G3, 2022, jkac193
https://doi.org/10.1093/g3journal/jkac193
Advance Access Publication Date: 5 August 2022
Genome Report

The genome sequence of the scarce swallowtail, *Iphiclides podalirius*

Alexander Mackintosh,[1,*] Dominik R. Laetsch ,[1] Tobias Baril ,[2] Sam Ebdon ,[1] Paul Jay ,[3] Roger Vila ,[4] Alex Hayward ,[2,‡] and Konrad Lohse [1,‡]

[1]Institute of Ecology and Evolution, University of Edinburgh, Edinburgh EH9 3FL, UK,
[2]Centre for Ecology and Conservation, University of Exeter, Penryn Campus, Cornwall TR10 9FE, UK,
[3]Ecologie Systématique Evolution, Bâtiment 360, CNRS, AgroParisTech, Université Paris-Saclay, 91400 Orsay, France,
[4]Institut de Biologia Evolutiva (CSIC—Universitat Pompeu Fabra), Barcelona 08003, Spain

*Corresponding author: Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3FL, UK. Email: a.j.f.mackintosh@sms.ed.ac.uk
‡These authors contributed equally to this work.

Abstract

The scarce swallowtail, *Iphiclides podalirius* (Linnaeus, 1758), is a species of butterfly in the family Papilionidae. Here, we present a chromosome-level genome assembly for *Iphiclides podalirius* as well as gene and transposable element annotations. We investigate how the density of genomic features differs between the 30 *Iphiclides podalirius* chromosomes. We find that shorter chromosomes have higher heterozygosity at four-fold-degenerate sites and a greater density of transposable elements. While the first result is an expected consequence of differences in recombination rate, the second suggests a counter-intuitive relationship between recombination and transposable element evolution. This high-quality genome assembly, the first for any species in the tribe Leptocircini, will be a valuable resource for population genomics in the genus *Iphiclides* and comparative genomics more generally.

Keywords: *Iphiclides podalirius*; genome assembly; genome annotation; chromosome length; heterozygosity

Introduction

The scarce swallowtail, *Iphiclides podalirius* (Linnaeus, 1758), is a widespread butterfly species in the family Papilionidae. The species is common in open habitats in the Palearctic, ranging from France to Western China, but is absent from Northern areas (e.g. Scandinavia and the British Isles) and some Mediterranean Islands (e.g. Sardinia, where only occasional records exist). *I. podalirius* is generally bivoltine, the larvae feed mainly on different species of *Prunus*, principally *P. spinosa*, and overwinter in the pupal stage.

The genus *Iphiclides* belongs to the tribe Leptocircini (kite swallowtails), which diverged from other taxa in the subfamily Papilioninae about 55 million years ago (Allio, Scornavacca *et al.* 2020), and only includes two other species: *I. podalirinus*, which is restricted to Central Asia and *Iphiclides feisthamelii*, the sister taxon of *I. podalirius*, which is found in Northern Africa and the Iberian Peninsula. A controversy about the taxonomic status of *I. feisthamelii*, which has been regarded as a subspecies of *I. podalirius* (Godart and Duponchel 1832; Tolman and Lewington 2009; Wiemers and Gottsberger 2010), has only recently been resolved; although the two species have no known differences in ecology or life history and share mitochondrial haplotypes (Dincă *et al.* 2015), Gaunet *et al.* (2019) show that they differ consistently in wing patterns (including UV reflectance of males), genital morphology and nuclear DNA and are separated by a narrow hybrid zone in Southern France (Descimon and Mallet 2009).

Although Allio, Scornavacca *et al.* (2020) have previously generated Illumina shotgun data for *I. podalirius*, the assemblies based on these data (Allio, Scornavacca *et al.* 2020; Ellis *et al.* 2021) are highly fragmented (with an N50 of 0.6 and 1.7 kb, respectively). More generally, while chromosome-level assemblies are available for several swallowtail butterflies in the genus *Papilio* (Lu *et al.* 2019), similarly contiguous genome assemblies are lacking for other tribes within Papilioninae.

Here, we present a chromosome-level genome assembly for *I. podalirius*, as well as gene and transposable element (TE) annotations. We use this assembly to investigate how heterozygosity in the reference individual varies both between genomic partitions and chromosomes.

Materials and methods

Sampling

Two male individuals (MO_IP_500 and MO_IP_504) were collected with a hand net at Le Moulin de Bertrand, Saint-Martin-de-Londres, Montpellier, France, and flash frozen in liquid nitrogen. High-molecular-weight (HMW) DNA was extracted from the thorax of one of these individuals (MO_IP_504) using a salting out extraction protocol as described in Mackintosh *et al.* (2022).

Received: March 28, 2022. **Accepted:** July 20, 2022

© The Author(s) 2022. Published by Oxford University Press on behalf of Genetics Society of America.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/g3journal/advance-article/doi/10.1093/g3journal/jkac193/6656352 by guest on 11 August 2022

Sequencing

A SMRTbell sequencing library was generated from the HMW extraction of MO_IP_504 by the Exeter Sequencing Service. This was sequenced on 3 SMRT cells on a Sequel I instrument to generate 24.0 Gb of Pacbio continuous long-read (CLR) data.

