Jul 29, 2025

Project Psyche sample collection to genome assembly publication workflow

DOI
dx.doi.org/10.17504/protocols.io.yxmvmbk1bg3p/v1

Karen Houliston[1], Caroline Howard[1], Kerstin Howe[1], Shane A. McCarthy[1], Jo Wood[1], Nancy Holroyd[1], Charlotte J. Wright[1], Joana I Meier[1]

[1]Tree of Life, Wellcome Sanger Institute, Hinxton, Cambridgeshire, CB10 1SA

project-psyche

Charlotte J. Wright
Wellcome Sanger Institute

DOI: https://dx.doi.org/10.17504/protocols.io.yxmvmbk1bg3p/v1

Protocol Citation: Karen Houliston, Caroline Howard, Kerstin Howe, Shane A. McCarthy, Jo Wood, Nancy Holroyd, Charlotte J. Wright, Joana I Meier 2025. Project Psyche sample collection to genome assembly publication workflow. **protocols.io** **https://dx.doi.org/10.17504/protocols.io.yxmvmbk1bg3p/v1**

Protocol status: Working
We use this protocol and it's working

Created: July 26, 2025

Last Modified: July 29, 2025

Protocol Integer ID: 223336

Keywords: Project Psyche, Lepidoptera, butterflies, moths, reference genomes, project psyche sample collection, project psyche, genome, assembly publication, lepidoptera, species of lepidoptera, chromosome, collaborative project, protocol summarises the workflow, workflow, species

Funders Acknowledgements:
Wellcome Trust
Grant ID: Wellcome Trust award 220540/Z/20/A 'Wellcome Sanger Institute Quinquennial Review 2021–2026'

Abstract

This protocol summarises the workflow employed by Project Psyche (**https://www.projectpsyche.org/**) – a collaborative project which aims to generate chromosome-level reference genomes for all 11,000 species of Lepidoptera found in Europe.

Troubleshooting

Sample collection

1 Samples are collected by Sample Collection Hubs (**https://www.projectpsyche.org/theorganisation/psyche-hubs/**) across Europe together with their collaborators and sub-hubs. All necessary permits and permissions are acquired before collection happens. Night-flying species are mostly collected with light traps in the evening and day-flying species with hand nets by day. Sampling aims for 1-3 adult females for larger, and 3-5 females for smaller species (<10 mm body size). Female specimens are preferred because females are the heterogametic sex in Lepidoptera, and thus allows identification and assembly of all sex chromosomes. Moths are typically identified in the field, if possible, and then put alive in vials, whereas butterflies are typically temporarily stored alive in glassine envelopes. In the lab or field lab, the specimens are photographed and processed in a standardised way following the Project Psyche standard operating procedures for sampling Lepidoptera.

The specimens are typically killed with CO_2 from dry ice or by putting them in the freezer. All specimens are stored in fluidX cryotubes provided by the Wellcome Sanger Institute (WSI). Each fluidX tube has a unique barcode that enables it to be tracked in databases. Small specimens (< 10 mm body size) are stored in a single FluidX cryotube, whereas larger specimens are dissected and stored in at least three FluidX cryotubes (thorax, head, abdomen) provided by the Wellcome Sanger Institute. For species that cannot be identified by morphological examination but can be identified through barcoding, a leg is put in a separate tube for barcoding. To avoid degradation of DNA and RNA, collected specimens are snap-frozen in liquid nitrogen or in -80°C freezers. Specimens are either immediately dissected on dry ice before snap-freezing, or this is done later on dry ice. Crucially, specimens remain deep-frozen until DNA extraction. The continuity of the cold chain is essential. If dry ice is not available, processing is performed very quickly after death and the tubes are immediately stored in a -80°C freezer or in a cryogenic dry shipper. A google spreadsheet with information on the specimen, such as species name, tissue type, sex, GSP collection location, people involved, etc. (hereafter called the “manifest”) is filled out for each tube. The specimens are photographed with a colour checker and next to the barcoded FluidX tube, where the head (or entire specimen) will be placed. Photos are uploaded to BioImage Archive (https://www.ebi.ac.uk/biostudies/bioimages/studies/S-BIAD1504). Each specimen is assigned a ToLID, a brief unique identifier in the format of ilGenSpec1, where i=insect, l=lepidoptera, Gen=first three letters of the genus, Spec=first four letters of the species name, 1=specimen number. These ToLIDs are used to track specimens throughout the production process at WSI. More information on how to use and assign ToLIDs (**https://id.tol.sanger.ac.uk/**).

