DATA NOTE

The genome sequence of the small white, *Pieris rapae* (Linnaeus, 1758) [version 1; peer review: awaiting peer review]

Konrad Lohse\textsuperscript{1}, Sam Ebdon\textsuperscript{1}, Roger Vila\textsuperscript{2},
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, Darwin Tree of Life Consortium

\textsuperscript{1}Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK
\textsuperscript{2}Institut de Biologia Evolutiva (CSIC - Universitat Pompeu Fabra), Barcelona, Spain

\textbf{V1} \textbf{First published:} 15 Oct 2021, \textbf{6:273}
https://doi.org/10.12688/wellcomeopenres.17288.1
\textbf{Latest published:} 15 Oct 2021, \textbf{6:273}
https://doi.org/10.12688/wellcomeopenres.17288.1

\textbf{Abstract}
We present a genome assembly from an individual female *Pieris rapae* (the small white; Arthropoda; Insecta; Lepidoptera; Pieridae). The genome sequence is 256 megabases in span. The majority of the assembly is scaffolded into 26 chromosomal pseudomolecules, with the W and Z sex chromosome assembled. Gene annotation of this assembly on Ensembl has identified 12,390 protein coding genes.

\textbf{Keywords}
Pieris rapae, small white, genome sequence, chromosomal

This article is included in the Tree of Life gateway.
Corresponding author: Darwin Tree of Life Consortium (mark.blaxter@sanger.ac.uk)

Author roles: Lohse K: Investigation, Resources, Writing – Review & Editing; Ebdon S: Writing – Original Draft Preparation, Writing – Review & Editing; Vila R: 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). KL is supported by a NERC fellowship (NE/L011522/1) and an ERC grant (ModelGenom Land 757648). SE is supported by a EastBio PhD studentship.

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

Copyright: © 2021 Lohse K 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: Lohse K, Ebdon S, Vila R et al. The genome sequence of the small white, Pieris rapae (Linnaeus, 1758) [version 1; peer review: awaiting peer review] Wellcome Open Research 2021, 6:273

https://doi.org/10.12688/wellcomeopenres.17288.1

First published: 15 Oct 2021, 6:273 https://doi.org/10.12688/wellcomeopenres.17288.1
Species taxonomy
Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta; Pterygota; Neoptera; Endopterygota; Lepidoptera; Glossata; Ditrysia; Papilionoidea; Pieridae; Pierinae; Pieris; Pieris rapae (Linnaeus, 1758) (NCBI:txid64459).

Introduction
*Pieris rapae*, commonly known as the small white or the small cabbage white, is a widespread butterfly found in Europe, north-west Africa, and Asia, as well as North America, Australia, and New Zealand where it has been introduced (and is known as the “imported cabbage worm”). Found throughout the British Isles, this bivoltine butterfly can be seen on the wing from spring until autumn. In warmer localities this species extends its flight period and the number of generations. The caterpillars feed on a range of brassicaceae and overwinter as pupae. *Pieris rapae* has a long history of consuming agricultural crops and has spread as a human commensal (Ryan *et al.*, 2019). Despite recent improvements, overall it has reduced in abundance and occurrence in the UK over the last 50 years (Fox *et al.*, 2015), but it is listed as Least Concern in the IUCN Red List (Europe) (van Swaay *et al.*, 2010). *Pieris rapae* has 25 pairs of chromosomes, a genome size of approximately 245.9 Mb (Shen *et al.*, 2016), and is female heterogametic (WZ). We note the recent production of a high-quality genome assembly for *P. rapae* (Shen *et al.*, 2016), and believe the sequence described here, generated as part of the Darwin Tree of Life project, will further aid understanding of the biology and ecology of this butterfly.

Genome sequence report
The genome was sequenced from a single female *P. rapae* collected from East Linton, East Lothian, Scotland, UK (latitude 55.977161, longitude -2.667545) (Figure 1). A total of 56-fold coverage in Pacific Biosciences single-molecule long reads (N50 14 kb) and 157-fold coverage in 10X Genomics read clouds were generated. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data. Manual assembly curation corrected five missing/misjoins and removed one haplotypic duplication, reducing the scaffold number by 9.30%.

