



DATA NOTE

The genome sequence of the Small Sallow Mining Bee, *Andrena praecox* (Scopoli, 1763)

[version 1; peer review: 4 approved]

Liam M. Crowley¹,
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

¹Department of Biology, University of Oxford, Oxford, England, UK

V1 First published: 02 Jun 2025, 10:290
<https://doi.org/10.12688/wellcomeopenres.24301.1>
Latest published: 02 Jun 2025, 10:290
<https://doi.org/10.12688/wellcomeopenres.24301.1>

Abstract

We present a genome assembly from a female specimen of *Andrena praecox* (Small Sallow Mining Bee; Arthropoda; Insecta; Hymenoptera; Andrenidae). The genome sequence has a total length of 563.05 megabases. The assembly is scaffolded into 7 chromosomal pseudomolecules. The mitochondrial genome has also been assembled, with a length of 20.36 kilobases. Gene annotation of this assembly on Ensembl identified 12,005 protein-coding genes.

Keywords

Andrena praecox, Small Sallow Mining Bee, genome sequence, chromosomal, Hymenoptera



This article is included in the [Tree of Life](#) gateway.

Open Peer Review

Approval Status

	1	2	3	4
version 1				
02 Jun 2025	view	view	view	view

1. **Pedro Henrique Pezzi** , University of Arkansas Fayetteville, Fayetteville, USA
2. **Jason L. Williams** , University of Florida, Florida, USA
3. **Paulo Cseri Ricardo** , University of Sao Paulo, São Paulo, Brazil
4. **Rodolpho Menezes** , State University of Santa Cruz, 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;

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.

Copyright: © 2025 Crowley LM *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

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 Small Sallow Mining Bee, *Andrena praecox* (Scopoli, 1763) [version 1; peer review: 4 approved]** Wellcome Open Research 2025, 10:290 <https://doi.org/10.12688/wellcomeopenres.24301.1>

First published: 02 Jun 2025, 10:290 <https://doi.org/10.12688/wellcomeopenres.24301.1>

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*; *Andrena incertae sedis*; *Andrena praecox* (Scopoli, 1763) (NCBI: txid1190803)

Background

We present a chromosomally complete genome sequence for *Andrena praecox*, based on a specimen from Wytham Woods, Oxfordshire, United Kingdom (Figure 1). It 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.

Genome sequence report

Sequencing data

The genome of a specimen of *Andrena praecox* (Figure 1) was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating 24.77 Gb from 2.64 million reads, which were used to assemble the genome. GenomeScope analysis estimated the haploid genome size at 517.47 Mb, with a heterozygosity of 0.30% and repeat content of 52.18%. These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 44 coverage. Hi-C sequencing produced 122.93 Gb from 814.11 million reads, and was used to scaffold the assembly. RNA sequencing data were also generated and are available in public sequence repositories. Table 1 summarises the specimen and sequencing details.

Assembly statistics

The primary haplotype was assembled, and contigs corresponding to an alternate haplotype were also deposited in INSDC



Figure 1. Photograph of the *Andrena praecox* (iyAndPrae1) specimen used for genome sequencing.

databases. The assembly was improved by manual curation, which corrected 98 misjoins or missing joins and removed 6 haplotypic duplications. These interventions decreased the scaffold count by 8.9% and increased the scaffold N50 by 51.21%. The final assembly has a total length of 563.05 Mb in 664 scaffolds, with 190 gaps, and a scaffold N50 of 62.44 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 (78.66%) was assigned to 7 chromosomal-level scaffolds. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 5; Table 3).

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 62.5. The *k*-mer recovery for the primary haplotype is 90.56%, and for the alternate haplotype 83.01%; the combined primary and alternate assemblies have a *k*-mer recovery of 96.56%. BUSCO v.5.5.0 analysis using the hymenoptera_odb10 reference set (*n* = 5,991) identified 96.7% of the expected gene set (single = 96.3%, duplicated = 0.3%).

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.

Genome annotation report

The *Andrena praecox* genome assembly (GCA_963932275.1) was annotated externally by Ensembl at the European Bioinformatics Institute (EBI). This annotation includes 20,407 transcribed mRNAs from 12,005 protein-coding and 1,897 non-coding genes. The average transcript length is 11,717.96 bp. There are 1.47 coding transcripts per gene and 6.12 exons per transcript. For further information about the annotation, please refer to <https://beta.ensembl.org/species/0a222fd5-5452-4530-b439-fdcbbf984ae1>.

