DATA NOTE

The genome sequence of a phantom cranefly, *Ptychoptera contaminata* (Linnaeus, 1758)

[version 1; peer review: 1 approved, 2 approved with reservations]

Liam M. Crowley[1],
University of Oxford and Wytham Woods Genome Acquisition Lab,
Natural History Museum Genome Acquisition Lab,
Darwin Tree of Life Barcoding collective,
Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team,
Wellcome Sanger Institute Scientific Operations: Sequencing Operations,
Wellcome Sanger Institute Tree of Life Core Informatics team,
Tree of Life Core Informatics collective, Darwin Tree of Life Consortium

[1]University of Oxford, Oxford, England, UK

v1 **First published:** 04 Jun 2025, **10**:308
https://doi.org/10.12688/wellcomeopenres.24276.1
Latest published: 04 Jun 2025, **10**:308
https://doi.org/10.12688/wellcomeopenres.24276.1

Abstract

We present a genome assembly from a male specimen of *Ptychoptera contaminata* (phantom cranefly; Arthropoda; Insecta; Diptera; Ptychopteridae). The genome sequence has a total length of 204.08 megabases. Most of the assembly (83.64%) is scaffolded into 6 chromosomal pseudomolecules, including the X and Y sex chromosomes. The mitochondrial genome has also been assembled, with a length of 17.64 kilobases. Gene annotation of this assembly on Ensembl identified 11,102 protein-coding genes.

Keywords

Ptychoptera contaminata, phantom cranefly, genome sequence, chromosomal, Diptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ✓ ? ?

	1	2	3
version 1 04 Jun 2025	✓ view	? view	? view

1. **Taro Nakamura**, National Institute for Basic Biology, Okazaki, Japan
2. **Christopher B. Cunningham**, University of Georgia, Georgia, USA
3. **Arun Arumugaperumal**, Rajalakshmi Engineering College, Chennai, India
Rajalakshmi Engineering College, Thandalam, Chennai, India

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: Crowley LM: Investigation, Resources;

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome through core funding to the Wellcome Sanger Institute (220540) and the Darwin Tree of Life Discretionary Award [218328, https://doi.org/10.35802/218328].
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Crowley LM, University of Oxford and Wytham Woods Genome Acquisition Lab, Natural History Museum Genome Acquisition Lab *et al.* **The genome sequence of a phantom cranefly, *Ptychoptera contaminata* (Linnaeus, 1758) [version 1; peer review: 1 approved, 2 approved with reservations]** Wellcome Open Research 2025, **10**:308 https://doi.org/10.12688/wellcomeopenres.24276.1

First published: 04 Jun 2025, **10**:308 https://doi.org/10.12688/wellcomeopenres.24276.1

Species taxonomy

Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Protostomia; Ecdysozoa; Panarthropoda; Arthropoda; Mandibulata; Pancrustacea; Hexapoda; Insecta; Dicondylia; Pterygota; Neoptera; Endopterygota; Diptera; Nematocera; Ptychopteromorpha; Ptychopteridae; Ptychopterinae; *Ptychoptera*; *Ptychoptera contaminata* (Linnaeus, 1758) (NCBI:txid1572538)

Background

As part of the Darwin Tree of Life Project – which aims to generate high-quality reference genomes for all named eukaryotic species in Britain and Ireland to support research, conservation, and the sustainable use of biodiversity – we present a chromosomally complete genome sequence for the phantom cranefly, *Ptychoptera contaminata*. This genome was assembled using the Tree of Life pipeline from a specimen collected in Wytham Woods, Oxfordshire, United Kingdom (Figure 1).

Genome sequence report

Sequencing data

The genome of a specimen of *Ptychoptera contaminata* (Figure 1) was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating 18.24 Gb (gigabases) from 1.98 million reads, which were used to assemble the genome. GenomeScope analysis estimated the haploid genome size at 228.51 Mb, with a heterozygosity of 0.76% and repeat content of 39.41%. These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 71 coverage. Hi-C sequencing produced 130.83 Gb from 866.43 million reads, used to scaffold the assembly. Table 1 summarises the specimen and sequencing details.

Assembly statistics

The primary haplotype was assembled, and contigs corresponding to an alternate haplotype were also deposited in INSDC databases. The assembly was improved by manual curation, which corrected 57 misjoins or missing joins and removed 2 haplotypic duplications. These interventions decreased the scaffold count by 11.11% and increased the scaffold N50 by 94.42%. The final assembly has a total length of 204.08 Mb in 335 scaffolds, with 153 gaps, and a scaffold N50 of 36.52 Mb (Table 2).

Figure 1. Photograph of the *Ptychoptera contaminata* (idPtyCont2) specimen used for genome sequencing.

The snail plot in Figure 2 provides a summary of the assembly statistics, indicating the distribution of scaffold lengths and other assembly metrics. Figure 3 shows the distribution of scaffolds by GC proportion and coverage. Figure 4 presents a cumulative assembly plot, with separate curves representing different scaffold subsets assigned to various phyla, illustrating the completeness of the assembly.

Most of the assembly sequence (83.64%) was assigned to 6 chromosomal-level scaffolds, representing 4 autosomes and the X and Y sex chromosomes. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 5; Table 3).

The mitochondrial genome was also assembled. This sequence is included as a contig in the multifasta file of the genome submission and as a standalone record.

Assembly quality metrics

The estimated Quality Value (QV) and *k*-mer completeness metrics, along with BUSCO completeness scores, were calculated for each haplotype and the combined assembly. The QV reflects the base-level accuracy of the assembly, while *k*-mer completeness indicates the proportion of expected *k*-mers identified in the assembly. BUSCO scores provide a measure of completeness based on benchmarking universal single-copy orthologues.

The combined primary and alternate assemblies achieve an estimated QV of 58.3. The *k*-mer completeness is 82.76% for the primary haplotype and 80.50% for the alternate haplotype; and 98.47% for the combined primary and alternate assemblies. BUSCO v.5.5.0 analysis using the diptera_odb10 reference set (n = 3,285) identified 93.3% of the expected gene set (single = 92.8%, duplicated = 0.5%).

