

## DATA NOTE

# The genome sequence of the red admiral, *Vanessa atalanta* (Linnaeus, 1758)

[version 1; peer review: 1 approved, 1 approved with reservations]

Konrad Lohse <sup>1</sup>, Aurora García-Berro <sup>2</sup>, Gerard Talavera <sup>2</sup>,  
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

<sup>1</sup>Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK

<sup>2</sup>Institut Botànic de Barcelona (IBB, CSIC-Ajuntament de Barcelona), Barcelona, Spain

**V1** First published: 21 Dec 2021, 6:356  
<https://doi.org/10.12688/wellcomeopenres.17524.1>  
 Latest published: 21 Dec 2021, 6:356  
<https://doi.org/10.12688/wellcomeopenres.17524.1>

## Abstract

We present a genome assembly from an individual female *Vanessa atalanta* (the red admiral; Arthropoda; Insecta; Lepidoptera; Nymphalidae). The genome sequence is 370 megabases in span. The majority of the assembly (99.44%) is scaffolded into 32 chromosomal pseudomolecules, with the W and Z sex chromosome assembled. Gene annotation of this assembly on Ensembl has identified 12,493 protein coding genes.

## Keywords

*Vanessa atalanta*, red admiral, genome sequence, chromosomal, Lepidoptera



This article is included in the [Tree of Life gateway](#).

## Open Peer Review

Approval Status

|                  | 1                    | 2                    |
|------------------|----------------------|----------------------|
| <b>version 1</b> |                      |                      |
| 21 Dec 2021      | <a href="#">view</a> | <a href="#">view</a> |

1. **Robert M Waterhouse** , University of Lausanne, Lausanne, Switzerland  
SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland
2. **John Sproul** , University of Nebraska at Omaha, Nebraska, USA

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](mailto:mark.blaxter@sanger.ac.uk))

**Author roles:** **Lohse K:** Investigation, Resources, Writing – Review & Editing; **García-Berro A:** Writing – Original Draft Preparation, Writing – Review & Editing; **Talavera G:** 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)

*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, García-Berro A, Talavera G *et al.* **The genome sequence of the red admiral, *Vanessa atalanta* (Linnaeus, 1758) [version 1; peer review: 1 approved, 1 approved with reservations]** Wellcome Open Research 2021, 6:356 <https://doi.org/10.12688/wellcomeopenres.17524.1>

**First published:** 21 Dec 2021, 6:356 <https://doi.org/10.12688/wellcomeopenres.17524.1>

## Species taxonomy

Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta; Pterygota; Neoptera; Endopterygota; Lepidoptera; Glossata; Ditrysia; Papilionoidea; Nymphalidae; Nymphalinae; Vanessa; *Vanessa atalanta* (Linnaeus, 1758) (NCBI:txid42275).

## Background

The red admiral, *Vanessa atalanta* (Linnaeus, 1758), earns its name due to the majesty of its colours: striking orange, dark brown and white. It has a [disjunct distribution in the Holarctic](#), occurring in the west Palaearctic (up to 85 degrees longitude approximately) and in North and Central America ([Williams, 1930](#)). The American populations differ slightly in appearance to the Eurasian populations, and are referred to subspecies *rubria* ([Fruhstorfer, 1909](#); [Vane-Wright & Hughes, 2007](#)). The sister species of *V. atalanta* is the kamehameha butterfly (*V. tameamea*), an endemic species from Hawaii ([Wahlberg & Rubinoff, 2011](#)). Red admirals are well known for their migratory movements: they migrate latitudinally in Europe and North America between the southern parts of their range, where the majority of individuals overwinter as adult and/or larva, and the northern areas, that are colonized during spring and summer ([Brattström et al., 2010](#); [Brattström et al., 2018](#); [Scott, 1992](#); [Walker, 2001](#); [Williams, 1930](#)). The species is listed as “Least concern” according to the IUCN Red List (Europe) ([van Swaay et al., 2010](#)). ([Roy & Sparks, 2000](#)) report changes in migration time and length due to global warming, which could be resulting in a northward shift of overwintering latitudes ([Fox et al., 2010](#)). The red admiral is polyvoltine across its migratory range. Females lay their eggs on nettles in the genera *Urtica*, *Boehmeria*, *Laportea*, and *Parietaria*. Males are markedly territorial, especially during late afternoon. The genome will further aid evolutionary studies of behavioral traits such as migration or male territoriality and the evolution of diapause. *Vanessa atalanta* has 31 pairs of chromosomes and an estimated genome size of 326 Mb ([Mackintosh et al., 2019](#)).