Table 1. Specimen and sequencing data for *Andrena praecox*.

Project information			
Study title	Andrena praecox (small sallow mining bee)		
BioProject	PRJEB57674		
Species	<i>Andrena praecox</i>		
BioSpecimen	SAMEA10166794		
NCBI taxonomy ID	1190803		
Specimen information			
Technology	ToLID	BioSample accession	Organism part
PacBio long read sequencing	iyAndPrae1	SAMEA10201054	thorax
Hi-C sequencing	iyAndPrae1	SAMEA10201053	head
RNA sequencing	iyAndPrae1	SAMEA10201055	abdomen
Sequencing information			
Platform	Run accession	Read count	Base count (Gb)
Hi-C Illumina NovaSeq 6000	ERR10501025	8.14e+08	122.93
PacBio Sequel Iie	ERR10499359	2.64e+06	24.77
RNA Illumina NovaSeq 6000	ERR11641113	7.02e+07	10.6

Table 2. Genome assembly data for *Andrena praecox*.

Genome assembly		
Assembly name	iyAndPrae1.1	
Assembly accession	GCA_963932275.1	
Alternate haplotype accession	GCA_963932255.1	
Assembly level for primary assembly	chromosome	
Span (Mb)	563.05	
Number of contigs	854	
Number of scaffolds	664	
Longest scaffold (Mb)	104.45	
Assembly metric	Measure	Benchmark
Contig N50 length	3.32 Mb	≥ 1 Mb
Scaffold N50 length	62.44 Mb	= chromosome N50
Consensus quality (QV)	Primary: 62.1; alternate: 62.8; combined: 62.5	≥ 40
k-mer completeness	Primary: 90.56%; alternate: 83.01%; combined: 96.56%	≥ 95%
BUSCO*	C:96.7%[S:96.3%,D:0.3%], F:0.6%,M:2.8%,n:5,991	S > 90%; D < 5%
Percentage of assembly mapped to chromosomes	78.66%	≥ 90%
Sex chromosomes	None	localised homologous pairs
Organelles	Mitochondrial genome: 20.36 kb	complete single alleles

* BUSCO scores based on the hymenoptera_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.

