Title: The genome of the Margined White butterfly (*Pieris macdunnoughii*): sex chromosome insights and the power of polishing with PoolSeq data.

Running title: Pieris macdunnoughii genome

Authors:

Rachel A. Steward 1

Yu Okamura 2

Carol L. Boggs 3,4,5

Heiko Vogel 2

Christopher W. Wheat 1

Affiliations:

1 Department of Zoology, Stockholm University, Stockholm, Sweden

2 Max Planck Institute for Chemical Ecology, Jena, Germany

3 School of the Earth, Ocean and Environment, University of South Carolina, SC, USA

4 Department of Biology, University of South Carolina, Columbia, SC, USA

5 Rocky Mountain Biological Laboratory, Crested Butte, CO, USA

***Author for Correspondence:** Rachel Steward, Department of Zoology, Stockholm University, Stockholm, Sweden, rachel.steward@zoologi.su.se

Abstract

We report a chromosome-level assembly for *Pieris macdunnoughii*, a North American butterfly whose involvement in an evolutionary trap imposed by an invasive Eurasian mustard has made it an emerging model system for studying maladaptation in plant-insect interactions. Assembled using nearly 100X coverage of Oxford Nanopore long reads, the contig-level assembly comprised 106 contigs totaling 316,549,294 bases, with an N50 of 5.2Mb. We polished the assembly with PoolSeq Illumina short-read data, demonstrating for the first time the comparable performance of individual and pooled short reads as polishing datasets. Extensive synteny between the reported contig-level assembly and a published, chromosome-level assembly of the European butterfly *Pieris napi* allowed us to generate a pseudo-chromosomal assembly of 47 contigs, placing 91.1% of our 317 Mbp genome into a chromosomal framework. Additionally, we found support for a Z chromosome arrangement in *P. napi*, showing that the fusion event leading to this rearrangement predates the split between European and North American lineages of *Pieris* butterflies. This genome assembly and its functional annotation lay the groundwork for future research into the genetic basis of adaptive and maladaptive egg-laying behavior by *P. macdunnoughii*, contributing to our understanding of the susceptibility and responses of insects to evolutionary traps.

Keywords: genome, long-read sequencing, polishing, PoolSeq, *Pieris*, evolutionary trap

Significance Statement

The North American butterfly *Pieris macdunnoughii* lays eggs on an invasive host plant that is lethal to its larvae and is emerging as a model system for the study of maladaptive responses to rapid environmental change. We constructed a high-quality genome for this butterfly using Nanopore long-read data. We further demonstrated the performance of pooled short-read data for genome polishing. This study demonstrates the performance of Nanopore-only assembly approaches, reveals a novel role for pooled sequencing data in genome assembly, and provides an important resource for advancing research on maladaptive plant-insect interactions.

Introduction

Anthropogenically induced rapid environmental change has led to many novel interactions between species and their biotic and abiotic environments. In some cases, historically adaptive behaviors have become evolutionary traps, wherein species continue to respond to cues that are no longer reliably linked to beneficial outcomes (Schlaepfer et al. 2002; Robertson et al. 2013). Thus, organisms might mistime life history events (van Dyck et al. 2015), fail to adjust to novel predators (Brown et al. 2018), or use poor, novel resources despite good, historical alternatives being available (Szaz et al. 2015). Identifying genetic variation involved in evolutionary traps can uncover how such maladaptive interactions arise, evolve, or persist through time, potentially informing conservation management (Robertson & Blumstein 2019; Supple & Shapiro 2018). In many cases, however, the tools needed to begin these investigations are missing; notably, high-quality genome assemblies.

P. macdunnoughii is one of at least five species (or semispecies) of North American *Pieris* butterflies (Chew & Watt 2006), resulting from the holarctic expansion of the *Pieris napi* species complex into North America (3-5 mya; Geiger & Shapiro 1992). This montane butterfly ranges across the southern Rocky Mountains and is a hostplant specialist, laying eggs and feeding upon plants in the Brassicaceae. Importantly, it is the focus of ongoing research into how specialization has resulted in maladaptive interactions with invasive Eurasian mustards, wherein females oviposit on invasive plants despite their lethality to caterpillars (Chew 1977, 1975; Nakajima et al. 2013; Nakajima & Boggs 2015; Steward & Boggs 2020; Steward et al. 2019). Given the extensive literature investigating interactions between *Pieris* species and their hostplants, and advances in understand these interactions at the genomic level (Edger et al.

2015), developing a functional genomics understanding of these interactions in a species such as *P. macdunnoughii* is a logical next step, but one that is hindered by the lack of genomic resources for this species.

Here, we present an annotated chromosome-level genome for *P. macdunnoughii,* generated using high-coverage Oxford Nanopore Technology (ONT) data. In addition, we explore assembly pipelines that leverage datasets often used for non-model organisms, such as PoolSeq data (i.e., a combination of individuals in a single sequencing library). Long-read genomes generally suffer from low accuracy and benefit from polishing with short reads (Watson & Warr 2019; Walker et al. 2014). We compared improvements to the assembly when polished with individual versus PoolSeq short read data, as PoolSeq data can also serve as an initial dataset quantifying genetic diversity within a population, greatly enriching the output of genome sequencing pipelines.

Finally, the recent publication of a chromosome-level genome for *P. napi* revealed unanticipated and extensive rearrangements compared to other Lepidoptera, despite no change in chromosomal count (Hill et al. 2019). Other high-quality genomes from *Pieris* are needed to discern whether the reciprocal translocation events leading to this rearrangement continue to shape the evolution of this lineage. Additionally, although the Z chromosome for *P. napi* was not completely assembled, it appears to have a novel fusion event (Hill et al. 2019; Pruisscher et al. 2021). By comparing our *P. macdunnoughii* assembly with the published *P. napi* genome, we shed light on evolutionary dynamics.