A chromium 10× library was prepared from the HMW extraction at the Cancer Research UK Cambridge Institute, UK. This library was sequenced by Edinburgh Genomics (EG) on a single HiSeqX lane, generating 25.3 Gb of paired-end reads after processing with Long Ranger v2.2.2 (Marks *et al.* 2019). However, the weighted mean molecule size of these data (12.86 kb) limited its use for scaffolding of the Pacbio assembly and the reads were therefore only used for polishing. Hereafter, these data are simply referred to as Illumina WGS.

In addition, tissue from MO_IP_500 was used for chromatin conformation capture (HiC) sequencing. The HiC reaction was performed using an Arima HiC kit, following the manufacturer's instructions for flash-frozen animal tissue. An NEBNext Ultra II library, prepared by EG, was sequenced on an Illumina MiSeq, generating 4.7 Gb of paired-end reads.

Paired-end RNA-seq data were generated for individual MO_IP_504. To obtain tissue for RNA extraction, the adult individual was divided bilaterally (including all parts of the body: head, thorax, and abdomen). For further details on RNA extraction, see Ebdon *et al.* (2021).

Genome assembly

Pacbio reads ≥2 kb (40.4× coverage) were assembled using wtdbg2.5 (Ruan and Li 2020) with the options: -L 2000 -x sq -g 400 m. Errors in the consensus sequence were corrected by three rounds of Pacbio CLR polishing and two rounds of Illumina WGS polishing using Racon v1.4.10 (Vaser *et al.* 2017) while retaining any unpolished sequences. Contigs belonging to nontarget organisms were identified using blobtools v1.1.1 (Laetsch and Blaxter 2017) and subsequently removed. Finally, duplicated regions and contigs with extremely low (<5×) or high (>200×) coverage were identified and removed with purge_dups v1.2.5 (Guan *et al.* 2020). Mapping of long reads and short reads for the above steps was performed with minimap2 v2.17 and bwa-mem v0.7.17, respectively (Li 2013, 2018).

The HiC and RNA-seq reads were adapter and quality trimmed with fastp v0.2.1 (Chen *et al.* 2018).

The trimmed HiC reads were aligned to the contig-level assembly with Juicer v1.6 (Durand *et al.* 2016). Scaffolding was performed with 3d-dna v180922 using default parameters (Dudchenko *et al.* 2017). The scaffolded assembly and HiC map generated by 3d-dna was visualized and manually curated in Juicebox v1.11.08 (Robinson *et al.* 2018). A total of 7 contigs had their orientation changed through manual curation.

Gene completeness was evaluated using BUSCO v5.0.0 with the lepidoptera_odb10 dataset ($n = 5{,}286$) (Manni *et al.* 2021). Kmer QV was calculated using Merqury v1.1 (Rhie *et al.* 2020). Genome size and heterozygosity were estimated from the Merqury kmer spectrum using Genomescope 2.0 (Ranallo-Benavidez *et al.* 2020).

The mitochondrial genome was assembled and annotated using the Mitofinder pipeline v1.4 (Allio, *et al.* 2020). Illumina WGS reads were assembled with metaSPAdes v3.14.1 (Nurk *et al.* 2017) and tRNAs were annotated with MiTFi (Jühling *et al.* 2012).

Genome annotation

TEs were annotated using the Earl Grey TE annotation pipeline (Jurka *et al.* 2005; Xu and Wang 2007; Rubino and Creevey 2014; Smit *et al.* 2015; Hubley *et al.* 2016; Platt *et al.* 2016; Ou and Jiang 2019; Wong and Simakov 2019; Flynn *et al.* 2020; Baril *et al.* 2022) as in Mackintosh *et al.* (2022).

Following gene annotation, 5′ and 3′ gene flanks were defined as those that were 20 kb upstream and downstream of genes. We define regions as intergenic space if they are neither genic (start/stop codons, exons, and introns) nor gene flanks. Bedtools intersect v2.27.1 (Quinlan and Hall 2010) was used to determine overlap (-wao) between TEs and genomic features. Following this, quantification and plotting were performed in R, using the tidyverse package (Wickham *et al.* 2019; RStudio Team 2020; R Core Team 2021).

The trimmed RNA-seq reads were mapped to the assembly with HISAT2 v2.1.0 (Kim *et al.* 2019). The softmasked assembly and RNA-seq alignments were used for gene prediction with braker2.1.5 (Stanke *et al.* 2006, 2008; Li *et al.* 2009; Barnett *et al.* 2011; Lomsadze *et al.* 2014; Buchfink *et al.* 2015; Hoff *et al.* 2016, 2019). Gene annotation statistics were calculated with GenomeTools v1.6.1 (Gremme *et al.* 2013).

Functional annotation was carried out using InterProScan v5.50-84.0 (Jones *et al.* 2014) and the Pfam-33.1 database (Mistry *et al.* 2021).