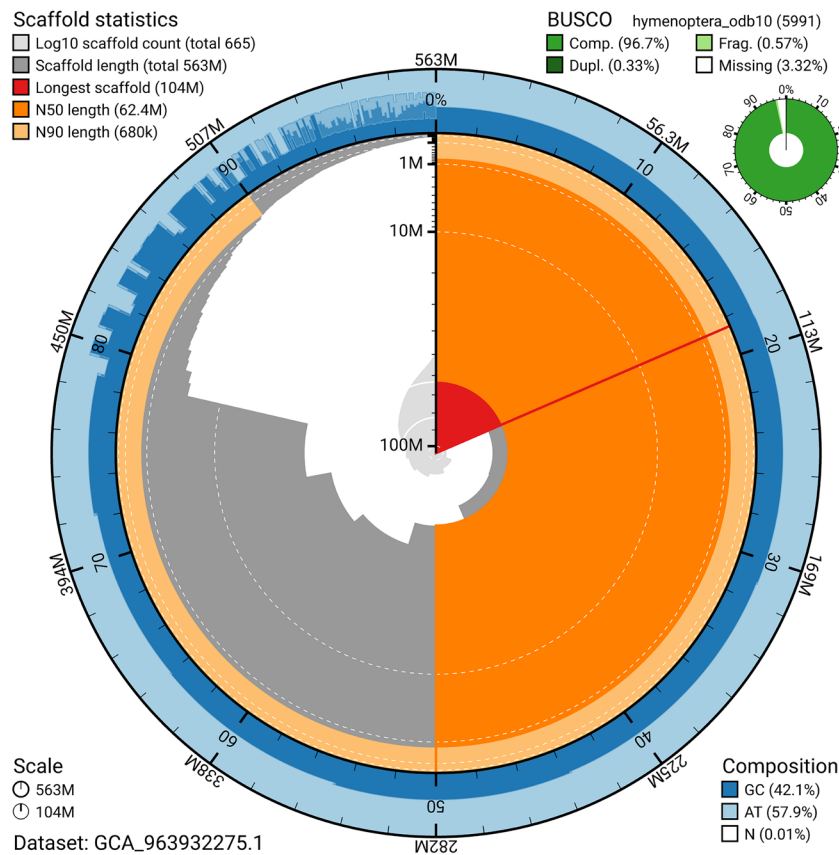


Figure 2. Genome assembly of *Andrena praecox*, iyAndPrael.1: metrics. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the circumference. 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 set is presented at the top right. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/GCA_963932275.1/dataset/GCA_963932275.1/snail.

Methods

Sample acquisition and DNA barcoding

The specimen used for genome sequencing was an adult female *Andrena praecox* (specimen ID Ox001317, ToLID iyAndPrael1), collected from Wytham Woods, Oxfordshire, United Kingdom (latitude 51.764, longitude -1.337) on 2021-04-23 by netting. 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 Core Laboratory includes a sequence of procedures: sample preparation and homogenisation, DNA extraction, fragmentation and purification. Detailed protocols are available on protocols.io (Denton *et al.*, 2023b).

The iyAndPrael 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

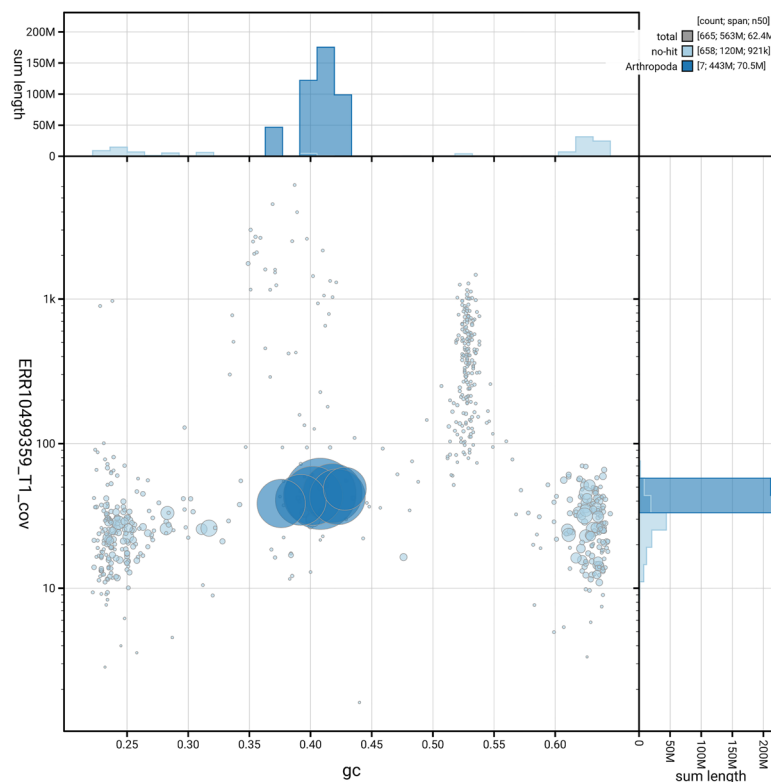


Figure 3. Genome assembly of *Andrena praecox*, iyAndPrael1.1: BlobToolKit GC-coverage plot. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. 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_963932275.1/dataset/GCA_963932275.1/blob.

disruptor (Denton *et al.*, 2023a). 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.

RNA was extracted from abdomen tissue of iyAndPrael1 in the Tree of Life Laboratory at the WSI using the RNA Extraction: Automated MagMax™ *mir*-Vana protocol (do Amaral *et al.*, 2023). 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.

Hi-C sample preparation and crosslinking

Hi-C data were generated from tissue from the head of the iyAndPrael1 sample using the Arima-HiC v2 kit (Arima Genomics). 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.

