DATA NOTE

The genome sequence of the wood white butterfly, *Leptidea sinapis* (Linnaeus, 1758) [version 1; peer review: 2 approved]

Konrad Lohse, Lars Höök, Karin Näsvall, Niclas Backström, Wellcome Sanger Institute Tree of Life programme, Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective, Tree of Life Core Informatics collective, Darwin Tree of Life Consortium

1Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK
2Evolutionary Biology Program, Department of Ecology and Genetics (IEG), Uppsala University, Uppsala, Sweden

**Abstract**
We present a genome assembly from an individual male *Leptidea sinapis* (the wood white; Arthropoda; Insecta; Lepidoptera; Pieridae). The genome sequence is 686 megabases in span. The majority (99.99%) of the assembly is scaffolded into 48 chromosomal pseudomolecules, with three Z sex chromosomes assembled. Gene annotation of this assembly on Ensembl has identified 14,800 protein coding genes.

**Keywords**
Leptidea sinapis, wood white, butterfly, 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; Papilionoidea; Pieridae; Dismorphiinae; *Leptidea; Leptidea sinapis* (Linnaeus, 1758) (NCBI:txid189913).

**Background**

The wood white butterfly (*Leptidea sinapis*) is recognized by its white wings with dark apical spots on the forewings and its distinctively slow flight (Thomas & Lewington, 2016). The preferred habitats are forest openings and meadows where herbaceous host plants from the family Fabaceae are present (Friberg et al., 2008; Wiklund, 1977). The distribution range covers a major part of the western Palearctic, the African continent excluded. Within Britain and Ireland, wood whites are restricted to fragmented, sheltered areas in southern Wales and England and a small region around Burren in western Ireland (Thomas & Lewington, 2016). As a consequence of considerable population declines over the last decades, the wood white was included in the UK Biodiversity Action Plan in 2007, but the species has likely been under-surveyed (Jefcoate & Joy, 2011).

The wood white has long been the subject of ecological studies, investigating, for example, interaction with recently discovered cryptic and sympatric sister species and habitat preference variation (Friberg et al., 2008; Friberg & Wiklund, 2009; Wiklund, 1977). Due to the presence of a striking chromosome number cline across the distribution range (Dincă et al., 2011; Lukhtanov et al., 2020), the wood white has also developed into a model species for understanding the mechanistic underpinnings and evolutionary consequences of rapid karyotype evolution (Lukhtanov et al., 2020; Šíchová et al., 2015; Talla et al., 2019). Previous genomic and cytogenetic research have revealed a drastically expanded and unusually repeat-rich genome compared to most studied butterflies (Talla et al., 2017), and the presence of an unexpected sex-chromosome system (Šíchová et al., 2015). Existing genomic resources have also paved way for investigating, for example, the genetic basis of local adaptation (Leal et al., 2018; Násvali et al., 2021) and expression dynamics of sex-linked and autosomal genes (Höök et al., 2019). We foresee that the Darwin Tree of Life assembly presented here will be an important tool for forthcoming research on chromosome number dynamics, the association between structural rearrangements and reproductive isolation, the genetic basis of adaptive traits and the mechanistic underpinnings of microevolutionary processes in butterflies.

**Genome sequence report**

The genome was sequenced from a single male *L. sinapis* (*Leptidea sinapis*; ilLepSina1) from Asturias, Spain, used to generate Pacific Biosciences, 10X genomics and Hi-C data. The final assembly has a total length of 686 Mb in 49 sequence scaffolds with a scaffold N50 of 14.4 Mb (Table 1). The