Results and Discussion

ONT assembly

Using only high-coverage (~90X) ONT long reads, we achieved a highly contiguous *Pieris macdunnoughii* draft assembly (v0.08) totaling 319,093,312 bases across 106 contigs with an N50 of 5.20Mbp (Fig.1A,B; SI: Table S1,S2). The assembly pipeline involved merging two long-read assemblies, one constructed with FLYE, the other with NECAT, each extensively polished and purged of duplicated haplotypes both before and after merging (v0.01 – v0.08; details available in SI: Table S1). Prior to polishing with additional short-read data, the v.0.08 draft assembly was already very complete, with 93.8% complete single copy Lepidoptera BUSCOs, with only 0.4% duplicated (Fig.1B).

Short-read polishing and annotation assessment

Illumina short reads from a single *P. macdunnoughii* female and from a pool of *P. macdunnoughii* females were both sufficient for polishing the v0.08 draft assembly, improving the assembly in comparable ways (Fig.1A-C; SI: Table S2). Polishing increased complete lepidopteran BUSCOs to 96.9% in the individual-polished (v0.09) and 96.7% in the pool-polished (v0.10) assemblies.

Our annotation identified 19,640 good transcripts (containing start and stop codons) in the unpolished assembly, representing 17,362 unique genes (Table 1). Annotations of the polished genomes identified fewer transcripts and genes (18,347 and 18,603 good transcripts, respectively). To evaluate the annotations' completeness, we compared them to a high-quality published annotation for *Bombyx mori* (Bombycidae) using an ortholog hit ratio (OHR) analysis

(modified from O'Neil et al. 2010). This analysis compares putative orthologs in two annotations (e.g., *B. mori* proteins vs. unpolished *P. macdunnoughii* proteins) by generating a ratio of the *P. macdunnoughii* protein length to the *B. mori* protein length (ratios approaching 1 indicate more complete recovery of the expected ortholog length). Illumina polishing improved the number and OHR of recovered orthologs in the *P. macdunnoughii* annotations, as expected (Miller et al. 2018), regardless of whether individual or PoolSeq reads were used (Fig.1C; Table 1).

Synteny and Z chromosome validation

There was a high degree of chromosome-scale synteny between the pool-polished assembly (v0.10) and the *P. napi* reference genome, validating our assembly accuracy at this scale. This synteny corroborates a massive chromosomal rearrangement in *Pieris napi* relative to model lepidopteran systems with chromosome-level assemblies, as previously reported by Hill et al. (2019). Combined with evidence that this rearrangement is also shared by the cabbage white butterfly *Pieris rapae* (Hill et al. 2019), our results suggest that despite the rampant rate of chromosomal rearrangements in the ancestral lineage of *Pieris* butterflies, no major fissions or fusions have subsequently occurred.

Extensive synteny supported superscaffolding the *P. macdunnoughii* genome using the *P. napi* chromosomes, thereby producing what we refer to as a pseudo-chromosomal assembly (v0.10_RagTag), which reduced our assembly to 47 contigs with an N50 of 13.0 Mbp (Fig.2C; SI: Table S2). Of the 106 contigs of the *P. macdunnoughii* assembly, 84 (288,535,466 bp, 91.1%) were aligned to chromosomes 1-25 of the *P. napi* assembly, while 22 contigs (28,013,828 bp) remained unplaced.

While previous work suggested that two unplaced *P. napi* scaffolds belonged on the Z chromosome (Pruisscher et al. 2021), our assembly of *P. macdunnoughii* indicates where they should be located (Fig. 2B). We validated that *P. macdunnoughii* Sc0000000/ Chromsome_1_RagTag is a contiguous Z chromosome by assessing the depth of PoolSeq reads across this chromosome, finding half as many reads mapped (30.2 ± 0.3) compared to the autosomes (51.2± 0.4; Fig.2D), consistent with females having only one Z chromosome. We then used our PoolSeq data to estimate nucleotide diversity (π), calculated across 50kbp sliding windows, identifying 4,876,412 variable sites, and estimating a genome-wide mean for autosomal π of 0.0060. Genetic diversity was on average 52.3% lower on the Z chromosome (Chromsome_1_RagTag) than on the autosomes. Across autosomes, estimates of π were relatively even and similar (ex. Chromosome_2_RagTag and Chromosome_3_RagTag, Fig. 2E), though we identified several regions of low genetic diversity (ex. Chromosome_23). This analysis illustrates the dual potential of PoolSeq data in a genome assembly pipeline, both as an effective dataset for short-read polishing and resource for population genetic inferences.

Conclusion

Our assembly for *P. macdunnoughii*, an emerging model system for studying maladaptation in plant-insect coevolutionary interactions, placed 91.1% of the 317 Mbp genome into a chromosomal framework and exhibited a high and accurate gene content using BUSCO metrics. We annotated 18,603 good transcripts for 16,496 genes. Our results provide an important validation of previously reported chromosomal rearrangements in *Pieris* butterflies and insights into their Z chromosome evolution, not only supporting the neo-Z structure of *P. napi* but

indicating that this fusion event happened before the split between European and North American lineages of *P. napi*. Further, we localized two large, previously unplaced scaffolds of *P. napi* into their proper locations on the Z chromosome. This genome assembly and its functional annotation lay the groundwork for identifying the genetic basis of persistent maladaptive egg-laying behavior by *P. macdunnoughii*, as well as the inability of larvae to eat invasive Eurasian mustards. At least two of the other North American *Pieris* species are involved in similar maladaptive host-plant interactions (Davis & Cipollini 2014; Keeler & Chew 2008). This genome will serve as an important resource for future research into susceptibility and responses to evolutionary traps.