## Genome sequence report

The genome was sequenced from a single female *V. atalanta* ([Figure 1A, B](#)) collected from Carrifran Wildwood, Dumfries and Galloway, Scotland (latitude 55.400132, longitude -3.3352). A total of 34-fold coverage in Pacific Biosciences single-molecule long reads (N50 11 kb) and 95-fold coverage in 10X Genomics read clouds were generated. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data. Manual assembly curation corrected 16 missing/misjoins and removed 59 haplotypic duplications, reducing the assembly size by 0.27% and scaffold number by 32.86%, and increasing the scaffold N50 by 1.71%.

The final assembly has a total length of 370 Mb in 142 sequence scaffolds with a scaffold N50 of 13 Mb ([Table 1](#)). Of the assembly sequence, 99.44% was assigned to 32 chromosomal-level scaffolds, representing 30 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 98.7%, duplicated 0.1%) 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.

## Genome sequence report

The ilVanAtal1.1 genome was annotated using the Ensembl annotation pipeline ([Table 1](#); [https://rapid.ensembl.org/Vanessa\\_atalanta\\_GCA\\_905147765.1/](https://rapid.ensembl.org/Vanessa_atalanta_GCA_905147765.1/)). The resulting annotation includes 57,591 transcribed mRNAs from 12,493 protein-coding and 2,614 non-coding genes. There are 2.25 transcripts per gene and 11.26 exons per transcript. The earlier version of the assembly, ilVanAtal1.1, was annotated, but the changes made to the assembly for ilVanAtal1.2 were minor and will not affect the annotation.

## Methods

### Sample acquisition and nucleic acid extraction

Two female *V. atalanta* specimens (ilVanAtal1 and ilVanAtal2; [Figure 1](#)) were collected from Carrifran Wildwood, Dumfries and Galloway, Scotland (latitude 55.400132, longitude -3.3352) by Konrad Lohse, University of Edinburgh, using a net. The samples were identified by the same individual and snap-frozen in liquid nitrogen.

DNA was extracted from whole organism tissue of ilVanAtal1 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 was extracted from whole organism tissue of ilVanAtal2 in the Tree of Life Laboratory at the WSI using TRIzol (Invitrogen), according to the manufacturer’s instructions. 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.

## Sequencing

Pacific Biosciences HiFi circular consensus and 10X Genomics Chromium read cloud 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. Sequencing was performed by the Scientific Operations core at the Wellcome Sanger Institute on Pacific Biosciences SEQUEL II (HiFi), Illumina HiSeq X (10X) and Illumina HiSeq 4000 (RNA-Seq) instruments. Hi-C data were generated from head tissue using the Arima v1 Hi-C 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 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



**Figure 1. Fore and hind wings of *Vanessa atalanta* specimens from which the genome was sequenced. (A)** Dorsal surface view of wings from specimen SC\_VA\_1209 (ilVanAtal1) from Carrifran Wildwood, Scotland, UK, used to generate Pacific Bioscience, 10X genomics and HiC data. **(B)** surface view of wings from specimen SC\_VA\_1209 (ilVanAtal1) from Carrifran Wildwood, Scotland, UK, used to generate Pacific Biosciences, 10X genomics and HiC data. **(C)** Dorsal surface view of wings from specimen SC\_VA\_1219 (ilVanAtal2) from Carrifran Wildwood, Scotland, UK, used to generate RNASeq data. **(D)** surface view of wings from specimen SC\_VA\_1219 (ilVanAtal2) from Carrifran Wildwood, Scotland, UK, used to generate RNASeq data.

**Table 1. Genome data for *Vanessa atalanta*, iVanAtal1.2.**

| <b>Project accession data</b>          |                                                  |
|----------------------------------------|--------------------------------------------------|
| Assembly identifier                    | iVanAtal1.2                                      |
| Species                                | <i>Vanessa atalanta</i>                          |
| Specimen                               | iVanAtal1 (HiFi, 10X, Hi-C); iVanAtal2 (RNAseq)  |
| NCBI taxonomy ID                       | NCBI:txid42275                                   |
| BioProject                             | PRJEB41956                                       |
| BioSample ID                           | SAMEA7523145                                     |
| Isolate information                    | Female, whole organism (iVanAtal1 and iVanAtal2) |
| <b>Raw data accessions</b>             |                                                  |
| PacificBiosciences SEQUEL II           | ERR6608650                                       |
| 10X Genomics Illumina                  | ERR6001611, ERR6002556, ERR6002557, ERR6003033   |
| Hi-C Illumina                          | ERR6002558, ERR600255, ERR6003034                |
| Illumina PolyA RNAseq                  | ERR6363249                                       |
| <b>Genome assembly</b>                 |                                                  |
| Assembly accession                     | GCA_905147765.2                                  |
| Accession of alternate haplotype       | GCA_905147705.2                                  |
| Span (Mb)                              | 370                                              |
| Number of contigs                      | 159                                              |
| Contig N50 length (Mb)                 | 12.2                                             |
| Number of scaffolds                    | 141                                              |
| Scaffold N50 length (Mb)               | 12.8                                             |
| Longest scaffold (Mb)                  | 15.0                                             |
| BUSCO* genome score                    | C:99.1%[S:98.9%,D:0.2%],F:0.2%,M:0.7%,n:1658     |
| <b>Genome annotation**</b>             |                                                  |
| Number of protein-coding genes         | 12,493                                           |
| Average length of coding sequence (bp) | 1,728                                            |
| Average number of exons per transcript | 9.45                                             |
| Average exon size (bp)                 | 203                                              |
| Average intron size (bp)               | 2,047                                            |