**Table 1. Genome data for Leptidea sinapis, ilLepSina1.1.**

| **Project accession data** |          |
|---------------------------|----------|
| Assembly identifier       | ilLepSina1.1 |
| Species                   | *Leptidea sinapis* |
| Specimen                  | ilLepSina1 (genome assembly, Hi-C); ilLepSina2 (RNA-Seq) |
| NCBI taxonomy ID          | 189913 |
| BioProject                | PRJEB43801 |
| BioSample ID              | SAMEA7523467 |
| Isolate information       | Male, whole organism (ilLepSina1; ilLepSina2) |

| **Raw data accessions** |          |
|-------------------------|----------|
| PacificBiosciences SEQUEL II | ERR6565941 |
| 10X Genomics Illumina    | ERR6054631-ERR6054634 |
| Hi-C Illumina           | ERR6054635 |
| PolyA RNA-Seq Illumina  | ERR6054636 |

| **Genome assembly** |          |
|---------------------|----------|
| Assembly accession  | GCA_905404315.1 |
| Accession of alternate haplotype | GCA_905404105.1 |
| Span (Mb)            | 686 |
| Number of contigs    | 62 |
| Contig N50 length (Mb) | 13.7 |
| Number of scaffolds  | 49 |
| Scaffold N50 length (Mb) | 14.4 |
| Longest scaffold (Mb) | 34.3 |

| BUSCO genome score | C:98.3%, S:97.6%, D:0.7%, F:0.5%, M:1.3%, n:5,268 |

*BUSCO scores based on the lepidoptera_odb10 BUSCO set using v5.1.2. 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/ilLepSina1.1/dataset/CAJQFP01/busco.*

---

**Figure 1. Fore and hind wings of the Leptidea sinapis specimen from which the genome was sequenced.** Dorsal (left) and ventral (right) surface view of wings from specimen SO_LS_389 (ilLepSina1) from Asturias, Spain, used to generate Pacific Biosciences, 10X genomics and Hi-C data.
majority, 99.99%, of the assembly sequence was assigned to 48 chromosomal-level scaffolds, representing 45 autosomes (numbered by sequence length) and three Z sex chromosomes (Figure 2–Figure 5; Table 2). The assembly has a BUSCO v5.1.2 (Manni et al., 2021) completeness of 98.3% (single 97.6%, duplicated 0.7%) 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.

**Genome annotation report**
The ilLepSina1.1 genome was annotated using the Ensembl rapid annotation pipeline (Table 1; https://rapid.ensembl.org/Leptidea_sinapis_GCA_905404315.1). The resulting annotation

![Graph showing scaffold statistics and BUSCO completeness](https://blobtoolkit.genomehubs.org/view/ilLepSina1.1/dataset/CAJQFP01/snail)

**Figure 2.** Genome assembly of *Leptidea sinapis*, ilLepSina1.1: 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 685,599,024 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 (36,552,532 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 chromosome lengths (14,447,461 and 9,623,130 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/ilLepSina1.1/dataset/CAJQFP01/snail.
includes 47,660 transcribed mRNAs from 14,800 protein-coding and 9,624 non-coding genes. There are 2.25 coding transcripts per gene and 8.29 exons per transcript.

Methods
Sample acquisition and nucleic acid extraction
Two male *L. sinapis* specimens (ilLepSina1, genome assembly, Hi-C; ilLepSina2, RNA-Seq) were collected from Somiedo, Pigueces, Asturias, Spain (latitude 43.1489, longitude -6.3127) using a net by Konrad Lohse, University of Edinburgh, who also identified the samples. The samples were frozen at -80°C.

DNA was extracted at the Scientific Operations Core, Wellcome Sanger Institute. The ilLepSina1 sample was weighed and dissected on dry ice with tissue set aside for Hi-C sequencing. Whole organism tissue was disrupted by manual grinding with a disposable 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 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 whole organism tissue of ilLepSina2 in the Tree of Life Laboratory at the WSI using TRIzol, according to the manufacturer’s instructions. RNA was then eluted in 50 μl RNase-free water and the RNA concentration assessed using a Nanodrop spectrophotometer and Qubit

**Figure 4. Genome assembly of Leptidea sinapis, ilLepSina1.1: 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/ilLepSina1.1/dataset/CAJQFP01/cumulative.
Fluorometer using the Qubit RNA Broad-Range (BR) Assay kit. Analysis of the integrity of the RNA was done using Agilent RNA 6000 Pico Kit and Eukaryotic Total RNA assay.