Materials and Methods

Genome sequencing

Pieris macdunnoughii individuals were collected near Gothic, CO, USA. The thorax of one female, collected in 2014, was sampled for genome sequencing. High molecular weight genomic DNA was isolated from the sample using paramagnetic nanodiscs (Nanobind Tissue Big DNA kit, Circulomics). The isolated DNA was processed using the Short Read Eliminator XS (Circulomics), selectively precipitating high molecular weight DNA based on polyethylene glycol (PEG) and sodium chloride concentrations. Isolated DNA was sequenced on a MinION platform, with fast base-calling using GUPPY (v.3.2.10) in the miniKNOW (v.3.6.5) software. A total of 3.82 million reads (28.17Gb) were generated from one flow cell (~90X coverage). The N50 length of the fast base-called subreads was 17.24 kb. We repeated base-calling with the

high-accuracy option in GUPPY (v.4.0.11) generating another set of raw reads with an N50 of 17.24Kb. Raw reads have been deposited (ENA accession: ERS5472768).

Assembly

From the raw reads we generated three assemblies, fast base-calling assembled with Flye (v.2.7, Kolmogorov et al. 2019), high-accuracy base-calling assembled with Flye and high-accuracy base-calling assembled with NECAT (v.0.0.1, Chen et al. 2020). For both Flye and NECAT, genome size was set as 300mb. Flye was run with the -meta option, which improves assemblies with biased read coverage (i.e., mitochondrial genome), and two polishing iterations. We ran four rounds of Racon (v.1.4.13, Vaser et al. 2017) on each assembly, followed by Medaka (v.1.0.3, https://nanoporetech.github.io/medaka/) to polish the assemblies with the nanopore data, using default settings (Latorre-Pérez et al. 2020; Huang et al. 2020).

We consolidated haplotype redundancies in each assembly using PURGEhaplotigs (v.1.0.3, default settings, Roach et al. 2018). Quickmerge (v.0.3, Chakraborty et al. 2016) was used to merge the polished and purged Flye (self_assembly, -l 197436) and NECAT (hybrid_assembly) assemblies. Finally, we used HaploMerger2 (v.20180603, Huang et al. 2017), to collapse duplicated haploid content in the Quickmerge assembly. Quality and completeness of these preliminary genome assemblies were assessed using SeqKit (v.0.12.1, Shen et al. 2016) and BUSCO using lepidoptera_odb10 (v.4.1.2, Seppey et al. 2019). BUSCO scores were visualized in R with scripts modified from the BUSCO output. All work in R was supported by the *tidyverse* (v.1.3.0, Wickham et al. 2019) and *ggpubr* (v.0.4.0, Kassambara 2020) packages.

Illumina sequencing and mapping

We produced two sets of Illumina short-read sequence data that were used to polish the draft assembly. The first set came from an individual *P. macdunnoughii* female collected in the Upper East River Valley, CO, USA in 2015. The second set came from pooled DNA (PoolSeq) from 18 adult *P. macdunnoughii* females reared in the lab in 2005. This lab population originated from the Upper East River Valley, but was likely more inbred than wild *P. macdunnoughii*. Extraction was per individual, using a commercial kit (cell and tissue DNA kit) for robotic extraction on a KingFisher Duo Prime purifier (ThermoFisher Scientific), following standard protocols with added RNAse A to remove RNA contamination. DNA concentration and purity were quantified using a Qubit 2.0 fluorometer (ThermoFisher Scientific) and nanodrop (ThermoFisher Scientific) and run on a 2% agarose gel stained with GelRed to visually ascertain that DNA fragmentation was minimal. Samples were then pooled using an equal amount of DNA from each individual and used for library preparation and sequencing (Illumina HiSeq), which was performed by SciLifeLab (Stockholm, Sweden), using 150-bp paired-end reads with 350 bp insert size.

Illumina sequencing datasets were filtered for PCR clones (Stacks 1.21; SI: Table S3), adapters trimmed using truseq and nextera reference databases, and filtered for read quality (bbmap 34.86, bbduk2.sh) with a base quality threshold of 20 and a minimum read length of 40. We mapped reads to the v0.08 draft assembly with NextGenMap (SI: Table S4, v.0.5.5, Sedlazeck et al. 2013) and assessed mapped reads using GoLeft (v.0.2.1, https://github.com/brentp/goleft/releases) and Qualimap (v.2.2.1, Okonechnikov et al. 2016).

Short-read polishing and annotation assessment

We polished the final unpolished genome with each set of the mapped Illumina short reads using Pilon (v.1.23, Walker et al. 2014). We compared the polished and unpolished genomes using SeqKit and BUSCO analyses. Higher read coverage can improve the quality of polishing by Pilon (Walker et al. 2014), so we used SeqKit to subsample $1.60x10^8$ of the individual sample reads. Polishing the draft genome with the subsample (coverage comparable to the PoolSeq data) had little impact when evaluated using BUSCOs (SI: Table S1).

We used the Braker2 automated (v.2.1.5, Brůna et al. 2020; Hoff et al. 2019, 2016; Stanke et al. 2008, 2006; Buchfink et al. 2015; Lomsadze et al. 2005; Ter-Hovhannisyan et al. 2008) pipeline to generate comprehensive annotations of all three assemblies (v0.08, v0.09, v0.10). Prior to running Braker2 we soft-masked the assembly (SI: Table S1,S2) with RED (v.05/22/2015, Girgis 2015) using the redmask.py wrapper (v0.0.2, https://github.com/nextgenusfs/redmask). We ran Braker2 in the genome and protein mode, using reference proteins from the Arthropoda section of OrthoDB (v.10). We also ran EggNOG on the final PoolSeq-polished genome annotation to generate a functional annotation.