The final assembly has a total length of 256 Mb in 40 sequence scaffolds with a scaffold N50 of 11 Mb (Table 1). Of the assembly sequence, 99.8% was assigned to 26 chromosomal-level scaffolds, representing 24 autosomes (numbered by sequence length), and the W and Z sex chromosome (Figure 2–Figure 5; Table 2). The assembly has a BUSCO (Simão *et al.*, 2015) v5.1.2 completeness of 98.8% (single 98.4%, duplicated 0.4%, fragmented 0.2%, missing 1.0%) using the lepidoptera_odb10 reference set. While not fully phased, the assembly deposited is of one haplotype. Contigs corresponding to the second haplotype have also been deposited.

Gene annotation
The Ensembl gene annotation system (Aken *et al.*, 2016) was used to generate annotation for the *Pieris rapae* assembly (GCA_905147795.1, see https://rapid.ensembl.org/Pieris_rapae_GCA_905147795.1/; Table 1). The 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) and OrthoDB (Kriventseva *et al.*, 2008). Prediction tools, CPC2 (Kang *et al.*, 2017) and RNAsamba (Camargo *et al.*, 2020), were used to aid determination of protein coding genes.

**Figure 1.** Fore and hind wings of *Pieris rapae* specimen from which the genome was sequenced. (A) Dorsal surface view of wings from specimen SC_PR_1370 (iPiPeRapa1) from East Linton, Scotland, UK, used to generate Pacific Biosciences, 10X genomics, Hi-C and RNASeq data. (B) Ventral surface view of wings from specimen SC_PR_1370 (iPiPeRapa1) from East Linton, Scotland, UK, used to generate Pacific Biosciences, 10X genomics, Hi-C and RNASeq data.
Methods

Sample acquisition and nucleic acid extraction

A single female *P. rapae* was collected from East Linton, Scotland (latitude 55.977161, longitude -2.667545) using a net by Konrad Lohse, University of Edinburgh, who also identified the sample. The sample was snap-frozen in liquid nitrogen.

DNA was extracted from the whole organism of ilPieRapa1 at the Wellcome Sanger Institute (WSI) Scientific Operations core from the whole organism using the Qiagen MagAttract HMW DNA kit, according to the manufacturer’s instructions. RNA (also from the whole organism) was extracted in the Tree of Life Laboratory at the WSI using TRIzol, according to Table 1. Genome data for *Pieris rapae*, ilPieRapa1.1.

| Project accession data                          |                   |
|------------------------------------------------|-------------------|
| Assembly identifier                             | ilPieRapa1.1      |
| Species                                        | *Pieris rapae*    |
| Specimen                                       | ilPieRapa1        |
| NCBI taxonomy ID                               | NCBI:txid64459    |
| BioProject                                     | PRJEB42143        |
| BioSample ID                                   | SAMEA7532735      |
| Isolate information                            | Female, whole organism |

| Raw data accessions                            |                   |
|-----------------------------------------------|-------------------|
| Pacific Biosciences SEQUEL II                 | ERR6606787        |
| 10X Genomics Illumina                         | ERR6002744-ERR6002747 |
| Hi-C Illumina                                 | ERR6002748        |
| Illumina PolyA RNAseq                         | ERR6286709        |

| Genome assembly                                |                   |
| Assembly accession                             | GCA_905147795.1   |
| Accession of alternate haplotype               | GCA_905147735.1   |
| Span (Mb)                                      | 256               |
| Number of contigs                              | 44                |
| Contig N50 length (Mb)                         | 10.7              |
| Number of scaffolds                            | 40                |
| Scaffold N50 length (Mb)                       | 10.7              |
| Longest scaffold (Mb)                          | 24                |
| BUSCO* genome score                            | C:98.8%[S:98.4%,D:0.4%],F:0.2%,M:1.0%,n:5286 |

| Gene annotation                                |                   |
| Number of protein coding genes                 | 12,390            |
| Average coding sequence length (bp)            | 3,149             |
| Average number of exons per transcript         | 10                |
| Average exon size (bp)                         | 365               |
| Average intron size (bp)                       | 1,667             |

*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/ilPieRapa1.1/dataset/CAJHWY01/busco.
RNA was then eluted in 50 μl RNAse-free water and its concentration assessed using a Nanodrop spectrophotometer and Qubit 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.