Sending specimens to a sequencing hub

2 In Phase 1 of Project Psyche, the WSI is the only sequencing hub. In later phases, the number of sequencing hubs will be expanded. Currently, only the Sample Collection Hub leaders can send specimens to the WSI. First, they fill out a collector onboarding form specifying general information of the next intended shipment (e.g. approximate number of species they intend to send and from which countries). The WSI sample management team then sends a manifest to be used to record the metadata for each specimen (see 'Sample Collection'). Once the manifest is completed by sample submitters, it is checked by the WSI sample management team to make sure that it contains all the required information and that it is all correctly formatted. The WSI compliance teams then check that all permits and legal contracts are in order and apply for export/import permits as required. The WSI sample management team then organises the shipment of the specimens on dry ice with a cold-chain specialist transportation company to ensure that the specimens are not at risk of DNA degradation in case of any delays during transport.

DNA extraction and sequencing

3 To generate each genome, we perform long-range sequencing (currently Hi-C), long-read sequencing (currently Pacific Biosciences, PacBio). RNA-sequencing data is also generated to facilitate gene annotation.

For tiny specimens (e.g. micromoths), an entire specimen is used for DNA extraction for long-read (PacBio) sequencing, another specimen is used for long-range (Hi-C) sequencing and a third one for
RNA extraction. For larger specimens, the thorax is used for DNA extraction, the head for Hi-C sequencing and the abdomen for RNA extraction. The workflow for high molecular weight (HMW) DNA extraction at the WSI Tree of Life Core Laboratory involves a sequence of procedures: sample preparation and homogenisation, DNA extraction, fragmentation and purification **(Howard et al. 2025)**. Detailed protocols are available on protocols.io (https://www.protocols.io/view/sanger-tree-of-life-wet-laboratory-protocol-collec-8epv5xxy6g1b/v2) and summarised below.

Sample preparation and HMW DNA extraction

4 Samples are prepared for DNA extraction by weighing and dissecting them on dry ice **(Jay et al. 2023)**. Tissue from the whole organism (tiny specimens) or thorax (larger specimens) is homogenised using a PowerMasher II tissue disruptor **(Denton et al. 2023)**. HMW DNA is extracted using the Automated MagAttract v2 protocol **(Oatley, Denton, et al. 2023)**. Where DNA yield and quality are sufficient for PacBio HiFi library prep, shearing is performed using the MegaRuptor method **(Bates et al. 2023)** to provide fragments of 12-22 kb. For very tiny specimens where ultra-low input (ULI) PacBio sequencing is required due to low DNA yield, fragmentation is conducted using the Covaris g-TUBE method **(Oatley, Sampaio, et al. 2023)** to provide fragments of 9-11 kb. Sheared DNA is purified by solid-phase reversible immobilisation, using AMPure PB

beads to eliminate shorter fragments and concentrate the DNA **(Oatley, Denton, et al. 2023)**. The concentration of the sheared and purified DNA is assessed using a Nanodrop spectrophotometer and Qubit Fluorometer using the Qubit dsDNA High Sensitivity Assay kit. Fragment size distribution is evaluated by running the sample on the FemtoPulse system.

PacBio library preparation and sequencing

5 For samples with total mass exceeding 400 ng, low-input PacBio libraries are prepared with the SMRTbell Prep Kit 3.0 protocol (Pacific Biosciences, California, USA) according to the manufacturer's instructions. The kit includes reagents for end repair/A-tailing, adapter ligation, post-ligation SMRTbell bead clean-up, and nuclease treatment. Size selection and clean-up are performed using diluted AMPure PB beads (Pacific Biosciences). DNA concentration is quantified using a Qubit Fluorometer v4.0 (ThermoFisher Scientific) and the Qubit 1X dsDNA HS assay kit. Final library fragment size is assessed with the Agilent Femto Pulse Automated Pulsed Field CE Instrument (Agilent Technologies) using the gDNA 55 kb BAC analysis kit.

