DATA NOTE

REVISED The genome sequence of the Anthracite Bee-fly, *Anthrax anthrax* (Schrank, 1781)

[version 2; peer review: 2 approved]

Liam M. Crowley[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]University of Oxford, Oxford, England, UK

v2 **First published:** 04 Jun 2025, **10**:306
https://doi.org/10.12688/wellcomeopenres.24305.1
Latest published: 26 Sep 2025, **10**:306
https://doi.org/10.12688/wellcomeopenres.24305.2

Abstract

We present a genome assembly from a male specimen of *Anthrax anthrax* (Anthracite Bee-fly; Arthropoda; Insecta; Diptera; Bombyliidae). The genome sequence has a total length of 334.62 megabases. The assembly is scaffolded into 9 chromosomal pseudomolecules, including the X and Y sex chromosomes. The mitochondrial genome has also been assembled, with a length of 20.34 kilobases. Gene annotation of this assembly on Ensembl identified 10,570 protein-coding genes.

Keywords

Anthrax anthrax, Anthracite Bee-fly, genome sequence, chromosomal, Diptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ✓ ✓

	1	2
version 2 (revision) 26 Sep 2025		✓ view
version 1 04 Jun 2025	✓ view	? view

1. **Anthony Bayega**, McGill University, Montreal, Canada
Canada's Michael Smith Genome Sciences Centre (Ringgold ID: 113898), Vancouver, Canada

2. **Arun Arumugaperumal**, Rajalakshmi Engineering College, Thandalam 602105, Chennai, India

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: Crowley LM: 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 (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: Crowley LM, University of Oxford and Wytham Woods Genome Acquisition Lab, Darwin Tree of Life Barcoding collective *et al.* **The genome sequence of the Anthracite Bee-fly, *Anthrax anthrax* (Schrank, 1781) [version 2; peer review: 2 approved]** Wellcome Open Research 2025, **10**:306 https://doi.org/10.12688/wellcomeopenres.24305.2

First published: 04 Jun 2025, **10**:306 https://doi.org/10.12688/wellcomeopenres.24305.1

REVISED **Amendments from Version 1**

In version 2 of this data note we have addressed peer reviewers' comments by correcting "54 coverage" to "54× coverage" in the sequence data section.

We have added a few sentences in the section "Assembly quality metrics", describing the Earth Biogenome Project summary metric.

We have also replaced Figure 5, now including a megabase scale on the bottom axis.

Any further responses from the reviewers can be found at the end of the article

Species taxonomy

Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Protostomia; Ecdysozoa; Panarthropoda; Arthropoda; Mandibulata; Pancrustacea; Hexapoda; Insecta; Dicondylia; Pterygota; Neoptera; Endopterygota; Diptera; Brachycera; Muscomorpha; Asiloidea; Bombyliidae; Anthracinae; Anthracini; *Anthrax*; *Anthrax anthrax* (Schrank, 1781) (NCBI:txid2725500)

Background

The genome of the Anthracite Bee-fly, *Anthrax anthrax*, 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 a chromosome-level genome sequence for *Anthrax anthrax*, based on a specimen from Wytham, Woods, Oxfordshire, England, UK, (Figure 1).

Genome sequence report

Sequencing data

The genome of a specimen of *Anthrax anthrax* (Figure 1) was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating 16.46 Gb from 2.16 million reads, which were used to assemble the genome. GenomeScope analysis estimated the haploid genome size at 291.10 Mb, with a heterozygosity of 1.35% and repeat content of 40.20%. These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 54x coverage. Hi-C sequencing produced 129.25 Gb from 855.96 million reads, and was used to scaffold the assembly. Table 1 summarises the specimen and sequencing details.

Figure 1. Photograph of the *Anthrax anthrax* (idAntAnth1) specimen used for genome sequencing.

Assembly statistics

The primary haplotype was assembled, and contigs corresponding to an alternate haplotype were also deposited in INSDC databases. The assembly was improved by manual curation, which corrected 82 misjoins or missing joins and removed two haplotypic duplications. These interventions decreased the scaffold count by 15.55% and increased the scaffold N50 by 50.67%. The final assembly has a total length of 334.62 Mb in 363 scaffolds, with 114 gaps, and a scaffold N50 of 30.14 Mb (Table 2).

The snail plot in Figure 2 provides a summary of the assembly statistics, indicating the distribution of scaffold lengths and other assembly metrics. Figure 3 shows the distribution of scaffolds by GC proportion and coverage. Figure 4 presents a cumulative assembly plot, with separate curves representing different scaffold subsets assigned to various phyla, illustrating the completeness of the assembly.

Most of the assembly sequence (87.88%) was assigned to 9 chromosomal-level scaffolds, representing 7 autosomes and the X and Y sex chromosomes. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 5; Table 3). During curation, we noted that the order and orientation of scaffolds on Chromosome 1 in the regions of approximately 21.6–27.1 Mb, 36.9–46.0 Mb and 54.7–60.5 Mb are uncertain. Chromosomes X and Y were identified based on read coverage.