Estimating heterozygosity

Heterozygosity was estimated within different partitions of the genome by first mapping the Illumina WGS reads to the assembly with bwa-mem, marking duplicated alignments with sambamba 0.8.1 (Tarasov *et al.* 2015), and calling variants with freebayes v1.3.2-dirty (Garrison and Marth 2012). Variants were normalized with bcftools v1.8 (Danecek *et al.* 2021), decomposed with vcfallelicprimitives (Garrison *et al.* 2021), filtered (QUAL>= 1), and subset to biallelic SNPs with bcftools. Callable sites were delimited with mosdepth v0.3.2 (Pedersen and Quinlan 2018), by applying a coverage threshold between 8 and 95 (inclusive). Callable sites were further restricted by removing all sites where there were indels or SNPs with QUAL < 1. Four-fold-degenerate (4D) and zero-fold-degenerate (0D) sites were identified using partition_cds.py v0.2 (see *Data availability*), based on the CDS BED file (where the 4th column contains transcript ID), the genome FASTA, and the VCF file. Bedtools v2.30.0 was used to intersect callable regions of the genome with intergenic, intronic, exonic, 4D, and 0D sites. Heterozygosity—the number of heterozygous bi-allelic SNPs divided by the number of callable sites—and callable sites of each genomic partition are listed in Table 1.

Table 1. Estimates of heterozygosity in different partitions of the genome.

Partition	Callable sites (Mb)	Heterozygosity
All	420.3	0.00598
Exonic	19.3	0.00295
Intronic	126.1	0.00615
Intergenic	275.1	0.00609
0D	12.4	0.00157
4D	3.1	0.00680

As a comparison, the kmer-based heterozygosity estimated from all Illumina reads is 0.00702.

Downloaded from https://academic.oup.com/g3journal/advance-article/doi/10.1093/g3journal/jkac193/6656352 by guest on 11 August 2022

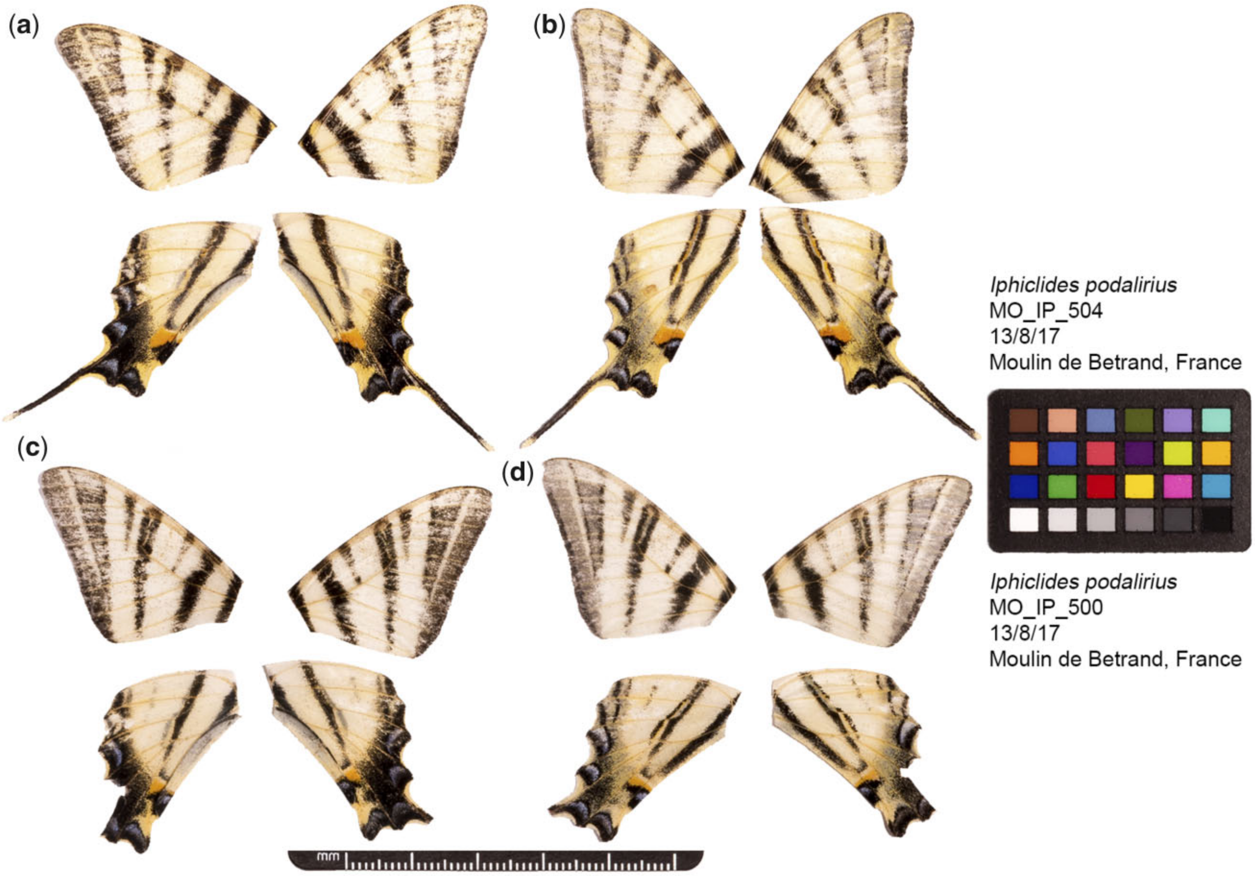

Fig. 1. Fore and hind wings of the two *I. podalirius* individuals used to generate the genome sequence. a) Dorsal and b) ventral surface view of wings of specimen MO_IP_504, used to generate Pacbio long-read, Illumina WGS short-read, and Illumina RNA-seq short-read data. c) Dorsal and d) ventral surface view of wings of specimen MO_IP_500, used to generate HiC data.