For samples with less than 400 ng, Ultra-low input libraries are prepared using PacBio SMRTbell® Express Template Prep Kit 2.0 and PacBio SMRTbell® gDNA Sample Amplification
Kit. Samples are normalised to 20 ng of DNA. Single-strand overhang removal, DNA damage repair, and end repair/A-tailing are performed according to the manufacturer's instructions. Amplification adapters are then ligated using the SMRTbell® gDNA Sample Amplification Kit. A 0.85x pre-PCR clean-up is performed using Promega ProNex beads, i.e. magnetic bead to DNA solution volume ratio is 0.85x. The sample is then split for two separate PCR reactions following the manufacturer's protocol. Each PCR reaction undergoes another 0.85x beads clean-up. DNA concentration is then quantified using the Qubit Fluorometer v2.0 (Thermo Fisher Scientific) and Qubit HS Assay Kit, and fragment size is assessed with the Agilent Femto Pulse Automated Pulsed Field CE Instrument (Agilent Technologies) using the gDNA 55 kb BAC Analysis Kit. If both PCR reactions are of sufficient quality, they are pooled, ensuring a combined DNA mass of ≥500 ng in 47.4 µl. The pooled sample then undergoes a second round of DNA damage repair, end repair/A-tailing, and ligation of hairpin adapters. A 1x clean-up is performed with ProNex beads. DNA concentration and fragment size are again evaluated using the Qubit and Femto Pulse systems. Size selection is performed using the Sage Sciences PippinHT system, with target fragment size (typically 4,000–9,000 bp) determined from Femto Pulse analysis. Lastly, the size-selected libraries undergo a final 1x ProNex clean-up.

Depending on the expected genome sizes as inferred by GoaT **(Challis et al. 2023)** (**https://goat.genomehubs.org/**), two to four samples are sequenced together on a PacBio Revio 25M SMRT cell (Pacific Biosciences, California, USA). Prepared libraries are normalised to 2 nM, and 15 µL is used for making complexes. Primers are annealed and polymerases are bound to create circularised complexes according to the manufacturer's instructions. Complexes are purified using 1.2x SMRTbell beads, then diluted to the Revio loading concentration (200–300 pM) and spiked with a Revio sequencing internal control. The SMRT link software, a PacBio web-based end-to-end workflow manager, is used to set up and monitor the run, and to carry out primary and secondary data analysis.

Hi-C preparation and sequencing

6 Tissue from the head (or whole specimen for tiny specimens) is processed for Hi-C sequencing using the Arima-HiC v2 kit. In brief, 20-50 mg of frozen tissue (stored at -70°C) is fixed, and the DNA is crosslinked using a TC buffer with a final formaldehyde concentration of 2%. After crosslinking, the tissue is homogenised using the Diagnocine Power Masher-II and BioMasher-II tubes and pestles. Following the Arima-HiC v2 kit manufacturer's instructions, crosslinked DNA is digested using a restriction enzyme master mix. The 5′-overhangs are filled in and labelled with biotinylated nucleotides and proximally ligated. An overnight incubation is carried out for enzymes to digest remaining proteins and for crosslinks to reverse. A clean-up is performed with SPRIselect beads prior to library preparation. The biotinylation percentage is estimated using the Qubit Fluorometer v4.0 (Thermo Fisher Scientific) with the Qubit HS Assay Kit and using Arima-HiC v2 QC beads.

For Hi-C library preparation, DNA is fragmented using the Covaris E220 sonicator (Covaris) and then size-selected using SPRISelect beads to a range of 400 to 600 bp. The DNA is then enriched using the Arima-HiC v2 kit Enrichment beads. Using the NEBNext Ultra II DNA Library Prep Kit (New England Biolabs) for end repair, A-tailing, and adapter ligation. This uses a custom protocol which resembles the standard NEBNext Ultra II DNA Library Prep protocol, but where library preparation occurs while DNA is bound to the Enrichment beads. For library amplification, 10 to 16 PCR cycles are required, determined by the sample biotinylation percentage. The Hi-C sequencing is performed using paired-end sequencing with a read length of 150 bp on an Illumina NovaSeqX instrument. Samples are multiplexed based on the genome size, with 10 samples of genome size <1.5 Gb run together on a lane of a NovaSeqX 25B lane (~1 Tbp data).

