DATA NOTE

The genome sequence of the chocolate mining bee, *Andrena scotica* Perkins, 1916 (Hymenoptera: Andrenidae)

[version 1; peer review: 2 approved]

Liam M. Crowley[1], Steven Falk[2], Jeyaraney Kathirithamby[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
[2]Independent researcher, Kenilworth, England, UK

v1 **First published:** 18 Mar 2026, **11**:170
https://doi.org/10.12688/wellcomeopenres.26130.1
Latest published: 18 Mar 2026, **11**:170
https://doi.org/10.12688/wellcomeopenres.26130.1

Abstract

We present a haploid genome assembly from an individual male *Andrena scotica* (chocolate mining bee; Arthropoda; Insecta; Hymenoptera; Andrenidae). The genome sequence has a total length of 446.96 megabases. Most of the assembly (80.96%) is scaffolded into 5 chromosomal pseudomolecules. The mitochondrial genome has also been assembled, with a length of 19.68 kilobases. This assembly was generated as part of the Darwin Tree of Life project, which produces reference genomes for eukaryotic species found in Britain and Ireland.

Keywords

Andrena scotica; chocolate mining bee; genome sequence; chromosomal; Hymenoptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ✓ ✓

	1	2
version 1 18 Mar 2026	✓ view	✓ view

1. **Eduardo Luís Menezes de Almeida**, Universidade Federal de Viçosa, Minas Gerais, Brazil

2. **Rodolpho S. T. Menezes**, State University of Santa Cruz, Ilhéus, Brazil
Universidade Estadual de Santa Cruz (Ringgold ID: 74361), Ilhéus, Brazil

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; **Falk S**: Investigation, Resources; **Kathirithamby J**: 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: Crowley LM, Falk S, Kathirithamby J *et al.* **The genome sequence of the chocolate mining bee, *Andrena scotica* Perkins, 1916 (Hymenoptera: Andrenidae) [version 1; peer review: 2 approved]** Wellcome Open Research 2026, **11**:170 https://doi.org/10.12688/wellcomeopenres.26130.1

First published: 18 Mar 2026, **11**:170 https://doi.org/10.12688/wellcomeopenres.26130.1

Species taxonomy

Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Protostomia; Ecdysozoa; Panarthropoda; Arthropoda; Mandibulata; Pancrustacea; Hexapoda; Insecta; Dicondylia; Pterygota; Neoptera; Endopterygota; Hymenoptera; Apocrita; Aculeata; Apoidea; Anthophila; Andrenidae; Andreninae; *Andrena*; *Hoplandrena*; *Andrena scotica* Perkins, 1916 (NCBI:txid2878414).

Background

Species of *Andrena* (mining bees) occur across much of the world (Michener, 2007), including the UK. Many *Andrena* species are solitary or weakly communal ground nesters; where communal nesting occurs, females may share a nest entrance but provision their brood cells independently (Paxton *et al.*, 1996). Females collect pollen and nectar to provision brood cells, and offspring develop within these cells.

The offspring complete their development in the brood cells in which they overwinter before emerging in the following spring as adults (Paxton *et al.*, 1996). Males emerge first and search for females to mate; after mating, males die. Mated females construct brood cells and lay one egg per cell, often totalling around five eggs per female. After hatching, larvae are fed on pollen and nectar provisioned by the female. Adults live for about 6–8 weeks (NatureScot, 2020). Stylopised male and female *Andrena* have been reported to emerge at the same time as unstylopised males, which has been suggested to reflect parasite-mediated changes in host behaviour (Hoffmann *et al.*, 2023; Saunders, 1850; Straka *et al.*, 2011).

Some *Andrena* species are hosts to Strepsiptera, including *Stylops* spp. In these associations, first-instar larvae (planidia) can be transported on flowers and transferred to foraging bees (phoresy) (Kathirithamby, 2025; Lähteenaro *et al.*, 2024). After transfer, planidia enter the host's brood cell and develop as endoparasites of the host egg or larva (Kathirithamby, 2025). Following host emergence, the strepsipteran male emerges as a free-living adult, whereas the female remains endoparasitic within the host (Kathirithamby, 2025). Stylopisation is associated with changes in host morphology and behaviour in some systems, and may affect seasonal timing of host activity, including early emergence in spring (Hoffmann *et al.*, 2023; Saunders, 1850; Straka *et al.*, 2011).

