



DATA NOTE

The genome sequence of the big-headed mining bee, *Andrena bucephala* (Stephens, 1846) [version 1; 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

¹University of Oxford, Oxford, England, UK

v1 First published: 01 Mar 2024, 9:111
<https://doi.org/10.12688/wellcomeopenres.21003.1>
Latest published: 01 Mar 2024, 9:111
<https://doi.org/10.12688/wellcomeopenres.21003.1>

Abstract

We present a genome assembly from an individual female *Andrena bucephala* (the Big-headed Mining Bee; Arthropoda; Insecta; Hymenoptera; Andrenidae). The genome sequence is 379.8 megabases in span. Most of the assembly is scaffolded into 5 chromosomal pseudomolecules. The mitochondrial genome has also been assembled and is 19.57 kilobases in length. Gene annotation of this assembly on Ensembl identified 12,022 protein coding genes.

Keywords

Andrena bucephala, big-headed 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		
01 Mar 2024	view	view

1. **Paulo Cseri Ricardo** , Universidade de São Paulo, São Paulo, Brazil
2. **Eduardo Zattara**, Instituto de Investigaciones en Biodiversidad y Medioambiente, CONICET-Universidad Nacional del Comahue, Bariloche, Argentina

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, Writing – Original Draft Preparation, Writing – Review & Editing;

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome through core funding to the Wellcome Sanger Institute (206194) and the Darwin Tree of Life Discretionary Award (218328).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2024 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 big-headed mining bee, *Andrena bucephala* (Stephens, 1846) [version 1; peer review: 2 approved]** Wellcome Open Research 2024, 9:111 <https://doi.org/10.12688/wellcomeopenres.21003.1>

First published: 01 Mar 2024, 9:111 <https://doi.org/10.12688/wellcomeopenres.21003.1>

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 bucephala* (Stephens, 1846) (NCBI:txid1542580).

Background

The Big-headed Mining Bee, *Andrena bucephala*, is a mid-sized (forewing length 7–10 mm) mining bee in the family Andrenidae. It occurs widely across Europe, although it is generally scarce and locally distributed. This is true in the UK also where it occurs locally across southern England and Wales, and it is designated as 'Nationally Scarce A' (Falk, 1991). Females have a sparse covering of hairs with denser clumps of pale hairs on the sides of the thorax, margins of the tergites and scopa, with yellow hairs on the hind tibiae. The metasoma is narrow and rounded, the facial foveae are pale, and the wings are distinctly tinted yellow. Males have large, square heads with long mandibles lacking a subapical tooth or emargination.

A. bucephala occurs in a wide range of habitats, including open deciduous woodland, calcareous grassland, scrub and gardens. It is univoltine, with a flight period from April to June. Whilst it is a solitary species, it nests in compact aggregations sharing a communal, single nest entrance (Perkins, 1917). Often nests occur in a steep bank or rabbit burrows, lasting for multiple years and may contain more than 200 females (O'Toole & Raw, 1991). It visits the flowers of a range of species, although pollen is mainly collected from common hawthorn, *Crataegus monogyna*, and field maple, *Acer campestre* (Falk & Lewington, 2019; Wood & Roberts, 2017).

The complete genome sequence for this species will facilitate studies into the evolution of sociality, reproductive systems and Hymenopteran taxonomy.

Genome sequence report

The genome was sequenced from one female *Andrena bucephala* (Figure 1) collected from Dry Sandford Pit Nature Reserve, Oxfordshire, UK (51.7, −1.32). A total of 51-fold coverage in Pacific Biosciences single-molecule HiFi long reads was generated. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data. Manual assembly curation corrected 102 missing joins or mis-joins and removed 2 haplotypic duplications, reducing the assembly length by 0.16% and the scaffold number by 85.26%, and increasing the scaffold N50 by 543.48%.

The final assembly has a total length of 379.8 Mb in 13 sequence scaffolds with a scaffold N50 of 110.2 Mb (Table 1). The snailplot in Figure 2 provides a summary of the assembly statistics, while the distribution of assembly scaffolds on GC proportion and coverage is shown in Figure 3. The cumulative assembly plot in Figure 4 shows curves for subsets of scaffolds assigned to different phyla. Most (99.93%) of the



Figure 1. Photograph of the *Andrena bucephala* (iyAndBuce1) specimen used for genome sequencing.

assembly sequence was assigned to 5 chromosomal-level scaffolds. Chromosome-scale scaffolds confirmed by the Hi-C data are named in order of size (Figure 5; Table 2). While not fully phased, the assembly deposited is of one haplotype. Contigs corresponding to the second haplotype have also been deposited. The mitochondrial genome was also assembled and can be found as a contig within the multifasta file of the genome submission.

