DATA NOTE

The genome sequence of the yellow-legged clearwing, *Synanthedon vespiformis* (Linnaeus, 1761) [version 1; peer review: 2 approved]

Douglas Boyes1+, 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, David Lees2, Darwin Tree of Life Consortium

1UK Centre for Ecology and Hydrology, Wallingford, Oxfordshire, UK
2Natural History Museum, London, UK

* Deceased author

First published: 14 Sep 2022, 7:233
https://doi.org/10.12688/wellcomeopenres.18109.1

Open Peer Review

Approval Status

1
version 1
14 Sep 2022
view

2
view

1. Jean-Michel Drezen1, CNRS - Université de Tours, Tours, France

2. Shuai Zhan1, University of Chinese Academy of Sciences, Beijing, China

Any reports and responses or comments on the article can be found at the end of the article.

Abstract
We present a genome assembly from an individual male *Synanthedon vespiformis* (the yellow-legged clearwing; Arthropoda; Insecta; Lepidoptera; Sesiidae). The genome sequence is 287 megabases in span. Of the assembly, 100% is scaffolded into 31 chromosomal pseudomolecules with the Z sex chromosome assembled. The complete mitochondrial genome was also assembled and is 17.3 kilobases in length.

Keywords
Synanthedon vespiformis, yellow-legged clearwing, genome sequence, chromosomal, Lepidoptera

This article is included in the Tree of Life gateway.
Species taxonomy
Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta; Pterygota; Neoptera; Endopterygota; Lepidoptera; Glossata; Ditrysia; Sesioida; Sesiidae; Sesiinae; Synanthedonini; Synanthedon; Synanthedon vespiformis (Linnaeus, 1761) (NCBI: txid1660703).

Background
The yellow-legged clearwing, Synanthedon vespiformis (Linnaeus, 1761), is a day flying, clearwing moth belonging to the family Sesiidae. Adults exhibit wasp mimicry, as with many others in the Sesiidae family. It is widespread in the Palearctic, including England and eastern Wales, but its range does not extend to the north of the British Isles or Ireland (Randle et al., 2019). Moths fly in April to September and were not historically believed to be abundant, with the species considered nationally scarce in the UK and rare in Sweden (Burman et al., 2016). More recently, research using pheromone lures indicates that it is much more widespread and easily observed in many countries (Burman et al., 2016). Despite this, there is limited evidence its distribution may have decreased in the UK since 1970, when accounting for increased sampling (Randle et al., 2019).

S. vespiformis is characterised as a woodland species, principally inhabiting the host-plant genus Quercus in northern latitudes, as well as other genera such as Populus, Aesculus and Salix (“Synanthedon Vespiformis (Linnaeus, 1761)” n.d.). The larvae are internal feeders, as with other Sesiidae, and are considered forestry pests in many of its southern ranges, including of the European chestnut, Castanea sativa, native to Turkey (Ülgentürk & Dokuyucu, 2019). S. vespiformis DNA has not previously been barcoded in the UK and the assembly could be used for further research into haplotype diversity from European DNA barcodes (BOLD:AAD7411). Loci could be evaluated for known traits, such as wasp mimicry.

Genome sequence report
The genome was sequenced from a single male S. vespiformis collected from Wytham Woods, Berkshire, UK (Figure 1). A total of 94-fold coverage in Pacific Biosciences single-molecule HiFi long reads and 129-fold coverage in 10X Genomics read clouds were generated. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data. Manual assembly curation corrected 6 missing/misjoins, reducing the assembly size by 0.01% and the scaffold number by 13.89%, and increasing the scaffold N50 by 1.03%.