Table 2 provides assembly metric benchmarks adapted from Rhie *et al.* (2021) and the Earth BioGenome Project Report on Assembly Standards September 2024. The primary assembly achieves the EBP reference standard of **6.7.Q57**.

Genome annotation report

The *Ptychoptera contaminata* genome assembly (GCA_963942525.1) was annotated externally by Ensembl at the European Bioinformatics Institute (EBI). This annotation includes 17,419 transcribed mRNAs from 11,102 protein-coding and 911 non-coding genes. The average transcript length is 6,733.49 bp. There are 1.45 coding transcripts per gene and 5.90 exons per transcript. For further information about the annotation, please refer to https://beta.ensembl.org/species/fd4aed48-b0ed-4076-addb-52e7dbdead79.

Table 1. **Specimen and sequencing data for *Ptychoptera contaminata*.**

Project information			
Study title	Ptychoptera contaminata		
Umbrella BioProject	PRJEB65380		
Species	*Ptychoptera contaminata*		
BioSpecimen	SAMEA112232596		
NCBI taxonomy ID	1572538		
Specimen information			
Technology	**ToLID**	**BioSample accession**	**Organism part**
PacBio long read sequencing	idPtyCont2	SAMEA112233062	whole organism
Hi-C sequencing	idPtyCont1	SAMEA7521282	whole organism
Sequencing information			
Platform	**Run accession**	**Read count**	**Base count (Gb)**
Hi-C HiSeq X Ten	ERR11904113	8.66e+08	130.83
PacBio Sequel IIe	ERR11892473	1.98e+06	18.24

Methods

Sample acquisition and DNA barcoding

The specimen used for genome sequencing was an adult male *Ptychoptera contaminata* (specimen ID Ox002377, ToLID idPtyCont2), collected from Wytham Woods, Oxfordshire, United Kingdom (latitude 51.772, longitude –1.338) on 2022-05-28 by netting. The specimen was collected and identified by Liam Crowley (University of Oxford) and preserved on dry ice.

A second specimen was used for Hi-C sequencing (specimen ID NHMUK014361433, ToLID idPtyCont1). It was a larval specimen collected from Snakeholme Pit, Lincolnshire, United Kingdom (latitude 53.2302, longitude –0.3303) on 2019-03-19, using a kick-net. The specimen was collected and identified by a team from the Environment Agency and preserved by snap freezing.

The initial identification was verified by an additional DNA barcoding process according to the framework developed by Twyford *et al.* (2024). A small sample was dissected from each specimen and stored in ethanol, while the remaining parts were shipped on dry ice to the Wellcome Sanger Institute (WSI) (Pereira *et al.*, 2022). The tissue was lysed, the COI marker region was amplified by PCR, and amplicons were sequenced and compared to the BOLD database, confirming the species identification (Crowley *et al.*, 2023). Following whole genome sequence generation, the relevant DNA barcode region was also used alongside the initial barcoding data for sample tracking at the WSI (Twyford *et al.*, 2024). The standard operating procedures for Darwin Tree of Life barcoding have been deposited on protocols.io (Beasley *et al.*, 2023).

Metadata collection for samples adhered to the Darwin Tree of Life project standards described by Lawniczak *et al.* (2022).

Nucleic acid extraction

The workflow for high molecular weight (HMW) DNA extraction at the Wellcome Sanger Institute (WSI) Tree of Life Core Laboratory includes a sequence of procedures: sample preparation and homogenisation, DNA extraction, fragmentation and purification (Howard *et al.*, 2025). Detailed protocols are available on protocols.io (Denton *et al.*, 2023b). The idPtyCont2 sample was prepared for DNA extraction by weighing and dissecting it on dry ice (Jay *et al.*, 2023). Tissue from the whole organism was homogenised using a PowerMasher II tissue disruptor (Denton *et al.*, 2023a). HMW DNA was extracted using the Automated MagAttract v2 protocol (Oatley *et al.*, 2023a). For ultra-low input (ULI) PacBio sequencing, DNA was fragmented using the Covaris g-TUBE method (Oatley *et al.*, 2023c). Sheared DNA was purified by solid-phase reversible immobilisation, using AMPure PB beads to eliminate shorter fragments and concentrate the DNA (Oatley *et al.*, 2023b). The concentration of the sheared and purified DNA was assessed using a Nanodrop spectrophotometer and Qubit Fluorometer using the Qubit dsDNA High Sensitivity Assay kit. Fragment size distribution was evaluated by running the

Table 2. Genome assembly data for *Ptychoptera contaminata*.

Genome assembly		
Assembly name	idPtyCont2.1	
Assembly accession	GCA_963942525.1	
Alternate haplotype accession	*GCA_963942475.1*	
Assembly level for primary assembly	chromosome	
Span (Mb)	204.08	
Number of contigs	488	
Number of scaffolds	335	
Longest scaffold (Mb)	52.71	
Assembly metric	**Measure**	***Benchmark***
Contig N50 length	2.63 Mb	*≥ 1 Mb*
Scaffold N50 length	36.52 Mb	*= chromosome N50*
Consensus quality (QV)	Primary: 57.1; alternate: 59.0; combined: 58.3	*≥ 40*
k-mer completeness	Primary: 82.76%; alternate: 80.50%; combined: 98.47%	*≥ 95%*
BUSCO*	C:93.3%[S:92.8%,D:0.5%], F:1.3%,M:5.4%,n:3,285	*S > 90%; D < 5%*
Percentage of assembly assigned to chromosomes	83.64%	*≥ 90%*
Sex chromosomes	X and Y	*localised homologous pairs*
Organelles	Mitochondrial genome: 17.64 kb	*complete single alleles*

* BUSCO scores based on the diptera_odb10 BUSCO set using version 5.5.0. C = complete [S = single copy, D = duplicated], F = fragmented, M = missing, n = number of orthologues in comparison.

sample on the FemtoPulse system. For this sample, the extracted DNA had a Qubit concentration of 3.04 ng/μL and a yield of 395.20 ng. Spectrophotometric measurements indicated 260/280 and 260/230 ratios of 2.66 and -1.4, respectively.