The mitochondrial genome was also assembled. This sequence is included as a contig in the multifasta file of the genome submission and as a standalone record.

Assembly quality metrics

The estimated Quality Value (QV) and *k*-mer completeness metrics, along with BUSCO completeness scores, were calculated for each haplotype and the combined assembly. The QV reflects the base-level accuracy of the assembly, while *k*-mer completeness indicates the proportion of expected *k*-mers identified in the assembly. BUSCO scores provide a measure of completeness based on benchmarking universal single-copy orthologues.

The combined primary and alternate assemblies achieve an estimated QV of 63.2. The *k*-mer recovery for the primary haplotype is 78.87%, and for the alternate haplotype 67.50%; the combined primary and alternate assemblies have a *k*-mer recovery of 99.31%. BUSCO v.5.5.0 analysis using the diptera_odb10 reference set (n = 3,285) identified 95.6% of the expected gene set (single = 94.8%, duplicated = 0.8%).

Table 1. **Specimen and sequencing data for *Anthrax anthrax*.**

Project information			
Study title	Anthrax anthrax (the anthracite bee-fly)		
Umbrella BioProject	PRJEB64720		
Species	*Anthrax anthrax*		
BioSpecimen	SAMEA112232691		
NCBI taxonomy ID	2725500		
Specimen information			
Technology	**ToLID**	**BioSample accession**	**Organism part**
PacBio long read sequencing	idAntAnth1	SAMEA112233170	thorax
Hi-C sequencing	idAntAnth1	SAMEA112233169	head
Sequencing information			
Platform	**Run accession**	**Read count**	**Base count (Gb)**
Hi-C Illumina NovaSeq 6000	ERR11814109	8.56e+08	129.25
PacBio Sequel IIe	ERR11809142	2.16e+06	16.46

Table 2. **Genome assembly data for *Anthrax anthrax*.**

Genome assembly		
Assembly name	idAntAnth1.1	
Assembly accession	GCA_963971155.1	
Alternate haplotype accession	*GCA_963971115.1*	
Assembly level for primary assembly	chromosome	
Span (Mb)	334.62	
Number of contigs	477	
Number of scaffolds	363	
Longest scaffold (Mb)	78.22	
Assembly metric	**Measure**	***Benchmark***
Contig N50 length	5.74 Mb	*≥ 1 Mb*
Scaffold N50 length	30.14 Mb	*= chromosome N50*
Consensus quality (QV)	Primary: 62.5; alternate: 63.8; combined: 63.2	*≥ 40*
k-mer completeness	Primary: 78.87%; alternate: 67.50%; combined: 99.31%	*≥ 95%*
BUSCO*	C:95.6%[S:94.8%,D:0.8%], F:0.9%,M:3.6%,n:3,285	*S > 90%; D < 5%*
Percentage of assembly assigned to chromosomes	87.88%	*≥ 90%*
Sex chromosomes	X and Y	*localised homologous pairs*
Organelles	Mitochondrial genome: 20.34 kb	*complete single alleles*

* BUSCO scores based on the diptera_odb10 BUSCO set using version 5.5.0. C = complete [S = single copy, D = duplicated], F = fragmented, M = missing, n = number of orthologues in comparison.

