



DATA NOTE

The genome sequence of the buff-tip, *Phalera bucephala*

(Linnaeus, 1758) [version 1; peer review: 1 approved]

Douglas Boyes¹⁺, Peter W.H. Holland ²,
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, Darwin Tree of Life Consortium

¹UK Centre for Ecology and Hydrology, Wallingford, Oxfordshire, UK

²Department of Zoology, University of Oxford, Oxford, UK

+ Deceased author

V1 First published: 27 Jan 2022, 7:28
<https://doi.org/10.12688/wellcomeopenres.17539.1>
Latest published: 27 Jan 2022, 7:28
<https://doi.org/10.12688/wellcomeopenres.17539.1>

Abstract

We present a genome assembly from an individual female *Phalera bucephala* (the buff-tip; Arthropoda; Insecta; Lepidoptera; Notodontidae). The genome sequence is 933 megabases in span. The majority of the assembly, 99.27%, is scaffolded into 31 chromosomal pseudomolecules, with the W and Z sex chromosome assembled.

Keywords

Phalera bucephala, buff-tip, genome sequence, chromosomal, Lepidoptera



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

Open Peer Review

Approval Status

1

version 1

27 Jan 2022



[view](#)

1. **William B Walker** , USDA Agricultural
Research Service, Beltsville, 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)

Author roles: Boyes D: Investigation, Resources; Holland PWH: Supervision, 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).

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

Copyright: © 2022 Boyes D *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: Boyes D, Holland PWH, University of Oxford and Wytham Woods Genome Acquisition Lab *et al.* **The genome sequence of the buff-tip, *Phalera bucephala* (Linnaeus, 1758) [version 1; peer review: 1 approved]** Wellcome Open Research 2022, 7:28 <https://doi.org/10.12688/wellcomeopenres.17539.1>

First published: 27 Jan 2022, 7:28 <https://doi.org/10.12688/wellcomeopenres.17539.1>

Species taxonomy

Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta; Pterygota; Neoptera; Endopterygota; Lepidoptera; Glossata; Ditrysia; Noctuoidea; Notodontidae; Phalerinae; Phalera; *Phalera bucephala* (Linnaeus, 1758) (NCBI:txid753216).

Background

Phalera bucephala (buff-tip) exhibits one of the most striking examples of camouflage amongst UK moths: the yellow-tipped forewings held tent-like along the body give the convincing appearance of a broken birch twig. The moth is nocturnal and found across the UK, mainland Europe and parts of Asia. The larvae are polyphagous, feeding on the leaves of several deciduous trees including birch, beech and oak. Ford (1967) comments that the larvae can produce a pungent smell, presumably as a defence mechanism. The species can become a transient pest; for example, defoliating trees along the Maidenhead bypass in the UK in the 1970s (Port & Thompson, 1980) and apple trees in Lithuania (Molis, 1970). The species has also been used in studies to assess the effect of multiple stressors (herbivores, powdery mildew and aphids) on oak trees, revealing complex plant-pathogen-insect interactions (van Dijk *et al.*, 2020).

The genome of *P. bucephala*, was sequenced as part of the Darwin Tree of Life Project, a collaborative effort to sequence all of the named eukaryotic species in the Atlantic Archipelago of Britain and Ireland. Here we present a chromosomally complete genome sequence for *P. bucephala*, based on one female specimen from Wytham Woods, Oxfordshire, UK.

Genome sequence report

The genome was sequenced from a single female *P. bucephala* (Figure 1) collected from Wytham Woods, Oxfordshire (biological vice-county: Berkshire), UK (latitude 51.764, longitude -1.327). A total of 34-fold coverage in Pacific Biosciences single-molecule circular consensus HiFi long reads (N50 15 kb) and 51-fold coverage in 10X Genomics read clouds were generated. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data. Manual assembly curation corrected 155 missing/misjoins and removed 4 haplotypic duplications, reducing the assembly size by 0.22% and scaffold number by 45.28%, and increasing the scaffold N50 by 40.20%.