Hi-C sample preparation and crosslinking

Hi-C data were generated from the whole organism of the idPtyCont1 sample using the Arima-HiC v2 kit (Arima Genomics) with 20–50 mg of frozen tissue (stored at –80 °C). As per manufacturer's instructions, tissue was fixed, and the DNA crosslinked using a TC buffer with 22% formaldehyde concentration, and a final formaldehyde concentration of 2%. The tissue was then homogenised using the Diagnocine Power Masher-II. The crosslinked DNA was digested using a restriction enzyme master mix, then biotinylated and ligated. A clean up was performed with SPRIselect beads prior to library preparation. DNA concentration was quantified using the Qubit Fluorometer v4.0 (Thermo Fisher Scientific) and Qubit HS Assay Kit, and sample biotinylation percentage was estimated using the Arima-HiC v2 QC beads.

Library preparation and sequencing

Library preparation and sequencing were performed at the WSI Scientific Operations core.

PacBio HiFi

The sample requires Covaris g-TUBE shearing to approximately 10 kb prior to library preparation. Ultra-low input libraries were prepared using PacBio SMRTbell® Express Template Prep Kit 2.0 and PacBio SMRTbell® gDNA Sample Amplification Kit. To begin, samples were normalised to 20 ng of DNA. Initial removal of single-strand overhangs, DNA damage repair, and end repair/A-tailing were performed per manufacturer's instructions. From the SMRTbell® gDNA Sample Amplification Kit, amplification adapters were then ligated. A 0.85X pre-PCR clean-up was performed with

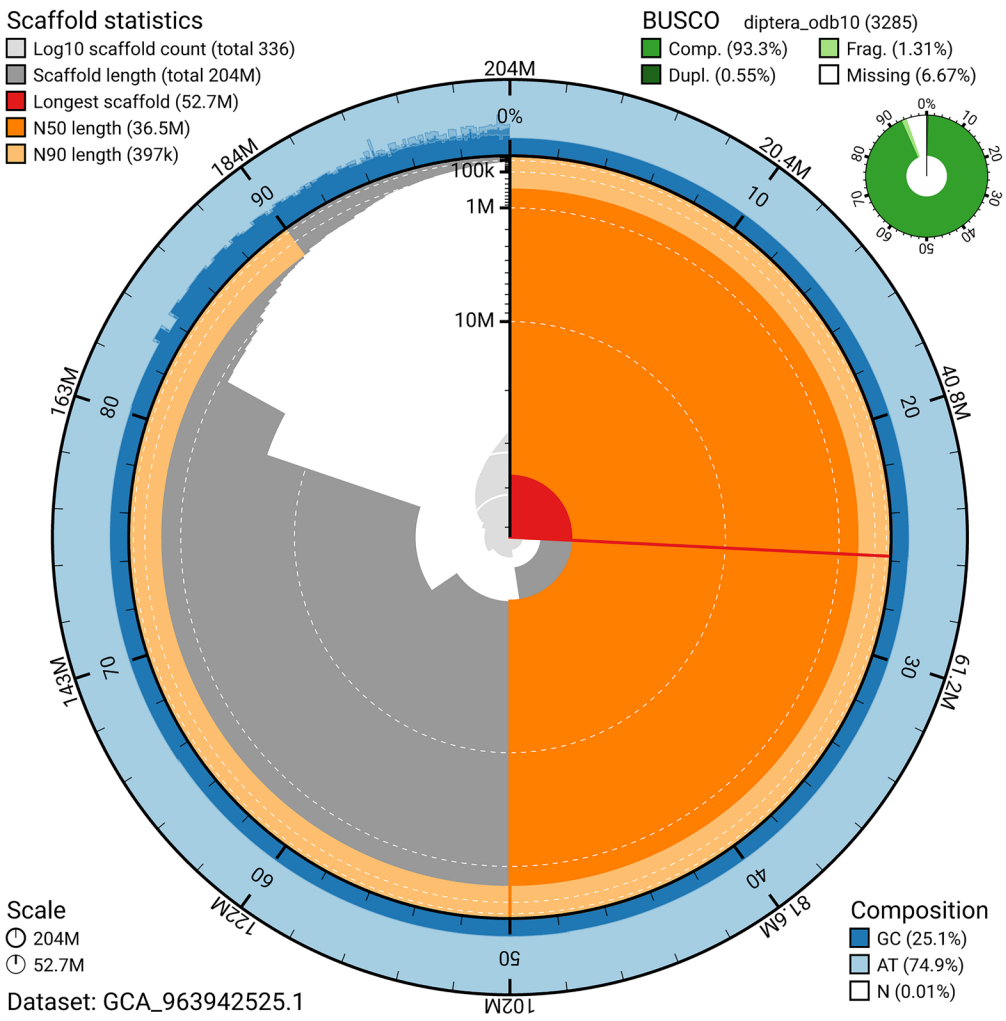

Figure 2. Genome assembly of *Ptychoptera contaminata*, idPtyCont2.1: metrics. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the circumference. The outermost blue tracks display the distribution of GC, AT, and N percentages across the bins. Scaffolds are arranged clockwise from longest to shortest and are depicted in dark grey. The longest scaffold is indicated by the red arc, and the deeper orange and pale orange arcs represent the N50 and N90 lengths. A light grey spiral at the centre shows the cumulative scaffold count on a logarithmic scale. A summary of complete, fragmented, duplicated, and missing BUSCO genes in the diptera_odb10 set is presented at the top right. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/GCA_963942525.1/dataset/GCA_963942525.1/snail.