Results and discussion

Genome assembly

We sequenced the genome of a male individual (Fig. 1. a and b) by generating both Pacbio continuous long reads (55.7× coverage) and Illumina paired-end short reads (58.8× coverage). From these data, we assembled a haploid representation of the genome consisting of 265 contigs with a total span of 430.6 Mb. The contig assembly is slightly larger than a genome size estimate based on the flow cytometry of male individuals (386.6 Mb, Mackintosh et al. 2019) and a previous assembly for this species based only on Illumina data (390.9 Mb, Allio, Scornavacca et al. 2020). However, it is smaller than an estimate of the genome size from kmers in the Illumina short reads (471 Mb, Supplementary Fig. 1), suggesting that the flow cytometry may have produced an underestimate and that the Illumina-based assembly contains collapsed repetitive regions. We scaffolded the contigs with Arima HiC data (11.0x coverage) generated from a different male individual collected at the same locality (Fig. 1, c and d). Scaffolding generated 30 chromosome-scale sequences, which together account for 99.5% of the total assembly length and range from 6.8 to 21.1 Mb in size (Supplementary Fig. 2). The assembly has a contig and scaffold N50 of 5.2 and 15.1 Mb, respectively.

The BUSCO analysis shows that the assembly contains the majority of expected single-copy orthologs with little duplication (S: 96.5%, D: 0.2%, F: 0.4%, M: 2.9%). The Phred quality of the consensus sequence, estimated using solid kmers in the short-read data, is 35.8.

We assembled a circularized mitochondrial genome of 15,396 bases, containing 13 protein-coding genes, 22 tRNA genes, and 2 rRNA genes. The sequence can be aligned colinearly with the mitochondrial genome of *Graphium doson* (Kong et al. 2019), another species in the tribe Leptocircini, demonstrating that these mitochondrial genomes have not undergone any rearrangements.

Genome annotation

TEs compromise 32.81% of the genome assembly (Supplementary Table 1 and Fig. 2a). The assembly contains representation from all major TE types (Supplementary Table 1): the most abundant TEs are long-interspersed nuclear elements (LINEs), which constitute 11.01% of the assembly and 33.56% of total TE sequence. Recent activity is high in LINEs and there is also evidence for a very recent increase in LTR element abundance (Fig. 2b).

Considering all TE classifications, most TEs (70.14%) are found in intergenic regions. As expected, TEs are largely absent from exons with only 0.07% of exonic sequence consisting of TEs, likely due to the deleterious effects of TE insertions in exons (Sultana et al. 2017; Bourque et al. 2018). In contrast, a substantial fraction of intronic sequence (31.47%) is comprised of TEs (Fig. 2c). The most abundant TEs in the assembly, LINEs, comprise 14.25% of intergenic space, 12.14% of gene flanks, 8.98% of intronic regions, and 0.69% of exonic regions (Fig. 2c).

We annotated the assembly with 17,826 protein-coding genes, coding for 20,222 transcripts (1.13 transcripts per gene). At least 1 Pfam domain was identified along proteins of 9,363 genes (52.5%)

Downloaded from https://academic.oup.com/g3journal/advance-article/doi/10.1093/g3journal/jkac193/6656352 by guest on 11 August 2022