\*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/iVanAtal1.2/dataset/CAJHWX02/busco>.

\*\*For assembly GCA\_905147765.1, iVanAtal1.1.

(Howe *et al.*, 2021). Manual curation was performed using gEVAL, HiGlass (Kerpedjiev *et al.*, 2018) and Pretext. The mitochondrial genome was assembled using MitoHiFi (Uliano-Silva *et al.*, 2021), which performed 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.

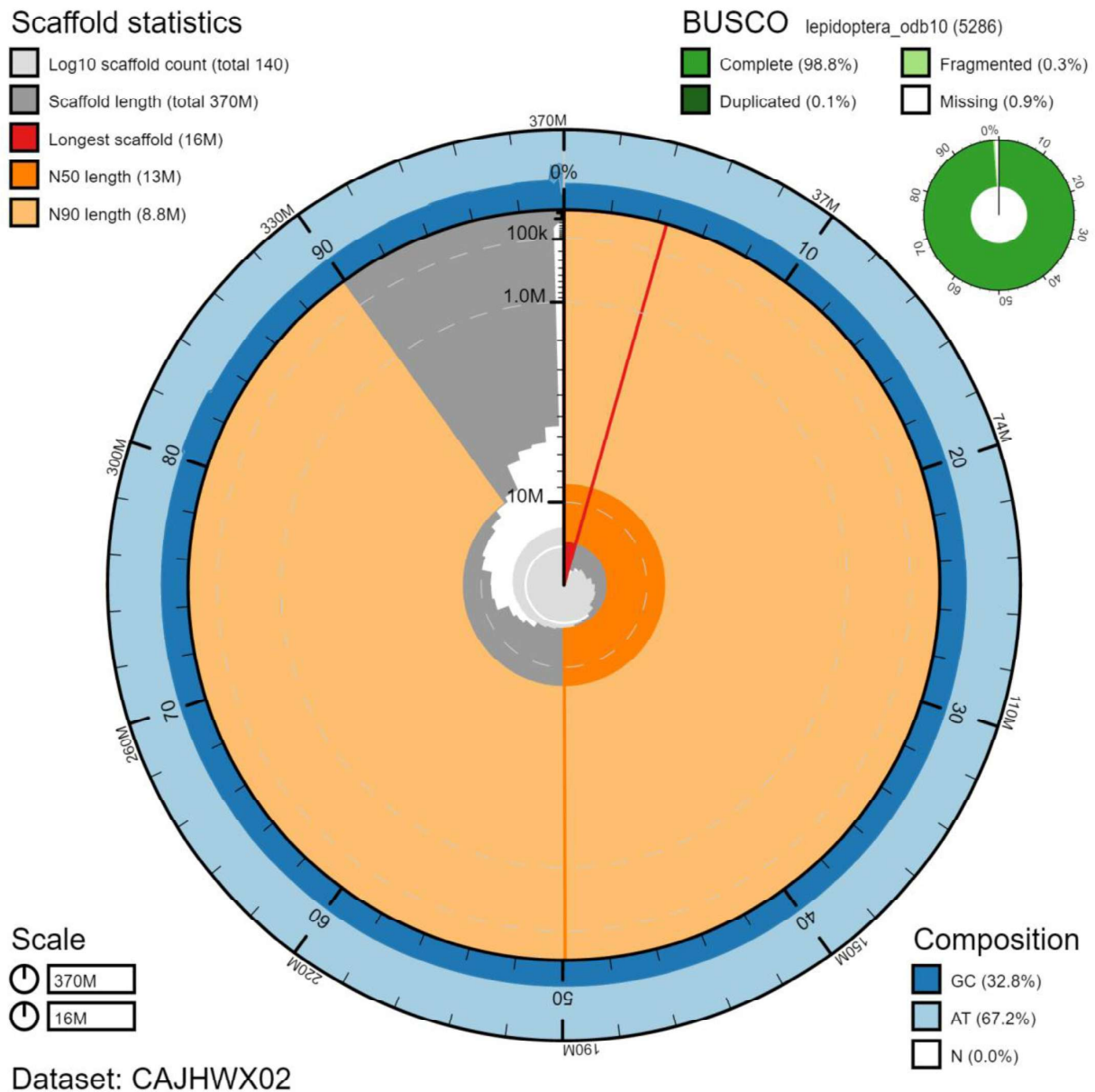


**Figure 2. Genome assembly of *Vanessa atalanta*, ilVanAtal1.2: metrics.** The BlobToolKit Snailplot shows N50 metrics and BUSCO gene completeness. The main plot is divided into 1,000 size-ordered bins with each bin representing 0.1% of the 370,436,055 bp assembly. The distribution of chromosome lengths is shown in dark grey with the plot radius scaled to the longest chromosome present in the assembly (16,450,097 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 chromosome lengths (12,792,537 and 8,781,517 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/Vanessa%20atalanta/dataset/CAJHWX02/snail>.

### Gene annotation

The Ensembl gene annotation system (Aken *et al.*, 2016) was used to generate annotation for version 1 of the *Vanessa atalanta*

assembly (GCA\_905147785.1). The annotation was created primarily through alignment of transcriptomic data to the genome, with gap filling via protein-to-genome alignments of a select