Promega ProNex beads and the sample was then divided into two for a dual PCR. PCR reactions A and B each followed the PCR programs as described in the manufacturer's protocol. A 0.85X post-PCR clean-up was performed with ProNex beads for PCR reactions A and B and DNA concentration was quantified using the Qubit Fluorometer v4.0 (Thermo Fisher Scientific) and Qubit HS Assay Kit and fragment size analysis was carried out using the Agilent Femto Pulse Automated Pulsed Field CE Instrument (Agilent Technologies) and gDNA 55kb BAC analysis kit. PCR reactions A and B were then pooled, ensuring the total mass was ≥500 ng in 47.4 μl. The pooled sample then repeated the process for DNA damage repair, end repair/A-tailing and additional hairpin adapter ligation. A 1X clean-up was performed with ProNex beads and DNA concentration was quantified using the Qubit and fragment size analysis was carried out using the Agilent Femto Pulse Automated Pulsed Field CE Instrument (Agilent Technologies). Size selection was performed using Sage Sciences' PippinHT system with target fragment size determined by analysis from the Femto Pulse, usually a value between 4000 and 9000 bp. Size selected libraries were then cleaned-up using1.0X ProNex beads and normalised to 2 nM before proceeding to sequencing.

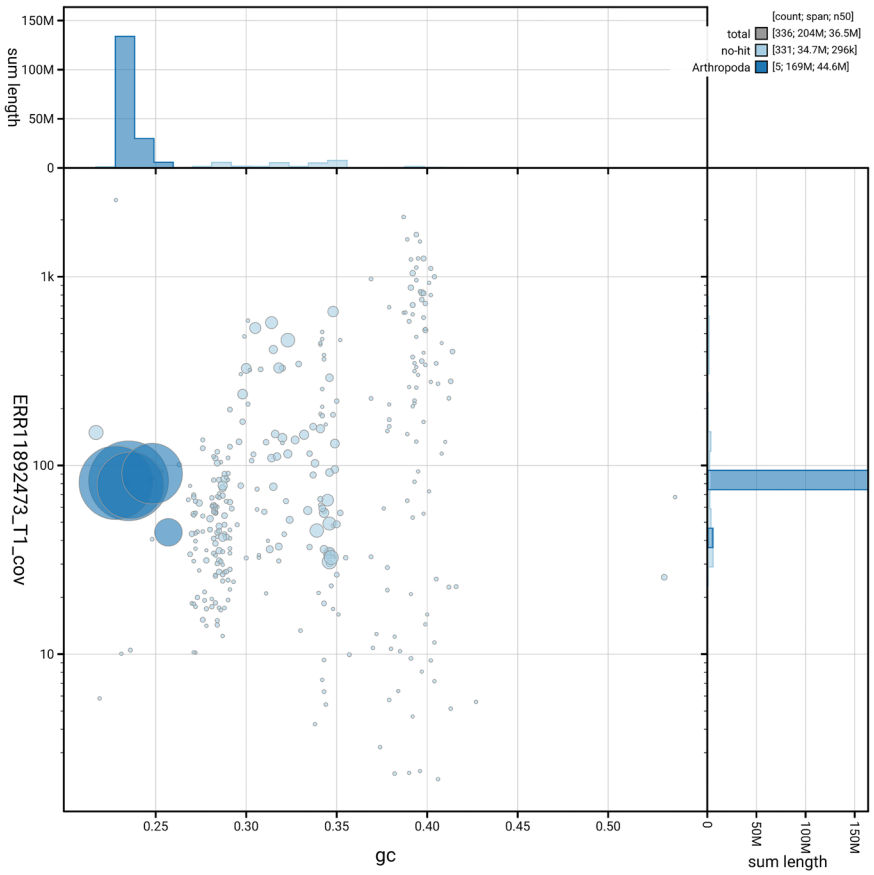

Figure 3. Genome assembly of *Ptychoptera contaminata*, idPtyCont2.1: BlobToolKit GC-coverage plot. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes display the total length of sequences distributed across different levels of coverage and GC content. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/GCA_963942525.1/dataset/GCA_963942525.1/blob.

Samples were sequenced using the Sequel IIe system (Pacific Biosciences, California, USA). The concentration of the library loaded onto the Sequel IIe was in the range 40–135 pM. The SMRT link software, a PacBio web-based end-to-end workflow manager, was used to set-up and monitor the run, as well as perform primary and secondary analysis of the data upon completion.

Hi-C

For Hi-C library preparation, the biotinylated DNA constructs were fragmented using a Covaris E220 sonicator and size-selected to 400–600 bp using SPRISelect beads. DNA was then enriched using Arima-HiC v2 Enrichment beads. The NEBNext Ultra II DNA Library Prep Kit (New England Biolabs) was used for end repair, A-tailing, and adapter ligation, following a modified protocol in which library preparation is carried out while the DNA remains bound to the enrichment beads. PCR amplification was performed using KAPA HiFi HotStart mix and custom dual-indexed adapters (Integrated DNA Technologies) in a 96-well plate format. Depending on sample concentration and biotinylation percentage determined at the crosslinking stage, samples were amplified for 10–16 PCR cycles. Post-PCR clean-up was carried out using SPRISelect beads. The libraries were quantified using the Accuclear Ultra High Sensitivity dsDNA Standards Assay kit (Biotium) and normalised to 10 ng/μL before sequencing. Hi-C sequencing was performed on the HiSeq X Ten instrument.

Genome assembly, curation and evaluation

Assembly

Prior to assembly of the PacBio HiFi reads, a database of *k*-mer counts (k = 31) was generated from the filtered reads using FastK. GenomeScope2 (Ranallo-Benavidez *et al.*, 2020) was used to analyse the *k*-mer frequency distributions, providing estimates of genome size, heterozygosity, and repeat content.