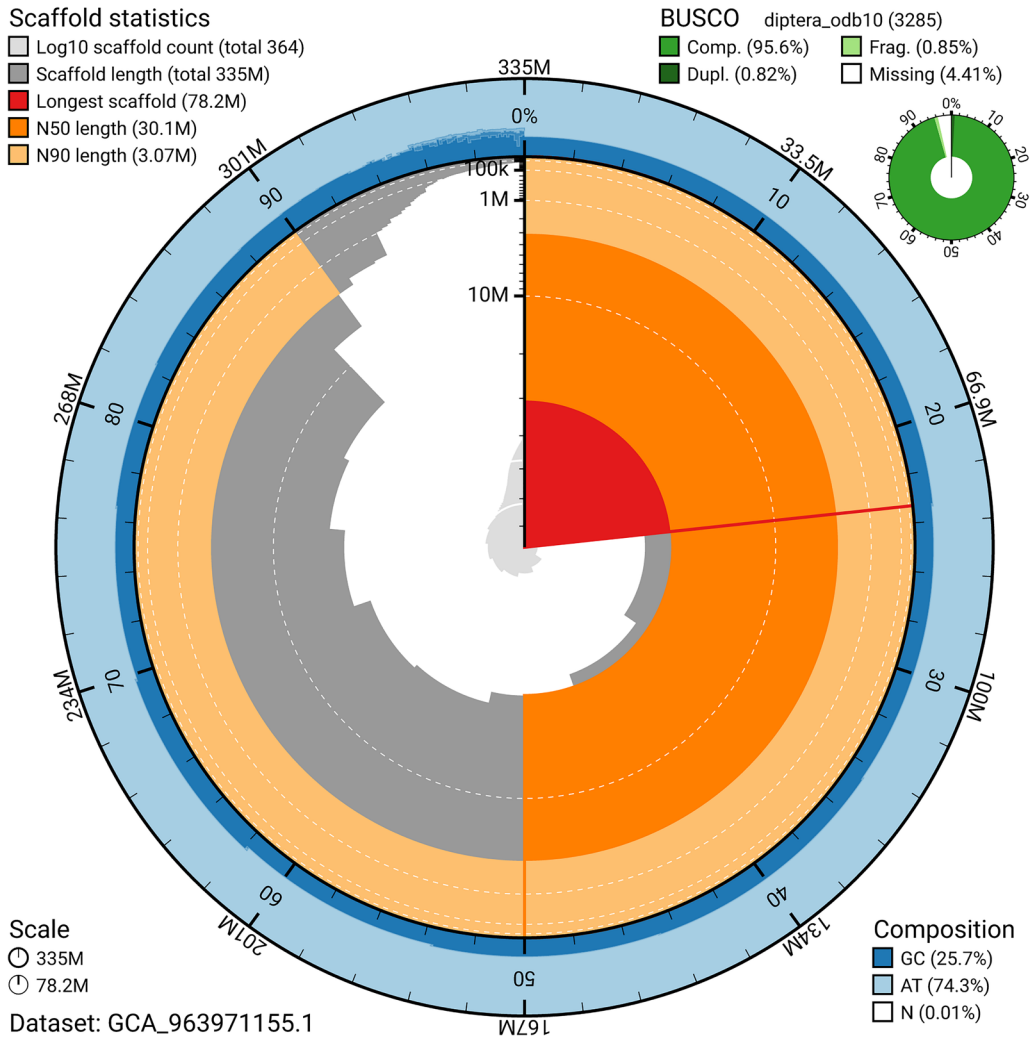

Figure 2. Genome assembly of *Anthrax anthrax*, idAntAnth1.1: metrics. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the circumference. 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 diptera_odb10 set is presented at the top right. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/GCA_963971155.1/dataset/GCA_963971155.1/snail.

Table 2 provides assembly metric benchmarks adapted from Rhie *et al.* (2021) and the Earth BioGenome Project Report on Assembly Standards September 2024. The assembly achieves the EBP reference standard of **6.7.Q62**. As less than 90% of the genome sequence was assembled onto chromosomes, with the remainder as unplaced scaffolds, the assembly does not meet the EBP chromosomal-scale ('C') criterion for the scaffold component (≥90% placed). The unplaced fraction includes highly repetitive sequence that cannot be confidently placed onto chromosomes.

Genome annotation report

The *Anthrax anthrax* genome assembly (GCA_963971155.1) was annotated externally by Ensembl at the European Bioinformatics Institute (EBI), using the Ensembl non-vertebrate genome annotation system. This annotation includes 15,982 transcribed mRNAs from 10,570 protein-coding and 589 non-coding genes. The average transcript length is 7,275.59 bp, with an average of 1.43 coding transcripts per gene and 5.85 exons per transcript. For further information about the annotation, please refer to https://beta.ensembl.org/species/0c17814a-08c5-469a-8cb3-e4ec17097691.

Methods

Sample acquisition and DNA barcoding

The specimen used for genome sequencing was an adult male *Anthrax anthrax* (specimen ID Ox002488, ToLID idAntAnth1), collected from Wytham Woods, Oxfordshire, United