Sequencing
Pacific Biosciences HiFi circular consensus and 10X Genomics read cloud 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), Illumina HiSeq X (10X) and Illumina HiSeq 4000 (RNA-Seq) instruments. Hi-C data were also generated from the whole organism of ilLepSina1 using the Arima v2 Hi-C kit and sequenced on an Illumina 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 and corrected using the gEVAL system (Chow et al., 2016) as described previously (Howe et al., 2021). Manual curation (Howe et al., 2021) was performed using gEVAL, 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.

Genome annotation
The Ensembl gene annotation system (Aken et al., 2016) was used to generate annotation for the Leptidea sinapis assembly (GCA_905404315.1). Annotation was created primarily
Table 2. Chromosomal pseudomolecules in the genome assembly of Leptidea sinapis, ilLepSina1.1.

Chromosomes Z2 and Z3 are listed as chromosomes 2 and 3 with INSDC.

| INSDC accession | Chromosome | Size (Mb) | GC%  |
|-----------------|------------|-----------|------|
| FR990154.1      | 1          | 34.33     | 35.5 |
| FR990155.1      | Z2         | 28.77     | 35.5 |
| FR990156.1      | Z3         | 18.68     | 35.4 |
| FR990157.1      | 4          | 17.20     | 35.6 |
| FR990158.1      | 5          | 16.73     | 35.7 |
| FR990159.1      | 6          | 16.62     | 36.0 |
| FR990160.1      | 7          | 16.18     | 35.9 |
| FR990161.1      | 8          | 15.95     | 35.0 |
| FR990162.1      | 9          | 15.83     | 35.8 |
| FR990163.1      | 10         | 15.81     | 35.4 |
| FR990164.1      | 11         | 15.48     | 35.2 |
| FR990165.1      | 12         | 15.36     | 35.3 |
| FR990166.1      | 13         | 15.14     | 35.8 |
| FR990167.1      | 14         | 14.85     | 36.0 |
| FR990168.1      | 15         | 14.70     | 35.4 |
| FR990169.1      | 16         | 14.65     | 35.9 |
| FR990170.1      | 17         | 14.56     | 35.7 |
| FR990171.1      | 18         | 14.45     | 35.6 |
| FR990172.1      | 19         | 14.27     | 36.6 |
| FR990173.1      | 20         | 14.10     | 35.8 |
| FR990174.1      | 21         | 14.01     | 35.7 |
| FR990175.1      | 22         | 13.72     | 35.5 |
| FR990176.1      | 23         | 13.66     | 36.0 |
| FR990153.1      | Z1         | 36.55     | 34.4 |

Table 3. Software tools used.

| Software tool     | Version | Source                                      |
|-------------------|---------|---------------------------------------------|
| Hifiasm           | 0.12    | 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          | 1       | Uliano-Silva et al., 2021                  |
| HiGlass           | 1.11.6  | Kerpedjiev et al., 2018                    |
| PretextView       | 0.1.x   | https://github.com/wtsi-hpag/PretextView   |
| BlobToolKit       | 2.6.4   | Challis et al., 2020                      |
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).

**Data availability**
European Nucleotide Archive: Leptidea sinapis (wood white).
Accession number PRJEB43801; https://identifiers.org/ena/emb/PRJEB43801 (Wellcome Sanger Institute, 2022)

The genome sequence is released openly for reuse. The L. sinapis 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 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.

**References**

Aken BL, Ayling S, Barrett D, et al.: The Ensembl Gene Annotation System. Database (Oxford). 2016; 2016; baw093.

Published Abstract | Publisher Full Text | Free Full Text

Alio 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.

Published 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): 1361–74.

Published Abstract | Publisher Full Text | Free Full Text

Cheng H, Concepcion GT, Feng X, et al.: Haplootype-Resolved de Nova Assembly Using Phased Assembly Graphs with Hifiasm. Nat Methods. 2021; 18(2): 170–75.

Published Abstract | Publisher Full Text | Free Full Text

Chow W, Brugger K, Caccamo M, et al.: gEVAL - a web-based browser for evaluating genome assemblies. Bioinformatics. 2016; 32(16): 2508–10.