Andrena are short-tongued bees and can be recognised by morphological characters including grooves below the antennal sockets (Wilson & Carril, 2016). *Andrena* are difficult to identify to species level, and stylopisation can further complicate host identification because it can alter external morphology. *Andrena scotica* (chocolate mining bee) is among the larger UK *Andrena* species, with a dark brown abdomen and contrasting dark upper and pale lower hairs on the hind-leg scopa. In the UK, adults are typically active from March to June.

We present a chromosome-level genome sequence for *Andrena scotica* from a stylopised specimen collected from Wytham Woods, Oxfordshire, UK. A *Stylops aterrimus* individual dissected from the same host was sequenced separately, and is described in a separate data note (Kathirithamby *et al.*, 2025).

Methods

Sample acquisition and DNA barcoding

The specimen used for genome sequencing was an adult male *Andrena scotica* (specimen ID Ox001251, ToLID iyAndCara2; Figure 1), collected from Wytham Woods, Oxfordshire, UK (latitude 51.786, longitude −1.317) on 2021-04-19. The specimen was collected by Liam Crowley and identified by Steven Falk. The same specimen was used for RNA sequencing.

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.

Nucleic acid extraction

Protocols for high molecular weight (HMW) DNA extraction developed at the Wellcome Sanger Institute (WSI) Tree of Life Core Laboratory are available on protocols.io (Howard *et al.*, 2025). The iyAndCara2 sample was weighed and triaged to determine the appropriate extraction protocol. Tissue from the thorax was homogenised by powermashing

Figure 1. Photograph of the *Andrena scotica* specimen used for genome sequencing.

using a PowerMasher II tissue disruptor. 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 2.92 ng/μL and a yield of 1 051.20 ng.

RNA was extracted from abdomen tissue of iyAndCara2 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 head tissue from the iyAndCara2 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 6000.

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 6000, 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.

The HiFi reads were assembled using Hifiasm (Cheng *et al.*, 2021) with the --primary and -l0 keys. 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 MERQURY.FK (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 one break and 27 joins. This reduced the scaffold count by 3.6% and increased the scaffold N50 by 161.9%. The curation process is described at https://gitlab.com/wtsi-grit/rapid-curation. PretextSnapshot was used to generate a Hi-C contact map of the final assembly.

Assembly quality assessment

The Merqury.FK tool (Rhie *et al.*, 2020) 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