The estimated Quality Value (QV) of the final assembly is 63.5 with *k*-mer completeness of 100.0%, and the assembly has a BUSCO v5.3.2 completeness of 96.7% (single = 96.4%, duplicated = 0.3%), using the hymenoptera_odb10 reference set (*n* = 5,991).

Metadata for specimens, barcode results, spectra estimates, sequencing runs, contaminants and pre-curation assembly statistics are given at <https://links.tol.sanger.ac.uk/species/1542580>.

Genome annotation report

The *Andrena bucephala* genome assembly (GCA_947577245.1) was annotated by the European Bioinformatics Institute (EBI) using the Ensembl rapid annotation pipeline (Table 1; https://rapid.ensembl.org/Andrena_bucephala_GCA_947577245.1/Info/Index). The resulting annotation includes 27,899 transcribed mRNAs from 12,022 protein-coding and 4,153 non-coding genes.

Methods

Sample acquisition and nucleic acid extraction

A female *Andrena bucephala* (specimen ID Ox001341, ToLID iyAndBuce1) was netted at Dry Sandford Pit Nature Reserve, Oxfordshire, UK (latitude 51.7, longitude −1.32) on 2021-05-10. The specimen was collected and identified by Liam Crowley (University of Oxford) and preserved on dry ice.

Table 1. Genome data for *Andrena bucephala*, iyAndBuce1.1.

Project accession data		
Assembly identifier	iyAndBuce1.1	
Species	<i>Andrena bucephala</i>	
Specimen	iyAndBuce1	
NCBI taxonomy ID	1542580	
BioProject	PRJEB55567	
BioSample ID	SAMEA10166816	
Isolate information	iyAndBuce1, female: thorax (DNA sequencing), head Hi-C sequencing), abdomen (RNA sequencing)	
Assembly metrics*		Benchmark
Consensus quality (QV)	63.5	≥ 50
k-mer completeness	100.0%	≥ 95%
BUSCO**	C:96.7%[S:96.4%,D:0.3%],F:0.9%,M:2.4%,n:5,991	C ≥ 95%
Percentage of assembly mapped to chromosomes	99.93%	≥ 95%
Sex chromosomes	None	localised homologous pairs
Organelles	Mitochondrial genome: 19.57 kb	complete single alleles
Raw data accessions		
PacificBiosciences SEQUEL II	ERR10115637	
Hi-C Illumina	ERR10123710	
PolyA RNA-Seq Illumina	ERR10890704	
Genome assembly		
Assembly accession	GCA_947577245.1	
Accession of alternate haplotype	GCA_947577235.1	
Span (Mb)	379.8	
Number of contigs	195	
Contig N50 length (Mb)	8.0	
Number of scaffolds	13	
Scaffold N50 length (Mb)	110.2	
Longest scaffold (Mb)	122.96	
Genome annotation		
Number of protein-coding genes	12,022	
Number of non-coding genes	4,153	
Number of gene transcripts	27,899	

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

** BUSCO scores based on the hymenoptera_odb10 BUSCO set using version 5.3.2. C = complete [S = single copy, D = duplicated], F = fragmented, M = missing, n = number of orthologues in comparison. A full set of BUSCO scores is available at <https://blobtoolkit.genomehubs.org/view/CANPUK01/dataset/CANPUK01/busco>.