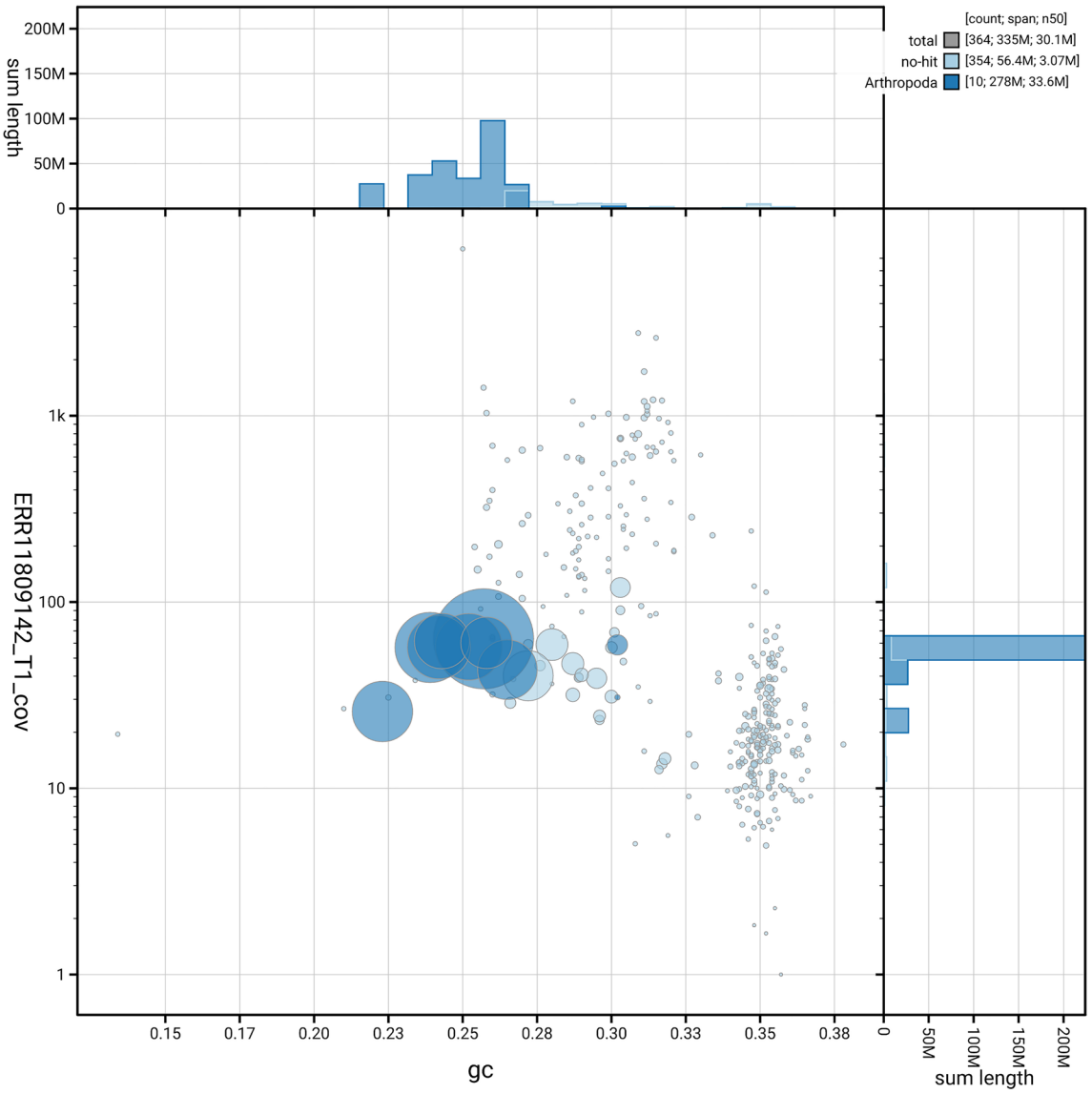

Figure 3. Genome assembly of *Anthrax anthrax*, idAntAnth1.1: BlobToolKit GC-coverage plot. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. 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 at https://blobtoolkit.genomehubs.org/view/GCA_963971155.1/dataset/GCA_963971155.1/blob.

Kingdom (latitude 51.783, longitude –1.317) on 2022-08-08 by potting. The specimen was collected and identified by Liam Crowley (University of Oxford) and preserved on dry ice.

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) (Pereira *et al.*, 2022). 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).

Metadata collection for samples adhered to the Darwin Tree of Life project standards described by Lawniczak *et al.* (2022).

Nucleic acid extraction

The workflow for high molecular weight (HMW) DNA extraction at the Wellcome Sanger Institute (WSI) Tree of Life