Annotations were assessed using the longest ortholog hit ratio (OHR), modified from O'Niel et al. (2010) for proteins. For each of the draft annotations (v0.08, v0.09, v0.10), we used CD-HIT (v.4.8.1, Li & Godzik 2006; Fu et al. 2012) to collapse protein clusters, from which we created blast databases (NCBI BLAST v. 2.5.0). The *Bombyx mori* protein set was accessed from NCBI (GCF_000151625.1_ASM15162v1). Each *B. mori* protein was blasted against the *P. macdunnoughii* databases. We calculated OHR as the proportion of a *B. mori* protein that was

overlapped by an orthologous alignment hit in the draft *P. macdunnoughii* annotations. In our analysis, we focus on the OHR of the longest overlapping sequence.

Synteny and pseudo-chromosomal assembly

We used *nucmer* (MUMmer4, v.4.0.0beta2, Marçais et al. 2018) to align the PoolSeq-polished *P. macdunnoughii* genome to the genome of the closely related Eurasian species *Pieris napi* (Hill et al. 2019). Due to considerable synteny between the two *Pieris* genomes, we used the *P. napi* genome as a reference to arrange the PoolSeq-polished *P. macdunnoughii* scaffolds into putative chromosomes for a pseudo-chromosomal assembly. To do this, RagTag (Alonge et al. 2019) was used to group and orient scaffolds against chromosomes 1 through 25 of the *P. napi* genome. We excluded unplaced *P. napi* scaffolds to target chromosomal placements of the *P. macdunnoughii* contigs, setting a lower grouping confidence threshold of 80% to avoid chimeric chromosomes. RagTag added 100 N between each joined scaffold, increasing the length of the genome by 5600bp. Synteny was visualized in R with the packages *circlize* (v 0.4.12, Gu 2014) and *RColorBrewer* (v.1.1-2, Neuwirth 2014).

Genome-wide variation

To illustrate their potential for subsequent population genetic inferences, the PoolSeq reads were mapped to the final pseudo-chromosomal assembly with NextGenMap and filtered for proper pairs. We generated a pileup file (Samtools v.1.0, Li 2011) and used PoPoolation (v.1.2.2, Kofler et al. 2011) to filter insertions and deletions before calculating nucleotide diversity (π) over 50,000bp windows in each pseudo-chromosomal contig.

Acknowledgements

We would like to thank K. Tunström and H. Dort for extensive feedback throughout the process of assembling, annotating, and evaluating the *P. macdunnoughii* genome. T. Karasov and C. Lemire helped collect and rear butterflies, funded by the Stanford Undergraduate Field Studies. Funding for this research was provided by the Max Planck Society, Carl Tryggers Stiftelse anslag (CTS 18:415 to CWW and RAS), Grant-in-Aid for Scientific Research from Japan Society for the Promotion of Science (nos: 202060676 to Yu Okamura), and the Swedish Research Council (2017-04386 to CWW).

Data availability statement

Final genome assemblies and annotations have been archived on ENA (https://www.ebi.ac.uk/ena/), under the project number PRJEB42400. Also available on ENA are the Oxford Nanopore Technology MinION fastq sequences used for the assembly (Accession number ERS5472768) and Illumina short-read whole genome sequences used for polishing (accession numbers ERS5520571, ERS5520572). Bash and R scripts for all steps described above are provided at https://github.com/rstewa03/Pieris_macdunnoughii_genome/SI. Original BRAKER2 annotations have been archived at https://github.com/rstewa03/Pieris_macdunnoughii_genome/data .

Author Contributions

YO and HV performed the long-read sequencing and preliminary assemblies. RAS and CWW completed the final assemblies and annotations, analyzed the data, and wrote the manuscript,

with input from YO, HV and CLB. RAS and CLB collected samples for sequencing. All authors reviewed the manuscript.

References

Alonge M, Soyk S, Ramakrishnan S, Want X, Goodwin S, Sedlazek FJ, Lippman ZB, Schatz MC. 2019. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. Genome Biology. 20:224.

Brown, TR, Coleman, RA, Swearer, SE, Hale, R. 2018. Behavioral responses to, and fitness consequences from, an invasive species are life-stage dependent in a threatened native fish. Biological conservation. 228:10-16.

Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. 2021. BRAKER2: Automatic Eukaryotic Genome Annotation with GeneMark-EP+ and AUGUSTUS Supported by a Protein Database. NAR Genomics and Bioinformatics. 3:lqaa108.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. Nature Methods 12:59–60.

Chakraborty M, Baldwin-Brown JG, Long AD, Emerson JJ. 2016. Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. Nucleic Acids Res. 44:e147–e147.

Chen Y et al. 2020. Fast and accurate assembly of Nanopore reads via progressive error correction and adaptive read selection. bioRxiv. 2020.02.01.930107.

Chew FS. 1975. Coevolution of Pierid butterflies and their cruciferous foodplants I. The relative quality of available resources. Oecologia.

Chew FS. 1977. Coevolution of Pierid butterflies and their cruciferous foodplants II. Distribution of eggs on potential foodplants. Evolution. 31:568–579.

Chew FS, Watt WB. 2006. The green-veined white (*Pieris napi* L.), its Pierine relatives, and the systematics dilemmas of divergent character sets (Lepidoptera, Pieridae). Biol J Linn Soc. 88:413–435.

Davis SL, Cipollini D. 2014. Do mothers always know best? Oviposition mistakes and resulting larval failure of *Pieris virginiensis* on *Alliaria petiolata*, a novel, toxic host. Biol Invasions. 16:1941–1950.