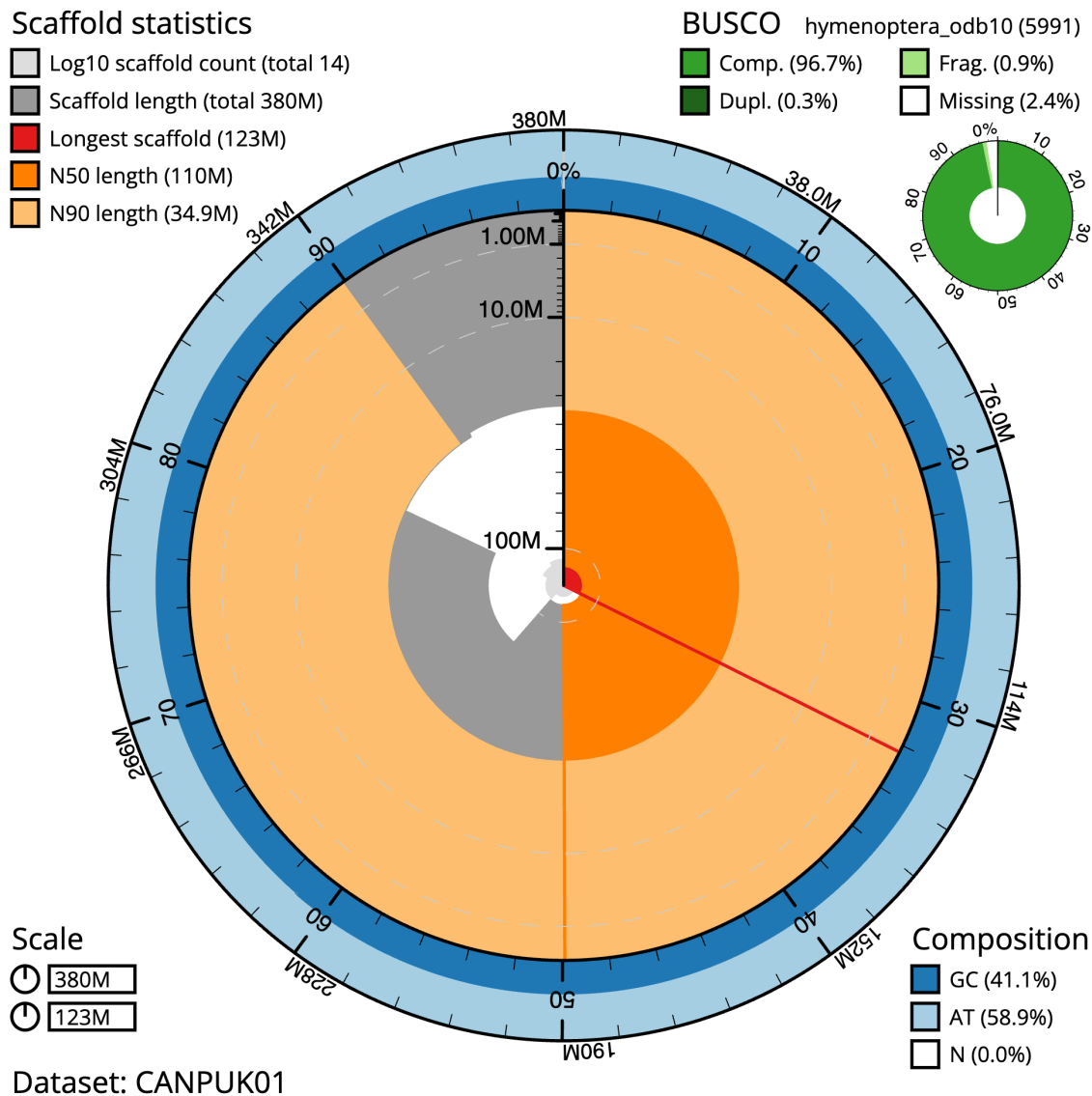


Figure 2. Genome assembly of *Andrena bucephala*, iyAndBuce1.1: metrics. The BlobToolKit Snailplot shows N50 metrics and BUSCO gene completeness. The main plot is divided into 1,000 size-ordered bins around the circumference with each bin representing 0.1% of the 379,790,978 bp assembly. The distribution of scaffold lengths is shown in dark grey with the plot radius scaled to the longest scaffold present in the assembly (122,962,308 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 scaffold lengths (110,195,567 and 34,891,842 bp), respectively. The pale grey spiral shows the cumulative scaffold count on a log scale with white scale lines showing successive orders of magnitude. The blue and pale-blue area around the outside of the plot shows the distribution of GC, AT and N percentages in the same bins as the inner plot. A summary of complete, fragmented, duplicated and missing BUSCO genes in the hymenoptera_odb10 set is shown in the top right. An interactive version of this figure is available at <https://blobtoolkit.genomehubs.org/view/CANPUK01/dataset/CANPUK01/snail>.

Protocols developed by the Wellcome Sanger Institute (WSI) Tree of Life core laboratory have been deposited on protocols.io (Denton *et al.*, 2023b). The workflow for high molecular weight (HMW) DNA extraction at the WSI includes a sequence of core procedures: sample preparation; sample homogenisation, DNA extraction, fragmentation, and clean-up. In sample preparation, the iyAndBuce1 sample was weighed and

dissected on dry ice (Jay *et al.*, 2023). Tissue from the thorax was homogenised using a PowerMasher II tissue 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). HMW DNA was sheared into an average fragment size of 12–20 kb in a Megaruptor 3 system with speed setting 31 (Bates *et al.*, 2023).