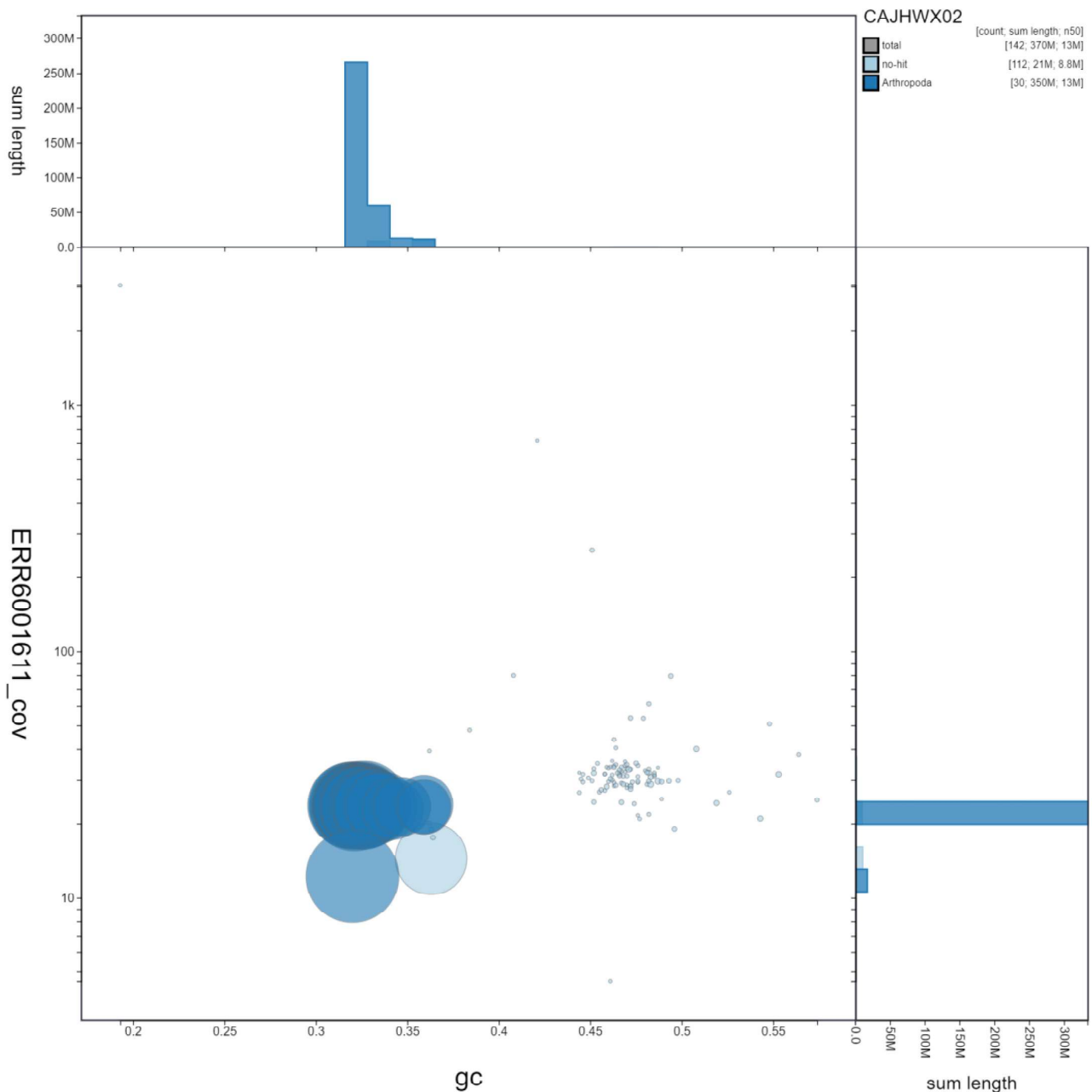


**Figure 3. Genome assembly of *Vanessa atalanta*, ilVanAtal1.2: GC coverage.** BlobToolKit GC-coverage plot. Scaffolds are coloured by phylum. Circles are sized in proportion to scaffold length. Histograms show the distribution of scaffold length sum along each axis. An interactive version of this figure is available at <https://blobtoolkit.genomehubs.org/view/Vanessa%20atalanta/dataset/CAJH WX02/blob>.

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. See [https://rapid.ensembl.org/Vanessa\\_atalanta\\_GCA\\_905147765.1/](https://rapid.ensembl.org/Vanessa_atalanta_GCA_905147765.1/) for further details.

#### Ethical/compliance issues

The materials that have contributed to this genome note were supplied by a Tree of Life collaborator. 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