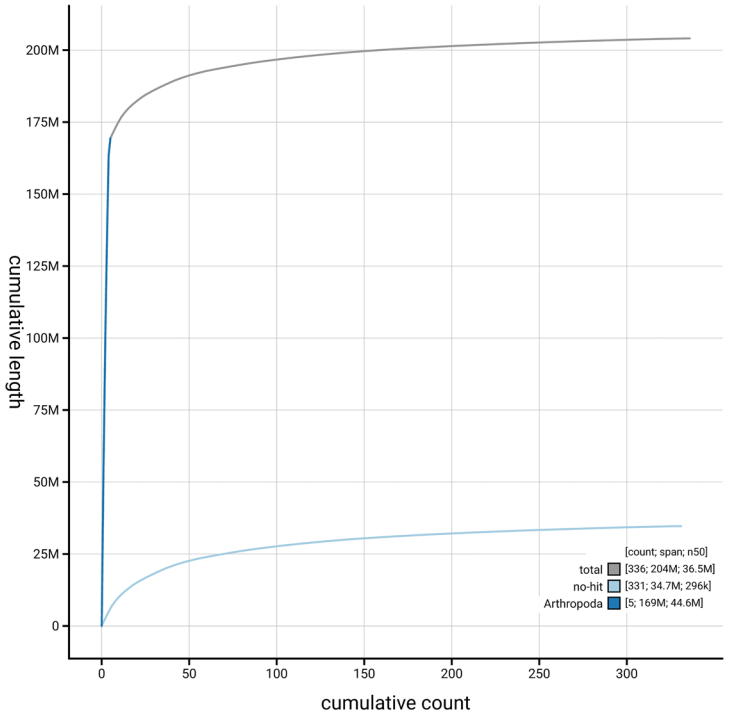

Figure 4. Genome assembly of *Ptychoptera contaminata*, idPtyCont2.1: BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/GCA_963942525.1/dataset/GCA_963942525.1/cumulative.

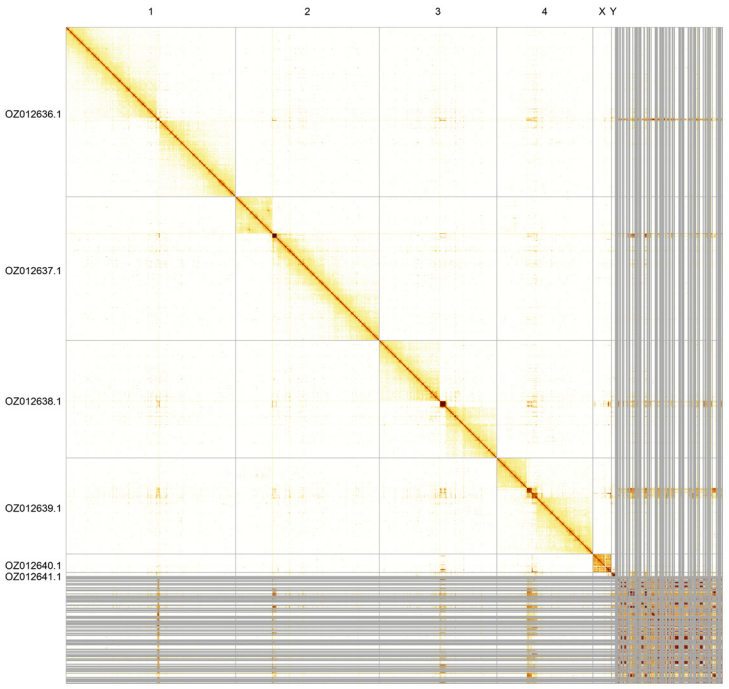

Figure 5. Genome assembly of *Ptychoptera contaminata*. Hi-C contact map of the idPtyCont2.1 assembly, generated using PretextSnapshot. Chromosomes are shown in order of size and labelled with chromosome numbers (top) and chromosome accession numbers (left).

Table 3. Chromosomal pseudomolecules in the genome assembly of *Ptychoptera contaminata*, idPtyCont2.

INSDC accession	Name	Length (Mb)	GC%
OZ012636.1	1	52.71	23.5
OZ012637.1	2	44.61	23
OZ012638.1	3	36.52	23.5
OZ012639.1	4	29.89	25
OZ012640.1	X	5.69	25.5
OZ012641.1	Y	1.26	21.5
OZ012642.1	MT	0.02	23

The HiFi reads were first assembled using Hifiasm (Cheng *et al.*, 2021) with the --primary option. Haplotypic duplications were identified and removed using purge_dups (Guan *et al.*, 2020). The Hi-C reads (Rao *et al.*, 2014) were mapped to the primary contigs using bwa-mem2 (Vasimuddin *et al.*, 2019), and the contigs were scaffolded in YaHS (Zhou *et al.*, 2023) using the --break option for handling potential misassemblies. The scaffolded assemblies were evaluated using Gfastats (Formenti *et al.*, 2022), BUSCO (Manni *et al.*, 2021) and MERQURY.FK (Rhie *et al.*, 2020).

The mitochondrial genome was assembled using MitoHiFi (Uliano-Silva *et al.*, 2023), which runs MitoFinder (Allio *et al.*, 2020) and uses these annotations to select the final mitochondrial contig and to ensure the general quality of the sequence.

Assembly curation

The assembly was decontaminated using the Assembly Screen for Cobionts and Contaminants (ASCC) pipeline. Flat files and maps used in curation were generated via the TreeVal pipeline (Pointon *et al.*, 2023). Manual curation was conducted primarily in PretextView (Harry, 2022) and HiGlass (Kerpedjiev *et al.*, 2018), with additional insights provided by JBrowse2 (Diesh *et al.*, 2023). Scaffolds were visually inspected and corrected as described by Howe *et al.* (2021). Any identified contamination, missed joins, and mis-joins were amended, and duplicate sequences were tagged and removed. The curation process is documented at https://gitlab.com/wtsi-grit/rapid-curation.

Assembly quality assessment

The Merqury.FK tool (Rhie *et al.*, 2020), run in a Singularity container (Kurtzer *et al.*, 2017), was used to evaluate *k*-mer completeness and assembly quality for the primary and alternate haplotypes using the *k*-mer databases (k = 31) computed prior to genome assembly. The analysis outputs included assembly QV scores and completeness statistics.