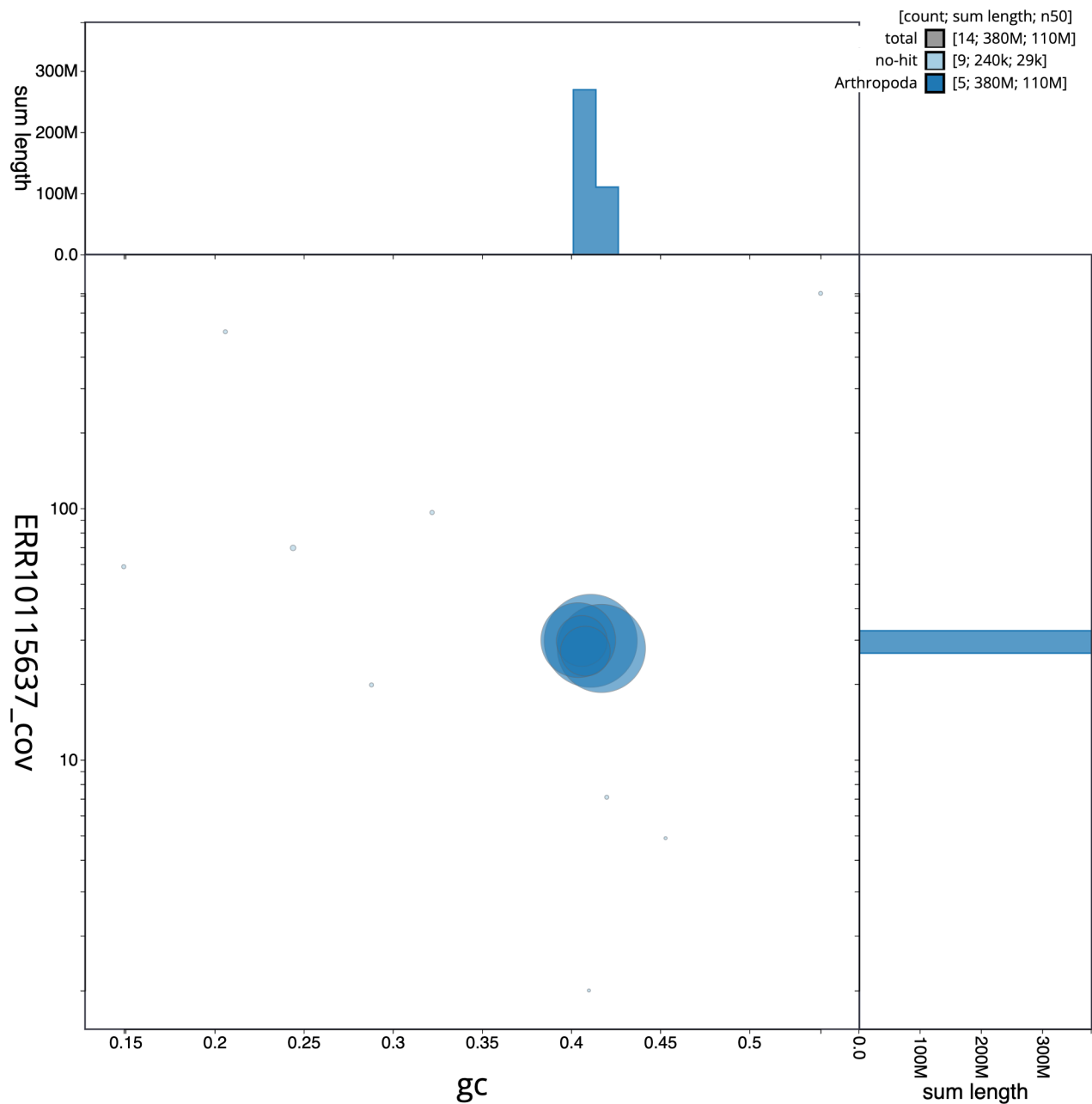


Figure 3. Genome assembly of *Andrena bucephala*, iyAndBuce1.1: BlobToolKit GC-coverage plot. Scaffolds are coloured by phylum. Circles are sized in proportion to scaffold length. Histograms show the distribution of scaffold length sum along each axis. An interactive version of this figure is available at <https://blobtoolkit.genomehubs.org/view/CANPUK01/dataset/CANPUK01/blob>.