The final assembly has a total length of 287 Mb in 31 sequence scaffolds with a scaffold N50 of 1.03%.

| Project accession data |  |
|------------------------|-----------------|
| Assembly identifier     | ilSynVesp1.1     |
| Species                | Synanthedon vespiformis |
| Specimen               | ilSynVesp1 (genome assembly, Hi-C) |
| NCBI taxonomy ID       | 1660703          |
| BioProject             | PRJEB46854       |
| BioSample ID           | SAMEA7701494     |

| Isolate information    | Male. Abdomen tissue (genome assembly); head/thorax tissue (Hi-C) |

| Raw data accessions    |  |
|------------------------|-----------------|
| PacificBiosciences SEQUEL II | ERR6939257 |
| 10X Genomics Illumina   | ERR6688650-ERR6688653 |
| Hi-C Illumina          | ERR6688654      |

| Genome assembly        |  |
|------------------------|-----------------|
| Assembly accession     | GCA_918317495.1 |
| Accession of alternate haplotype | GCA_918305855.1 |
| Span (Mb)              | 287             |
| Number of contigs      | 39              |
| Contig N50 length (Mb) | 10.3            |
| Number of scaffolds    | 31              |
| Scaffold N50 length (Mb)| 10.4            |
| Longest scaffold (Mb)  | 12.8            |
| BUSCO* genome score    | C:98.0%; [S: 97.4%; D: 0.6%]; F: 0.4%; M: 1.5%; n: 5,286 |

*BUSCO scores based on the lepidoptera_odb10 BUSCO set using v5.3.2. 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/IlSynVesp1.1/dataset/ilSynVesp1.1.busco.

Figure 1. Image of the Synanthedon vespiformis specimen taken prior to preservation and processing.

Table 1. Genome data for Synanthedon vespiformis, iILVesp1.1.
Figure 2. Genome assembly of *Synanthedon vespiformis*, ilSynVesp1.1: 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 287,400,900 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 (12,814,766 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 chromosome lengths (10,410,606 and 6,963,860 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 lepidoptera_odb10 set is shown in the top right. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/ilSynVesp1.1/dataset/ilSynVesp1_1/snail.

The assembly has a BUSCO v5.3.2 (Manni et al., 2021) completeness of 98.0% (single 97.4%, duplicated 0.6%) using the lepidoptera_odb10 reference set (n=5,286). 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 nucleic acid extraction**

A single male *S. vespiformis* specimen (ilSynVesp1) was collected using a light trap from Wytham Woods, Berkshire, UK (latitude 51.772, longitude -1.338) by Douglas Boyes (University of Oxford). The specimen was identified by Douglas Boyes and snap-frozen on dry ice.

DNA was extracted at the Tree of Life laboratory, Wellcome Sanger Institute. The ilSynVesp1 sample was weighed and dissected on dry ice with tissue set aside for Hi-C sequencing. Abdomen tissue was disrupted using a Nippi Powermasher fitted with a BioMasher pestle. Fragment size analysis of 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.
Figure 5. Genome assembly of Synanthedon vespiformis, iISynVesp1.1: Hi-C contact map. Hi-C contact map of the iISynVesp1.1 assembly, visualised in HiGlass. Chromosomes are arranged in size order from left to right and top to bottom. The interactive Hi-C map can be viewed at https://genome-note-higlass.tol.sanger.ac.uk/l/?d=LL7j52AQYegMKQJa8JRmg.

Table 2. Chromosomal pseudomolecules in the genome assembly of Synanthedon vespiformis, iISynVesp1.1.

| INSDC accession | Chromosome | Size (Mb) | GC%  |
|-----------------|------------|-----------|------|
| OU906945.1      | 1          | 12.81     | 34.2 |
| OU906946.1      | 2          | 12.71     | 34.5 |
| OU906948.1      | 3          | 12.11     | 34.6 |
| OU906949.1      | 4          | 11.96     | 34   |
| OU906950.1      | 5          | 11.92     | 34.5 |
| OU906951.1      | 6          | 11.64     | 33.8 |
| OU906952.1      | 7          | 11.31     | 34.2 |
| OU906953.1      | 8          | 11.06     | 33.8 |
| OU906954.1      | 9          | 11.06     | 33.9 |
| OU906955.1      | 10         | 10.74     | 34.1 |
| OU906956.1      | 11         | 10.66     | 34.6 |
| OU906957.1      | 12         | 10.41     | 34.4 |
| OU906958.1      | 13         | 10.3      | 34.2 |
| OU906959.1      | 14         | 10.26     | 34.2 |
| OU906960.1      | 15         | 9.89      | 34.3 |
| OU906961.1      | 16         | 9.89      | 34.2 |
| OU906962.1      | 17         | 9.59      | 34.8 |
| OU906963.1      | 18         | 9.19      | 34.3 |
| OU906964.1      | 19         | 8.65      | 34.2 |
| OU906965.1      | 20         | 8.33      | 34.6 |
| OU906966.1      | 21         | 8.1       | 34.7 |
| OU906967.1      | 22         | 7.06      | 34.4 |
| OU906968.1      | 23         | 7.04      | 35.3 |
| OU906969.1      | 24         | 7         | 34.6 |
| OU906970.1      | 25         | 6.96      | 35.2 |
| OU906971.1      | 26         | 6.11      | 35.1 |
| OU906972.1      | 27         | 5.19      | 35.6 |
| OU906973.1      | 28         | 4.56      | 35.7 |
| OU906974.1      | 29         | 4.28      | 36.8 |
| OU906975.1      | 30         | 4.07      | 36.6 |
| OU906947.1      | Z          | 12.51     | 34.1 |
| OU906976.1      | MT         | 0.02      | 20.2 |
Fluorometer and Qubit dsDNA High Sensitivity Assay kit. Fragment size distribution was evaluated by running the sample on the FemtoPulse system.