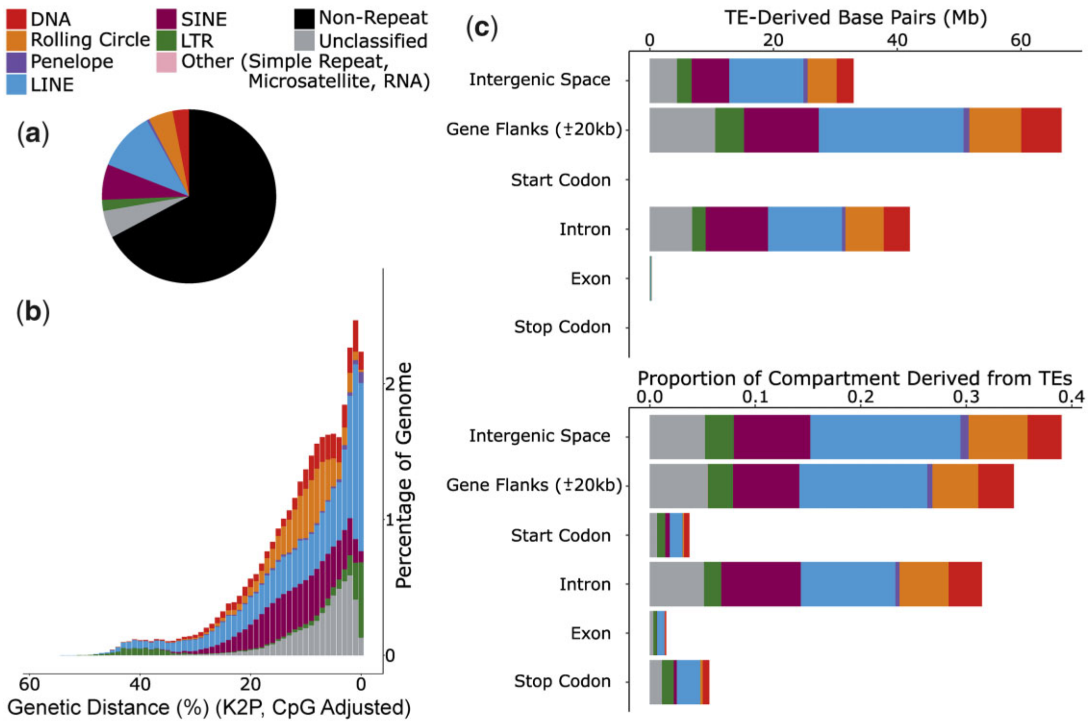

Fig. 2. TEs within the genome assembly of *I. podalirius*. a) The proportion of the assembly comprised of the main TE classifications. b) A repeat landscape plot illustrating the proportion of repeats in the genome at different genetic distances (%) to their respective RepeatModeler consensus sequence. Genetic distance is calculated under a Kimura 2 parameter model with correction for CpG site hypermutability. Lower genetic distances suggest more recent activity. c) The abundance of TEs in different partitions of the genome, shown in bases and as a proportion of the partition.

and start codons were found in genes coding for 20,163 proteins (99.71%). The BUSCO score of the protein sequences (S: 66.9%, D: 12.8%, F: 5.5%, M: 14.8%) is lower than the score of the genome sequence (see above) with an expected increase in duplicated BUSCOs.

The median length of genes is 4.0 kb, with the majority (51.7%) consisting of 4 exons or fewer. The number of gene predictions is higher than in genome annotations for species in the genus Papilio, such as *Papilio dardanus* (12,795, Timmermans *et al.* 2020) and Papilio *bianor* (15,375, Lu *et al.* 2019), but far lower than in the annotation of the *Parnassius apollo* genome (28,344, Podsiadlowski *et al.* 2021).

The density of annotated genomic features differs across chromosomes (Fig. 3 and Supplementary Table 2). For example, the proportion of sequence made up of TEs ranges from 24.4% on chromosome 7 to 39.4% on chromosome 29. Similarly, exon density also ranges approximately two-fold across chromosomes, from 3.3% on chromosome 25 to 6.4% on chromosome 30. The density of TEs is negatively correlated with chromosome length (Spearman's rank correlation, $\rho_{(28)} = -0.520$, $P = 0.004$), whereas exon density and chromosome length are uncorrelated (Fig. 3).

Genome-wide heterozygosity

Across the genome assembly, we identified 2,514,242 heterozygous SNPs in the reference individual MO_IP_504, which is equivalent to a per-base heterozygosity (across all sites) of 0.00598 (Table 1). As expected, given selective constraint, heterozygosity in exons is less than half of that in introns and intergenic regions (Table 1). Likewise, within coding sequence heterozygosity is highest for 4D sites and lowest for 0D sites (Table 1), which is

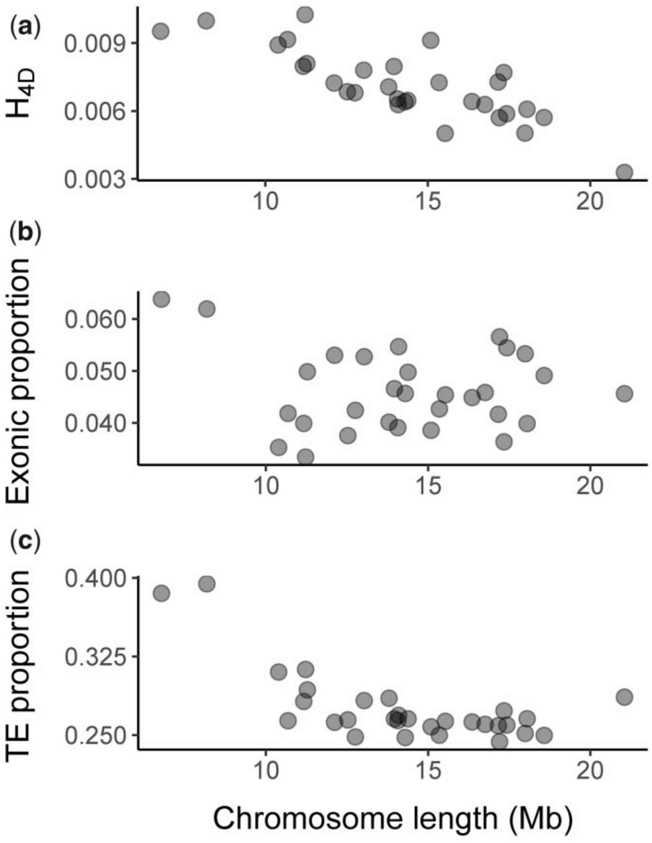

Fig. 3. The relationship between chromosome length and a) heterozygosity at 4D sites, b) exon density, and c) TE density.

