly***

A Hi-C map for the final assembly was produced using bwa-mem2 (Vasimuddin *et al.*, 2019) in the Cooler file format (Abdennur & Mirny, 2020). To assess the assembly metrics, the *k*-mer completeness and QV consensus quality values were calculated in Merqury (Rhie *et al.*, 2020). This work was done using Nextflow (Di Tommaso *et al.*, 2017) DSL2 pipelines "sanger-tol/readmapping" (Surana *et al.*, 2023a) and "sanger-tol/genomenote" (Surana *et al.*, 2023b). The genome was analysed within the BlobToolKit environment (Challis *et al.*, 2020) and BUSCO scores (Manni *et al.*, 2021; Simão *et al.*, 2015) were calculated.

The genome evaluation pipelines were developed using the nf-core tooling (Ewels *et al.*, 2020), use MultiQC (Ewels *et al.*, 2016), and make extensive use of the Conda package manager, the Bioconda initiative (Grüning *et al.*, 2018), the Biocontainers infrastructure (da Veiga Leprevost *et al.*, 2017), and the Docker (Merkel, 2014) and Singularity (Kurtzer *et al.*, 2017) containerisation solutions.

Table 4 contains a list of relevant software tool versions and sources.

## Genome annotation

The BRAKER2 pipeline (Brůna *et al.*, 2021) was used in the default protein mode to generate annotation for the *Crambus lathoniellus* assembly (GCA_949710035.1) in Ensembl Rapid Release at the EBI.

## 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 here. By agreeing with and signing up to the Sampling Code of Practice, the Darwin Tree of Life Partner agrees they will meet the legal and ethical requirements and standards set out within this document in respect of all samples acquired for, and supplied to, the Darwin Tree of Life Project.

Further, the Wellcome Sanger Institute employs a process whereby due diligence is carried out proportionate to the nature of the materials themselves, and the circumstances under which they have been/are to be collected and provided for use. The purpose of this is to address and mitigate any potential legal and/or ethical implications of receipt and use of the materials as part of the research project, and to ensure that in doing so we align with best practice wherever possible. The overarching areas of consideration are:

- Ethical review of provenance and sourcing of the material
- Legality of collection, transfer and use (national and international)

Each transfer of samples is further undertaken according to a Research Collaboration Agreement or Material Transfer Agreement entered into by the Darwin Tree of Life Partner, Genome Research Limited (operating as the Wellcome

**Table 4. Software tools: versions and sources.**

| Software tool | Version | Source |
|---|---|---|
| BlobToolKit | 4.2.1 | https://github.com/blobtoolkit/blobtoolkit |
| BUSCO | 5.3.2 | https://gitlab.com/ezlab/busco |
| Hifiasm | 0.16.1-r375 | https://github.com/chhylp123/hifiasm |
| HiGlass | 1.11.6 | https://github.com/higlass/higlass |
| Merqury | MerquryFK | https://github.com/thegenemyers/MERQURY.FK |
| MitoHiFi | 2 | https://github.com/marcelauliano/MitoHiFi |
| PretextView | 0.2 | https://github.com/wtsi-hpag/PretextView |
| purge_dups | 1.2.3 | https://github.com/dfguan/purge_dups |
| sanger-tol/genomenote | v1.0 | https://github.com/sanger-tol/genomenote |
| sanger-tol/readmapping | 1.1.0 | https://github.com/sanger-tol/readmapping/tree/1.1.0 |
| YaHS | 1.1a.2 | https://github.com/c-zhou/yahs |

Sanger Institute), and in some circumstances other Darwin Tree of Life collaborators.

## Data availability

European Nucleotide Archive: *Crambus lathoniellus* (hook-streak grass-veneer). Accession number PRJEB57115; https://identifiers.org/ena.embl/PRJEB57115 (Wellcome Sanger Institute, 2023). The genome sequence is released openly for reuse. The *Crambus lathoniellus* genome sequencing initiative is part of the Darwin Tree of Life (DToL) project. All raw sequence data and the assembly have been deposited in INSDC databases. Raw data and assembly accession identifiers are reported in Table 1 and Table 2.

### Author information

Members of the University of Oxford and Wytham Woods Genome Acquisition Lab are listed here: https://doi.org/10.5281/zenodo.12157525.

Members of the Darwin Tree of Life Barcoding collective are listed here: https://doi.org/10.5281/zenodo.12158331

Members of the Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team are listed here: https://doi.org/10.5281/zenodo.12162482.

Members of Wellcome Sanger Institute Scientific Operations: Sequencing Operations are listed here: https://doi.org/10.5281/zenodo.12165051.

Members of the Wellcome Sanger Institute Tree of Life Core Informatics team are listed here: https://doi.org/10.5281/zenodo.12160324.

Members of the Tree of Life Core Informatics collective are listed here: https://doi.org/10.5281/zenodo.12205391.