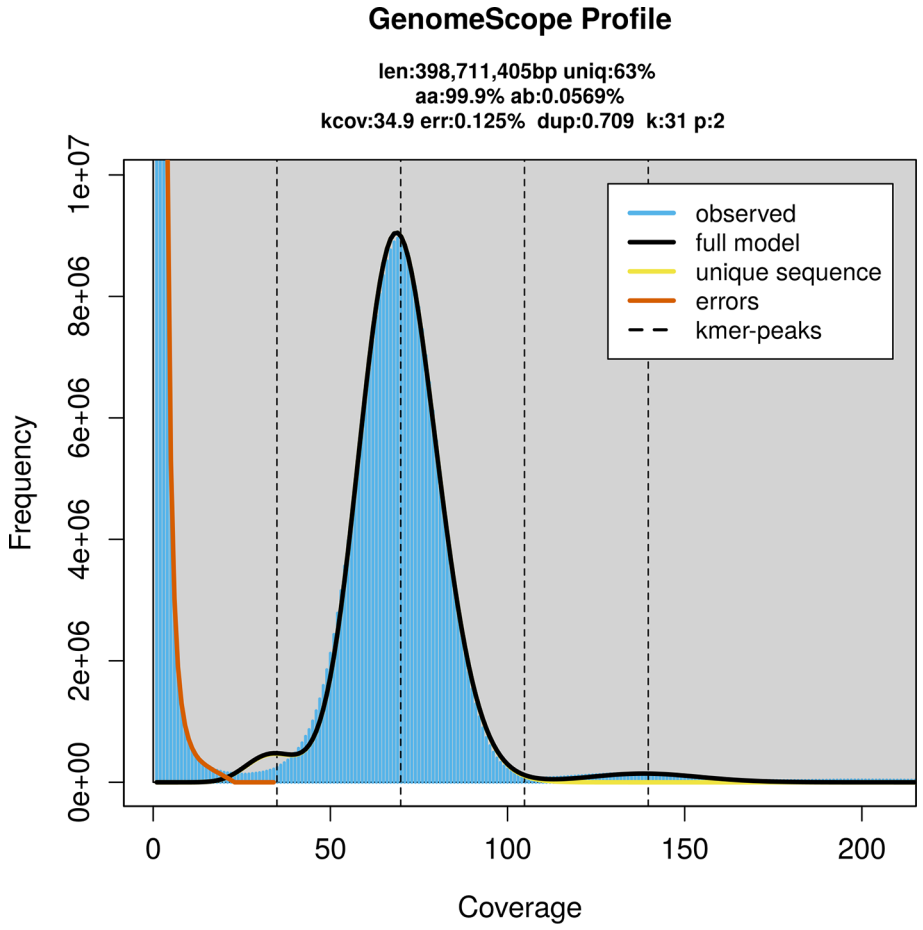

Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2. The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

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 *Andrena scotica* specimen generated 29.03 Gb (gigabases) from 2.49 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 399.00 Mb, with a heterozygosity of 0.06% and repeat content of 37.01% (Figure 2). These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 70× coverage. Hi-C sequencing produced 90.90 Gb from 601.99 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.

Table 1. Specimen and sequencing data for BioProject PRJEB58246.

Platform	PacBio HiFi	Hi-C	RNA-seq
ToLID	iyAndCara2	iyAndCara2	iyAndCara2
Specimen ID	Ox001251	Ox001251	Ox001251
BioSample (source individual)	SAMEA10166736	SAMEA10166736	SAMEA10166736
BioSample (tissue)	SAMEA10200944	SAMEA10200943	SAMEA10200945
Tissue	thorax	head	abdomen
Instrument	Sequel IIe	Illumina NovaSeq 6000	Illumina NovaSeq 6000
Run accessions	ERR10677852	ERR10684079	ERR11242514
Read count total	2.49 million	601.99 million	82.00 million
Base count total	29.03 Gb	90.90 Gb	12.38 Gb

Table 2. Genome assembly statistics.

Assembly name	iyAndCara2.1
Assembly accession	GCA_952773225.1
Assembly level	chromosome
Span (Mb)	446.96
Number of chromosomes	5
Number of contigs	769
Contig N50	5.67 Mb
Number of scaffolds	668
Scaffold N50	86.76 Mb
Organelles	Mitochondrion: 19.68 kb