Downloaded from https://academic.oup.com/g3journal/advance-article/doi/10.1093/g3journal/jkac193/6656352 by guest on 11 August 2022

expected given that the latter are under strong evolutionary constraint (Sawyer et al. 1987).

We note that our estimate of 4D site heterozygosity is comparable, but slightly higher, than π_{4D} estimates previously reported for *I. podalirius* based on transcriptome assemblies and data from two individuals (0.0052, 0.0057) (Mackintosh et al. 2019; Ebdon et al. 2021). This difference most likely reflects the fact that transcriptome assemblies are biased toward highly expressed transcripts, which experience greater indirect effects of purifying selection (Marek and Tomala 2018).

Heterozygosity at 4D sites (H_{4D}) varies by chromosome (Fig. 3 and Supplementary Table 2): it is lowest on chromosome 1 (0.00328, the putative Z chromosome) and highest on chromosome 25 (0.01026, an autosome). We find a significant negative correlation between chromosome length and H_{4D} (Spearman's rank correlation, $\rho_{(28)} = -0.787$, $p = 2 \times 10^{-6}$) (Fig. 3a).

Conclusions

We describe a chromosome-level genome assembly for the scarce swallowtail butterfly *I. podalirius*, with gene and repeat annotations that are similar to previously published *Papilio* butterfly genomes. By contrast, the number of gene predictions and TEs in the *P. apollo* genome assembly is far greater (Podsiadlowski et al. 2021), suggesting gene and repeat expansions in the subfamily Parnassiinae.

When comparing heterozygosity in the reference individual both between chromosomes and between genomic partitions, we recover two well-documented patterns of genome-wide sequence variation, which result from the direct and indirect effects of selection, respectively: (1) stark differences in genetic diversity between genomic partitions reflecting differences in selective constraint and (2) a negative correlation between chromosome length and heterozygosity at putatively neutral 4D sites. The latter has been described for several species (including butterflies Martin et al. 2016) and is an expected consequence of the fact that the indirect effect of selection on nearby neutral sites depends on the rate of recombination (which tends to be greater for short chromosomes, Haenel et al. 2018).

We also find a negative relationship between chromosome length and TE density. This is surprising given that increased recombination rates on shorter chromosomes are expected to result in more efficient selection against TE insertions (Langley et al. 1988). Despite this, the observation that smaller chromosomes contain a greater density of repetitive elements has also been reported in *Heliconius* and *Melitaea* butterflies (Ahola et al. 2014; Cicconardi et al. 2021), suggesting that this may be a general feature of Lepidopteran genomes.

The *I. podalirius* genome will be valuable resource—not only for genomic analyses that investigate the forces driving genome evolution in the long term—but will also allow for detailed studies of the population-level processes that lead to the accumulation of barriers between species still experiencing gene flow.

Data availability

The genome assembly, gene annotation, and raw sequence data can be found at the European Nucleotide Archive under project accession PRJEB51340. The script used for calculating the site degeneracy (partition_cds.py) and the script used for visualizing HiC contacts (HiC_view.py) can be found at the following github repository: https://github.com/A-J-F-Mackintosh/Mackintosh_et_al_2022_Ipod. The mitochondrial genome sequence and the TE annotation can be found at the same repository.

Supplemental material is available at *G3* online.

Acknowledgments

We would like to thank Marian Thompson for preparing the Pacbio sequencing libraries and Andrés de la Filia and Katy MacDonald for help in the molecular lab. We thank Alvaro Martinez Barrio (10× Genomics) and Paul Coupland (Cancer Research UK Cambridge Institute) for help with generating the 10× data and Charlotte Wright for helpful comments on an earlier draft of this manuscript.

Funding

AM is supported by an E4 PhD studentship from the Natural Environment Research Council (NE/S007407/1). KL is supported by a fellowship from the Natural Environment Research Council (NERC, NE/L011522/1). RV is supported by Grant PID2019-107078GB-I00 funded by Ministerio de Ciencia e Innovación and Agencia Estatal de Investigación (MCIN/AEI/10.13039/501100011033). This work was supported by a European Research Council starting grant (ModelGenomLand 757648) to KL and a David Phillips Fellowship (BB/N020146/1) by the Biotechnology and Biological Sciences Research Council (BBSRC) to AH.

Conflicts of interest

None declared.

Literature cited

Ahola V, Lehtonen R, Somervuo P, Salmela L, Koskinen P, Rastas P, Välimäki N, Paulin L, Kvist J, Wahlberg N, et al. The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera. Nat Commun. 2014;5:4737.

Allio R, Schomaker-Bastos A, Romiguier J, Prosdocimi F, Nabholz B, Delsuc F. Mitofinder: efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. Mol Ecol Resour. 2020;20(4):892–905.

Allio R, Scornavacca C, Nabholz B, Clamens AL, Sperling FA, Condamine FL. Whole genome shotgun phylogenomics resolves the pattern and timing of swallowtail butterfly evolution. Syst Biol. 2020;69(1):38–60.