Edger PP, Heidel-Fischer HM, Bakaert M, Rota J, Glöckner G, Platts AE, Heckel DG, Der JP, Wafula EK, Tang M, Hofberger JA, Smithson A, Jall JC, Blanchette M, Bureau TE, Wright SI, dePamphilis CW, Schranz EM, Barker MS, Conant GC, Wahlberg N, Vogel H, Pires JC, Wheat CW. 2015, The butterfly plant arms-race escalated by gene and genome duplications. Proceedings of the National Academy of Sciences. 112:8362:8366.

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 28:3150–3152.

Geiger H, Shapiro AM. 1992. Genetics, systematics and evolution of holarctic *Pieris napi* species group populations (Lepidoptera, Pieridae). Journal of Zoological Systematics and Evolutionary Research. 30:100–122.

Girgis HZ. 2015. Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. BMC Bioinformatics. 16:227.

Gu Z. 2014. circlize implements and enhances circular visualization in R. Bioinformatics. 30:2811-2.

Hill J, Pasi R, Hornett EA, Neethiraj R, Clark N, Morehouse N, Celorio-Mancera MP, Cols JC, Dircksen H, Meslin C, Keehnen N, Pruisscher P, Sikkink K, Vives M, Vogel H, Wiklund K, Woronik A, Boggs CL, Nylin S, Wheat CW. 2019. Unprecedented reorganization of holocentric chromosomes provides insights into the enigma of lepidopteran chromosome evolution. Science Advances. 5:eaau3648.

Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. 2016. BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. Bioinformatics. 32:767–769.

Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. 2019. Whole-Genome Annotation with BRAKER. In: Kollmar M (ed) Gene Prediction: Methods and Protocols. Springer, New York, NY, pp 65–95.

Huang S, Kang M, Xu A. 2017. HaploMerger2: rebuilding both haploid sub-assemblies from high-heterozygosity diploid genome assembly. Bioinformatics. 33:2577–2579.

Huang Y-T, Liu P-Y, Shih P-W. 2020. High-Quality Genomes of Nanopore Sequencing by Homologous Polishing. bioRxiv. 2020.09.19.304949.

Kassambara A (2020). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0. https://CRAN.R-project.org/package=ggpubr

Keeler MS, Chew FS. 2008. Escaping an evolutionary trap: preference and performance of a native insect on an exotic invasive host. Oecologia. 156:559–568.

Kofler R et al. 2011. PoPoolation: A Toolbox for Population Genetic Analysis of Next Generation Sequencing Data from Pooled Individuals. PLOS ONE. 6:e15925.

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. Nature Biotechnology. 37:540–546.

Latorre-Pérez A, Villalba-Bermell P, Pascual J, Vilanova C. 2020. Assembly methods for nanopore-based metagenomic sequencing: a comparative study. Scientific Reports. 10:13588.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 27: 2987-93.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 22:1658–1659.

Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. 2005. Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Research. 33:6494–6506.

Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. 2018. MUMmer4: A fast and versatile genome alignment system. PLOS Computational Biology. 14:e1005944.

Miller DE, Staber C, Zeitlinger J, Hawley RS. 2018. Highly Contiguous Genome Assemblies of 15 Drosophila Species Generated Using Nanopore Sequencing. G3: Genes, Genomes, Genetics. 8:3131–3141.

Nakajima M, Boggs CL. 2015. Fine-grained distribution of a non-native can alter the population dynamics of a native consumer. PLoS ONE. 10:e0143052.

Nakajima M, Boggs CL, Bailey S, Reithel J, Paape T. 2013. Fitness costs of butterfly oviposition on a lethal non-native plant in a mixed native and non-native plant community. Oecologia. 172:823–832.

Neuwirth E. 2014. RColorBrewer: ColorBrewer Palettes. R package version 1.1-2.

https://CRAN.R-project.org/package=RColorBrewer

Okonechnikov K, Conesa A, García-Alcalde F. 2016. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics. 32:292–294.

O'Neil ST, Dzurisin JDK, Carmichael RD, Lobo NF, Hellman, JJ. 2010. Population-level transcriptome sequencing of nonmodel organisms *Erynnis propertius* and *Papilio zelicaon*. BMC Genomics. 11:310.

Pruisscher P, Nylin S, Wheat CW, Gotthard K. A region of the sex chromosome associated with population differences in diapause induction contains highly divergent alleles at clock genes. 2021. Evolution. Unassigned.

Roach MJ, Schmidt SA, Borneman AR. 2018. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics. 19:460.

Robertson, BA, Rehage, JS, Sih, JS. 2013. Ecological novelty and the emergence of evolutionary traps. Trends in Ecology & Evolution. 28:552-560.

Robertson, BA, Blumstein, DT. 2019. How to disarm an evolutionary trap. Conservation Science and practice. 1:e116.

Schlaepfer, MA, Runge, MC, Sherman, PW. 2002. Ecological and evolutionary traps. Trends in Ecology & Evolution. 17:474-480.

Sedlazeck FJ, Rescheneder P, von Haeseler A. 2013. NextGenMap: fast and accurate read mapping in highly polymorphic genomes. Bioinformatics. 29:2790–2791.

Seppey M, Manni M, Zdobnov EM. 2019. BUSCO: Assessing Genome Assembly and Annotation Completeness. In: Gene Prediction: Methods and Protocols. Kollmar, M, editor. Methods in Molecular Biology Springer: New York, NY pp. 227–245.

Shen W, Le S, Li Y, Hu F. 2016. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. PLOS ONE. 11:e0163962.

Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics 24:637–644.

Stanke M, Schöffmann O, Morgenstern B, Waack S. 2006. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics. 7:62.