Sheared DNA was purified by solid-phase reversible immobilisation (Strickland *et al.*, 2023): in brief, the method employs a 1.8X ratio of AMPure PB beads to sample to eliminate shorter fragments and concentrate the DNA. The concentration of the sheared and purified DNA was assessed using a Nanodrop spectrophotometer and Qubit Fluorometer and 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 iyAndBuce1 in the Tree of Life Laboratory at the WSI using the RNA Extraction: Automated MagMax™ mirVana 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.

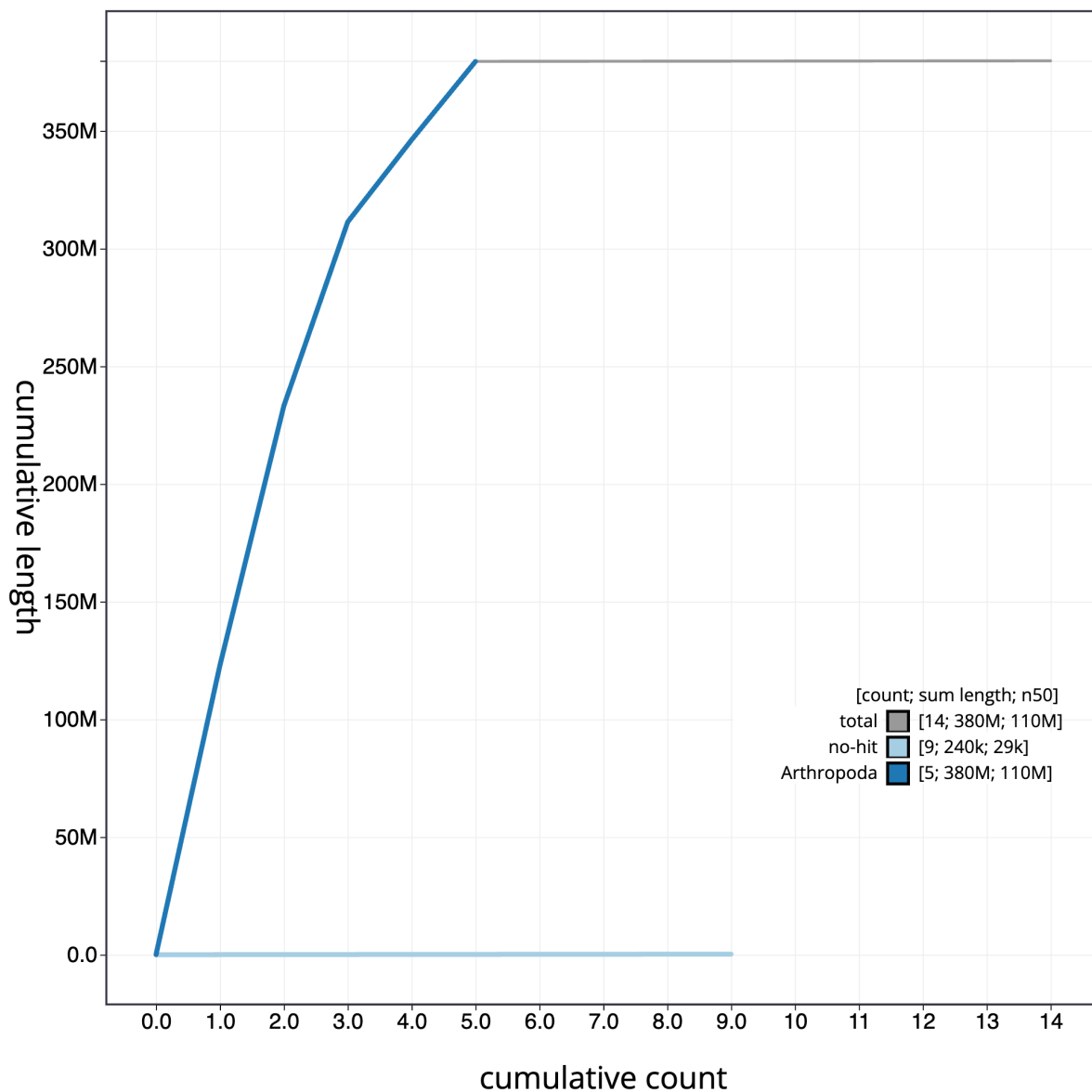


Figure 4. Genome assembly of *Andrena bucephala*, iyAndBuce1.1: BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the busco genes taxrule. An interactive version of this figure is available at <https://blobtoolkit.genomehubs.org/view/CANPUK01/dataset/CANPUK01/cumulative>.