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 NovaSeq 6000 (10X) instruments. Hi-C data were generated in the Tree of Life laboratory from head and thorax tissue of iSynVesp1 using the Arima v2 kit and sequenced on a NovaSeq 6000 instrument.

Genome assembly
Assembly was carried out with Hifiasm (Cheng et al., 2021); haplotypic duplication was identified and removed with 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.

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 |
| 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 |
| BlobToolKit   | 3.2.6   | Challis et al., 2020 |

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: Synanthedon vespiformis (yellow-legged clearwing). Accession number PRJEB46854; https://identifiers.org/ena.embl/PRJEB46854.

The genome sequence is released openly for reuse. The S. vespiformis 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.6418202.

Members of the Darwin Tree of Life Barcoding collective are listed here: https://doi.org/10.5281/zenodo.6418156.

Members of the Wellcome Sanger Institute Tree of Life programme are listed here: https://doi.org/10.5281/zenodo.6866293.

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.6418363.
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
Burman J, Westerberg L, Ostrow S, et al.: Revealing Hidden Species Distribution with Pheromones: The Case of Synanthedon Vesiformis (Lepidoptera: Sesiidae) in Sweden. J Insect Conserv. 2016; 1(20): 11–21. Publisher Full Text
Challis R, Richards E, Rajan J, et al.: BlobToolKit-Interactive Quality Assessment of Genome Assemblies. G3 (Bethesda). 2020; 10(4): 1361–74. 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): 170–75. PubMed Abstract | Publisher Full Text | Free Full Text
Garrison E, Marth G: Haplotype-Based Variant Detection from Short-Read Sequencing. 2012; arXiv: 1207.3907.
Garrison E, Marth G: Integrating Hi-C Links with Assembly Graphs for Chromosome-Scale Assembly. PloS Comput Biol. 2019; 15(8): e1007273. PubMed Abstract | Publisher Full Text | Free Full Text
Gui D, McCarthy SA, Wood J, et al.: Identifying and Removing Haplotypic Duplication in Primary Genome Assemblies. Bioinformatics. 2020; 36(9): 2896-98. PubMed Abstract | Publisher Full Text | Free Full Text
Howe K, Chow W, Collins J, et al.: Significantly Improving the Quality of Genome Assemblies through Curation. GigaScience. 2021; 10(1): giaa153. PubMed Abstract | Publisher Full Text | Free Full Text
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, Seppmy 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): 4647-54. PubMed Abstract | Publisher Full Text | Free Full Text
Randle Z, Evans-Hill LJ, Parsons MS, et al.: Atlas of Britain & Ireland's Larger Moths. Pisces Publications, 2019. Reference Source
Rao SS, 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): 1665-80. PubMed Abstract | Publisher Full Text | Free Full Text
Synanthedon Vesiformis (Linnaeus, 1761). n.d. NBN Atlas. Accessed July 17, 2022. Reference Source
Ulgentürk S, Dokuyucu O: Pest Species of Coccoidea (Hemiptera; Coccomorpha) in Forest of Turkey. Turkish Journal of Forestry. 2019; 20(4): 482-91. Publisher Full Text
Uliano-Silva M, Nunes JGF, Krasheninnikova K, et al.: marcelauliano/MitoHiFi: mitohifi_v2.0. 2021. Publisher Full Text
Open Peer Review

Current Peer Review Status:  

Version 1

Reviewer Report 06 March 2023

https://doi.org/10.21956/wellcomeopenres.20080.r55144

© 2023 Zhan S. 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.