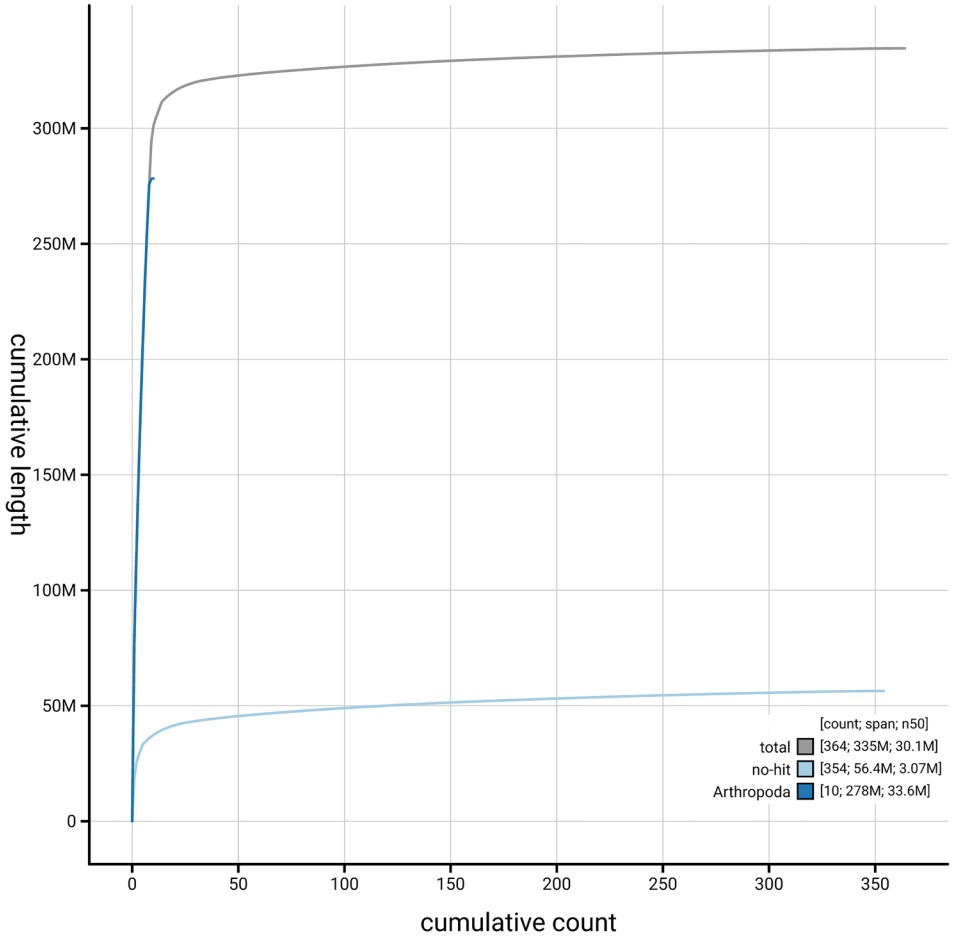

Figure 4. Genome assembly of *Anthrax anthrax,* idAntAnth1.1: BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/GCA_963971155.1/dataset/GCA_963971155.1/cumulative.

Core Laboratory includes a sequence of procedures: sample preparation and homogenisation, DNA extraction, fragmentation and purification. Detailed protocols are available on protocols.io (Howard *et al.*, 2025). The idAntAnth1 sample was prepared for DNA extraction by weighing and dissecting it on dry ice (Jay *et al.*, 2023). Tissue from the thorax was homogenised using a PowerMasher II tissue disruptor (Denton *et al.*, 2023). HMW DNA was extracted in 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, using AMPure PB beads to eliminate shorter fragments and concentrate the DNA (Strickland *et al.*, 2023). 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.

Hi-C sample preparation and crosslinking

Hi-C data were generated from the head of the idAntAnth1 sample using the Arima-HiC v2 kit (Arima Genomics) with 20–50 mg of frozen tissue (stored at –80 °C). As per manufacturer's instructions, tissue was fixed, and the DNA crosslinked using a TC buffer with 22% formaldehyde concentration, and a final formaldehyde concentration of 2%. The tissue was then homogenised using the Diagnocine Power Masher-II. The crosslinked DNA was digested using a restriction enzyme master mix, then biotinylated and ligated. A clean up was performed with SPRIselect beads prior to library preparation. DNA concentration was quantified using the Qubit Fluorometer v4.0 (Thermo Fisher Scientific) and Qubit HS Assay Kit, and sample biotinylation percentage was estimated using the Arima-HiC v2 QC beads.

Library preparation and sequencing

Library preparation and sequencing were performed at the WSI Scientific Operations core.