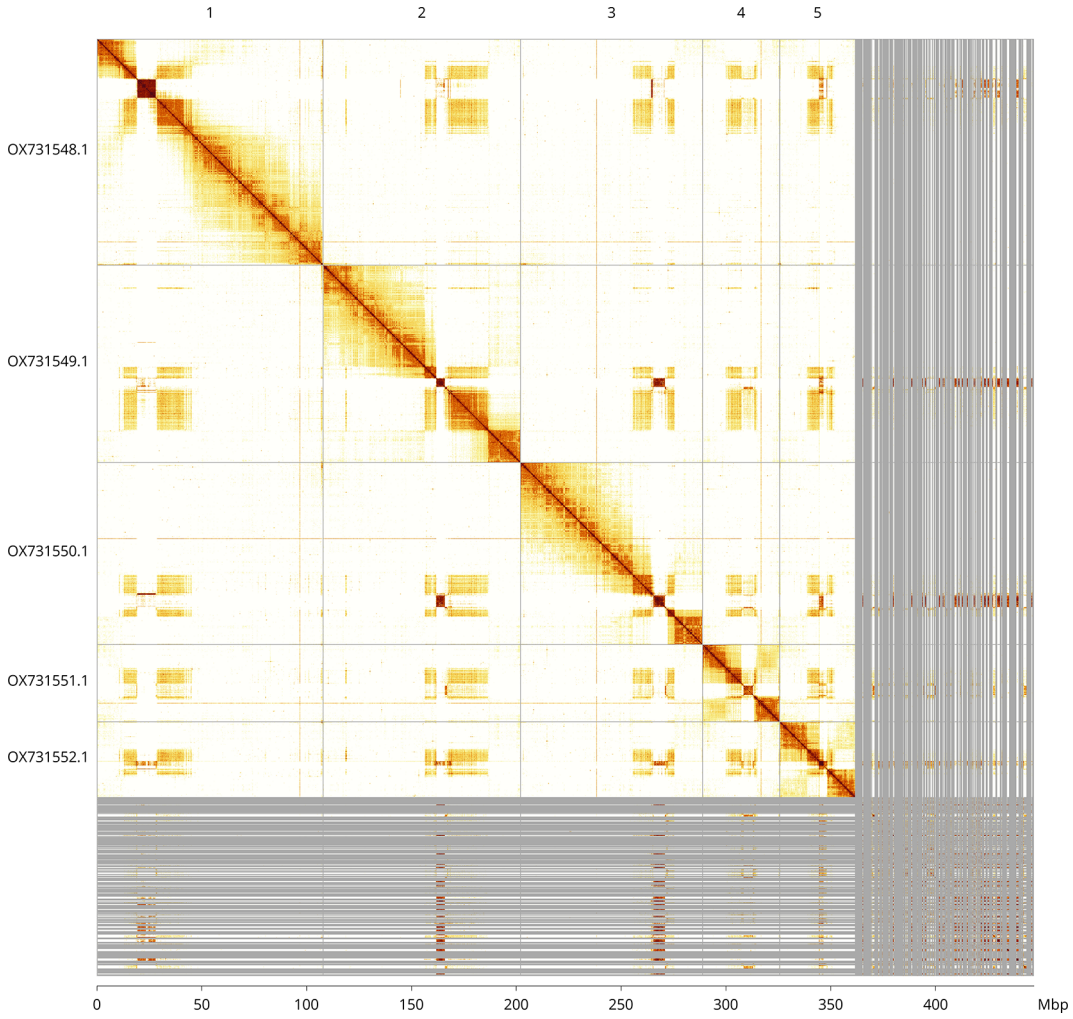

Figure 3. Hi-C contact map of the *Andrena scotica* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes, with a megabase scale shown below. The plot was generated using PretextSnapshot.

Assembly statistics

This is a haploid assembly from a male hymenopteran. The final assembly has a total length of 446.96 Mb in 668 scaffolds, with a scaffold N50 of 86.76 Mb (Table 2).

Most of the assembly sequence (80.96%) was assigned to 5 chromosomal-level scaffolds. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 3; Table 3).

The mitochondrial genome was also assembled (length 19.68 kb, OX731553.1). This sequence is included as a contig in the multifasta file of the genome submission and as a standalone record.

Assembly quality metrics

The *k*-mer completeness is 98.90% for the haploid assembly (Figure 4). BUSCO v. 5.3.2 analysis using the reference set (n = 5 991) identified 96.9% of the expected gene set (single = 96.7%, duplicated = 0.3%). 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.

Table 3. Chromosomal pseudomolecules in the primary genome assembly of *Andrena scotica* iyAndCara2.

INSDC accession	Molecule	Length (Mb)	GC%
OX731548.1	1	107.87	42.50
OX731549.1	2	94.24	42
OX731550.1	3	86.76	42
OX731551.1	4	36.79	39
OX731552.1	5	36.19	42.50