The genome was analysed in the blobtoolkit pipeline, a Nextflow (Di Tommaso *et al.*, 2017) port of the previous Snakemake Blobtoolkit pipeline (Challis *et al.*, 2020). It aligns the PacBio reads in SAMtools (Danecek *et al.*, 2021) and minimap2 (Li, 2018) and generates coverage tracks for regions of fixed size. In parallel, it queries the GoaT database (Challis *et al.*, 2023) to identify all matching BUSCO lineages to run BUSCO (Manni *et al.*, 2021). For the three domain-level BUSCO lineages, the pipeline aligns the BUSCO genes to the UniProt Reference Proteomes database (Bateman *et al.*, 2023) with DIAMOND blastp (Buchfink *et al.*, 2021). The genome is also divided into chunks according to the density of the BUSCO genes from the closest taxonomic lineage, and each chunk is aligned to the UniProt Reference Proteomes database using DIAMOND blastx. Genome sequences without a hit are chunked using seqtk and aligned to the NT database with blastn (Altschul *et al.*, 1990). The blobtools suite combines all these outputs into a blobdir for visualisation.

The blobtoolkit pipeline was developed using nf-core tooling (Ewels *et al.*, 2020) and MultiQC (Ewels *et al.*, 2016), relying on the Conda package manager, the Bioconda initiative (Grüning *et al.*, 2018), the Biocontainers infrastructure (da Veiga Leprevost *et al.*, 2017), as well as the Docker (Merkel, 2014) and Singularity (Kurtzer *et al.*, 2017) containerisation solutions.

Table 4 contains a list of relevant software tool versions and sources.

Wellcome Sanger Institute – Legal and Governance

The materials that have contributed to this genome note have been supplied by a Darwin Tree of Life Partner. The submission of materials by a Darwin Tree of Life Partner is subject to the **'Darwin Tree of Life Project Sampling Code of Practice'**, which can be found in full on the Darwin Tree of Life website here. By agreeing with and signing up to the Sampling Code of Practice, the Darwin Tree of Life Partner agrees they will meet the legal and ethical requirements and standards set out within this document in respect of all samples acquired for, and supplied to, the Darwin Tree of Life Project.

Further, 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 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

Table 4. Software tools: versions and sources.

Software tool	Version	Source
BLAST	2.14.0	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/
BlobToolKit	4.3.9	https://github.com/blobtoolkit/blobtoolkit
BUSCO	5.5.0	https://gitlab.com/ezlab/busco
bwa-mem2	2.2.1	https://github.com/bwa-mem2/bwa-mem2
DIAMOND	2.1.8	https://github.com/bbuchfink/diamond
fasta_windows	0.2.4	https://github.com/tolkit/fasta_windows
FastK	666652151335353eef2fcd58880bcef5bc2928e1	https://github.com/thegenemyers/FASTK
Gfastats	1.3.6	https://github.com/vgl-hub/gfastats
GoaT CLI	0.2.5	https://github.com/genomehubs/goat-cli
Hifiasm	0.19.5-r587	https://github.com/chhylp123/hifiasm
HiGlass	44086069ee7d4d3f6f3f0012569789ec138f42b84 aa44357826c0b6753eb28de	https://github.com/higlass/higlass
MerquryFK	d00d98157618f4e8d1a9190026b19b471055b22e	https://github.com/thegenemyers/MERQURY.FK
Minimap2	2.24-r1122	https://github.com/lh3/minimap2
MitoHiFi	3	https://github.com/marcelauliano/MitoHiFi
MultiQC	1.14, 1.17, and 1.18	https://github.com/MultiQC/MultiQC
Nextflow	23.04.1	https://github.com/nextflow-io/nextflow
PretextView	0.2.5	https://github.com/sanger-tol/PretextView
PretextSnapshot	-	https://github.com/sanger-tol/PretextSnapshot
purge_dups	1.2.5	https://github.com/dfguan/purge_dups
samtools	1.19.2	https://github.com/samtools/samtools
sanger-tol/ascc	0.1.0	https://github.com/sanger-tol/ascc
sanger-tol/ blobtoolkit	0.4.0	https://github.com/sanger-tol/blobtoolkit
Seqtk	1.3	https://github.com/lh3/seqtk
Singularity	3.9.0	https://github.com/sylabs/singularity
TreeVal	1.2.0	https://github.com/sanger-tol/treeval
YaHS	1.2a.2	https://github.com/c-zhou/yahs

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

Data availability

European Nucleotide Archive: Ptychoptera contaminata. Accession number PRJEB65380; https://identifiers.org/ena.embl/PRJEB65380. The genome sequence is released openly for reuse. The *Ptychoptera contaminata* genome sequencing

initiative is part of the Darwin Tree of Life Project (PRJEB40665) and Sanger Institute Tree of Life Programme (PRJEB43745). All raw sequence data and the assembly have been deposited in INSDC databases. Raw data and assembly accession identifiers are reported in Table 1 and Table 2.

Author information

Members of the University of Oxford and Wytham Woods Genome Acquisition Lab are listed here: https://doi.org/10.5281/zenodo.12157525.

Members of the Natural History Museum Genome Acquisition Lab are listed here: https://doi.org/10.5281/zenodo.12159242.

Members of the Darwin Tree of Life Barcoding collective are listed here: https://doi.org/10.5281/zenodo.12158331.

Members of the Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team are listed here: https://doi.org/10.5281/zenodo.12162482.

Members of Wellcome Sanger Institute Scientific Operations: Sequencing Operations are listed here: https://doi.org/10.5281/zenodo.14870789.

Members of the Wellcome Sanger Institute Tree of Life Core Informatics team are listed here: https://doi.org/10.5281/zenodo.12160324.

Members of the Tree of Life Core Informatics collective are listed here: https://doi.org/10.5281/zenodo.12205391.

Members of the Darwin Tree of Life Consortium are listed here: https://doi.org/10.5281/zenodo.4783558.