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

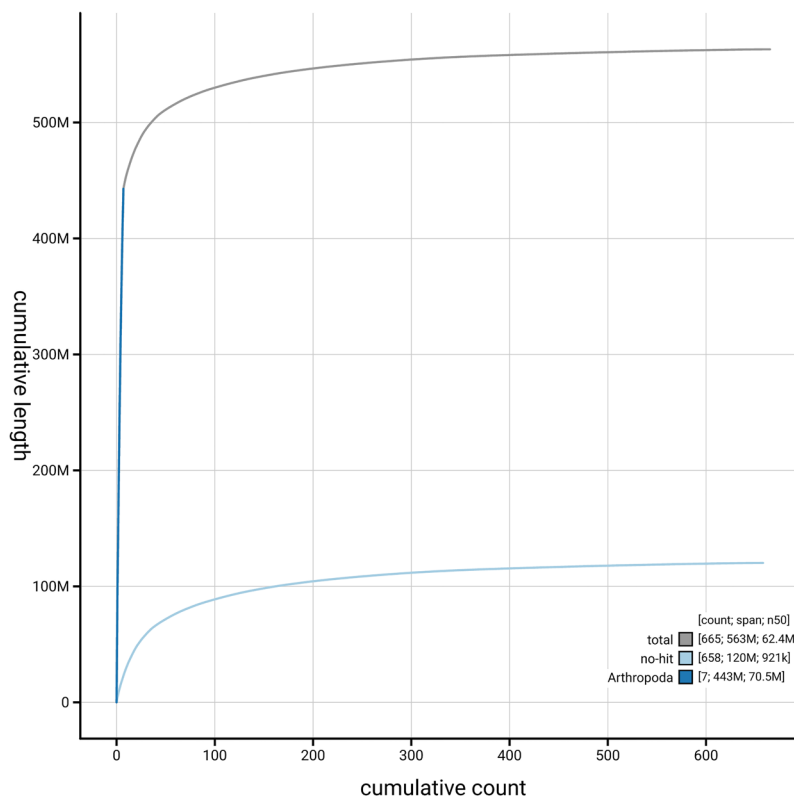


Figure 4. Genome assembly of *Andrena praecox*, iyAndPrae1.1: BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/GCA_963932275.1/dataset/GCA_963932275.1/cumulative.

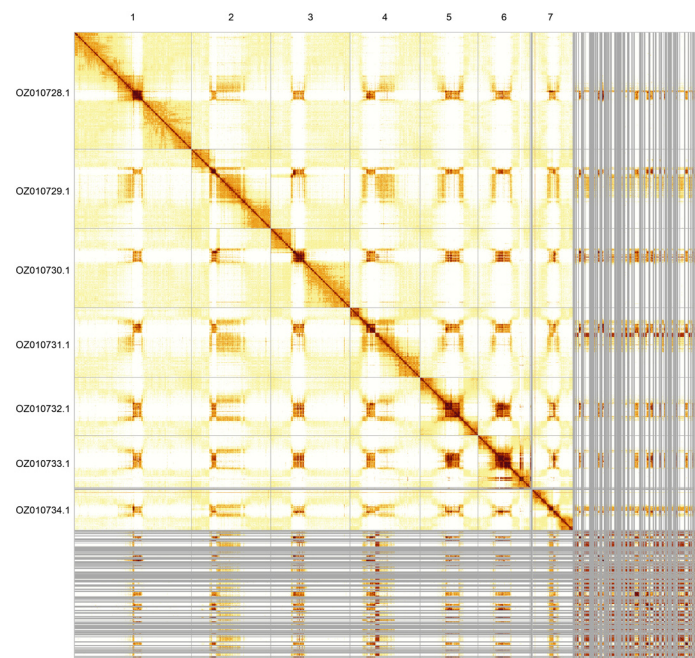


Figure 5. Hi-C contact map of the iyAndPrae1.1 assembly. The map was generated using PretextSnapshot. Chromosomes are shown in order of size and labelled with chromosome accession numbers (left) and names (top).

Table 3. Chromosomal pseudomolecules in the genome assembly of *Andrena praecox*, iyAndPrae1.

INSDC accession	Name	Length (Mb)	GC%
OZ010728.1	1	104.45	41
OZ010729.1	2	70.88	42
OZ010730.1	3	70.53	40
OZ010731.1	4	62.44	42
OZ010732.1	5	51.62	39
OZ010733.1	6	46.59	37.5
OZ010734.1	7	36.36	43
OZ010735.1	MT	0.02	23

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.