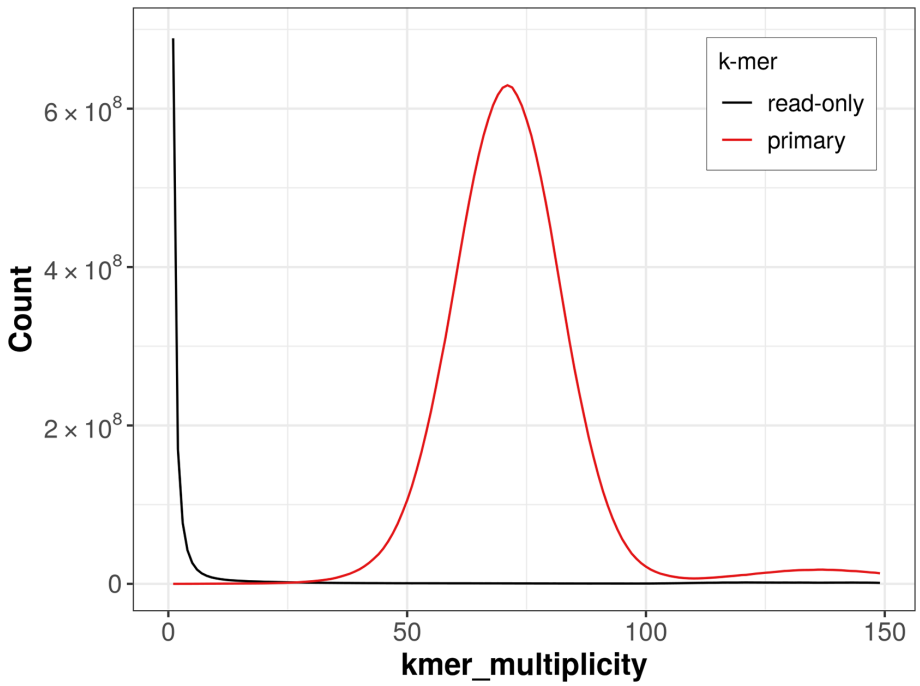

Figure 4. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers 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.

Table 4 lists the assembly metric benchmarks adapted from Rhie *et al.* (2021) and the Earth BioGenome Project Report on Assembly Standards September 2024. The EBP metric, calculated for the primary assembly, is **6.7.Q66**.