Shuai Zhan
University of Chinese Academy of Sciences,, Beijing, China

The authors present a chromosomal-level genome assembly for the yellow-legged clearwing Synanthedon vespiformis. Presented features support that the genome reference is relatively complete in terms of gene content. The manuscript was written clearly. I have minor concerns for the authors to improve the manuscript.

1. The method for genome assembly was not fully described. Detailed commands and parameters should be provided. More importantly, how the genome assembly was manually corrected should be described rather than linking to a previous manuscript.

2. It's unclear to me how Chr.Z was assigned. Based on the sequencing ratio between male and female, or the cross-species synteny? Related methods and evidence should be provided.

3. (Optional) I note that the reported genome reference has not been annotated. A fully-annotated reference genome would be much more helpful for the community.

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

Are the protocols appropriate and is the work technically sound?  
Yes

Are sufficient details of methods and materials provided to allow replication by others?  
Partly

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

Competing Interests: No competing interests were disclosed.
Reviewer Expertise: 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 23 February 2023

https://doi.org/10.21956/wellcomeopenres.20080.r54250

© 2023 Drezen 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.

Jean-Michel Drezen
CNRS - Université de Tours,, Tours, France

The genome report of the Lepidoptera synanthedon vespiformis begins with an introduction describing the natural history of the species, followed by technical aspects of genome sequencing performed to obtain high-quality genome chromosome scale assembly as in other data notes from the Darwinian tree of life project. The introduction is well written and will be of interest for a broad audience, whereas the detailed protocol and figures will be understood by specialists in genomics. The genome obtained using a combination of the best available approaches is of high quality and at chromosomal scale. The values assessing quality are all very good. This genome will provide a strong reference for future studies provided that some effort is put on the study of the genetic of mimicry that makes this and related lepidopteran species look like wasps which is quite fascinating. It will also be of interest to check for horizontal transfers of genes from their parasites, that have been shown to impact many lepidopteran genomes1,2.

By searching on the internet to find information potentially missing I did not find much to add: the scientific literature is scarce on this species. The only criticism I can provide is on the rather obscure sentence on whether the species distribution may have decreased in the UK since 1970 since as stated the sampling method now using pheromones is completely different from that used for old records. The reference cited for this sentence is not "open access" which prevent a better understanding of this point. Concerning the pest status one can also mention that it appears to be a pest also for stone fruits in Israel3, which is probably one of the reasons why pheromone lures have been developed. On the other hand, the adult has probably a beneficial role as pollinator, as most Lepidoptera, as suggested by many pictures showing them foraging on flowers, but I did not find any reference on this potential role. There might be also sexual dimorphism interesting to mention but again I did not find a reference on that topic.

References
1. Gasmi L, Boulain H, Gauthier J, Hua-Van A, et al.: Recurrent Domestication by Lepidoptera of Genes from Their Parasites Mediated by Bracoviruses. PLoS Genet. 2015; 11 (9): e1005470 PubMed Abstract | Publisher Full Text
2. Heisserer C, Muller H, Jouan V, Musset K, et al.: Somatic chromosomal integration of polydnavirus during parasitism triggered their germline infiltration in multiple lepidopteran families. bioRxiv. 2022. Publisher Full Text
3. Levi-Zada A, Ben-Yehuda S, Dunkelblum E, Gindin G, et al.: Identification and field bioassays of the sex pheromone of the yellow-legged clearwing Synanthedon vespiformis (Lepidoptera: Sesiidae). Chemoecology. 2011; 21 (4): 227-233 Publisher Full Text

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

**Are the protocols appropriate and is the work technically sound?**
Yes

**Are sufficient details of methods and materials provided to allow replication by others?**
Yes

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

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** wasp genomics, insect viruses, horizontal transfer

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.