DATA NOTE

The genome sequence of the ferruginous bee-grabber, *Sicus ferrugineus* (Linnaeus, 1761) [version 1; peer review: 2 approved]

University of Oxford and Wytham Woods Genome Acquisition Lab, Darwin Tree of Life Barcoding collective, Wellcome Sanger Institute Tree of Life programme, Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective, Tree of Life Core Informatics collective, Will Nash, Darwin Tree of Life Consortium

1Earlham Institute, Norwich, UK

**Abstract**

We present a genome assembly from an individual male *Sicus ferrugineus* (the ferruginous bee-grabber; Arthropoda; Insecta; Diptera; Conopidae). The genome sequence is 312 megabases in span. The majority of the assembly (99.67%) is scaffolded into 5 chromosomal pseudomolecules, with the X and Y sex chromosomes assembled. The complete mitochondrial genome was also assembled and is 16.9 kilobases in length.

**Keywords**

Sicus ferrugineus, ferruginous bee-grabber, genome sequence, chromosomal, Diptera

This article is included in the Tree of Life gateway.
Species taxonomy
Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta; Pterygota; Neoptera; Endopterygota; Diptera; Brachycera; Muscomorpha; Conopoidea; Conopidae; Myopinae; Sicus; *Sicus ferrugineus* (Linnaeus, 1761) (NCBI:txid1219204).

Background
The ferruginous bee-grabber, *Sicus ferrugineus* (Linnaeus 1761), is the commonest and most abundant member of the dipteran family Conopidae (thick-headed flies) found in the British Isles (Conopid Recording Scheme of Britain & Ireland, personal communication). Widespread throughout Europe (Stuke, 2017), these enigmatic flies inhabit grassland, woodland, hedgerow and garden habitats. Often seen resting on and around flowering plants, on which they feed, adults hold their elongated abdomens curled under the body (Smith, 1969). At such food sources, *S. ferrugineus* females can readily be seen ‘grabbing’ several species of bumblebee (*Bombus* spp.) both in the air and on surfaces (Schmid-Hempel & Schmid-Hempel, 1996b). *S. ferrugineus* is an endoparasite of these bees, and these lunging ‘grabs’ are usually the point of egg delivery. Female *S. ferrugineus* bear specialised abdominal structures used to seize and inject a single egg into the body cavity of the target bee. These include the theca, a grasping structure under sternite 5. The theca is notably smaller in *S. ferrugineus* than that of the only other recorded British *Sicus* species, *S. abdominalis* (Smith, 1969), and is the primary morphological structure used for identifying species of this genus. *S. ferrugineus* eggs bear a hooked micropyle (Kotrba, 2011; Smith, 1969) and, once hatched, the resulting larva feeds on the haemolymph of the host, reaching pupal stage in around 11 days (Schmid-Hempel & Schmid-Hempel, 1996a). The high-quality *S. ferrugineus* reference genome presented here is the first full genome sequence of a conopid fly and presents a unique opportunity to better understand the fascinating parasitic ecology of this species.

Genome sequence report
The genome was sequenced from a single male *S. ferrugineus* collected from Wytham Great Wood, Oxfordshire (Biological vice-county: Berkshire), UK (latitude 51.770, longitude -1.339) (Figure 1). A total of 49-fold coverage in Pacific Biosciences single-molecule HiFi long reads and 83-fold coverage in 10X Genomics read clouds were generated. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data. Manual assembly curation corrected 103 missing/misjoins, reducing the assembly size by 0.16% and the scaffold number by 65.88%, and increasing the scaffold N50 by 86.05%.

The final assembly has a total length of 312 Mb in 29 sequence scaffolds with a scaffold N50 of 44.9 Mb (Table 1).
The majority, 99.67%, of the assembly sequence was assigned to 7 chromosomal-level scaffolds, representing 5 autosomes (numbered by sequence length), and the X and Y sex chromosomes (Figure 2–Figure 5; Table 2). Orientation and location of some pieces of heterochromatic repeat is less clear than for other regions of the assembly, particularly for heterochromatic regions in chromosome 2. The assembly contains many regions of pentameric repeat that cause problems with Hi-C mapping and are visible as drops in association.

The assembly has a BUSCO v5.1.2 (Manni et al., 2021) completeness of 95.4% (single 94.3%, duplicated 1.0%) using the diptera_odb10 reference set (n=3285). While not fully phased, the assembly deposited is of one haplotype. Contigs corresponding to the second haplotype have also been deposited.

### Methods

#### Sample acquisition and DNA extraction

One male *S. ferrugineus* sample, idSicFerr1, was collected from Wytham Great Wood, Oxfordshire, (Biological vice-county: Berkshire), UK (latitude 51.770, longitude -1.339) by Liam Crowley, University of Oxford, on 15 June 2020. The specimen was caught in grassland with a net, identified by the same individual, snap-frozen on dry ice and stored using a CoolRack.