The final assembly has a total length of 933 Mb in 116 sequence scaffolds with a scaffold N50 of 34 Mb (Table 1). Of the assembly sequence, 99.27% was assigned to 31 chromosomal-level scaffolds, representing 29 autosomes (numbered by sequence length), and the W and Z sex chromosome (Figure 2–Figure 5; Table 2). The assembly has a BUSCO v5.1.2 (Manni *et al.*, 2021) completeness of 98.9% (single 97.8%, duplicated 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.

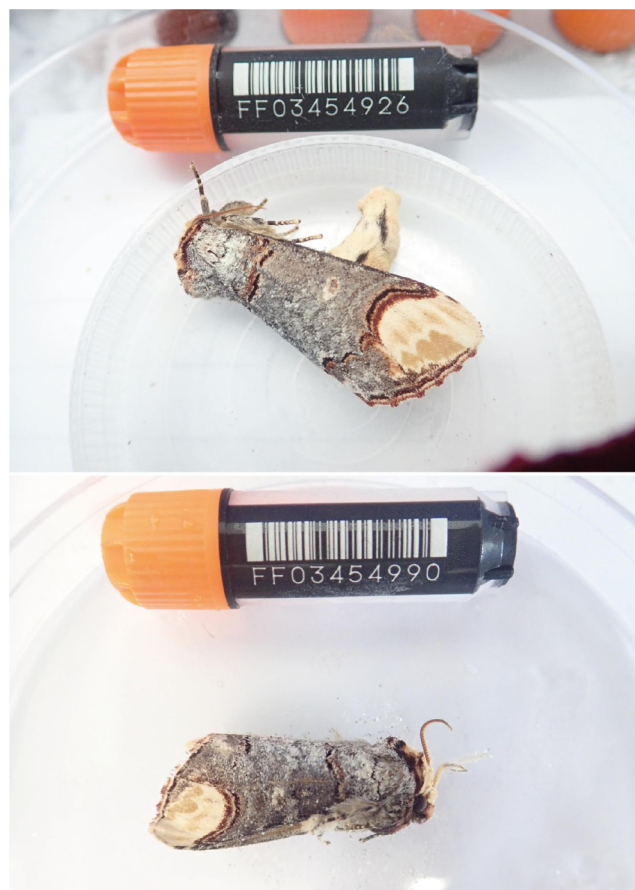


Figure 1. Image of the *Phalera bucephala* specimens taken prior to preservation and processing. Above, ilPhaBuce1, used for genome and Hi-C sequencing; below, ilPhaBuce2, used for RNA-Seq.

Methods

Sample acquisition and nucleic acid extraction

A female *P. bucephala* (ilPhaBuce1) and a second specimen of unknown sex (ilPhaBuce2) were collected from Wytham Woods, Oxfordshire (biological vice-county: Berkshire), UK (latitude 51.764, longitude -1.327) by Douglas Boyes, UKCEH, using a net. The samples were identified by the same individual and snap-frozen on dry ice.

DNA was extracted from whole organism tissue of ilPhaBuce1 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 thorax/abdomen tissue of ilPhaBuce2 in the Tree of Life Laboratory at the WSI using TRIzol (Invitrogen), according to the manufacturer's instructions.

Table 1. Genome data for *Phalera bucephala*, ilPhaBuce1.2.

Project accession data	
Assembly identifier	ilPhaBuce1.2
Species	<i>Phalera bucephala</i>
Specimen	ilPhaBuce1
NCBI taxonomy ID	NCBI:txid753216
BioProject	PRJEB42140
BioSample ID	SAMEA7519921
Isolate information	Female, head/abdomen/thorax
Raw data accessions	
PacificBiosciences SEQUEL II	ERR6594494, ERR6594495
10X Genomics Illumina	ERR6002720-ERR6002727
Hi-C Illumina	ERR6002728-ERR6002730
Illumina polyA RNA-Seq	ERR6002731
Genome assembly	
Assembly accession	GCA_905147815.2
Accession of alternate haplotype	GCA_905147805.2
Span (Mb)	933
Number of contigs	295
Contig N50 length (Mb)	8.5
Number of scaffolds	116
Scaffold N50 length (Mb)	34.1
Longest scaffold (Mb)	43.5
BUSCO* genome score	C:98.9%[S:97.8%,D:1.0%],F:0.3%,M:0.8%,n:5286

*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/ilPhaBuce1.2/dataset/CAJHXA02/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.