Steward RA, Boggs CL. 2020. Experience may outweigh cue similarity in maintaining a persistent host-plant-based evolutionary trap. Ecological Monographs. 90:e01412.

Steward RA, Fisher LM, Boggs CL. 2019. Pre- and post-ingestive defenses affect larval feeding on a lethal invasive host plant. Entomologia Experimentalis et Applicata. 167:292–305.

Supple MA, Shapiro B. 2018. Conservation of biodiversity in the genomics era. Genome Biology. 19:131.

Szaz, D, Horvath, G, Barta, A, Robertson, BA, Egri, A, Tarjanyi, N, Racz, G, Kriska, G. Lamp-Lit Bridges as Dual Light-Traps for the Night-Swarming Mayfly, *Ephoron virgo*: Interaction of Polarized and Unpolarized Light Pollution. PLoS ONE. 10: e0121194.

Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M. 2008. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. Genome Research 18:1979–1990.

van Dyck, H, Bonte, D, Puls, Rik, Gotthard, K, Maes, D. 2015. The lost generation hypothesis: could climate change drive ectotherms into a developmental trap? Oikos. 124:54-61.

Vaser R, Sović I, Nagarajan N, Šikić M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 27:737–746.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q Wortman J, Young SK, Earl AM. 2014. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLOS ONE. 9:e112963.

Watson M, Warr A. 2019. Errors in long-read assemblies can critically affect protein prediction. Nature Biotechnology. 37:124–126.

Wickham H, Averick, M, Bryan, J, Chang W, McGowen LD, Francois R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H. 2019. Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Tables

Table 1: Genes identified in Braker2 annotations of P. macdunnoughii assemblies and ortholog hit ratio (OHR) analysis with *Bombyx mori*.

	Annotation			
	B. mori	Unpolished	Individual polished	Pool polished
Good transcripts	NA	19640	18347	18603
Genes	14802	17362	16251	16496
Clustered proteins (90%)	14439	17550	16260	16501

Total *B. mori* orthologs	NA	13599	13637	13669
Median OHR of longest hit	NA	0.96	0.98	0.98
Longest hits with OHR > 0.95	NA	7062	8203	8233
Median identity of longest hit	NA	63.1%	64.6%	64.6%
Longest hits with identity > 95%	NA	240	277	283
Hits in both the unpolished and polished annotation	NA	NA	13524	13550

NA: not applicable

Figure captions

Figure 1. Genome assembly pipeline and metrics for 10 genomes show dramatic improvement during refinement steps. (A) Progressive increase in N50 and decrease in total contigs during polishing and merging of the nanopore assembly. Steps to refine the assembly included polishing with Illumina whole genome short reads (WGS) from a single individual and a pool of 18 individuals. (B) Assessment of the content and quality of 5286 lepidopteran single copy orthologs shows complete, duplicated, fragmented or missing BUSCOs across the 10 assemblies (v.01-v.10). (C) Assessment of changes in genome quality using whole genome annotations show similar effects of polishing using individual or PoolSeq Illumina data. (C) Polishing with Illumina short reads improved the ortholog hit ratio (OHR, values closer to 1 indicate a higher quality annotation).

Figure 2. Chromosome-level assessment of synteny, read depth and genetic variation in *P. macdunnoughii.* (A) A circle plot showing each contig of the *P. macdunnoughii* v0.10 (non-colored scaffolds) and *P. napi* (colored scaffolds, showing scaffolds > 1Mb representing 90.9%

of the 318bp assembly) assemblies, with lines between them showing aligned genomic regions of greater than 5000 bp and 90% identity (for example, *P. napi* Chromosome2 is covered by four *P. macdunnoughii* scaffolds, Sc0000044, Sc0000004, Sc0000047, and Sc0000054). (B) Detailed assessment of the alignments between the Z chromosome of *P. macdunnoughii* (Sc0000000) and aligned, unplaced *P. napi* scaffolds modScaffold_17_1 and modScaffold_95_1, supporting their inclusion on the Z chromosome between 3003162 - 5135852bp and 10002530 – 10466150bp, respectively. (C) The final pseudo-chromosomal *P. macdunnoughii* assembly (v0.10_RagTag) aligned with *P. napi.* (D) Consistently lower read depths of PoolSeq reads mapped to the pseudo-chromosomal assembly support conclusions about the Z chromosome (Pmac_chromosome_1_RagTag; colors as in B). (E) Nucleotide diversity (π), varied across the genome, as seen in representative autosomes 2, 3, and 23.

Figures

Figure 1.