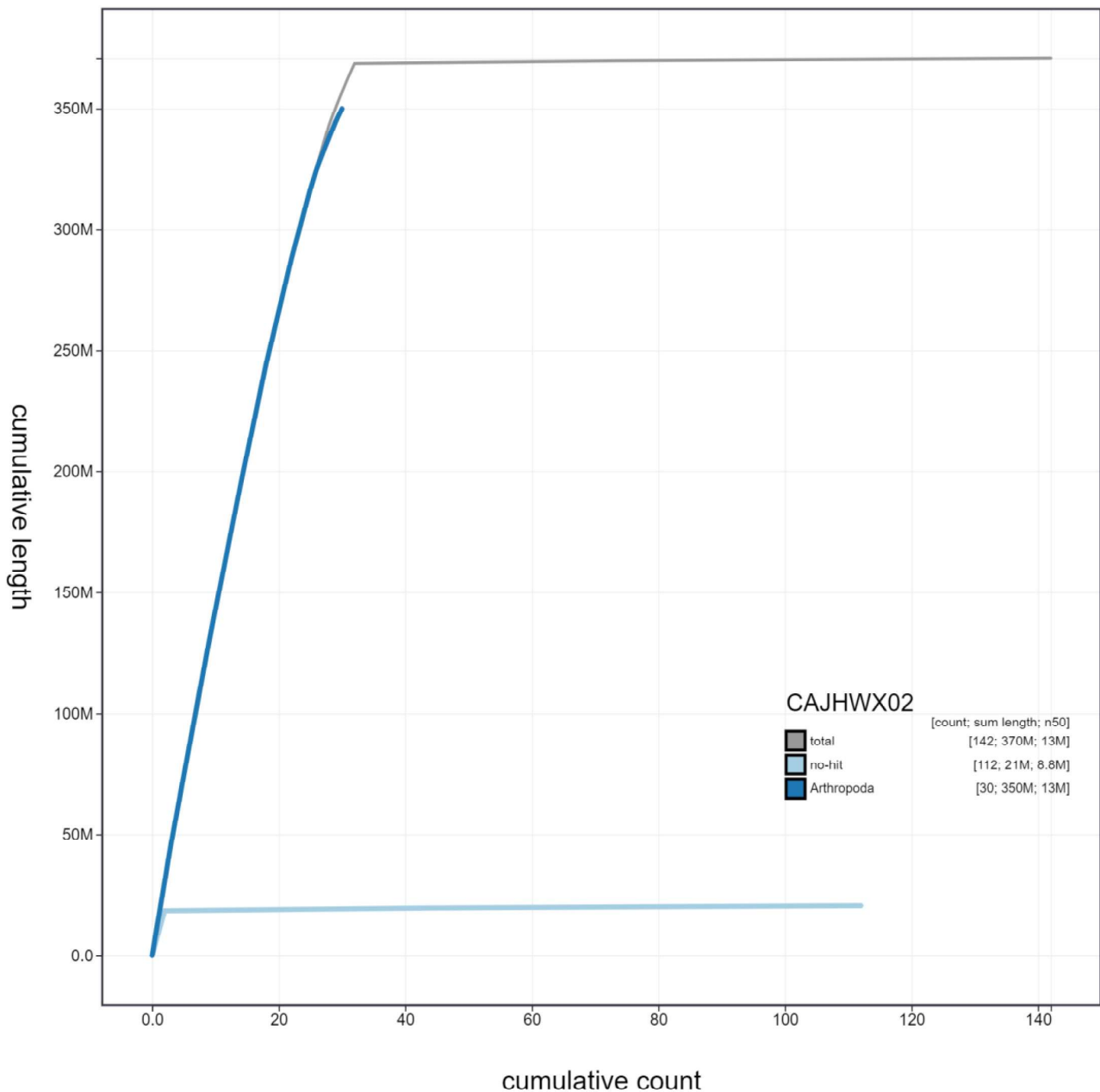


**Figure 4. Genome assembly of *Vanessa atalanta*, iVanAtal1.2: cumulative sequence.** BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at <https://blobtoolkit.genomehubs.org/view/Vanessa%20atalanta/dataset/CAJHWX02/cumulative>.

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).

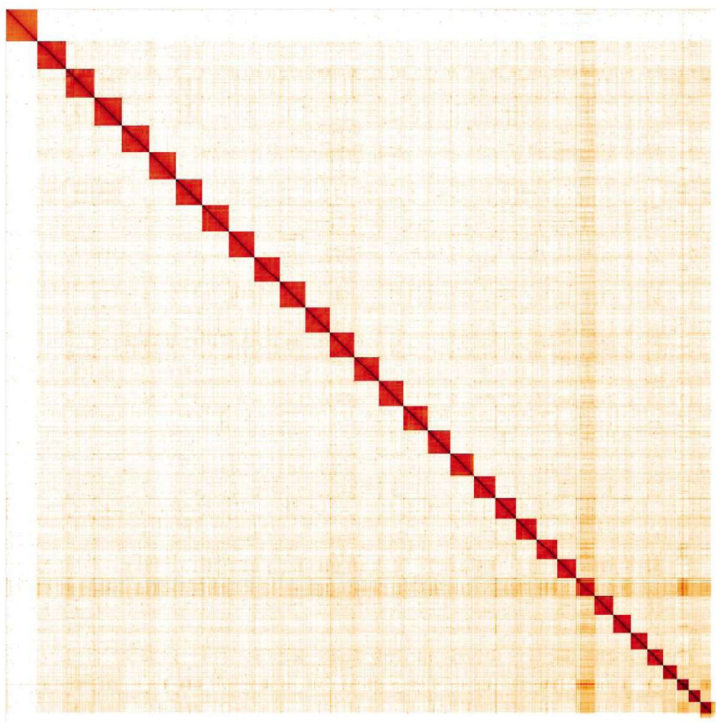


**Figure 5. Genome assembly of *Vanessa atalanta*, ilVanAtal1.2: Hi-C contact map.** Hi-C contact map of the ilVanAtal1.2 assembly, visualised in HiGlass.

**Table 2. Chromosomal pseudomolecules in the genome assembly of *Vanessa atalanta*, ilVanAtal1.2.**

| INSDC accession | Chromosome | Size (Mb) | GC%  |
|-----------------|------------|-----------|------|
| LR990550.1      | 1          | 15.00     | 32.1 |
| LR990551.1      | 2          | 14.66     | 32.6 |
| LR990552.2      | 3          | 14.42     | 32.5 |
| LR990553.2      | 4          | 14.40     | 32.0 |
| LR990554.1      | 5          | 14.01     | 31.9 |
| LR990555.1      | 6          | 13.78     | 32.6 |
| LR990556.1      | 7          | 13.67     | 32.0 |
| LR990557.1      | 8          | 13.65     | 32.3 |
| LR990558.1      | 9          | 13.29     | 32.2 |
| LR990559.2      | 10         | 13.09     | 32.1 |
| LR990560.1      | 11         | 12.98     | 32.1 |
| LR990561.1      | 12         | 12.80     | 32.5 |
| LR990562.1      | 13         | 12.79     | 32.4 |
| LR990563.1      | 14         | 12.69     | 32.2 |
| LR990564.1      | 15         | 12.58     | 32.4 |
| LR990565.1      | 16         | 12.15     | 32.4 |