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 Qiagen EpiTect 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

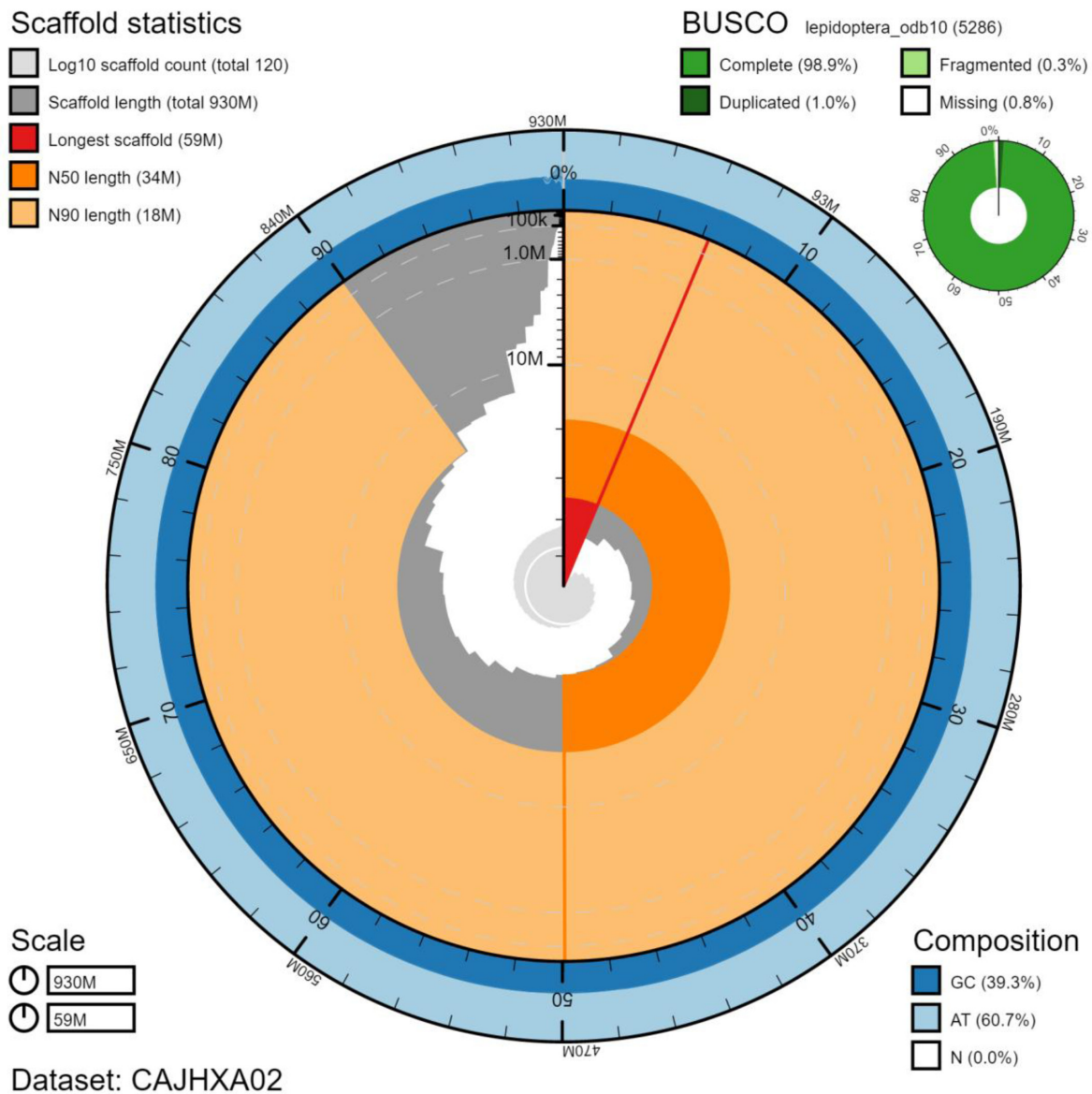


Figure 2. Genome assembly of *Phalera bucephala*, ilPhaBuce1.2: metrics. The BlobToolKit Snailplot shows N50 metrics and BUSCO gene completeness. The main plot is divided into 1,000 size-ordered bins around the circumference with each bin representing 0.1% of the 933,147,695 bp assembly. The distribution of scaffold lengths is shown in dark grey with the plot radius scaled to the longest scaffold present in the assembly (59,027,677 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 scaffold lengths (34,116,407 and 18,324,721 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/ilPhaBuce1.2/dataset/CAJHXA02/snail>.