DNA was extracted at the Tree of Life laboratory, Wellcome Sanger Institute. The idSicFerr1 sample was weighed and dissected on dry ice with tissue set aside for Hi-C sequencing, thorax/abdomen tissue was cryogenically disrupted to a fine powder using a Covaris cryoPREP Automated Dry Pulveriser, receiving multiple impacts. Fragment size analysis of
Figure 2. Genome assembly of *Sicus ferrugineus*, idSicFerr1.1: 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 311,933,319 bp assembly. The distribution of chromosome lengths is shown in dark grey with the plot radius scaled to the longest chromosome present in the assembly (76,899,290 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 chromosome lengths (44,927,291 and 26,327,110 bp), respectively. The pale grey spiral shows the cumulative chromosome 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 diptera_odb10 set is shown in the top right. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/idSicFerr1.1/dataset/CAKLPJ01/snail.
0.01-0.5 ng of DNA was then performed using an Agilent FemtoPulse. High molecular weight (HMW) DNA was extracted using the Qiagen MagAttract HMW DNA extraction kit. Low molecular weight DNA was removed from a 200-ng aliquot of extracted DNA using 0.8X AMPure XP purification kit prior to 10X Chromium sequencing; a minimum of 50 ng DNA was submitted for 10X sequencing. HMW DNA was sheared into an average fragment size between 12–20 kb in a Megaruptor 3 system with speed setting 30. Sheared DNA was purified by solid-phase reversible immobilisation using AMPure PB beads with a 1.8X ratio of beads to sample to remove the shorter fragments and concentrate the DNA sample. 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.
Figure 4. Genome assembly of *Sicus ferrugineus*, idSicFerr1.1: cumulative sequence. 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/idSicFerr1.1/dataset/CAKLPJ01/cumulative.

**Sequencing**

Pacific Biosciences HiFi circular consensus and 10X Genomics Chromium read cloud sequencing libraries were constructed according to the manufacturers’ instructions. Sequencing was performed by the Scientific Operations core at the Wellcome Sanger Institute on Pacific Biosciences SEQUEL II (HiFi) and Illumina HiSeq X (10X) instruments. Hi-C data were generated in the Tree of Life laboratory from head tissue of idSicFerr1 using the Arima v2 kit and sequenced on a HiSeq X instrument.

**Genome assembly**

Assembly was carried out with Hifiasm (*Cheng et al.,* 2021); haplotypic duplication was identified and removed with
**Figure 5.** Genome assembly of *Sicus ferrugineus*, idSicFerr1.1: Hi-C contact map. Hi-C contact map of the idSicFerr1.1 assembly, visualised in HiGlass. Chromosomes are arranged in size order from left to right and top to bottom.

**Table 2.** Chromosomal pseudomolecules in the genome assembly of *Sicus ferrugineus*, idSicFerr1.1.

| INSDC accession | Chromosome | Size (Mb) | GC%  |
|-----------------|------------|-----------|------|
| OV277348.1      | 1          | 76.90     | 41.5 |
| OV277349.1      | 2          | 67.51     | 40.8 |
| OV277350.1      | 3          | 44.93     | 40.3 |
| OV277352.1      | 4          | 39.27     | 40.4 |
| OV277354.1      | 5          | 15.24     | 39.2 |
| OV277351.1      | X          | 40.60     | 41.3 |
| OV277353.1      | Y          | 26.33     | 37.5 |
| OV277355.1      | MT         | 0.02      | 27.1 |
| -               | Unplaced   | 1.14      | 45.3 |
purge_dups (Guan et al., 2020). One round of polishing was performed by aligning 10X Genomics read data to the assembly with longranger align, calling variants with freebayes (Garrison & Marth, 2012). The assembly was then scaffolded with Hi-C data (Rao et al., 2014) using SALSA2 (Ghurye et al., 2019). The assembly was checked for contamination as described previously (Howe et al., 2021). Manual curation was performed using HiGlass (Kerpedjiev et al., 2018) and Pretext. The mitochondrial genome was assembled using MitoHiFi (Uliano-Silva et al., 2021), which performs annotation using MitoFinder (Allio et al., 2020). The genome was analysed and BUSCO scores generated within the BlobToolKit environment (Challis et al., 2020). Table 3 contains a list of all software tool versions used, where appropriate.

### Ethics/compliance issues

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. 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. 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: Sicus ferrugineus (ferruginous bumble-grabber). Accession number PRJEB48117; https://identifiers.org/ena.embl/PRJEB48117.

The genome sequence is released openly for reuse. The *S. ferrugineus* 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. The genome will be annotated and presented through the Ensembl pipeline at the European Bioinformatics Institute. 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.5746938.

Members of the Darwin Tree of Life Barcoding collective are listed here: https://doi.org/10.5281/zenodo.5744972.

Members of the Wellcome Sanger Institute Tree of Life programme are listed here: https://doi.org/10.5281/zenodo.6125027.

Members of Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective are listed here: https://doi.org/10.5281/zenodo.5746904.

Members of the Tree of Life Core Informatics collective are listed here: https://doi.org/10.5281/zenodo.6125046.

Members of the Darwin Tree of Life Consortium are listed here: https://doi.org/10.5281/zenodo.5638618.

### Acknowledgments

We would like to thank David Clements for useful comments on the introductory text.