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, 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 using 150 bp paired-end reads.

RNA

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 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). 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. 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 using the BlobToolKit pipeline, a Nextflow (Di Tommaso *et al.*, 2017) implementation of the earlier Snakemake BlobToolKit pipeline (Challis *et al.*, 2020). The pipeline aligns PacBio reads using minimap2 (Li, 2018) and SAMtools (Daneczek *et al.*, 2021) to generate coverage tracks. Simultaneously, it queries the GoAT database (Challis *et al.*, 2023) to identify relevant BUSCO lineages and runs BUSCO (Manni *et al.*, 2021). For the three domain-level BUSCO lineages, BUSCO genes are aligned to the UniProt Reference Proteomes database (Bateman *et al.*, 2023) using DIAMOND blastp (Buchfink *et al.*, 2021). The genome is divided into chunks based on the density of BUSCO genes from the closest taxonomic lineage,

and each chunk is aligned to the UniProt Reference Proteomes database with DIAMOND blastx. Sequences without hits are chunked using seqtk and aligned to the NT database with blastn (Altschul *et al.*, 1990). The BlobToolKit suite consolidates all outputs into a blobdir for visualisation.

The BlobToolKit pipeline was developed using nf-core tooling (Ewels *et al.*, 2020) and MultiQC (Ewels *et al.*, 2016), with package management via Conda and Bioconda (Grüning *et al.*, 2018), and containerisation through Docker (Merkel, 2014) and Singularity (Kurtzer *et al.*, 2017).

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 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

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	666652151335353eef2fcd58880bcef5bc2928e1	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
Goat CLI	0.2.5	https://github.com/genomehubs/goat-cli
Hifiasm	0.16.1-r375	https://github.com/chhyllp123/hifiasm
HiGlass	44086069ee7d4d3f6f3f0012569789ec138f42b84aa44357826c0b6753eb28de	https://github.com/higlass/higlass
MerquyFK	d00d98157618f4e8d1a9190026b19b471055b22e	https://github.com/thegenemyers/MERQUERY.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

Software tool	Version	Source
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	yahs-1.2a	https://github.com/c-zhou/yahs

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

Data availability

European Nucleotide Archive: *Andrena praecox* (small sawfly mining bee). Accession number PRJEB57674; <https://identifiers.org/ena.embl/PRJEB57674>. The genome sequence is released openly for reuse. The *Andrena praecox* 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. 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](#)

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.* 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](#)

do Amaral RJV, Bates A, Denton A, *et al.*: **Sanger Tree of Life RNA extraction: automated MagMax™ mirVana.** *protocols.io.* 2023.
[Publisher Full Text](#)

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

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

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

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](#)

Howe K, Chow W, Collins J, *et al.*: **Significantly improving the quality of genome assemblies through curation.** *GigaScience.* 2021; **10**(1): g1aa153.
[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-Conceicao 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, Seppely 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.*: **Merquy: 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, Krashenninnikova 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 1

Reviewer Report 23 August 2025

<https://doi.org/10.21956/wellcomeopenres.26800.r126862>

© 2025 Menezes R. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Rodolpho Menezes 

State University of Santa Cruz, Ilhéus, Brazil

The manuscript by Crowley presents a genome assembly from a female *Andrena praecox* (Scopoli, 1763), resulting in a final assembly size of 563 megabases. This assembly represents a valuable resource for comparative genomics, as well as for studies on bee evolution and conservation. The methodology is thorough and clearly described, the figures are well constructed, and the manuscript is generally well written.

However, I recommend that the author compare this assembly with those of other bee species to contextualize the results more effectively. Additionally, it would be valuable to include an analysis of repetitive elements, including the proportion of the genome they occupy and the types of repeats that are most prevalent.