Sequencing

Pacific Biosciences HiFi circular consensus DNA sequencing libraries were constructed according to the manufacturers' instructions. Poly(A) RNA-Seq libraries were constructed using the NEB Ultra II RNA Library Prep kit. DNA and RNA sequencing was performed by the Scientific Operations core at the WSI on Pacific Biosciences SEQUEL II (HiFi) and Illumina NovaSeq 6000 (RNA-Seq) instruments. Hi-C data were also generated from head tissue of iyAndBuce1 using the Arima2 kit and sequenced on the Illumina NovaSeq 6000 instrument.

Genome assembly, curation and evaluation

Assembly was carried out with Hifiasm (Cheng *et al.*, 2021) and haplotypic duplication was identified and removed with purge_dups (Guan *et al.*, 2020). The assembly was then scaffolded with Hi-C data (Rao *et al.*, 2014) using YaHS (Zhou *et al.*, 2023). The assembly was checked for contamination and corrected as described previously (Howe *et al.*, 2021). Manual curation was performed using HiGlass (Kerpedjiev *et al.*, 2018) and Pretext (Harry, 2022). The mitochondrial genome was assembled using MitoHiFi (Uliano-Silva *et al.*, 2023), which runs MitoFinder (Allio *et al.*, 2020) or MITOS (Bernt *et al.*, 2013).

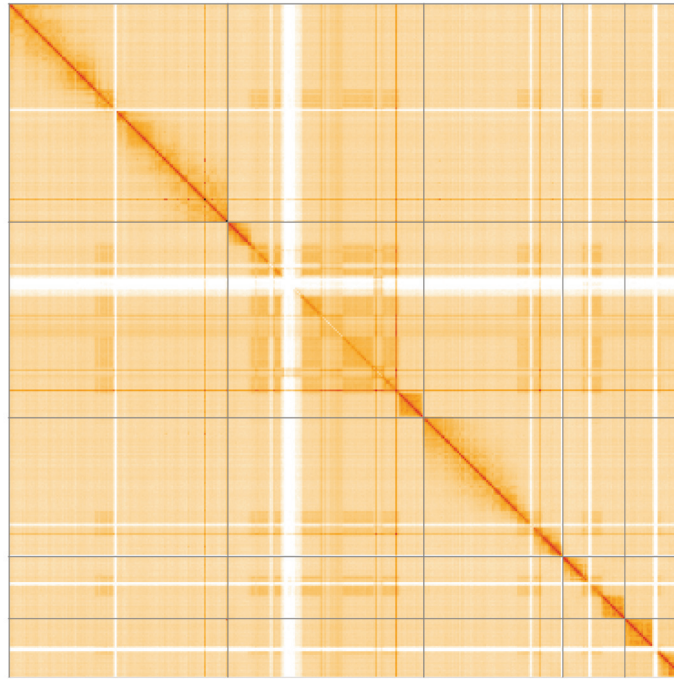


Figure 5. Genome assembly of *Andrena bucephala*, iyAndBuce1.1: Hi-C contact map of the iyAndBuce1.1 assembly, visualised using HiGlass. Chromosomes are shown in order of size from left to right and top to bottom. An interactive version of this figure may be viewed at <https://genome-note-higlass.tol.sanger.ac.uk/1/?d=ZfRikMozSyiKke1P8NWdfA>.

Table 2. Chromosomal pseudomolecules in the genome assembly of *Andrena bucephala*, iyAndBuce1.

INSDC accession	Chromosome	Length (Mb)	GC%
OX387715.1	1	122.96	41.0
OX387716.1	2	110.2	41.5
OX387717.1	3	78.17	40.5
OX387718.1	4	34.89	40.5
OX387719.1	5	33.33	41.0
OX387720.1	MT	0.02	20.5

and uses these annotations to select the final mitochondrial contig and to ensure the general quality of the sequence.

A Hi-C map for the final assembly was produced using bwa-mem2 (Vasimuddin *et al.*, 2019) in the Cooler file format (Abdennur & Mirny, 2020). To assess the assembly metrics, the *k*-mer completeness and QV consensus quality values were calculated in Merqury (Rhie *et al.*, 2020). This work

was done using Nextflow (Di Tommaso *et al.*, 2017) DSL2 pipelines “sanger-tol/readmapping” (Surana *et al.*, 2023a) and “sanger-tol/genomenote” (Surana *et al.*, 2023b). The genome was analysed within the BlobToolKit environment (Challis *et al.*, 2020) and BUSCO scores (Manni *et al.*, 2021; Simão *et al.*, 2015) were calculated.