RNA extraction and sequencing

7 RNA is extracted from abdomen tissue of larger species or a whole specimen of smaller species **(do Amaral et al. 2023)**. The RNA concentration is assessed using a Nanodrop spectrophotometer and a Qubit Fluorometer using the Qubit RNA Broad-Range Assay kit. RNA libraries are prepared using the NEBNext® Ultra™ II Directional RNA Library Prep Kit for Illumina (New England Biolabs), following the manufacturer's instructions. Poly(A)

mRNA in the total RNA solution is isolated using oligo(dT) beads, converted to cDNA, and uniquely indexed; 14 PCR cycles are performed. Libraries are size-selected to produce fragments between 100–300 bp. Libraries are quantified, normalised, pooled to a final concentration of 2.8 nM, and diluted to 150 pM for loading. Specimens are sequenced on an Illumina NovaSeqX 25B lane

Data QC

8 Analyse all generated data and describe it in ToLQC (https://tolqc.cog.sanger.ac.uk/). The default requirement for a diploid genome is 25x coverage in PacBio Hifi reads and 50x coverage in Hi-C reads. Once sufficient data has been generated, start the automated genome assembly process.

Genome assembly

9 Prior to assembly of the PacBio HiFi reads, a database of *k*-mer counts (*k* = 31) is generated from the filtered reads using FastK (https://github.com/thegenemyers/FASTK). The *k*-mer frequency distributions are analysed through GenomeScope2 **(Ranallo-Benavidez et al. 2020)**, providing estimates of the genome size, heterozygosity, and repeat content.

The HiFi reads are first assembled into contigs using Hifiasm **(Cheng et al. 2021)** in Hi-C mode enabling phasing of the contigs. Hi-C reads are then mapped to the contigs using bwa-mem2 **(Vasimuddin et al. 2019)**, and the contigs are scaffolded in YaHS **(Zhou et al. 2023)**, using the --break option to handle potential mis-assemblies. The scaffolded assemblies are evaluated using Gfastats **(Formenti et al. 2022)**, BUSCO **(Manni et al. 2021)** and **MERQURY.FK** **(Rhie et al. 2020)**.

The mitochondrial genome is 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 ensure the overall quality of the sequence.

Manual genome curation

10 The assembly is decontaminated using the WSI Assembly Screen for Cobionts and Contaminants (ASCC) pipeline (https://github.com/sanger-tol/ascc). Both haplotypes are combined for curation. Sequence data is remapped to a curated assembly using the WSI manual curation pipeline (https://github.com/sanger-tol/curationpretext). Flat files and maps used in curation are generated via the TreeVal pipeline (https://zenodo.org/records/10047654) (**Pointon et al., 2023**). Manual curation is conducted primarily in PretextView (https://github.com/sanger-tol/PretextView) and HiGlass **(Kerpedjiev et al. 2018)**. Scaffolds are visually inspected and corrected as described by **(Howe et al. 2021)**. Any identified contamination, missed joins, and misjoins

are manually amended in PretextView, and duplicate sequences are tagged and removed. Sex chromosomes are identified based on coverage in females (Z and W chromosomes are at half coverage compared to autosomes in females) and Merian element painting **(Wright et al. 2024)**, where a male has been sequenced (the canonical Z chromosome is highly conserved in Lepidoptera). All autosomes are labelled according to size. The full manual curation process is documented in detail at https://gitlab.com/wtsi-grit/rapid-curation.