**Figure 2.** Genome assembly of *Pieris rapae*, ilPieRapa1.1: 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 256,370,197 bp assembly. The distribution of scaffold lengths is shown in dark grey with the plot radius scaled to the longest chromosome present in the assembly (12,960,541 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 chromosome lengths (10,651,053 and 8,441,140 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 lepidoptera_odb10 set is shown in the top right. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/ilPieRapa1.1/dataset/CAJHWY01/snail.
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 using the Arima v1.0 kit and sequenced on HiSeq X.

Genome assembly
Assembly was carried out with HiCanu (Nurk et al., 2020). Haplotypic duplication was identified and removed with purge_dups (Guan et al., 2020). One round of polishing was

Figure 3. Genome assembly of *Pieris rapae*, ilPieRapa1.1: GC coverage. 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/ilPieRapa1.1/dataset/CAJHWY01/blob.
Figure 4. Genome assembly of *Pieris rapae*, ilPieRapa1.1: 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.genomeshubs.org/view/ilPieRapa1.1/dataset/CAJHWY01/cumulative.

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 was performed using gEVAL, HiGlass (Kerpedjiev et al., 2018) and Precis. 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.

Ethical/compliance issues
The materials that have contributed to this genome note were supplied by a Tree of Life collaborator. The Wellcome Sanger
**Figure 5.** Genome assembly of *Pieris rapae*, ilPieRapa1.1: Hi-C contact map. Hi-C contact map of the ilPieRapa1.1 assembly, visualised in HiGlass. Chromosomes are given in size order from left to right and top to bottom.

**Table 2.** Chromosomal pseudomolecules in the genome assembly of *Pieris rapae*, ilPieRapa1.1.

| INSDC accession | Chromosome | Size (Mb) | GC%  |
|-----------------|------------|-----------|------|
| LR990582.1      | 1          | 12.96     | 33.2 |
| LR990584.1      | 2          | 11.92     | 33.1 |
| LR990585.1      | 3          | 11.71     | 32.8 |
| LR990586.1      | 4          | 11.41     | 33.4 |
| LR990587.1      | 5          | 11.34     | 33   |
| LR990588.1      | 6          | 11.08     | 32.8 |
| LR990589.1      | 7          | 11.00     | 33.1 |
| LR990590.1      | 8          | 11.00     | 33   |
| LR990591.1      | 9          | 10.90     | 32.7 |
| LR990592.1      | 10         | 10.89     | 33   |
| LR990593.1      | 11         | 10.65     | 33.2 |
| LR990594.1      | 12         | 10.48     | 33.2 |
| LR990595.1      | 13         | 10.42     | 32.9 |
|                 |            |           |      |
| LR990596.1      | 14         | 10.42     | 32.9 |
| LR990597.1      | 15         | 10.32     | 33.1 |
| LR990598.1      | 16         | 10.18     | 33.2 |
| LR990599.1      | 17         | 9.91      | 33.1 |
| LR990600.1      | 18         | 9.38      | 33   |
| LR990601.1      | 19         | 9.22      | 32.8 |
| LR990602.1      | 20         | 8.93      | 33.2 |
| LR990603.1      | 21         | 8.44      | 33.1 |
| LR990604.1      | 22         | 7.59      | 33.5 |
| LR990605.1      | 23         | 5.39      | 34.6 |
| LR990606.1      | 24         | 5.22      | 34.9 |
| LR990607.1      | W          | 3.18      | 38   |
| LR990583.1      | Z          | 12.02     | 33.5 |
| LR990608.1      | MT         | 0.02      | 19.9 |
|                 | -          | 0.40      | 46.7 |
Table 3. Software tools used.

| Software tool   | Version | Source                                      |
|-----------------|---------|---------------------------------------------|
| HiCanu          | 1.0     | Nurk et al., 2020                          |
| purge_dups      | 1.2.3   | Guan et al., 2020                          |
| SALSA2          | 2.2     | Ghurye et al., 2019                        |
| lonranger 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          |
| gEVAL           | N/A     | Chow et al., 2016                          |
| HiGlass         | 1.11.6  | Kerpedjiev et al., 2018                    |
| PretextView     | 0.1.x   | https://github.com/wtsi-hpag/PretextView   |
| BlobToolKit     | 2.6.2   | Challis et al., 2020                       |

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 undertaken according to a Research Collaboration Agreement or Material Transfer Agreement entered into by the Tree of Life collaborator, Genome Research Limited (operating as the Wellcome Sanger Institute) and in some circumstances other Tree of Life collaborators.