Table 3 contains a list of relevant software tool versions and sources.

Genome annotation

The EBI used Ensembl Genebuild (Aken *et al.*, 2016) to generate annotation for the *Andrena bucephala* assembly (GCA_947577245.1). Annotation was created primarily through alignment of transcriptomic data to the genome, with gap filling via protein-to-genome alignments of a select set of proteins from UniProt (UniProt Consortium, 2019).

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](https://www.darwintreeoflife.org/). By agreeing with and signing up to the Sampling Code of Practice, the Darwin Tree of Life

Table 3. Software tools: versions and sources.

Software tool	Version	Source
BlobToolKit	4.1.7	https://github.com/blobtoolkit/blobtoolkit
BUSCO	5.3.2	https://gitlab.com/ezlab/busco
Hifiasm	0.16.1-r375	https://github.com/chhylp123/hifiasm
HiGlass	1.11.6	https://github.com/higlass/higlass
Merqury	MerquryFK	https://github.com/thegenemyers/MERQURY.FK
MitoHiFi	2.3	https://github.com/marcelauliano/MitoHiFi
PretextView	0.2	https://github.com/wtsi-hpag/PretextView
purge_dups	1.2.3	https://github.com/dfguan/purge_dups
sanger-tol/genomenote	v1.0	https://github.com/sanger-tol/genomenote
sanger-tol/readmapping	1.1.0	https://github.com/sanger-tol/readmapping/tree/1.1.0
YaHS	yahs-1.1.91eebc2	https://github.com/c-zhou/yahs

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 bucephala*. Accession number PRJEB55567; <https://identifiers.org/ena.embl/PRJEB55567> (Wellcome Sanger Institute, 2022). The genome sequence is released openly for reuse. The *Andrena bucephala*

genome sequencing initiative is part of the Darwin Tree of Life (DTOL) project. All raw sequence data and the assembly have been deposited in INSDC databases. Raw data and assembly accession identifiers are reported in Table 1.

Author information

Members of the University of Oxford and Wytham Woods Genome Acquisition Lab are listed here: <https://doi.org/10.5281/zenodo.7125292>.

Members of the Darwin Tree of Life Barcoding collective are listed here: <https://doi.org/10.5281/zenodo.4893703>.

Members of the Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team are listed here: <https://doi.org/10.5281/zenodo.10066175>.

Members of Wellcome Sanger Institute Scientific Operations: Sequencing Operations are listed here: <https://doi.org/10.5281/zenodo.10043364>.

Members of the Wellcome Sanger Institute Tree of Life Core Informatics team are listed here: <https://doi.org/10.5281/zenodo.10066637>.

Members of the Tree of Life Core Informatics collective are listed here: <https://doi.org/10.5281/zenodo.5013541>.