Assembly quality assessment

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

The genomes are analysed using the BlobToolKit pipeline, a Nextflow **(Di Tommaso et al. 2017)** implementation of the earlier Snakemake BlobToolKit pipeline **(Challis et al. 2020)**. The pipeline aligns PacBio reads using minimap2 **(Li 2018)** and SAMtools **(Danecek et al. 2021)** to generate coverage tracks. Simultaneously, it queries the GoaT database **(Challis et al. 2023)** to identify relevant BUSCO lineages and runs BUSCO **(Manni et al. 2021)**. For the three domain-level BUSCO lineages, BUSCO genes are aligned to the UniProt Reference Proteomes database **(UniProt Consortium 2023)** using DIAMOND blastp **(Buchfink et al. 2021)**. The genome is divided into chunks based on the density of BUSCO genes from the closest taxonomic lineage, and each chunk is aligned to the UniProt Reference Proteomes database with DIAMOND blastx. Sequences without hits are chunked using seqtk and aligned to the NT database with blastn **(Altschul et al. 1990)**.

The BlobToolKit suite consolidates all 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)**, with package management via Conda (**https://docs.conda.io/**) and Bioconda **(Grüning et al. 2018)**, and containerisation through Docker **(Merkel 2014)** and Singularity **(Kurtzer et al. 2017)**.

Data release and genome notes

12 All curated assemblies, including metadata, the mitochondrial genome, raw sequencing data and potential co-biont assemblies, are publicly released to the European Nucleotide Archive (ENA) following the model of BioProject organisation (https://www.earthbiogenome.org/report-on-assembly-standards) recommended by the Earth Biogenome Project. The BioProject for Project Psyche is PRJEB71705.Genomes can be downloaded from ENA (**https://www.ebi.ac.uk/ena/**) or NCBI

> (**https://www.ncbi.nlm.nih.gov/**).
>
> Every genome is accompanied by a brief “genome note” that introduces the genome sequence(s) and acknowledges all contributors. Genome notes are published in Wellcome Open Research under an open peer-review policy, at which point they are citable and receive a doi. All Project Psyche genome notes are available in the Project Psyche gateway in Wellcome Open Research (https://wellcomeopenresearch.org/gateways/treeoflife/projectpsyche).
>
> Genome notes include a brief introduction, a description of the methods and links to protocols, an automatically generated summary of the assembly quality, and full accession information. Specimen collectors are listed as first author(s), and if a different Psyche member writes the introduction, they are listed as the second author. If the genome was curated by a member of the Project Psyche community rather than a member of the WSI curation team, then the curator is also acknowledged as an author.
>
> After publication, the journal invites peer reviewers. Peer review comments and the names of the peer reviewers are published alongside the article. An article has passed peer review when it receives two “approved”, or one “approved” and two “approved with reservations” ratings. New versions of the data note may be submitted with a description of all changes made. All articles appear in Google Scholar. When a data note passes peer review, it will be indexed in PubMed, PubMed Central, Europe PMC, Scopus, etc.

Protocol references

Allio, R. et al., 2020. MitoFinder: Efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. *Molecular ecology resources*, 20(4), pp.892–905.

Altschul, S.F. et al., 1990. Basic local alignment search tool. *Journal of molecular biology*, 215(3), pp.403–410.

do Amaral, R.J.V. et al., 2023. Sanger Tree of Life RNA Extraction: Automated MagMaxTM mirVana. protocols.io. Available at: https://www.protocols.io/view/sanger-tree-of-life-rna-extraction-automated-magma-6qpvr36n3vmk/v1 [Accessed May 3, 2024].

Bates, A., Clayton-Lucey, I. Howard, C., 2023. Sanger Tree of Life HMW DNA Fragmentation: Diagenode Megaruptor®3 for LI PacBio. protocols.io. Available at: https://www.protocols.io/view/sanger-tree-of-life-hmw-dna-fragmentation-diagenod-81wgbxzq3lpk/v1 [Accessed May 3, 2024].

Buchfink, B., Reuter, K. Drost, H.-G., 2021. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature methods*, 18(4), pp.366–368.

Challis, R. et al., 2020. BlobToolKit - Interactive Quality Assessment of Genome Assemblies. *G3*, 10(4), pp.1361–1374.

Challis, R. et al., 2023. Genomes on a Tree (GoaT): A versatile, scalable search engine for genomic and sequencing project metadata across the eukaryotic tree of life. *Wellcome Open Research*, 8(24). Available at: https://wellcomeopenresearch.org/articles/8-24/v1 [Accessed June 19, 2024].