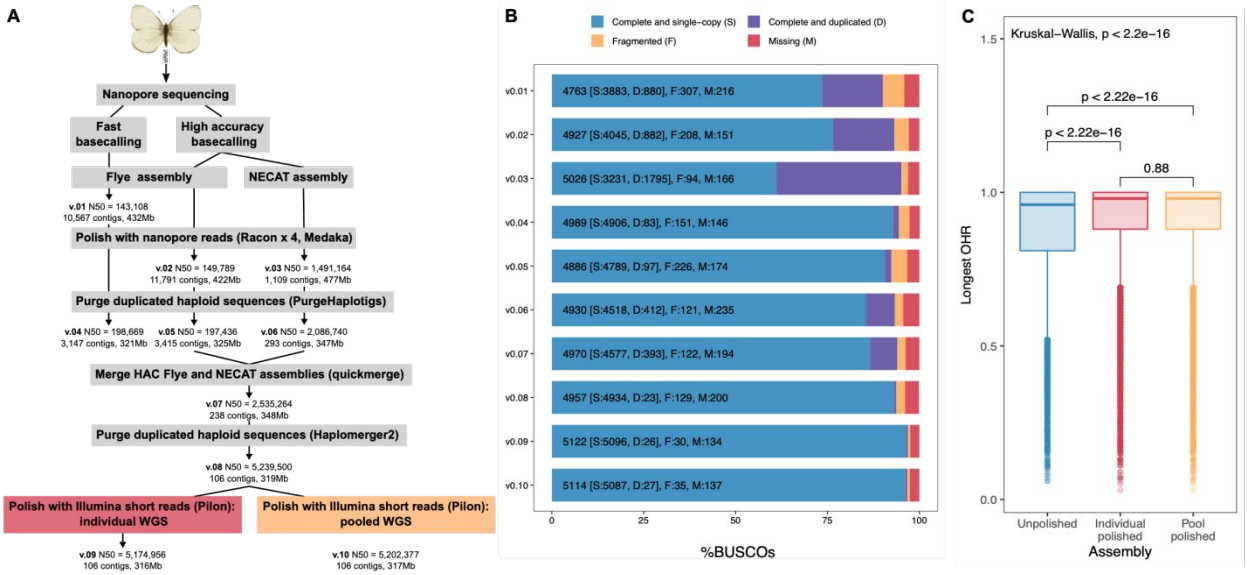
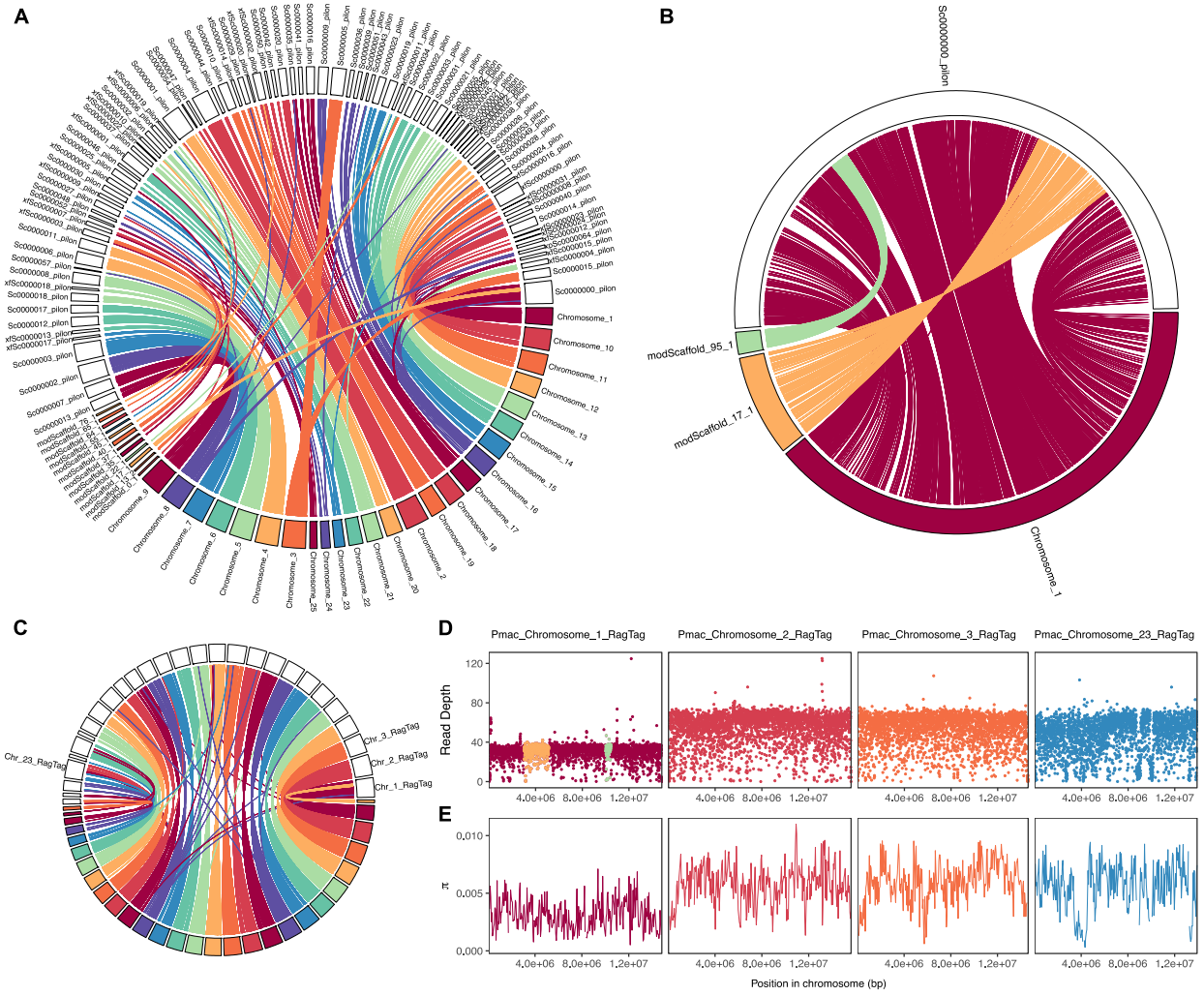


Figure 2.