(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

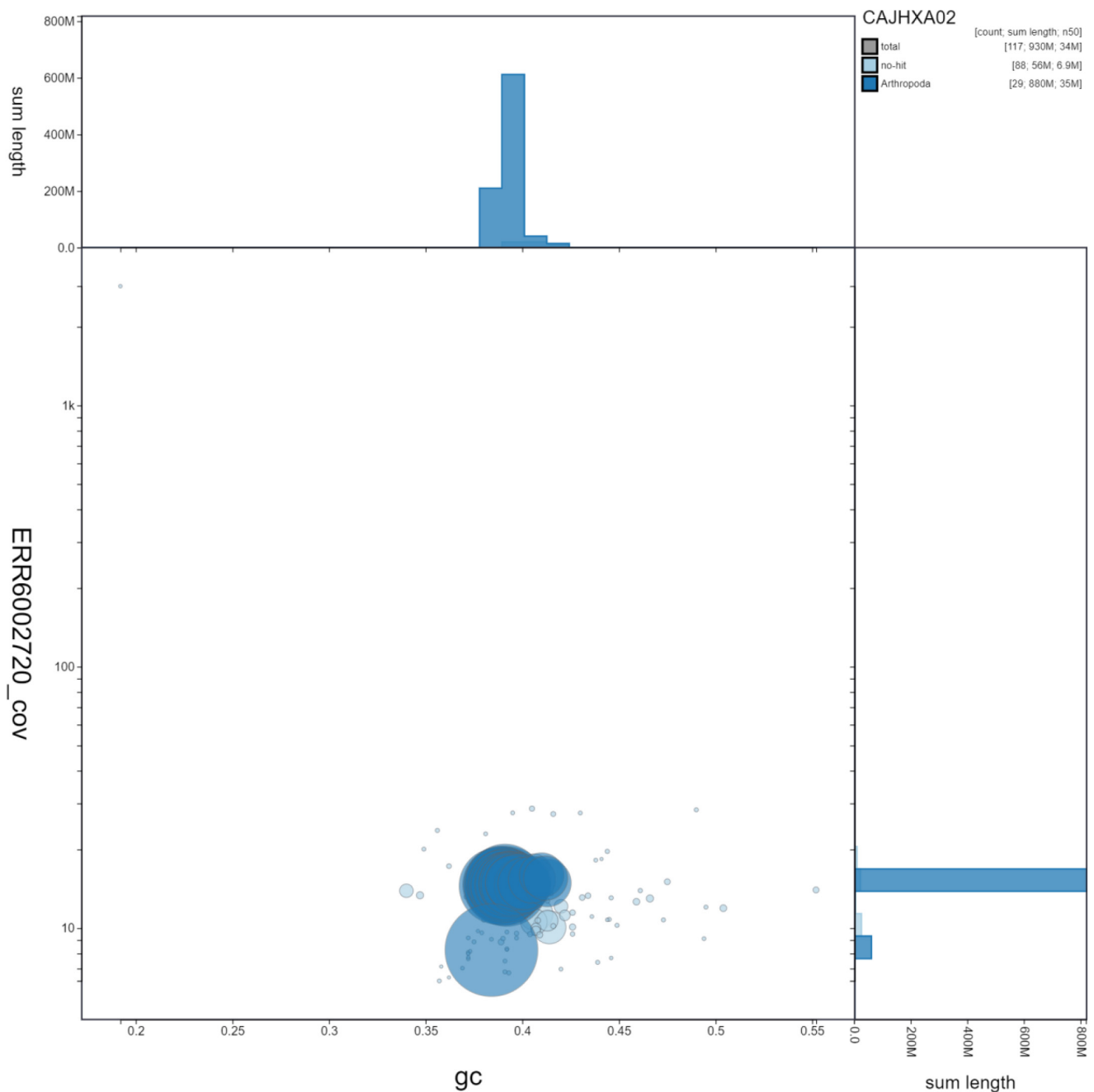


Figure 3. Genome assembly of *Phalera bucephala*, ilPhaBuce1.2: GC coverage. BlobToolKit GC-coverage plot. Scaffolds are coloured by phylum. Circles are sized in proportion to scaffold length. Histograms show the distribution of scaffold length sum along each axis. An interactive version of this figure is available at <https://blobtoolkit.genomehubs.org/view/ilPhaBuce1.2/dataset/CAJHXA02/blob>.