### Table 3. Software tools used.

| Software tool | Version | Source |
|---------------|---------|--------|
| Hifiasm       | 0.15.3  | Cheng et al., 2021 |
| purge_dups    | 1.2.3   | Guan et al., 2020 |
| SALSA2        | 2.2     | Ghurye et al., 2019 |
| longranger align | 2.2.2 | [https://support.10xgenomics.com/genome-exome/software/pipelines/latest/advanced/other-pipelines](https://support.10xgenomics.com/genome-exome/software/pipelines/latest/advanced/other-pipelines) |
| freebayes     | 1.3.1-17-gaa2ace8 | Garrison & Marth, 2012 |
| MitoHiFi      | 2.0     | Uliano-Silva et al., 2021 |
| HiGlass       | 1.11.6  | Kerpedjiev et al., 2018 |
| PretextView   | 0.2.x   | [https://github.com/wtsi-hpag/PretextView](https://github.com/wtsi-hpag/PretextView) |
| BlobToolKit   | 2.6.4   | Challis et al., 2020 |

### 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](https://pubmed.ncbi.nlm.nih.gov/32008501/) | [Publisher Full Text](https://doi.org/10.1111/1755-0998.13201) | [Free Full Text](https://onlinelibrary.wiley.com/doi/10.1111/1755-0998.13201)

Challis R, Richards E, Rajan J, et al.: BlobToolKit - Interactive Quality Assessment of Genome Assemblies. G3 (Bethesda). 2020; 10(4): 1361–74. [PubMed Abstract](https://pubmed.ncbi.nlm.nih.gov/33020020/) | [Publisher Full Text](https://doi.org/10.1534/g3.120.203582) | [Free Full Text](https://doi.org/10.1534/g3.120.203582)

Cheng H, Concepcion GT, Feng X, et al.: Haplotype-Resolved de Novo Assembly Using Phased Assembly Graphs with Hifiasm. Nat Methods. 2021; 18(2): 170–75. [PubMed Abstract](https://pubmed.ncbi.nlm.nih.gov/32896726/) | [Publisher Full Text](https://doi.org/10.1038/s41592-020-0944-8) | [Free Full Text](https://doi.org/10.1038/s41592-020-0944-8)
Open Peer Review

Current Peer Review Status: ✅ ✅

Version 1

Reviewer Report 07 June 2023

https://doi.org/10.21956/wellcomeopenres.19642.r58063

© 2023 Fuhrmann N. 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.

Nico Fuhrmann
Biogeography, University of Trier, Trier, Rhineland-Palatinate, Germany

The article presents the first chromosome-level genome assembly for the ferruginous bee-grabber, *Sicus ferrugineus*. It is the first publicly available genome sequence of an abundant member of the dipteran family Conopidae. Extracted from a single male individual, the authors chose to sequence the genome on the PacBio SEQUEL II and Illumina HiSeq X. The initial assembly was done with HiFiasm and followed by short-read polishing. Using Hi-C data for the genome scaffolding reduced the scaffold numbers significantly. The resulting assembly consists of five autosomes, the X- and Y-chromosome and the mitochondrion of the species. The quality analyses of the assembly indicates large scaffolds and 95% found complete dipteran BUSCOs. This high-quality genome sequence is available on the nucleotide database ENA. A genome annotation is still missing. However, the authors mention that the genome will be annotated in the future and available through Ensembl.

The authors utilize state-of-the-art methods and shows that high afford results in high-quality genomic sequences. The detailed laboratory performance description makes replication very easy. The authors do provide a list of the utilized software and version numbers (table 3), which is important and helpful information. Publishing the genome annotation alongside the genomic sequence would have been a great addition to the article.

The last sentence of the background section hints towards the importance of the genomic sequence. The authors explain the species biology very well. *S. ferrugineus*’ bumblebee parasitism renders it an important target for biomonitoring and conservation affords of key pollinator species in the ecosystem. The article should directly mention application possibilities of the assembly.

One more open question is about an unplaced fraction of about 1 Mb in size, about 0.3% of the total assembly. It is worth investigating where these fragments originate or belong. A possibility is using BLAST to identify potential contaminations (e.g. endosymbionts, parasites, human). Please add the information if these fragments are not part of the *S. ferrugineus* genome.
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: Evolution and population 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 08 March 2023

https://doi.org/10.21956/wellcomeopenres.19642.r55183

© 2023 Gomulski L. 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.

Ludvik M Gomulski
University of Pavia, Pavia, Italy

The article describes the sequencing and assembly of the first full genome of a conopid fly, the ferruginous bee-grabber, *Sicus ferrugineus*. The species is an endoparasite of bees. The female grabs and injects an egg into a target bee and the resulting larva feeds on the haemolymph of its host. The description of the methods is detailed and clear and the resulting genome covering the 5 autosomes and the X and Y sex chromosomes appears to be of excellent quality. Both the genome sequence and the complete mitochondrial genome were assembled. The genome sequence has not been annotated.

The species author is missing for *Sicus abdominalis*.

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:** Insect functional 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.