References

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

Altschul SF, Gish W, Miller W, *et al.*: **Basic Local Alignment Search Tool.** *J Mol Biol.* 1990; **215**(3): 403–410.
PubMed Abstract | Publisher Full Text

Bateman A, Martin MJ, Orchard S, *et al.*: **UniProt: the Universal Protein Knowledgebase in 2023.** *Nucleic Acids Res.* 2023; **51**(D1): D523–D531.
PubMed Abstract | Publisher Full Text | Free Full Text

Beasley J, Uhl R, Forrest LL, *et al.*: **DNA barcoding SOPs for the Darwin Tree of Life project.** *protocols.io.* 2023; [Accessed 25 June 2024].
Publisher Full Text

Buchfink B, Reuter K, Drost HG: **Sensitive protein alignments at Tree-of-Life scale using DIAMOND.** *Nat Methods.* 2021; **18**(4): 366–368.
PubMed Abstract | Publisher Full Text | Free Full Text

Challis R, Kumar S, Sotero-Caio C, *et al.*: **Genomes on a Tree (GoaT): a versatile, scalable search engine for genomic and sequencing project metadata across the eukaryotic Tree of Life [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2023; **8**: 24.
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): 1361–1374.
PubMed Abstract | Publisher Full Text | Free Full Text

Cheng H, Concepcion GT, Feng X, *et al.*: **Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm.** *Nat Methods.* 2021; **18**(2): 170–175.
PubMed Abstract | Publisher Full Text | Free Full Text

Crowley L, Allen H, Barnes I, *et al.*: **A sampling strategy for genome sequencing the British terrestrial arthropod fauna [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2023; **8**: 123.
PubMed Abstract | Publisher Full Text | Free Full Text

da Veiga Leprevost F, Grüning BA, Alves Aflitos S, *et al.*: **BioContainers: an open-source and community-driven framework for software standardization.** *Bioinformatics.* 2017; **33**(16): 2580–2582.
PubMed Abstract | Publisher Full Text | Free Full Text

Danecek P, Bonfield JK, Liddle J, *et al.*: **Twelve years of SAMtools and BCFtools.** *GigaScience.* 2021; **10**(2): giab008.
PubMed Abstract | Publisher Full Text | Free Full Text

Denton A, Oatley G, Cornwell C, *et al.*: **Sanger Tree of Life sample homogenisation: PowerMash.** *protocols.io.* 2023a.
Publisher Full Text

Denton A, Yatsenko H, Jay J, *et al.*: **Sanger Tree of Life wet laboratory protocol collection V.1.** *protocols.io.* 2023b.
Publisher Full Text

Di Tommaso P, Chatzou M, Floden EW, *et al.*: **Nextflow enables reproducible computational workflows.** *Nat Biotechnol.* 2017; **35**(4): 316–319.
PubMed Abstract | Publisher Full Text

Diesh C, Stevens GJ, Xie P, *et al.*: **JBrowse 2: a modular genome browser with views of synteny and structural variation.** *Genome Biol.* 2023; **24**(1): 74.
PubMed Abstract | Publisher Full Text | Free Full Text

Ewels P, Magnusson M, Lundin S, *et al.*: **MultiQC: summarize analysis results for multiple tools and samples in a single report.** *Bioinformatics.* 2016; **32**(19): 3047–3048.
PubMed Abstract | Publisher Full Text | Free Full Text

Ewels PA, Peltzer A, Fillinger S, *et al.*: **The nf-core framework for community-curated bioinformatics pipelines.** *Nat Biotechnol.* 2020; **38**(3): 276–278.
PubMed Abstract | Publisher Full Text

Formenti G, Abueg L, Brajuka A, *et al.*: **Gfastats: conversion, evaluation and manipulation of genome sequences using assembly graphs.** *Bioinformatics.* 2022; **38**(17): 4214–4216.
PubMed Abstract | Publisher Full Text | Free Full Text

Grüning B, Dale R, Sjödin A, *et al.*: **Bioconda: sustainable and comprehensive software distribution for the life sciences.** *Nat Methods.* 2018; **15**(7): 475–476.
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): 2896–2898.
PubMed Abstract | Publisher Full Text | Free Full Text

Harry E: **PretextView (Paired REad TEXTure Viewer): a desktop application for viewing pretext contact maps**. 2022.
Reference Source

Howard C, Denton A, Jackson B, *et al.*: **On the path to reference genomes for all biodiversity: lessons learned and laboratory protocols created in the Sanger Tree of Life core laboratory over the first 2000 species.** *bioRxiv.* 2025.
Publisher 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

Jay J, Yatsenko H, Narváez-Gómez JP, *et al.*: **Sanger Tree of Life sample preparation: triage and dissection.** *protocols.io.* 2023.
Publisher 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

Kurtzer GM, Sochat V, Bauer MW: **Singularity: scientific containers for mobility of compute.** *PLoS One.* 2017; **12**(5): e0177459.
PubMed Abstract | Publisher Full Text | Free Full Text

Lawniczak MKN, Davey RP, Rajan J, *et al.*: **Specimen and sample metadata standards for biodiversity genomics: a proposal from the Darwin Tree of Life project [version 1; peer review: 2 approved with reservations].** *Wellcome Open Res.* 2022; **7**: 187.
Publisher Full Text

Li H: **Minimap2: pairwise alignment for nucleotide sequences.** *Bioinformatics.* 2018; **34**(18): 3094–3100.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Manni M, Berkeley MR, Seppey 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): 4647–4654.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Merkel D: **Docker: lightweight Linux containers for consistent development and deployment.** *Linux J.* 2014; **2014**(239): 2, [Accessed 2 April 2024].
Reference Source

Oatley G, Denton A, Howard C: **Sanger Tree of Life HMW DNA extraction: automated MagAttract v.2.** *protocols.io.* 2023a.
Publisher Full Text

Oatley G, Sampaio F, Howard C: **Sanger Tree of Life fragmented DNA clean up: automated SPRI.** *protocols.io.* 2023b.
Publisher Full Text

Oatley G, Sampaio F, Kitchin L, *et al.:* **Sanger Tree of Life HMW DNA Fragmentation: Covaris g-TUBE for ULI PacBio.** *protocols.io.* 2023c.
Publisher Full Text

Pereira L, Sivell O, Sivess L, *et al.*: **DToL taxon-specific standard operating procedure for the terrestrial and freshwater arthropods working group.** 2022.
Publisher Full Text

Pointon DL, Eagles W, Sims Y, *et al.*: **sanger-tol/treeval v1.0.0 – Ancient Atlantis.** 2023.
Publisher Full Text