Baril TJ, Imrie RM, Hayward A. Earl Grey: a fully automated user-friendly transposable element annotation and analysis pipeline. bioRxiv, 2022. https://doi.org/10.1101/2022.06.30.498289

Barnett DW, Garrison EK, Quinlan AR, Strömberg MP, Marth GT. BamTools: a C++ API and toolkit for analyzing and managing BAM files. Bioinformatics. 2011;27(12):1691–1692.

Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, Imbeault M, Izsvák Z, Levin HL, Macfarlan TS, et al. Ten things you should know about transposable elements. Genome Biol. 2018;19(1):199.

Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12(1):59–60.

Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884–i890.

Cicconardi F, Lewis JJ, Martin SH, Reed RD, Danko CG, Montgomery SH. Chromosome fusion affects genetic diversity and

Downloaded from https://academic.oup.com/g3journal/advance-article/doi/10.1093/g3journal/jkac193/6656352 by guest on 11 August 2022

evolutionary turnover of functional loci but consistently depends on chromosome size. Mol Biol Evol. 2021;38(10):4449–4462.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, *et al.* Twelve years of SAMtools and BCFtools. GigaScience. 2021;10(2): giab008.

Descimon H, Mallet J. Bad species. In: Settele J, Shreeve TG, Konvicka M, Dyck VH (editors). Ecology of Butterflies in Europe. Cambridge: Cambridge University Press; 2009. p. 219–249.

Dincă V, Montagud S, Talavera G, Juan HR, Munguira ML, García-Barros E, Hebert PDN, Vila R. DNA barcode reference library for Iberian butterflies enables a continental-scale preview of potential cryptic diversity. Sci Rep. 2015;5:12395–12322.

Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP, *et al.* De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science. 2017;356(6333):92–95.

Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, Aiden EL. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst. 2016;3(1):95–98.

Ebdon S, Laetsch DR, Dapporto L, Hayward A, Ritchie MG, Dincă V, Vila R, Lohse K. The Pleistocene species pump past its prime: evidence from European butterfly sister species. Mol Ecol. 2021; 30(14):3575–3589.

Ellis EA, Storer CG, Kawahara AY. De novo genome assemblies of butterflies. GigaScience. 2021;10(6):giab041.

Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. Repeatmodeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A. 2020;117(17): 9451–9457.

Garrison E, Kronenberg ZN, Dawson ET, Pedersen BS, Prins P. Vcflib and tools for processing the VCF variant call format. bioRxiv, 2021. https://doi.org/10.1101/2021.05.21.445151

Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv e-prints, 2012. https://doi.org/10.48550/arXiv.1207.3907

Gaunet A, Dincă V, Dapporto L, Montagud S, Vodă R, Schär S, Badiane A, Font E, Vila R. Two consecutive wolbachia-mediated mitochondrial introgressions obscure taxonomy in palearctic swallowtail butterflies (Lepidoptera, Papilionidae). Zool Scr. 2019; 48(4):507–519.

Godart JB, Duponchel PAJ. *Histoire naturelle des lépidoptères ou papillons de France, par M. J.-B. Godart. Supplément*, Vol. 1. Paris: Méquignon-Marvis, Libraire-Éditeur; 1832. https://www.biodiversitylibrary.org/bibliography/9257.

Gremme G, Steinbiss S, Kurtz S. Genometools: a comprehensive software library for efficient processing of structured genome annotations. IEEE ACM Trans Comput Biol Bioinform. 2013;10(3): 645–656.

Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics. 2020;36(9):2896–2898.

Haenel Q, Laurentino TG, Roesti M, Berner D. Meta-analysis of chromosome-scale crossover rate variation in eukaryotes and its significance to evolutionary genomics. Mol Ecol. 2018;27(11): 2477–2497.

Hoff K, Lange S, Lomsadze A, Borodovsky M, Stanke M. BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics. 2016;32(5): 767–769.

Hoff K, Lomsadze A, Borodovsky M, Stanke M. Whole-genome annotation with BRAKER. In: M Kollmar, editor. Gene Prediction: Methods and Protocols. New York: Springer; 2019. p. 65–95.

Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, Smit AF, Wheeler TJ. The Dfam database of repetitive DNA families. Nucleic Acids Res. 2016;44(D1):D81–D89.

Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, *et al.* InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30(9): 1236–1240.

Jühling F, Pütz J, Bernt M, Donath A, Middendorf M, Florentz C, Stadler PF. Improved systematic tRNA gene annotation allows new insights into the evolution of mitochondrial tRNA structures and into the mechanisms of mitochondrial genome rearrangements. Nucleic Acids Res. 2012;40(7):2833–2845.

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 2005;110(1–4): 462–467.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37(8):907–915.

Kong W, Hou Y, Zhang G, Yang T, Xiao R. Complete mitochondrial genome of *Graphium doson* (Papilioninae: Leptocircini). Mitochondrial DNA B. 2019;4(1):698–699.

Laetsch D, Blaxter M. Blobtools: interrogation of genome assemblies [version 1; peer review: 2 approved with reservations]. F1000Research. 2017;6:1287.

Langley CH, Montgomery E, Hudson R, Kaplan N, Charlesworth B. On the role of unequal exchange in the containment of transposable element copy number. Genet Res. 1988;52(3):223–235.

Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997v2 [q-bio.GN]. 2013.

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

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–2079.

Lomsadze A, Burns PD, Borodovsky M. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. Nucleic Acids Res. 2014;42(15):e119.

Lu S, Yang J, Dai X, Xie F, He J, Dong Z, Mao J, Liu G, Chang Z, Zhao R, *et al.* Chromosomal-level reference genome of Chinese peacock butterfly (*Papilio bianor*) based on third-generation DNA sequencing and Hi-C analysis. GigaScience. 2019;8(11):giz128.

Mackintosh A, Laetsch DR, Baril T, Foster RG, Dincă V, Vila R, Hayward A, Lohse K. The genome sequence of the lesser marbled fritillary, *Brenthis ino*, and evidence for a segregating neo-Z chromosome. G3 (Bethesda). 2022;12:jkac069.

Mackintosh A, Laetsch DR, Hayward A, Charlesworth B, Waterfall M, Vila R, Lohse K. The determinants of genetic diversity in butterflies. Nat Commun. 2019;10(1):3466.

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 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.

Marek A, Tomala K. The contribution of purifying selection, linkage, and mutation bias to the negative correlation between gene expression and polymorphism density in yeast populations. Genome Biol Evol. 2018;10(11):2986–2996.

Marks P, Garcia S, Barrio AM, Belhocine K, Bernate J, Bharadwaj R, Bjornson K, Catalanotti C, Delaney J, Fehr A, *et al.* Resolving the

Downloaded from https://academic.oup.com/g3journal/advance-article/doi/10.1093/g3journal/jkac193/6656352 by guest on 11 August 2022

full spectrum of human genome variation using linked-reads. Genome Res. 2019;29(4):635–645.

Martin SH, Möst M, Palmer WJ, Salazar C, McMillan WO, Jiggins FM, Jiggins CD. Natural selection and genetic diversity in the butterfly *Heliconius melpomene*. Genetics. 2016;203(1):525–541.

Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, *et al.* Pfam: the protein families database in 2021. Nucleic Acids Res. 2021;49(D1):D412–D419.

Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. Genome Res. 2017;27(5):824–834.

Ou S, Jiang N. LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. Mob DNA. 2019;10:48.

Pedersen BS, Quinlan AR. Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics. 2018;34(5):867–868.

Platt RNI, Blanco-Berdugo L, Ray DA. Accurate transposable element annotation is vital when analyzing new genome assemblies. Genome Biol Evol. 2016;8(2):403–410.

Podsiadlowski L, Tunström K, Espeland M, Wheat CW. The genome assembly and annotation of the apollo butterfly *Parnassius apollo*, a flagship species for conservation biology. Genome Biol Evol. 2021;13(8):evab122.

Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–842.

R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. 2022.

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.

Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. 2020;21(1):245.

Robinson JT, Turner D, Durand NC, Thorvaldsdóttir H, Mesirov JP, Aiden EL. Juicebox.js provides a cloud-based visualization system for Hi-C data. Cell Syst. 2018;6(2):256–258.e1.

RStudio Team. Rstudio: Integrated Development Environment for R. Rstudio, PBC, Boston, MA. 2020.

Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. Nat Methods. 2020;17(2):155–158.

Rubino F, Creevey C. MGkit: metagenomic framework for the study of microbial communities. Figshare. Poste, 2014.

Sawyer SA, Dykhuizen DE, Hartl DL. Confidence interval for the number of selectively neutral amino acid polymorphisms. Proc Natl Acad Sci U S A. 1987;84(17):6225–6228.

Smit A, Hubley R, Green P. Repeatmasker open-4.0; 2015. http://www.repeatmasker.org.

Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics. 2008;24(5):637–644.

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

Sultana T, Zamborlini A, Cristofari G, Lesage P. Integration site selection by retroviruses and transposable elements in eukaryotes. Nat Rev Genet. 2017;18(5):292–308.

Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. Bioinformatics. 2015;31(12):2032–2034.

Timmermans MJTN, Srivathsan A, Collins S, Meier R, Vogler AP. Mimicry diversification in *Papilio dardanus* via a genomic inversion in the regulatory region of *engrailed–invected*. Proceedings of the Royal Society. Proc Biol Sci. 2020;287(1926):20200443.

Tolman T, Lewington R. Collins Butterfly Guide. London: Harper Collins; 2009.

Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 2017;27(5):737–746.

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, *et al.* Welcome to the tidyverse. JOSS. 2019;4(43):1686.

Wiemers M, Gottsberger B. Discordant patterns of mitochondrial and nuclear differentiation in the Scarce Swallowtail *Iphiclides podalirius feisthamelii* (Duponchel, 1832) (Lepidoptera: Papilionidae). Entomol Z. 2010;120:111–115.

Wong WY, Simakov O. RepeatCraft: a meta-pipeline for repetitive element de-fragmentation and annotation. Bioinformatics. 2019;35(6):1051–1052.

Xu Z, Wang H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. 2007;35(Web Server issue):W265–W268.

Communicating editor: M. Joron

Downloaded from https://academic.oup.com/g3journal/advance-article/doi/10.1093/g3journal/jkac193/6656352 by guest on 11 August 2022