(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

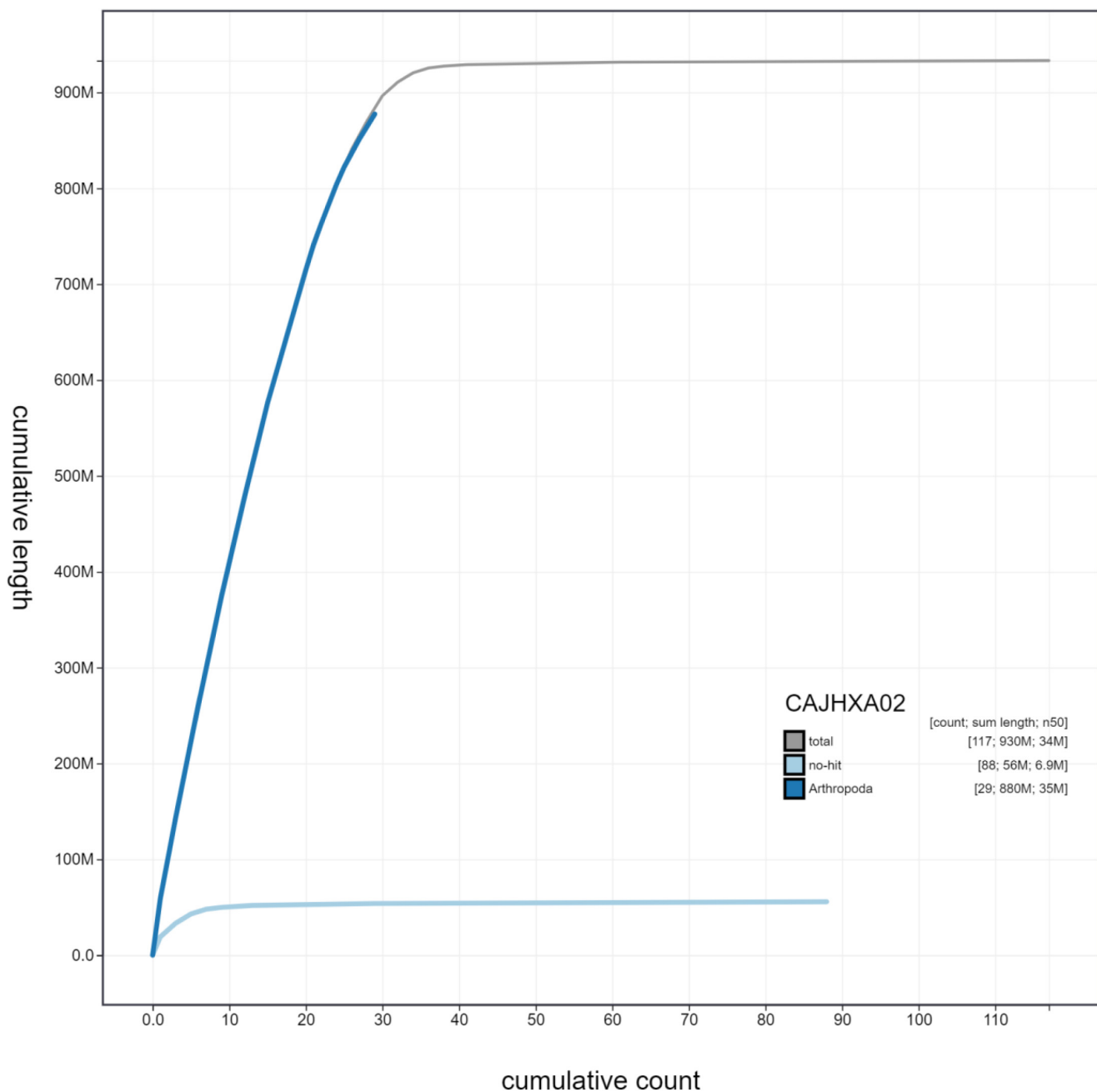


Figure 4. Genome assembly of *Phalera bucephala*, ilPhaBuce1.2: cumulative sequence. BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at <https://blobtoolkit.genomehubs.org/view/ilPhaBuce1.2/dataset/CAJHXA02/cumulative>.

gEVAL, HiGlass (Kerpedjiev *et al.*, 2018) and Pretext. 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.

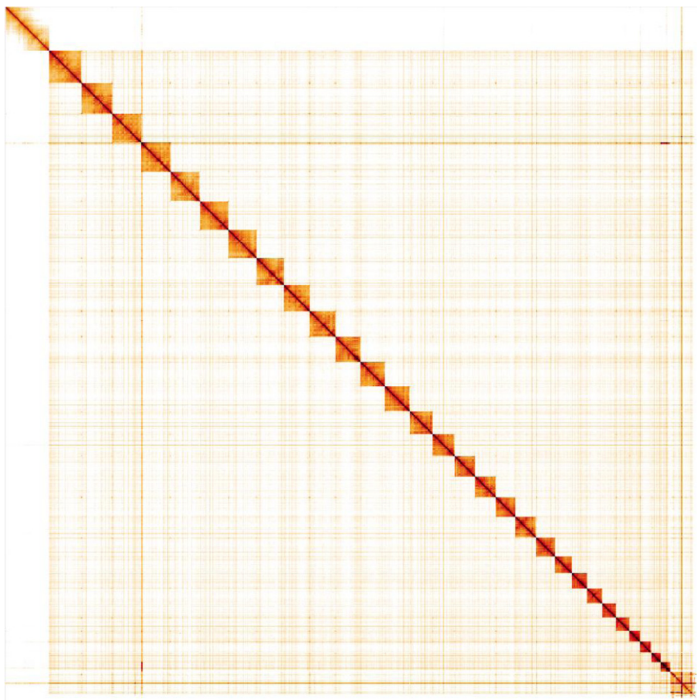


Figure 5. Genome assembly of *Phalera bucephala*, ilPhaBuce1.2: Hi-C contact map. Hi-C contact map of the ilPhaBuce1.2 assembly, visualised in HiGlass. Chromosomes are given in order of size from left to right and top to bottom.

Table 2. Chromosomal pseudomolecules in the genome assembly of *Phalera bucephala*, ilPhaBuce1.2.

INSDC accession	Chromosome	Size (Mb)	GC%
LR990610.1	1	43.49	39.2
LR990611.1	2	40.87	39.1
LR990612.1	3	39.81	38.9
LR990613.1	4	39.67	39
LR990614.1	5	39.31	38.7
LR990615.1	6	38.13	39.1
LR990616.1	7	37.74	38.9
LR990617.1	8	37.02	39.2
LR990618.1	9	34.85	39.1
LR990619.1	10	34.54	39.3
LR990620.1	11	34.12	38.9
LR990621.1	12	33.31	39.1
LR990622.1	13	33.06	39
LR990623.1	14	31.29	39.1
LR990624.1	15	29.29	39.3
LR990625.1	16	27.79	39.2