Members of the Darwin Tree of Life Consortium are listed here: https://doi.org/10.5281/zenodo.4783558.

## References


Abdennur N, Mirny LA: **Cooler: scalable storage for Hi-C data and other genomically labeled arrays.** *Bioinformatics.* 2020; **36**(1): 311–316.
PubMed Abstract | Publisher Full Text | Free Full Text

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

Bates A, Clayton-Lucey I, Howard C: **Sanger Tree of Life HMW DNA fragmentation: diagenode megaruptor®3 for LI PacBio.** *protocols.io.* 2023.
Publisher Full Text

Beasley J, Uhl R, Forrest LL, *et al.*: **DNA barcoding SOPs for the Darwin Tree of Life project.** *protocols.io.* 2023.
Publisher Full Text

Brůna T, Hoff KJ, Lomsadze A, *et al.*: **BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database.** *NAR Genom Bioinform.* 2021; **3**(1): lqaa108.
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 [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2023; **8**: 123.
PubMed Abstract | Publisher Full Text | Free Full Text

da Veiga Leprevost F, Grüning BA, Alves Aflitos S, *et al.*: **BioContainers: an open-source and community-driven framework for software standardization.** *Bioinformatics.* 2017; **33**(16): 2580–2582.
PubMed Abstract | Publisher Full Text | Free Full Text

Denton A, Oatley G, Cornwell C, *et al.*: **Sanger Tree of Life sample homogenisation: PowerMash.** *protocols.io.* 2023a.
Publisher Full Text

Denton A, Yatsenko H, Jay J, *et al.*: **Sanger Tree of Life wet laboratory protocol collection V.1.** *protocols.io.* 2023b.
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

Diesh C, Stevens GJ, Xie P, *et al.*: **JBrowse 2: a modular genome browser with views of synteny and structural variation.** *Genome Biol.* 2023; **24**(1): 74.
PubMed Abstract | Publisher Full Text | Free Full Text

Ewels P, Magnusson M, Lundin S, *et al.*: **MultiQC: summarize analysis results for multiple tools and samples in a single report.** *Bioinformatics.* 2016; **32**(19): 3047–3048.
PubMed Abstract | Publisher Full Text | Free Full Text

Ewels PA, Peltzer A, Fillinger S, *et al.*: **The nf-core framework for community-curated bioinformatics pipelines.** *Nat Biotechnol.* 2020; **38**(3): 276–278.
PubMed Abstract | Publisher Full Text

Formenti G, Abueg L, Brajuka A, *et al.*: **Gfastats: conversion, evaluation and manipulation of genome sequences using assembly graphs.** *Bioinformatics.* 2022; **38**(17): 4214–4216.
PubMed Abstract | Publisher Full Text | Free Full Text

Grüning B, Dale R, Sjödin A, *et al.*: **Bioconda: sustainable and comprehensive software distribution for the life sciences.** *Nat Methods.* 2018; **15**(7): 475–476.
PubMed Abstract | Publisher Full Text | Free Full Text

Guan D, McCarthy SA, Wood J, *et al.*: **Identifying and removing haplotypic duplication in primary genome assemblies.** *Bioinformatics.* 2020; **36**(9): 2896–2898.
PubMed Abstract | Publisher Full Text | Free Full Text

Harry E: **PretextView (Paired REad TEXTure Viewer): a desktop application for viewing pretext contact maps**. 2022; [Accessed 19 October 2022].
Reference Source

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

Jay J, Yatsenko H, Narváez-Gómez JP, *et al.*: **Sanger Tree of Life sample preparation: triage and dissection.** *protocols.io.* 2023.
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

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

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

Oatley G, Denton A, Howard C: **Sanger Tree of Life HMW DNA extraction: automated MagAttract v.2.** *protocols.io.* 2023.
Publisher Full Text

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

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

Simão FA, Waterhouse RM, Ioannidis P, *et al.*: **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.** *Bioinformatics.* 2015; **31**(19): 3210–3212.
PubMed Abstract | Publisher Full Text

Strickland M, Cornwell C, Howard C: **Sanger Tree of Life fragmented DNA clean up: manual SPRI.** *protocols.io.* 2023.
Publisher Full Text

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

Surana P, Muffato M, Sadasivan Baby C: **sanger-tol/genomenote (v1.0.dev).** *Zenodo.* 2023b.
Publisher Full Text

Twyford AD, Beasley J, Barnes I, *et al.*: **A DNA barcoding framework for taxonomic verification in the Darwin Tree of Life project [version 1; peer review: awaiting peer review].** *Wellcome Open Res.* 2024; **9**: 339.
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

Wellcome Sanger Institute: **The genome sequence of the Hook-streak Grass-veneer moth, *Crambus lathoniellus* (Zincken, 1817).** European Nucleotide Archive. [dataset], accession number PRJEB57115, 2023.

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