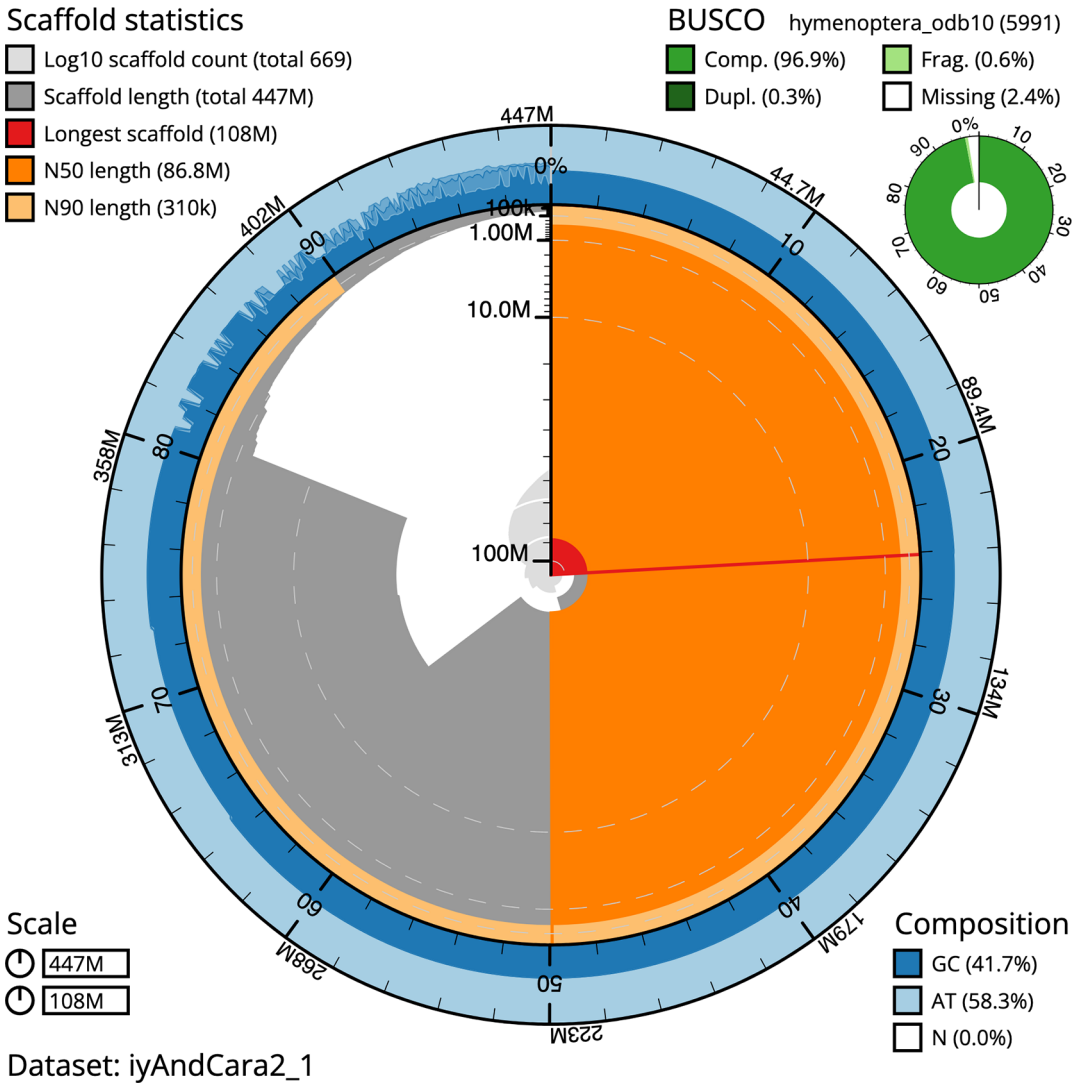

Figure 5. Assembly metrics for iyAndCara2.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1 000 bins around the circumference. The outermost blue tracks display the distribution of GC, AT, and N percentages across the bins. Scaffolds are arranged clockwise from longest to shortest and are depicted in dark grey. The longest scaffold is indicated by the red arc, and the deeper orange and pale orange arcs represent the N50 and N90 lengths. A light grey spiral at the centre shows the cumulative scaffold count on a logarithmic scale. A summary of complete, fragmented, duplicated, and missing BUSCO genes in the set is presented at the top right. An interactive version of this figure can be accessed on the BlobToolKit viewer.