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

References

- Abdennur N, Mirny LA: **Cooler: Scalable storage for Hi-C data and other genomically labeled arrays.** *Bioinformatics*. 2020; **36**(1): 311–316.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Aken BL, Ayling S, Barrell D, *et al.*: **The Ensembl gene annotation system.** *Database (Oxford)*. 2016; **2016**: baw093.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Allio R, Schomaker-Bastos A, Romiguier J, *et al.*: **MitoFinder: Efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics.** *Mol Ecol Resour*. 2020; **20**(4): 892–905.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Bates A, Clayton-Lucey I, Howard C: **Sanger Tree of Life HMW DNA Fragmentation: Diagenode Megaruptor[®]3 for LI PacBio.** *protocols.io*. 2023.
[Publisher Full Text](#)
- Bernt M, Donath A, Jühling F, *et al.*: **MITOS: Improved *de novo* metazoan mitochondrial genome annotation.** *Mol Phylogenet Evol*. 2013; **69**(2): 313–319.
[PubMed Abstract](#) | [Publisher 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](#)
- 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](#)
- Falk S: **A review of the scarce and threatened flies of Great Britain.** Nature Conservancy Council, 1991.
[Reference Source](#)
- Falk S, Lewington R: **Field Guide to the Bees of Great Britain and Ireland.** London: Bloomsbury, 2019.
- Guan D, McCarthy SA, Wood J, *et al.*: **Identifying and removing haplotypic duplication in primary genome assemblies.** *Bioinformatics*. 2020; **36**(9): 2896–2898.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Harry E: **PretextView (Paired Read TEXTure Viewer): A desktop application for viewing pretext contact maps.** 2022; [Accessed 19 October 2022].
[Reference Source](#)
- Howe K, Chow W, Collins J, *et al.*: **Significantly improving the quality of genome assemblies through curation.** *GigaScience*. Oxford University Press, 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](#)
- 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](#)
- Oatley G, Denton A, Howard C: **Sanger Tree of Life HMW DNA Extraction: Automated MagAttract v.2.** *protocols.io*. 2023.
[Publisher Full Text](#)
- O'Toole C, Raw A: **Bees of the world.** New York: Blandford Press, 1991.
- Perkins RCL: **Notes on the collection of British Hymenoptera (Aculeata) formed by F. Smith.** *Entomologist's Monthly Magazine*. 1917; **53**: 71–76.
- 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](#)
- Simão FA, Waterhouse RM, Ioannidis P, *et al.*: **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.** *Bioinformatics*. 2015; **31**(19): 3210–3212.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Strickland M, Cornwell C, Howard C: **Sanger Tree of Life Fragmented DNA clean up: Manual SPRI.** *protocols.io*. 2023.
[Publisher Full Text](#)
- Surana P, Muffato M, Qi G: **sanger-tol/readmapping: sanger-tol/readmapping v1.1.0 - Hebridean Black (1.1.0).** *Zenodo*. 2023a; (Accessed: 17 April 2023).
[Publisher Full Text](#)
- Surana P, Muffato M, Sadasivan Baby C: **sanger-tol/genomenote (v1.0.dev).** *Zenodo*. 2023b; (Accessed: 17 April 2023).
[Publisher Full Text](#)
- Uliano-Silva M, Ferreira GJR, 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](#)
- UniProt Consortium: **UniProt: a worldwide hub of protein knowledge.** *Nucleic Acids Res*. 2019; **47**(D1): D506–D515.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Vasimuddin Md, Misra S, Li H, *et al.*: **Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems.** In: *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. IEEE, 2019; 314–324.
[Publisher Full Text](#)
- Wellcome Sanger Institute: **The genome sequence of the big-headed mining bee, *Andrena bucephala* (Stephens, 1846).** European Nucleotide Archive, [dataset], accession number PRJEB55567. 2022.
- Wood TJ, Roberts SPM: **An assessment of historical and contemporary diet breadth in polylectic *Andrena* bee species.** *Biol Conserv*. 2017; **215**: 72–80.
[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 13 August 2024

<https://doi.org/10.21956/wellcomeopenres.23238.r90509>

© 2024 Zattara E. 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.



Eduardo Zattara

Instituto de Investigaciones en Biodiversidad y Medioambiente, CONICET-Universidad Nacional del Comahue, Bariloche, Argentina

This short and concise data note presents the genome assembly, annotation and preliminary content analysis for a species of mining bee (Andrenidae). It is clearly written, and adequately justifies the rationale for generating the data, and describes the currently standard benchmarks for assembly quality and content. The data are valuable and should be instrumental both in comparative studies of bee genomes and research on this specific species of bee.

My only observation is that the authors might want to deposit an annotated version of the specific scripts and commands used throughout their assembly and analytical pipelines, and also during the curation process, both to allow for full replicability of the study, but also (and perhaps, more important) to provide other researchers with a template they can use to analyse their own datasets. Such custom code is often hosted at sites like Zenodo or GitHub, which ideally allow version control, support updates and documentation.

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?

Partly

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

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Insect Ecological 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.

Reviewer Report 01 August 2024

<https://doi.org/10.21956/wellcomeopenres.23238.r88900>

© 2024 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 

Departamento de Genética e Biologia Evolutiva – Instituto de Biociências, Universidade de São Paulo, São Paulo, Brazil

The article presents data on the assembly of the genome of the bee species *Andrena bucephala*, which is widely distributed throughout Europe. The main result is a final assembly comprising 379.8 Mb, distributed across 13 scaffolds, with five scaffolds accounting for 99.93% of the assembled sequence, in addition to a 19.57 Kb mitogenome. The assembly metrics indicate high quality (QV) and completeness (k-mer completeness and BUSCO). The methods employed were appropriate, utilizing HiFi long reads in conjunction with a Hi-C map to achieve chromosome-scale scaffolds. The protocols are well-described or referenced to ensure reproducibility. Overall, the manuscript is written clearly and objectively, presenting highly relevant information for research on the evolution of hymenopteran species, particularly bees. I believe it is suitable for indexing in its current format.

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, genomics, and evolution 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.