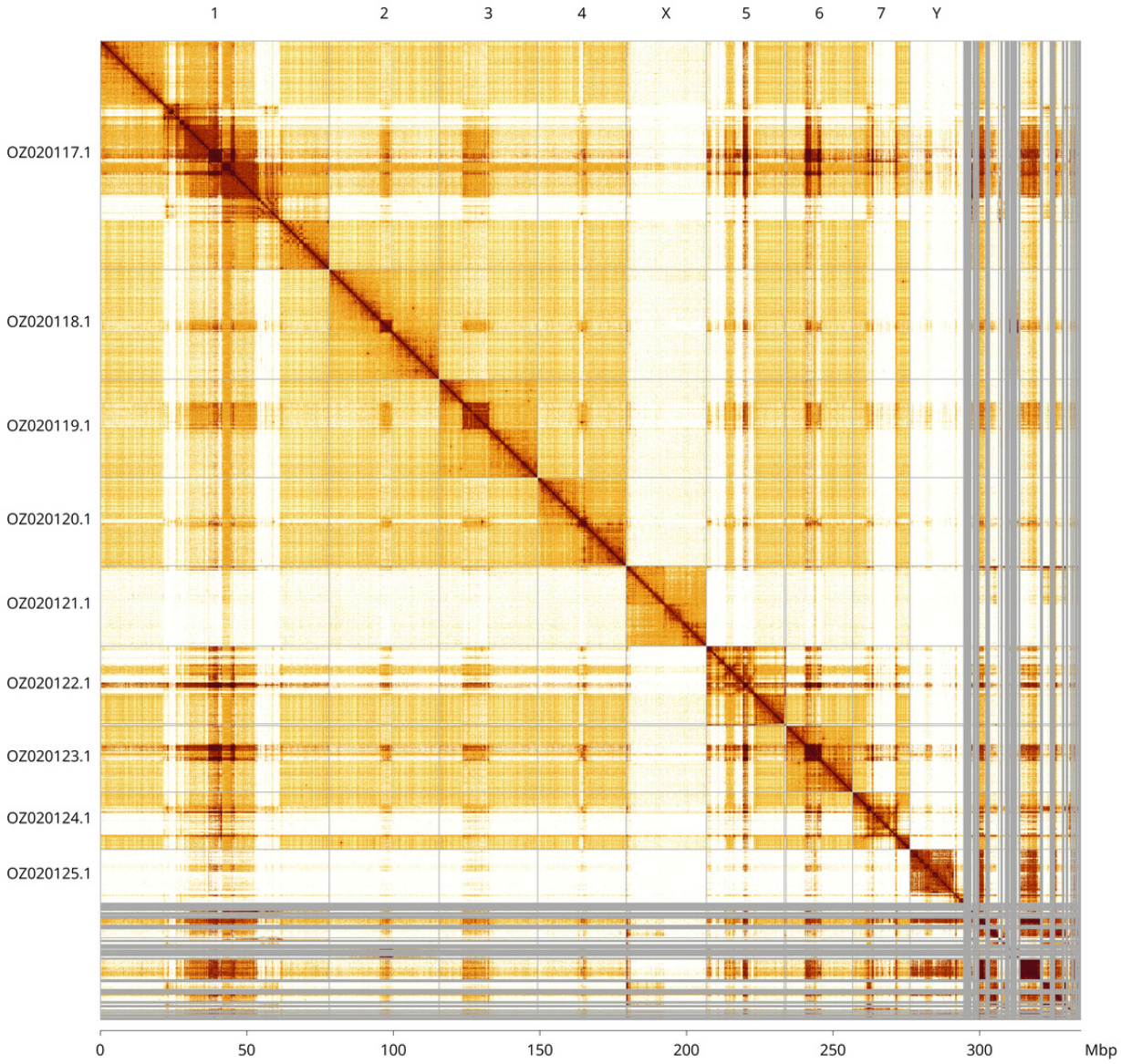

Figure 5. Genome assembly of idAntAnth1.1: Hi-C contact map generated using PretextSnapshot. Chromosomes are shown in order of size and labelled with chromosome numbers (top) and chromosome accession numbers (left). A megabase scale is shown on the bottom axis.

Table 3. Chromosomal pseudomolecules in the genome assembly of *Anthrax anthrax*, idAntAnth1.

INSDC accession	Name	Length (Mb)	GC%
OZ020117.1	1	78.22	25.5
OZ020118.1	2	37.45	24
OZ020119.1	3	33.6	25
OZ020120.1	4	30.14	24
OZ020122.1	5	26.61	26.5
OZ020123.1	6	22.7	24.5
OZ020124.1	7	19.54	26
OZ020121.1	X	27.42	22.5
OZ020125.1	Y	18.39	27
OZ020126.1	MT	0.02	25

PacBio HiFi

At a minimum, samples were required to have an average fragment size exceeding 8 kb and a total mass over 400 ng to proceed to the low-input SMRTbell Prep Kit 3.0 protocol (Pacific Biosciences), depending on genome size and sequencing depth required. Libraries were prepared using the SMRTbell Prep Kit 3.0 as per the manufacturer's instructions. The kit includes the reagents required for end repair/A-tailing, adapter ligation, post-ligation SMRTbell bead cleanup, and nuclease treatment. Size-selection and clean-up were carried out using diluted AMPure PB beads (Pacific Biosciences). DNA concentration was quantified using the Qubit Fluorometer v4.0 (ThermoFisher Scientific) with Qubit 1X dsDNA HS assay kit and the final library fragment size analysis was carried out using the Agilent Femto Pulse Automated Pulsed Field CE Instrument (Agilent Technologies) and the gDNA 55kb BAC analysis kit.

Samples were sequenced using the Sequel IIe system (Pacific Biosciences, California, USA). The concentration of the library loaded onto the Sequel IIe was 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, as well as perform primary and secondary analysis of the data upon completion.

Hi-C

For Hi-C library preparation, the biotinylated DNA constructs were fragmented using a Covaris E220 sonicator and size-selected to 400–600 bp using SPRISelect beads. DNA was then enriched using Arima-HiC v2 Enrichment beads. The NEBNext Ultra II DNA Library Prep Kit (New England Biolabs) was used for end repair, A-tailing, and adapter ligation, following a modified protocol in which library preparation is carried out while the DNA remains bound to the enrichment beads. PCR amplification was performed using KAPA HiFi HotStart mix and custom dual-indexed adapters (Integrated DNA Technologies) in a 96-well plate format. Depending on sample concentration and biotinylation percentage determined at the crosslinking stage, samples were amplified for 10–16 PCR cycles. Post-PCR clean-up was carried out using SPRISelect beads. The libraries were quantified using the Accuclear Ultra High Sensitivity dsDNA Standards Assay kit (Biotium) and normalised to 10 ng/μL before sequencing. Hi-C sequencing was performed on the Illumina NovaSeq 6000 instrument.