Cheng, H. et al., 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature methods*, 18(2), pp.170–175.

Danecek, P. et al., 2021. Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2). Available at: http://dx.doi.org/10.1093/gigascience/giab008.

Denton, A. et al., 2023. Sanger Tree of Life Sample Homogenisation: PowerMash. protocols.io. Available at: https://www.protocols.io/view/sanger-tree-of-life-sample-homogenisation-powermas-5qpvo3r19v4o/v1 [Accessed November 15, 2024].

Di Tommaso, P. et al., 2017. Nextflow enables reproducible computational workflows. *Nature biotechnology*, 35(4), pp.316–319.

Ewels, P. et al., 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19), pp.3047–3048.

Ewels, P.A. et al., 2020. The nf-core framework for community-curated bioinformatics pipelines. *Nature biotechnology*, 38(3), pp.276–278.

Formenti, G. et al., 2022. Gfastats: conversion, evaluation and manipulation of genome sequences using assembly graphs. *Bioinformatics (Oxford, England)*, 38(17), pp.4214–4216.

Grüning, B. et al., 2018. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature methods*, 15(7), pp.475–476.

Howard, C. et al., 2025. 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*, p.2025.04.11.648334. Available at: https://www.biorxiv.org/content/10.1101/2025.04.11.648334v1.abstract [Accessed July 26, 2025].

Howe, K. et al., 2021. Significantly improving the quality of genome assemblies through curation. *GigaScience*, 10(1). Available at: http://dx.doi.org/10.1093/gigascience/giab153.

Jay, J. et al., 2023. Sanger Tree of Life Sample Preparation: Triage and Dissection. *protocols.io*. Available at: https://www.protocols.io/view/sanger-tree-of-life-sample-preparation-triage-and-ctzex6je [Accessed July 26, 2025].

Kerpedjiev, P. et al., 2018. HiGlass: web-based visual exploration and analysis of genome interaction maps. *Genome biology*, 19(1), p.125.

Kurtzer, G.M., Sochat, V. & Bauer, M.W., 2017. Singularity: Scientific containers for mobility of compute. *PloS one*, 12(5), p.e0177459.

Li, H., 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* , 34(18), pp.3094–3100.

Manni, M. et al., 2021. BUSCO update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Molecular biology and evolution*, 38(10), pp.4647–4654.

Merkel, D., 2014. Docker: lightweight Linux containers for consistent development and deployment. *Linux Journal*, 2014, p.2.

Oatley, G., Sampaio, F., et al., 2023. Sanger Tree of Life HMW DNA Fragmentation: Covaris g-TUBE for ULI PacBio. *protocols.io*. Available at: https://www.protocols.io/view/sanger-tree-of-life-hmw-dna-fragmentation-covaris-cztpx6mn.pdf [Accessed September 9, 2024].

Oatley, G., Denton, A. & Howard, C., 2023. Sanger Tree of Life HMW DNA Extraction: Automated MagAttract v.2. *protocols.io*. Available at: https://www.protocols.io/view/sanger-tree-of-life-hmw-dna-extraction-automated-mkxygx3y4dg8j/v1 [Accessed May 3, 2024].

Ranallo-Benavidez, T.R., Jaron, K.S. & Schatz, M.C., 2020. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature communications*, 11(1), p.1432.

Rhie, A. et al., 2020. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome biology*, 21(1), p.245.

Uliano-Silva, M. et al., 2023. MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio high fidelity reads. *BMC bioinformatics*, 24(1), p.288.

UniProt Consortium, 2023. UniProt: The universal protein knowledgebase in 2023. *Nucleic acids research*, 51(D1), pp.D523–D531.

Vasimuddin, M. et al., 2019. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. In *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. IEEE, pp. 314–324.

Wright, C.J. et al., 2024. Comparative genomics reveals the dynamics of chromosome evolution in Lepidoptera. *Nature ecology & evolution*, 8(4), pp.777–790.

Zhou, C., McCarthy, S.A. & Durbin, R., 2023. YaHS: yet another Hi-C scaffolding tool. *Bioinformatics* , 39(1). Available at: http://dx.doi.org/10.1093/bioinformatics/btac808.