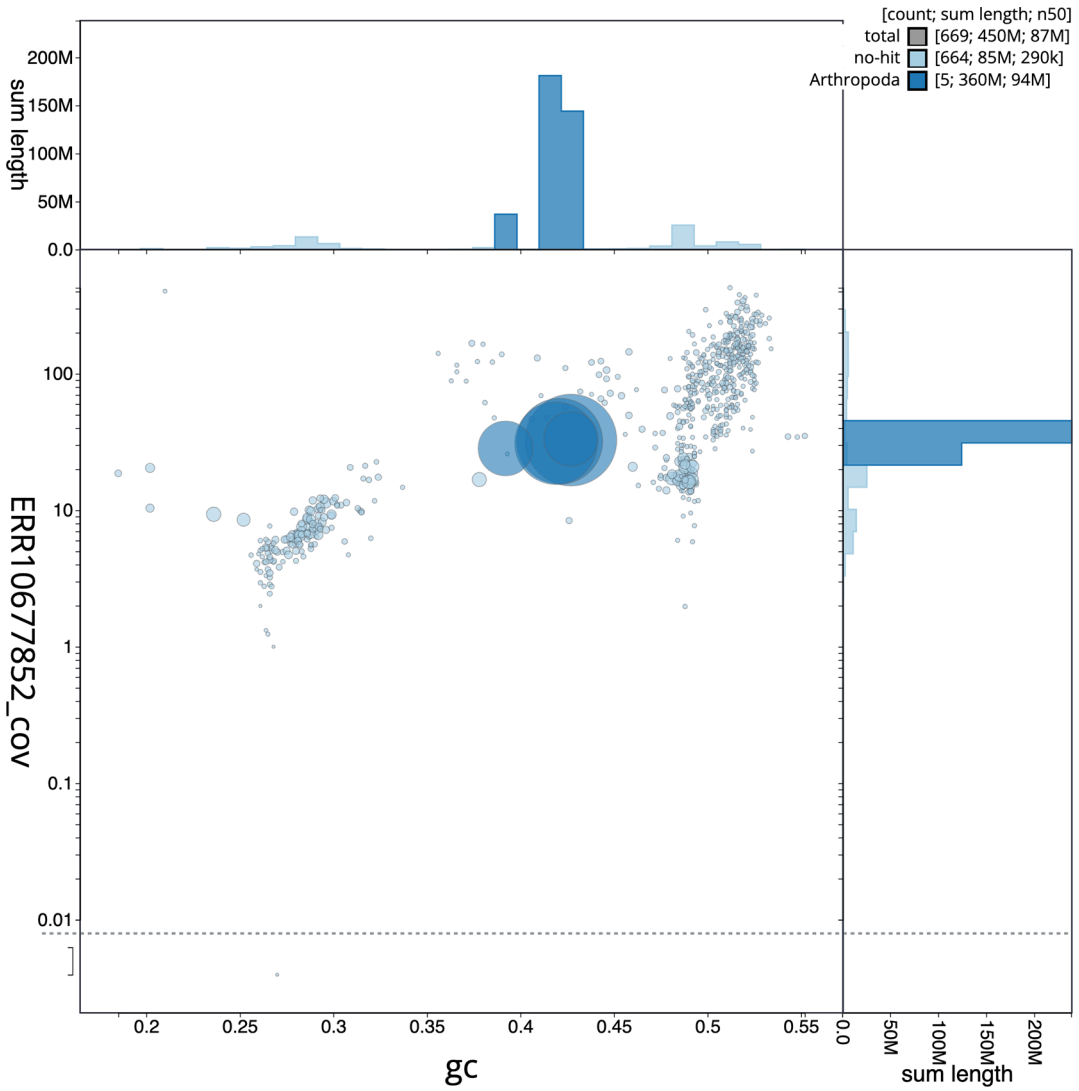

Figure 6. BlobToolKit blob plot for iyAndCara2.1. The plot shows base coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes display the total 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 *Andrena scotica* assembly.

Measure	Value	Benchmark
EBP summary	6.7.Q66	6.C.Q40
Contig N50 length	5.67 Mb	≥ 1 Mb
Scaffold N50 length	86.76 Mb	= chromosome N50
Consensus quality (QV)	66.1	≥ 40
k-mer completeness	98.90%	≥ 95%
BUSCO	C:96.9%[S:96.7%,D:0.3%], F:0.6%,M:2.4%,n:5991	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	80.96%	≥ 90%

Notes: EBP summary uses log10(Contig N50); chromosome-level (C) or log10(Scaffold N50); Q (Merqury QV). BUSCO: C = complete; S = single-copy; D = duplicated; F = fragmented; M = missing; n = orthologues using the hymenoptera_odb10 reference set.

Author information

Contributors are listed at the following links:

- Members of the University of Oxford and Wytham Woods 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: Andrena scotica. Accession number PRJEB58246. The genome sequence is released openly for reuse. The *Andrena scotica* genome sequencing initiative is part of the Darwin Tree of Life Project (PRJEB40665) and the Sanger Institute Tree of Life Programme (PRJEB43745). 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
BEDTools	2.30.0	https://github.com/arq5x/bedtools2
BLAST	2.14.0	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/
bwa-mem2	2.2.1	https://github.com/bwa-mem2/bwa-mem2
Cooler	0.8.11	https://github.com/open2c/cooler
fasta_windows	0.2.4	https://github.com/tolkit/fasta_windows

Table 5. *Continued*

Software	Version	Source
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	0.16.1-r375	https://github.com/chhylp123/hifiasm
HiGlass	1.13.4	https://github.com/higlass/higlass
MerquryFK	1.1.2	https://github.com/thegenemyers/MERQURY.FK
MitoHiFi	2	https://github.com/marcelauliano/MitoHiFi
MultiQC	1.14; 1.17 and 1.18	https://github.com/MultiQC/MultiQC
PretextSnapshot	0.0.5	https://github.com/sanger-tol/PretextSnapshot
PretextView	1.0.3	https://github.com/sanger-tol/PretextView
sanger-tol/ascc	0.1.0	https://github.com/sanger-tol/ascc
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	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.
Publisher Full Text

Challis R, Richards E, Rajan J, *et al.*: **BlobToolKit – interactive quality assessment of genome assemblies.** *G3 Genes|Genomes|Genetics.* 2020; **10**(4): 1361–1374.
Publisher 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.
Publisher 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.
Publisher Full Text

Danecek P, Bonfield JK, Liddle J, *et al.*: **Twelve years of SAMtools and BCFtools.** *Gigascience.* 2021; **10**(2).
Publisher Full Text

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

Hoffmann M, Gardein H, Greil H, *et al.*: **Anatomical, phenological and genetic aspects of the host–parasite relationship between *Andrena vaga* (Hymenoptera) and *Stylops ater* (Strepsiptera).** *Parasitology.* 2023; **150**: 744–753.
Publisher Full Text

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

Kathirithamby J: *Insect from Outer Space: Biology of Strepsiptera.* London: Wiley; 2025.