Genome assembly, curation and evaluation

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 first assembled using Hifiasm (Cheng *et al.*, 2021) with the --primary option. Haplotypic duplications were identified and removed using purge_dups (Guan *et al.*, 2020). 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 using 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.

Assembly curation

The assembly was decontaminated using the Assembly Screen for Cobionts and Contaminants (ASCC) pipeline. Flat files and maps used in curation were generated via the TreeVal pipeline (Pointon *et al.*, 2023). Manual curation was conducted primarily in PretextView (Harry, 2022) and HiGlass (Kerpedjiev *et al.*, 2018), with additional insights provided by JBrowse2 (Diesh *et al.*, 2023). Scaffolds were visually inspected and corrected as described by Howe *et al.* (2021). Any identified contamination, missed joins, and mis-joins were amended, and duplicate sequences were tagged and removed. Sex chromosomes were identified by read coverage analysis. The curation process is documented at https://gitlab.com/wtsi-grit/rapid-curation.

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 the primary and alternate haplotypes using the *k*-mer databases (k = 31) computed prior to genome assembly. The analysis outputs included assembly QV scores and completeness statistics.

The genome was analysed in the blobtoolkit pipeline, a Nextflow (Di Tommaso *et al.*, 2017) port of the previous Snakemake Blobtoolkit pipeline (Challis *et al.*, 2020). It aligns the PacBio reads in SAMtools (Danecek *et al.*, 2021) and minimap2 (Li, 2018) and generates coverage tracks for regions of fixed size. In parallel, it queries the GoaT database (Challis *et al.*, 2023) to identify all matching BUSCO lineages to run BUSCO (Manni *et al.*, 2021). For the three domain-level BUSCO lineages, the pipeline aligns the BUSCO genes to the UniProt Reference Proteomes database (Bateman *et al.*, 2023) with DIAMOND blastp (Buchfink *et al.*, 2021). The genome is also divided into chunks according to the density of the BUSCO genes from the closest taxonomic lineage, and each chunk is aligned to the UniProt Reference Proteomes database using DIAMOND blastx. Genome sequences without a hit are chunked using seqtk and aligned to the NT database with blastn (Altschul *et al.*, 1990). The blobtools suite combines all these 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), relying on the Conda package manager, the Bioconda initiative (Grüning *et al.*, 2018), the Biocontainers infrastructure

(da Veiga Leprevost *et al.*, 2017), as well as the Docker (Merkel, 2014) and Singularity (Kurtzer *et al.*, 2017) containerisation solutions.

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

Wellcome Sanger Institute – Legal and Governance

The materials that have contributed to this genome note have been supplied by a Darwin Tree of Life Partner. The submission of materials by a Darwin Tree of Life Partner is subject to the **'Darwin Tree of Life Project Sampling Code of 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

Table 4. Software tools: versions and sources.

Software tool	Version	Source
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
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
Gfastats	1.3.6	https://github.com/vgl-hub/gfastats
GoaT CLI	0.2.5	https://github.com/genomehubs/goat-cli
Hifiasm	0.16.1-r375	https://github.com/chhylp123/hifiasm
HiGlass	1.13.4	https://github.com/higlass/higlass
MerquryFK	1.1	https://github.com/thegenemyers/MERQURY.FK
Minimap2	2.24-r1122	https://github.com/lh3/minimap2
MitoHiFi	2	https://github.com/marcelauliano/MitoHiFi
MultiQC	1.14, 1.17, and 1.18	https://github.com/MultiQC/MultiQC
Nextflow	23.04.1	https://github.com/nextflow-io/nextflow
PretextView	0.2.5	https://github.com/sanger-tol/PretextView
purge_dups	1.2.3	https://github.com/dfguan/purge_dups
samtools	1.19.2	https://github.com/samtools/samtools
sanger-tol/ascc	0.1.0	https://github.com/sanger-tol/ascc
sanger-tol/blobtoolkit	0.4.0	https://github.com/sanger-tol/blobtoolkit
Seqtk	1.3	https://github.com/lh3/seqtk
Singularity	3.9.0	https://github.com/sylabs/singularity
TreeVal	1.2.0	https://github.com/sanger-tol/treeval
YaHS	1.1a.2	https://github.com/c-zhou/yahs

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: Anthrax anthrax (the anthracite bee-fly). Accession number PRJEB64720; https://identifiers.org/ena.embl/PRJEB64720. The genome sequence is released openly for reuse. The *Anthrax anthrax* genome sequencing initiative is part of the Darwin Tree of Life Project (PRJEB40665) and Sanger Institute Tree of Life Programme (PRJEB43745). 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

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

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

Beasley J, Uhl R, Forrest LL, *et al.*: **DNA barcoding SOPs for the Darwin Tree of Life project.** *protocols.io.* 2023; [Accessed 25 June 2024].
Publisher Full Text

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, Kumar S, Sotero-Caio C, *et al.*: **Genomes on a Tree (GoaT): a versatile, scalable search engine for genomic and sequencing project metadata across the eukaryotic Tree of Life [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2023; **8**: 24.
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