INSDC accession	Chromosome	Size (Mb)	GC%
LR990626.1	17	27.27	39.4
LR990627.1	18	27.25	39.4
LR990628.1	19	26.99	39.5
LR990629.1	20	26.05	39.8
LR990630.1	21	22.17	39.6
LR990631.1	22	20.78	40
LR990632.1	23	20.08	39.5
LR990633.1	24	19.01	40
LR990634.1	25	18.32	40.1
LR990635.1	26	14.81	41.3
LR990636.1	27	14.56	40.5
LR990637.1	28	12.96	41
LR990638.1	29	12.86	41.2
LR990639.1	W	7.37	40.7
LR990609.1	Z	59.03	38.4
LR990640.1	MT	0.02	19.3
-	Unplaced	29.32	40.9

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
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
gEVAL	N/A	Chow et al., 2016
PretextView	0.1.x	https://github.com/wtsi-hpag/PretextView
HiGlass	1.11.6	Kerpedjiev et al., 2018
BlobToolKit	2.6.4	Challis et al., 2020

Data availability

European Nucleotide Archive: *Phalera bucephala* (buff-tip) genome assembly, ilPhaBuce1. Accession number [PRJEB42140](#); <https://identifiers.org/ena.embl/PRJEB42140>.

The genome sequence is released openly for reuse. The *P. bucephala* genome sequencing initiative is part of the Darwin Tree of Life (DTOL) project. All raw sequence data and the assembly have been deposited in INSDC databases. The genome will be annotated and presented through the [Ensembl](#) pipeline at the European Bioinformatics Institute. Raw data and assembly accession identifiers are reported in [Table 1](#).

Author information

Members of the University of Oxford and Wytham Woods Genome Acquisition Lab are listed here: <https://doi.org/10.5281/zenodo.5746938>.

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

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

- 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](#)
- van Dijk LJA, Ehrlén J, Tack AJM: **The Timing and Asymmetry of Plant-pathogen-insect Interactions**. *Proc Biol Sci*. 2020; **287**(1935): 20201303. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Ford EB: **Moths, No. 30, New Naturalist Series**. Collins, London, 1967. [Reference Source](#)
- Garrison E, Marth G: **Haplotype-Based Variant Detection from Short-Read Sequencing**. arXiv: 1207.3907. 2012. [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.** *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, Seppely 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](#)

Molis S: **The Buff-Tip Moth (Phalera Bucephala L.)-a Pest of Apple Trees.** *Acta Entomologica Lituanica.* 1970; **1**: 182–83.

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](#)

Port GR, Thompson JR: **Outbreaks of Insect Herbivores on Plants Along Motorways in the United Kingdom.** *J Appl Ecol.* 1980; **17**(3): 649–56.

[Publisher 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](#)

Open Peer Review

Current Peer Review Status: 

Version 1

Reviewer Report 15 February 2023

<https://doi.org/10.21956/wellcomeopenres.19395.r54445>

© 2023 Walker W. 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.



William B Walker 

USDA Agricultural Research Service, Beltsville, MD, USA

The authors present a chromosomally complete genome sequence for the buff-tip moth, *Phalera bucephala*. The genome is sequenced from a single female, with a total of 34X coverage from HiFi long reads, 51X coverage from 10X Genomics read clouds and chromosome confirmation Hi-C data. Manual assembly curations are performed to improve the assembly. Standard bioinformatics pipelines are followed, with all softwares clearly indicated.

The main issue with the report stems from a comment mentioned on page 4 that "Poly(A) RNA-Seq libraries were constructed", suggesting that multiple libraries were generated and sequenced, however in Table 1, it is apparent that there was only one library sequenced.

Following up on this, it is not clear, in the methods describing "Genome Assembly" if the RNA-Seq library was at all utilized during assembly process, for example, during manual curation or otherwise.

Finally, a cosmetic note concerning Figure 5, it may be useful to the reader if Chromosome labels are indicated, at least for the W and Z chromosomes as these are specifically mentioned in the abstract, and there are otherwise no clear indications in the figure which plots correlate specifically to these chromosomes. (This is mentioned because in another similar Data Note in this journal, there is an a similar Hi-C contact plot with an online interactive version, in which the chromosome accessions are clearly shown for each box).

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

Yes

Are the protocols appropriate and is the work technically sound?

Yes

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

Yes

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

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: insect molecular biology, genetics and genomics/transcriptomics

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.