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: Cytogenetics, Phylogeography and Phylogenomics

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 19 August 2025

<https://doi.org/10.21956/wellcomeopenres.26800.r126858>

© 2025 Ricardo P. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**Paulo Cseri Ricardo** 

University of Sao Paulo, São Paulo, Brazil

This article presents data on the genome assembly of the bee species *Andrena praecox*. A robust approach was used to generate chromosome-level scaffolds, combining PacBio HiFi long reads with Hi-C mapping. Assembly metrics indicated high quality (QV) and completeness (k-mer and BUSCO). Most protocols were clearly described or appropriately referenced to ensure reproducibility. The final assembly spans 563.05 Mb across 664 scaffolds, with seven scaffolds accounting for 78.66% of the total sequence (chromosomal pseudomolecules), in addition to a 20.36 Kb mitogenome. Notably, a portion of the assembly did not align with the pseudomolecules and, for the most part, lacked hits to known taxa in BlobToolKit analyses. Examining these smaller scaffolds could clarify whether they stem from contaminants, low-complexity sequences, or assembly artifacts. Overall, the manuscript is clearly and objectively written and presents relevant data for studies on hymenopteran genome evolution, particularly in bees. I consider it suitable for indexing in its current form.

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:** Transcriptomics and genomics applied to the evolution and molecular ecology of bee species.**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 12 August 2025

<https://doi.org/10.21956/wellcomeopenres.26800.r126861>

© 2025 Williams J. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Jason L. Williams 

University of Florida, Florida, USA

In this manuscript, the authors present a chromosomally complete genome sequence of the small sallow mining bee, *Andrena praecox*, supported by detailed sequencing, assembly statistics, and thorough quality assessment and annotation. This high-quality genome represents a valuable resource for investigating the evolution, ecology, and conservation of wild pollinators and provides a solid foundation for comparative genomics within *Andrena* and across Hymenoptera.

Overall, the assembly is supported by rigorous methods and demonstrates strong completeness and accuracy, as evidenced by high contiguity, base-level quality, and robust BUSCO and k-mer metrics. The annotated gene set enhances the utility of the genome for downstream analyses, and the authors ensure transparency and reusability by making all data openly accessible.

Despite the overall high quality of the genome assembly, a few areas could benefit from clarification or future improvement. These include slightly sub-threshold k-mer completeness in the primary assembly, incomplete chromosomal assignment, absence of identifiable sex chromosomes, and lack of functional annotation summary.

Suggestions for improvement:

- Notably, there is no discussion section in the manuscript. Perhaps one could be included to offer further clarification of some points laid out below.
- Although combined k-mer completeness is over 96%, the primary assembly on its own only accounts for 90.56%. This is not an issue, but it might be worth elaborating in the manuscript on what genomic regions (e.g., repetitive sequences, heterochromatic regions, etc.) might be under-represented or missing and why.
- Only 78.66% of the genome is scaffolded into chromosomes, well below the 90% benchmark cited in Table 2. Have any additional attempts been made to help reach the benchmark? Would additional Hi-C rounds or linkage mapping help close this gap? At the very least, some explanation about this gap in scaffolding completeness is warranted.
- The assembly lacks identification of sex chromosomes, which are biologically important in bees and other Hymenoptera due to haplodiploid sex determination. It is not uncommon to face difficulty in identifying sex chromosomes of Hymenoptera, but the authors should briefly discuss this point and provide some interpretation as to why.
- While gene annotation is included, no summary of functional categories (e.g., GO terms or pathways) is provided. Including or referencing such data would increase the genome's utility for downstream analyses.
- Finally, the manuscript would benefit from a brief explanation of the practical significance of sequencing *Andrena praecox*. Clarifying why this species was selected beyond the broad goal of sequencing all named eukaryotes in the Atlantic Archipelago and highlighting potential applications in ecology, conservation, or pollination biology would enhance the work's impact and relevance.

Is the rationale for creating the dataset(s) clearly described?

Partly

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: Insect biodiversity, taxonomy, systematics, phylogenomics, bioinformatics.

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 26 July 2025

<https://doi.org/10.21956/wellcomeopenres.26800.r126855>

© 2025 Pezzi P. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Pedro Henrique Pezzi 

Biological Sciences, University of Arkansas Fayetteville, Fayetteville, Arkansas, USA

This paper generated a high-quality chromosome-level genome assembly of *Andrena praecox*, the small sallow mining bee, which represents a valuable resource for future research. The authors followed the standard Darwin Tree of Life pipeline, resulting in a genome with high BUSCO scores, contiguity and coverage. They also generated RNA data to help with genome annotation, making this a valuable resource for studies on bees.

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: Phylogenetics, genomics, evolution, hybridization, population genetics.

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.