Data availability

European Nucleotide Archive: Pieris rapae (small white). Accession number PRJEB42143; https://identifiers.org/ena.embl/PRJEB42143.

The genome sequence is released openly for reuse. The *P. rapae* 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.

Acknowledgements

Members of the Darwin Tree of Life Barcoding collective are listed here: https://doi.org/10.5281/zenodo.4893704.

Members of the Wellcome Sanger Institute Tree of Life programme collective are listed here: https://doi.org/10.5281/zenodo.5377053.

Members of Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective are listed here: https://doi.org/10.5281/zenodo.4790456.

Members of the Tree of Life Core Informatics collective are listed here: https://doi.org/10.5281/zenodo.5013542.

Members of the Darwin Tree of Life Consortium are listed here: https://doi.org/10.5281/zenodo.4783559.

References

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
Camargo AP, Sourkov V, Pereira GAC, et al.: *RNAsamba: Neural Network-Based Assessment of the Protein-Coding Potential of RNA Sequences*. NAR Genom Bioinform. 2020; 2(1): lq024. PubMed Abstract | Publisher Full Text | Free Full Text
Challis R, Richards E, Rajan J, et al.: *BlobToolKit-Interactive Quality Assessment of Genome Assemblies*. G3 (Bethesda). 2020; 10(4): 1361–74. PubMed 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. PubMed Abstract | Publisher Full Text | Free Full Text

Fox R, Brereton TM, Asher J, et al.: The State of the UK’s Butterflies 2015. 2015. Reference Source

Garrison E, Marth G: Haplotype-Based Variant Detection from Short-Read Sequencing. 2012; arXiv: 1207.3907. Reference Source

Ghurye J, Rhie A, Walenz BP , et al.: 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

Guan 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. Gigascale, 2021; 10(1): giaa153. PubMed Abstract | Publisher Full Text | Free Full Text

Kang VY, Yang DC, Kong L, et al.: CPC2: A Fast and Accurate Coding Potential Calculator Based on Sequence Intrinsic Features. Nucleic Acids Res. 2017; 45(W1): W12–16. 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

Kriventseva EV, Rahman N, Espinosa O, et al.: OrthoDB: The Hierarchical Catalog of Eukaryotic Orthologs. Nucleic Acids Res. 2008; 36(Database issue): D271–75. PubMed Abstract | Publisher Full Text | Free Full Text

Nurk S, Walenz BP, Rhie A, et al.: HiCanu: Accurate Assembly of Segmental Duplications, Satellites, and Allelic Variants from High-Fidelity Long Reads. Genome Res. 2020; 30(9): 1291–1305. PubMed 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; 159(7): 1665–80. PubMed Abstract | Publisher Full Text | Free Full Text

Ryan SF, Lombaert E, Espeset A, et al.: Global Invasion History of the Agricultural Pest Butterfly Pieris Rapae Revealed with Genomics and Citizen Science. Proc Natl Acad Sci U S A. 2019; 116(40): 20015–24. PubMed Abstract | Publisher Full Text | Free Full Text

Shen J, Cong Q, Kinch LN, et al.: Complete genome of Pieris rapae, a resilient alien, a cabbage pest, and a source of anti-cancer proteins [version 1; peer review: 2 approved]. F1000Res. 2016; 5: 2631. 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): 3210–12. PubMed Abstract | Publisher Full Text

UniProt Consortium: UniProt: A Worldwide Hub of Protein Knowledge. Nucleic Acids Res. 2019; 47(D1): D566–15. PubMed Abstract | Publisher Full Text | Free Full Text

van Swaay C, Wynhoff I, Verovnik R, et al.: Pieris rapae. The IUCN Red List of Threatened Species 2010: e.T174353A7056010. Downloaded on 12 October 2021, 2010.