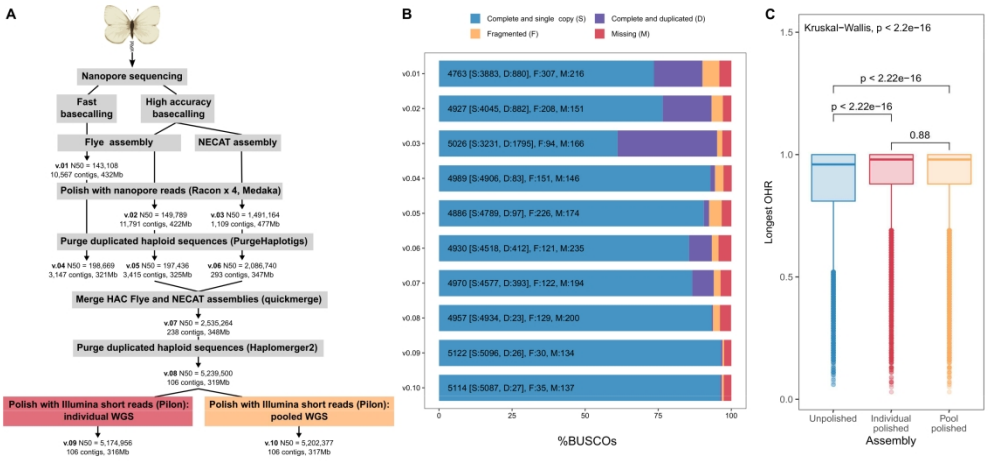


Figure 1. Genome assembly pipeline and metrics for 10 genomes show dramatic improvement during refinement steps. (A) Progressive increase in N50 and decrease in total contigs during polishing and merging of the nanopore assembly. Steps to refine the assembly included polishing with Illumina whole genome short reads (WGS) from a single individual and a pool of 18 individuals. (B) Assessment of the content and quality of 5286 lepidopteran single copy orthologs shows complete, duplicated, fragmented or missing BUSCOs across the 10 assemblies (v.01-v.10). (C) Assessment of changes in genome quality using whole genome annotations show similar effects of polishing using individual or PoolSeq Illumina data. (C) Polishing with Illumina short reads improved the ortholog hit ratio (OHR, values closer to 1 indicate a higher quality annotation).

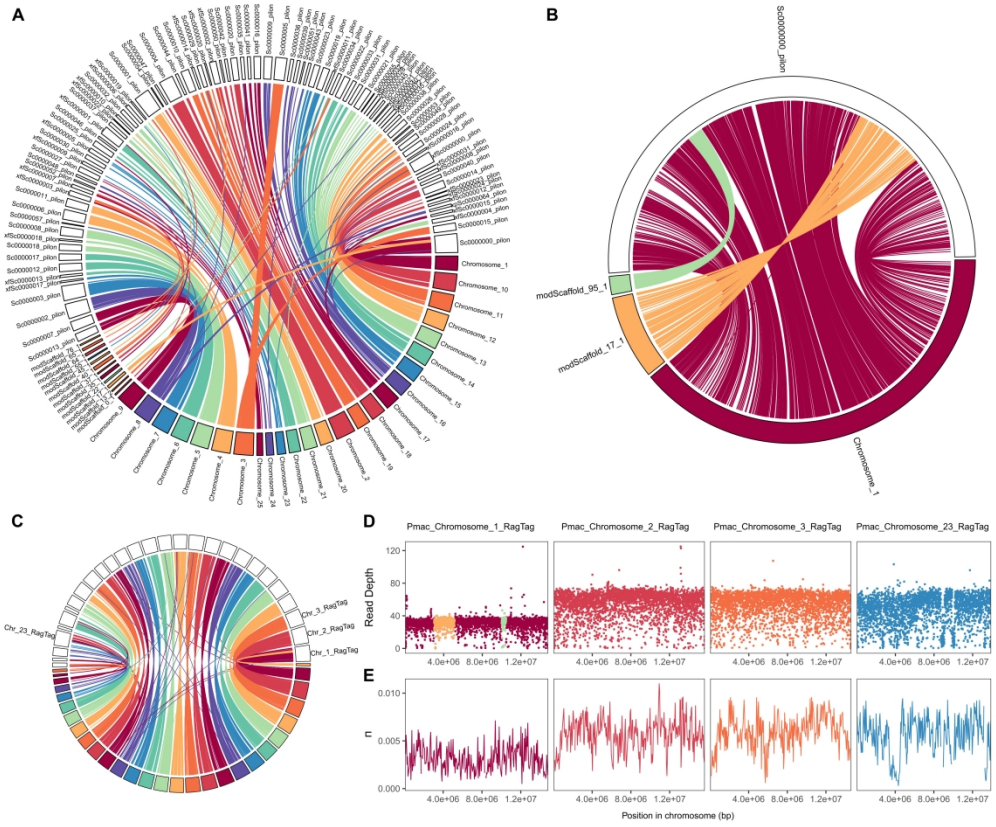


Figure 2. Chromosome-level assessment of synteny, read depth and genetic variation in P. macdunnoughii. (A) A circle plot showing each contig of the P. macdunnoughii v0.10 (non-colored scaffolds) and P. napi (colored scaffolds, showing scaffolds > 1Mb representing 90.9% of the 318bp assembly) assemblies, with lines between them showing aligned genomic regions of greater than 5000 bp and 90% identity (for example, P. napi Chromosome2 is covered by four P. macdunnoughii scaffolds, Sc0000044, Sc0000004, Sc0000047, and Sc0000054). (B) Detailed assessment of the alignments between the Z chromosome of P. macdunnoughii (Sc0000000) and aligned, unplaced P. napi scaffolds modScaffold_17_1 and modScaffold_95_1, supporting their inclusion on the Z chromosome between 3003162 - 5135852bp and 10002530 – 10466150bp, respectively. (C) The final pseudo-chromosomal P. macdunnoughii assembly (v0.10_RagTag) aligned with P. napi. (D) Consistently lower read depths of PoolSeq reads mapped to the pseudo-chromosomal assembly support conclusions about the Z chromosome (Pmac_chromosome_1_RagTag; colors as in B). (E) Nucleotide diversity (□), varied across the genome, as seen in representative autosomes 2, 3, and 23.