| INSDC accession | Chromosome | Size (Mb) | GC%  |
|-----------------|------------|-----------|------|
| LR990566.1      | 17         | 11.86     | 32.5 |
| LR990567.2      | 18         | 11.60     | 33.0 |
| LR990568.1      | 19         | 10.76     | 32.7 |
| LR990569.1      | 20         | 10.62     | 33.0 |
| LR990570.1      | 21         | 10.59     | 33.1 |
| LR990571.1      | 22         | 9.89      | 32.8 |
| LR990572.1      | 23         | 9.62      | 32.9 |
| LR990573.1      | 24         | 9.26      | 33.6 |
| LR990574.1      | 25         | 8.78      | 33.2 |
| LR990575.1      | 26         | 8.43      | 33.4 |
| LR990577.1      | 27         | 6.87      | 34.2 |
| LR990578.1      | 28         | 6.38      | 35.9 |
| LR990579.1      | 29         | 6.22      | 34.7 |
| LR990580.1      | 30         | 5.42      | 35.9 |
| LR990576.2      | W          | 9.65      | 36.3 |
| LR990549.1      | Z          | 16.45     | 32.0 |
| LR990581.1      | MT         | 0.02      | 19.6 |
|                 | Unplaced   | 2.08      | 46.9 |

**Table 3. Software tools used.**

| Software tool    | Version           | Source                                                                                                                                                                                                        |
|------------------|-------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| HiCanu           | 1.0               | <a href="#">Nurk et al., 2020</a>                                                                                                                                                                             |
| purge_dups       | 1.2.3             | <a href="#">Guan et al., 2020</a>                                                                                                                                                                             |
| SALSA2           | 2.2               | <a href="#">Ghurye et al., 2019</a>                                                                                                                                                                           |
| longranger align | 2.2.2             | <a href="https://support.10xgenomics.com/genome-exome/software/pipelines/latest/advanced/other-pipelines">https://support.10xgenomics.com/genome-exome/software/pipelines/latest/advanced/other-pipelines</a> |
| freebayes        | 1.3.1-17-gaa2ace8 | <a href="#">Garrison &amp; Marth, 2012</a>                                                                                                                                                                    |
| gEVAL            | N/A               | <a href="#">Chow et al., 2016</a>                                                                                                                                                                             |
| PretextView      | 0.1.x             | <a href="https://github.com/wtsi-hpag/PretextView">https://github.com/wtsi-hpag/PretextView</a>                                                                                                               |
| HiGlass          | 1.11.6            | <a href="#">Kerpedjiev et al., 2018</a>                                                                                                                                                                       |
| BlobToolKit      | 2.6.2             | <a href="#">Challis et al., 2020</a>                                                                                                                                                                          |

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: Vanessa atalanta (red admiral) Accession number [PRJEB42064](#); <https://identifiers.org/ena.embl/PRJEB42064>.

The genome sequence is released openly for reuse. The *V. atalanta* 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 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.5744840>.

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.5743293>.

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

### 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](#)
- 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](#)
- Brattström O, Åkesson S, Bensch S: **AFLP Reveals Cryptic Population Structure in Migratory European Red Admirals (*Vanessa Atalanta*). Ecol Entomol.** 2010; **35**(2): 248–52.  
[Publisher Full Text](#)
- Brattström O, Shapoval A, Wassenaar LI, et al.: **Geographic Origin and Migration Phenology of European Red Admirals (*Vanessa Atalanta*) as Revealed by Stable Isotopes. Movement Ecology.** 2018; **6**: 25.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Camargo AP, Sourkov V, Pereira GAG, et al.: **RNAse: Neural Network-Based Assessment of the Protein-Coding Potential of RNA Sequences.**

- NAR Genom Bioinform.* 2020; **2**(1): lqz024.  
[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, Dennis RLH and Others: **Winter Survival of Vanessa Atalanta (Linnaeus, 1758)(Lepidoptera: Nymphalidae): A New Resident Butterfly for Britain and Ireland?. Entomologist's Gazette.** 2010; **61**(2): 94.  
[Reference Source](#)
- Fruhstorfer H: **Neue Vanessa- Und Pyrameis-Rassen. Internationale Entomologische Zeitschrift.** 1909; **3**(94).
- Garrison E, Marth G: **Haplotype-Based Variant Detection from Short-Read Sequencing.** July. 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**(1): 2896–2898.

[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](#)

Kang YJ, 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](#)

Mackintosh A, Laetsch DR, Hayward A, *et al.*: **The Determinants of Genetic Diversity in Butterflies.** *Nat Commun.* 2019; **10**(1): 3466.

[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](#)

Roy DB, Sparks TH: **Phenology of British Butterflies and Climate Change.** *Glob Change Biol.* 2000; **6**(4): 407–16.

[Publisher Full Text](#)

Scott JA: **The Butterflies of North America: A Natural History and Field Guide.** Stanford University Press, 1992.

[Reference Source](#)

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](#)

Uliano-Silva M, Nunes JGF, Krashenninnikova K, *et al.*: **marcelauliano/MitoHiFi: mitohifi\_v2.0.** 2021.