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

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

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

Lawniczak MKN, Pereira-da-Conceicoa LL, Blaxter ML, *et al.*: **Specimen and sample metadata standards for biodiversity genomics: a proposal from the Darwin Tree of Life project [version 1; peer review: 2 approved with reservations].** *Wellcome Open Res.* 2022; **7**: 187.
Publisher 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): 2, [Accessed 2 April 2024].
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

Pereira L, Sivell O, Sivess L, *et al.*: **DToL taxon-specific standard operating procedure for the terrestrial and freshwater arthropods working group.** 2022.
Publisher Full Text

Pointon DL, Eagles W, Sims Y, *et al.*: **sanger-tol/treeval v1.0.0 – Ancient Atlantis.** 2023.
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): 245.
PubMed Abstract | Publisher Full Text | Free Full Text

Strickland M, Cornwell C, Howard C: **Sanger Tree of Life fragmented DNA clean up: manual SPRI.** *protocols.io.* 2023.
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: 2 approved].** *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.** In: *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS).* IEEE, 2019; 314–324.
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 2

Reviewer Report 01 October 2025

https://doi.org/10.21956/wellcomeopenres.27529.r134445

✔ **Arun Arumugaperumal**
Department of Biotechnology, Rajalakshmi Engineering College, Thandalam 602105, Chennai, Tamil Nadu, India

The authors have carried out the correction. The article can be indexed.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics; Genomics

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

Version 1

Reviewer Report 18 September 2025

https://doi.org/10.21956/wellcomeopenres.26805.r130912

? **Arun Arumugaperumal**
Department of Biotechnology, Rajalakshmi Engineering College, Thandalam 602105, Chennai, Tamil Nadu, India

The data note describes the genome sequence of an Anthracite Bee-fly, *Anthrax anthrax*. The photograph of the specimen used for sequencing is shown in Figure 1. The author has used

advanced sequencing technologies to arrive at the chromosome-level genome assembly. The assembly size reported is 334.62 Mb scaffolded into 9 chromosomes, including the XY sex chromosomes. The mitochondrial genomic DNA of size 20.34 kb has also been provided separately. A total of 10,570 protein-coding genes have been identified. These values go hand in hand with those of a similar fly, *Bombylius major*, which is from the same family [1]. The contig and scaffold N50 values were above the threshold values. BUSCO analysis has also revealed that the genome is a near-complete one as evident from the 95.6% completeness. The GC% plot shows more than one distinct blob. As the colour difference of the blob is not deviating much, it can be taken as the absence of any contaminants. The links and accession numbers for raw data and annotation have been provided. The article can be indexed after amending the minor corrections.

Minor corrections are:
1. Change '54 coverage' to '54X coverage' in Genome Sequence Report
2. The quality metrics according to the Earth Biogenome Project guidelines can be mentioned.

References

1. Lawniczak M, Crowley L, McAlister E, University of Oxford and Wytham Woods Genome Acquisition Lab, et al.: The genome sequence of the dark-edged bee fly, Bombylius major (Linnaeus, 1758). *Wellcome Open Research*. 2023; **8**. Publisher Full Text

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: Bioinformatics; Genomics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 18 Sep 2025

Tree of Life Team Sanger

Thank you for reviewing this data note. We have submitted a revised version addressing your comments. 1) Change “54 coverage” to “54× coverage” in the Genome sequence report. Corrected. [If the journal prefers ‘X’, we have used “54X”.] 2) The quality metrics

according to the Earth BioGenome Project guidelines can be mentioned. These metrics were already included in version 1 (see “Assembly quality metrics”). We have added brief interpretation of the values.

Competing Interests: No competing interests were disclosed.

Reviewer Report 25 June 2025

https://doi.org/10.21956/wellcomeopenres.26805.r125240

Anthony Bayega
[1] McGill University, Montreal, Canada
[2] Canada's Michael Smith Genome Sciences Centre (Ringgold ID: 113898), Vancouver, British Columbia, Canada

Crowley al. provides a high quality chromosomal-level genome assembly of *Anthrax anthrax*, a commendable effort. My comments follow below:

1. In Figure 3, the label on the Y-axis which reads “ERR1...” could be changed to something more informative to help the reader make sense of this figure.
2. In the “Genome annotation report” section, the authors report an average transcript length of 7,275.59 bp. It would be helpful to the reader to know if this is the spliced or unspliced length.
3. In this particular effort the group has not attempted to decontaminate the assembly of cobionts and other contaminants. This could be highlighted or even explained.

Overall, the authors provide a good-quality genome and also assign 87.88% of it to chromosomes. Although much work remains to order the scaffolds, fully phase the contigs and scaffolds, and complete the gaps and also structurally and functionally annotate the genome, the current work will indeed be valuable to the whole community. I therefore recommend the indexing of this genome so that this resource becomes widely accessible to the scientific community.

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: Human genetics and genomics

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