Kathirithamby J, Crowley LMUniversity of Oxford and Wytham Woods Acquisition Lab: **The genome sequence of *Stylops aterrimus* Newport, 1851 (Strepsiptera: Stylopidae).** *Wellcome Open Res.* 2025; **10**: 684.
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.
Publisher Full Text

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

Lähteenaro M, Benda D, Straka J, *et al.*: **Phylogenetic analysis of *Stylops* reveals the evolutionary history of a Holarctic Strepsiptera radiation parasitizing wild bees.** *Mol. Phylogenet. Evol.* 2024; **195**: 108068.
Publisher Full Text

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

Michener CD: *The Bees of the World.* Johns Hopkins University Press; 2007.
Publisher Full Text

NatureScot: **Bees, gardens, Insects, National Nature Reserves, St Cyrus National Nature Reserve.** 2020.
Reference Source

Paxton RJ, Tengö J, Hedström L: **Dipteran parasites and other associates of a communal bee, *Andrena scotica* (Hymenoptera: Apoidea), on Öland, SE Sweden.** *Entomol Tidskr.* 1996; **117**(4): 165–178.

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.
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.
Publisher 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.
Publisher 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

Saunders SS: **Descriptions of two new Strepsipteran insect species from Albania, parasitical of bees of the genus *Hylaeus*, with an account of their habits and metamorphosis.** *Trans Entomol Soc Lond.* 1850; **1**: 43–59.
Publisher 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

Straka J, Rezkova K, Batelka J, *et al.*: **Early nest emergence of females parasitised by Strepsiptera in protandrous bees (Hymenoptera Andrenidae).** *Ethol. Ecol. Evol.* 2011; **23**(2): 97–109.
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.** *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.
Publisher 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; 314–24.
Publisher Full Text

Wilson JS, Carril OM: *The Bees in Your Backyard.* Princeton University Press; 2016.

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

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 1

Reviewer Report 07 April 2026

https://doi.org/10.21956/wellcomeopenres.28777.r151128

✔ **Rodolpho S. T. Menezes**
[1] State University of Santa Cruz, Ilhéus, Brazil
[2] Universidade Estadual de Santa Cruz (Ringgold ID: 74361), Ilhéus, State of Bahia, Brazil

The data note by Crowley et al. presents a genome assembly for the bee *Andrena scotica* Perkins, 1916 (Hymenoptera: Andrenidae). The assembled genome has a total length of 446.96 megabases. This genomic resource represents a valuable contribution to future comparative genomics studies.
Overall, the methodology is robust and clearly described, the figures are well constructed, and the manuscript is well written.
However, the background section would benefit from further development. In particular, the authors should expand on the broader significance of this study and provide a more detailed comparison of their results with those from other bee genomes, including species within the genus *Andrena*.

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: Phylogenomics, Cytogenetics, and phylogeography.

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 01 April 2026

https://doi.org/10.21956/wellcomeopenres.28777.r151125

Eduardo Luís Menezes de Almeida
Universidade Federal de Viçosa, Minas Gerais, Brazil

General comments:
The data comprises a high-quality chromosome level assembly of *Andrena scotica* as part of the Darwin Tree of Life project. The methods and assembly data is well presented and consistent. Few points remain as described below.
Specific comments:
Background:
The authors provide a good background for *Andrena* bees. I suggest also adding a brief paragraph regarding other *Andrena* genome sequences (if available) and common genetic characteristics (if available) like chr number, ploidy, genome size..., which were might also have been expected along the sequencing effort.
Genome sequence report:
Figure 2: I could not recover the repeat content in this figure. Could the authors elaborate on that?
Assembly statistics: What could be responsible for the difference in size between the assembled genome and the prediction by GenomeScope2?

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: I work with bioinformatics and computational biology, including genomics, transcriptomics, and metabolic modeling of prokaryotes and eukaryotes.

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.