Published Abstract | Publisher Full Text | Free Full Text

Dincă V, Lukhtanov VA, Talavera G, et al.: Unexpected Layers of Cryptic Diversity in Wood White Leptidea Butterflies. Nat Commun. 2011; 2: 324.

Published Abstract | Publisher Full Text | Free Full Text

Friberg M, Bergman M, Kullberg J, et al.: Host Plant Diet Affects Growth and Induces Altered Gene Expression and Microbiome Composition in the Wood White (Leptidea sinapis) Butterfly. Mol Ecol. 2021; 30(2): 499–516.

Published Abstract | Publisher Full Text | Free Full Text

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; 157(7): 1665–80.

Published Abstract | Publisher Full Text | Free Full Text

Šíchová J, Voleníková A, Dincă V, et al.: Dynamic Karyotype Evolution and Unique Sex Determination Systems in Leptidea Wood White Butterflies. BMC Evol Biol. 2015; 15: 89.

Published Abstract | Publisher Full Text | Free Full Text

Talla V, Johansson A, Dincă V, et al.: Lack of Gene Flow: Narrow and Dispersed Differentiation Islands in a Triplet of Leptidea Butterfly Species. Mol Ecol. 2019; 28(16): 3756–70.

Published Abstract | Publisher Full Text | Free Full Text

Thomas J, Lewington R: The Butterflies of Britain and Ireland. Bloomsbury Publishing, 2016.

Reference Source

Ullano-Silva M, Nunes JG, Krasheninnikova K, et al.: marceliaulano/MitoHifi: v2.0.2021. Publisher Full Text

UniProt Consortium: UniProt: A Worldwide Hub of Protein Knowledge. Nucleic Acids Res. 2019; 47(D1): D506–15.

Published Abstract | Publisher Full Text | Free Full Text

Wellcome Sanger Institute: The genome sequence of the wood white butterfly, Leptidea sinapis (Linnaeus, 1758), European Nucleotide Archive [dataset]. 2022; accession number PRJEB43801.

Published Abstract | Publisher Full Text | Free Full Text

Wellcome Sanger Institute: The genome sequence of the wood white butterfly, Leptidea sinapis (Linnaeus, 1758), European Nucleotide Archive [dataset]. 2022; accession number PRJEB43801.

Published Abstract | Publisher Full Text | Free Full Text

Wellcome Sanger Institute: The genome sequence of the wood white butterfly, Leptidea sinapis (Linnaeus, 1758), European Nucleotide Archive [dataset]. 2022; accession number PRJEB43801.
Open Peer Review

Current Peer Review Status: ✔️ ✔️

Version 1

Reviewer Report 03 April 2023

https://doi.org/10.21956/wellcomeopenres.20089.r54367

© 2023 Smolander O. 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.

Olli-Pekka Smolander
Department of Chemistry and Biotechnology, Tallinn University of Technology, Tallinn, Estonia

The authors justify and describe well the production of the genome sequence of wood white butterfly and the related data. The procedure includes information on sampling, DNA isolation, sequencing methods as well as assembly and annotation. The results are described in good manner.

The only point that I feel needs to be made is regarding the used software tools. It would be good if there was mention whether default parameters were used or if there were modifications to those.

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.
Rhys Parry
The University of Queensland, Saint Lucia, Queensland, Australia

The data note presents a genome assembly of a male wood white butterfly, *Leptidea sinapis*, with a genome sequence of 686 megabases from Pacific Biosciences, 10X genomics and Hi-C data.

The assembly is scaffolded into 48 chromosomal pseudomolecules, including three Z sex chromosomes, and gene annotation has identified 14,800 protein-coding genes. The methods are clearly defined, reproducible and appropriate.

Given the species has been the subject of ecological studies and is a model species for understanding karyotype evolution, the genome assembly is expected to be a reasonable genomic resource for future butterfly evolution and adaptation research.

**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:** Virology, genomics, genome assemblies, transcriptomics. Small RNA analysis and annotation.

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.