[Publisher Full Text](#)

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

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

van Swaay C, Wynhoff I, Verovnik R, *et al.*: **IUCN Red List of Threatened Species: *Vanessa atalanta*.** *IUCN Red List of Threatened Species.* 2010.

[Reference Source](#)

Vane-Wright RI, Hughes HWD: **Did a Member of the *Vanessa indica* Complex (*Nymphalidae*) Formerly Occur in North America?** *J Lepid Soc.* 2007; **61**(4): 199.

[Reference Source](#)

Wahlberg N, Rubinoff D: **Vagility across *Vanessa* (Lepidoptera: Nymphalidae): Mobility in Butterfly Species Does Not Inhibit the Formation and Persistence of Isolated Sister Taxa.** *Syst Entomol.* 2011; **36**(2): 362–70.

[Publisher Full Text](#)

Walker TJ: **Butterfly Migrations in Florida: Seasonal Patterns and Long-Term Changes.** *Environ Entomol.* 2001; **30**(6): 1052–60.

[Publisher Full Text](#)

Williams CB: **The Migration of Butterflies.** *The Migration of Butterflies.* 1930.

[Reference Source](#)

# Open Peer Review

Current Peer Review Status:



Version 1

Reviewer Report 06 February 2023

<https://doi.org/10.21956/wellcomeopenres.19376.r53123>

© 2023 Sproul 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.



**John Sproul** 

University of Nebraska at Omaha, Nebraska, USA

In this data note, the authors present the genome assembly of *Vanessa atalanta*. The assembly is highly contiguous (i.e., resolves chromosomal pseudomolecules) and the authors present a summary of gene annotations.

My only criticism of the work is that the authors do not provide any description of repetitive DNA components of the assembly, or reference any analysis that was done to characterize repetitive elements. Similar to providing a description of genic content, a summary of the repetitive composition is of basic importance in summarizing a new assembly. Ideally any library of repetitive element sequences produced would also be submitted to a public repository of repetitive elements (e.g., Dfam) so the genomics community can benefit from basic datasets generated as part of the project.

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

Yes

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

Partly

**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:** biodiversity genomics, rapid genome evolution

**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, however I have significant reservations, as outlined above.**

Reviewer Report 30 January 2023

<https://doi.org/10.21956/wellcomeopenres.19376.r54170>

© 2023 Waterhouse R. 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.



**Robert M Waterhouse** 

<sup>1</sup> University of Lausanne, Lausanne, Switzerland

<sup>2</sup> SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland

This Data Note for the red admiral butterfly presents a clear and comprehensive description of all the steps taken to generate the *Vanessa atalanta* genome assembly spanning 370 Mb with >99% assigned to 32 chromosomal-level scaffolds including the W and Z sex chromosomes. As an important model for studying migration, and with now concerns over how climate change could impact migration, the rationale for building these resources is clear. The described data collection and analysis methods follow the best practices in the field and have delivered a high-quality complete and accurate chromosome-level reference genome. The genome assembly is of one haplotype, with contigs corresponding to the second haplotype also deposited. Gene annotations are provided with the genome assembly.

**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:** arthropod evolutionary and functional genomics, bioinformatics

**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.**

## Comments on this article

**Version 1**

Reader Comment 10 Jan 2022

**Jacques Dainat**, CNRS, France

I don't want to throw you the stone because all papers describing genome annotations do the same, but I find pity that BUSCO results of the assembly completeness are always mentioned while annotation completeness (gene predictions) is never. Based on the translation of the protein coding gene, the BUSCO annotation completeness is very useful to give a picture of how complete is the annotation. The comparison between the assembly completeness score and the annotation completeness score shows the potential room for improvement of the annotation. It's easy to see the proportion of genes present in the assembly that are finally not present in the final gene build. In your case, the BUSCO for the assembly is 98.8% but the BUSCO for the annotation might be around 50%. That type of result might be afraid and it is why researchers do not like to show it, but such a result does not mean that the annotation is bad. It reflects a choice in the annotation approach type. Evidence based annotations can provide really good gene models (low number of false positive and a high number of high-confidence gene models (protein coding genes as well as non-coding genes), while can miss a lot of genes. Other approaches like *ab-initio* approaches might miss fewer genes but can, at the same time, predict a lot of false positives. Some people do not care about false positives and want to know if they can trust the blastP on the proteome when it says that the gene/protein is missing. In such case, if we had the BUSCO annotation score, we would know if were better to tBlastn on the assembly instead of blastP the proteome.

I think that the Darwin Tree of Life Projects should include that information by default.

Best regards,

Jacques Dainat Ph.D.

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