Ranallo-Benavidez TR, Jaron KS, Schatz MC: **GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes.** *Nat Commun.* 2020; **11**(1): 1432.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Rao SSP, 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): 1665–1680.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Rhie A, McCarthy SA, Fedrigo O, *et al.*: **Towards complete and error-free genome assemblies of all vertebrate species.** *Nature.* 2021; **592**(7856): 737–746.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Rhie A, Walenz BP, Koren S, *et al.*: **Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies.** *Genome Biol.* 2020; **21**(1): 245.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Twyford AD, Beasley J, Barnes I, *et al.*: **A DNA barcoding framework for taxonomic verification in the Darwin Tree of Life project [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2024; **9**: 339.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Uliano-Silva M, Ferreira JGRN, Krasheninnikova K, *et al.*: **MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio high fidelity reads.** *BMC Bioinformatics.* 2023; **24**(1): 288.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Vasimuddin M, Misra S, Li H, *et al.*: **Efficient architecture-aware acceleration of BWA-MEM for multicore systems.** In: *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS).* IEEE, 2019; 314–324.
Publisher Full Text

Zhou C, McCarthy SA, Durbin R: **YaHS: Yet another Hi-C Scaffolding tool.** *Bioinformatics.* 2023; **39**(1): btac808.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Open Peer Review

Current Peer Review Status: ✔ ? ?

Version 1

Reviewer Report 14 August 2025

https://doi.org/10.21956/wellcomeopenres.26774.r125906

?

Arun Arumugaperumal
[1] Rajalakshmi Engineering College, Chennai, India
[2] Department of Biotechnology, Rajalakshmi Engineering College, Thandalam, Chennai, Tamil Nadu, India

The genome sequence of *Ptychoptera contaminata*, one of the phantom crane flies, is being discussed in this data note. The photograph of the specimen is wonderfully presented as figure 1. The size of the assembly presented is 204.08 Mb spanning 6 chromosomes. In the case of *P. albimana*, which is a related species, there were 7 chromosomes [1]. The authors can comment on this. Consider changing '71 coverage' to '71X coverage'.

References

1. Sivell O, Webb J, Mitchell R, Sivell D, et al.: The genome sequence of a fold-wing cranefly, Ptychoptera albimana (Fabricius, 1787). *Wellcome Open Research*. 2024; **9**. Publisher Full Text

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

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

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

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

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics; 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, however I have significant reservations, as outlined above.

Reviewer Report 18 July 2025

https://doi.org/10.21956/wellcomeopenres.26774.r125911

?

Christopher B. Cunningham
University of Georgia, Georgia, USA

The project continues to produce quality resources for the insect genomics community. The methods are sound and relatively well documented. The article needs has enough details to be reasonably reproducible. The data supporting the publication are publicly available.

Abstract. There is no justification within the abstract, just results. At least a sentence that this was part of a large survey of insects is needed. And one sentence for its possible value.

Background. Background. The stated motivation for this study in the introduction is fine – part of a very large survey. However, it would be nice for the reader to understand if this is the first of its genus, is of particular interest for ecological or genetics studies, or something unusual about its biology. Helping the reader understand what this genome might help with investigation in.

Results. Table 1 is not really needed. They are not details that the reader needs to under the work and they can all be found under the NCBI BioProject ID or are repeated from the methods paragraph.

BUSCO completeness needs to be assessed with the updated OrthoDB version 12. Difference from 10 to 12 have made a big difference for several insects I work with and colleagues have reported the same. The paper should not be published until this is updated. The new version has been for 6 months.

Figure 4. Fig 4 has very little information beyond what is found elsewhere in the manuscript and can easily be summarized as a sentence in the results.

Figure 3 & 5. Suggest lots of assembly dead-ends. I appreciate that they have no hits to anything other than Arthropoda but still the variation in coverage, variation in GC content, and strong Hi-C links among themselves is very odd. It suggests the QC of the genome draft assembly is incomplete.

The Genome annotation should give the reader some indicator of the quality of the annotation. Nothing is known to the reader other than it's a reasonable amount of genes. No BUSCO, no

Ontology, no percentage of hits to other Dipteran proteins, etc. It was produced by a standing pipeline but the reader does not even have the slightest hint of how well that really worked.

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

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: Genetics, Genomics, Epigenetics, Insects, Behavior, Reproduction.

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 04 July 2025

https://doi.org/10.21956/wellcomeopenres.26774.r125913

Taro Nakamura
National Institute for Basic Biology, Okazaki, Japan

This Data Note reports the chromosomal-level genome assembly of the phantom cranefly, Ptychoptera contaminata, as part of the Darwin Tree of Life project. The work represents a valuable contribution to insect genomics and biodiversity genomics efforts. The manuscript is clear, and provides sufficient methodological detail for reproducibility. The data have been deposited in public repositories, ensuring open access and facilitating downstream research.

Sequencing was performed using state-of-the-art PacBio HiFi long-read technology and Hi-C scaffolding, yielding a high-quality genome (204.08 Mb, scaffold N50 = 36.52 Mb, 83.64% of the assembly assigned to chromosomes).
Assembly quality is well-documented using metrics such as BUSCO (93.3% complete), QV (58.3), and k-mer completeness.
Annotation via Ensembl has identified 11,102 protein-coding genes and 911 non-coding genes.

The manuscript thoroughly documents the collection, DNA extraction, sequencing, assembly, and annotation protocols, including software versions and quality control procedures.

Suggestions for Improvement:
While the inclusion of the Hi-C contact map (Figure 5) is highly valuable, the figure's current presentation—with relatively thick grid lines—makes it difficult to discern finer contact patterns, especially in regions with dense signals. For improved clarity and accessibility, I recommend refining the figure as follows:
Use thinner grid lines or reduce the opacity of grid overlays so as not to overpower the actual Hi-C signal.

Conclusion and Recommendation:
The manuscript provides a high-quality genomic resource with clear, well-documented methods and public data availability. The work meets the standards for a data note, and the minor suggestions above are meant to further enhance clarity and user-friendliness. If the authors can address the suggestions above, this manuscript will be ready for acceptance.

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: Developmental Biology

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.