Improved chromosome-level genome assembly of the Glanville fritillary butterfly (Melitaea cinxia) integrating Pacific Biosciences long reads and a high-density linkage map

Smolander, Olli-Pekka

2022-01-12

Improved chromosome-level genome assembly of the Glanville fritillary butterfly (*Melitaea cinxia*) integrating Pacific Biosciences long reads and a high-density linkage map

Olli-Pekka Smolander [1,2,†], Daniel Blande [3,†], Virpi Ahola [3,4], Pasi Rastas [1], Jaakko Tanskanen[5], Juhana I. Kammonen [1], Vicencio Oostra [3,6], Lorenzo Pellegrini [1,7], Suvi Ikonen[3], Tad Dallas [8], Michelle F. DiLeo [3], Anne Duplouy [3,9], Ilhan Cem Duru [1], Pauliina Halimaa[10,11], Aapo Kahilainen [3], Suyog S. Kuwar [12,13], Sirpa O. Kärenlampi[10,11], Elvira Lafuente [14], Shiqi Luo [15], Jenny Makkonen [10,11], Abhilash Nair [3], Maria de la Paz Celorio-Mancera [16], Ville Pennanen [17], Annukka Ruokolainen[3], Tarja Sundell[1], Arja I. Tervahauta [10], Victoria Twort [9,18], Erik van Bergen [3], Janina Österman-Udd [3], Lars Paulin [1], Mikko J. Frilander [1], Petri Auvinen [1,*] and Marjo Saastamoinen [3,19,*]

[1]Institute of Biotechnology, University of Helsinki, 00790 Helsinki, Finland
[2]Department of Chemistry and Biotechnology, Tallinn University of Technology, 12618 Tallinn, Estonia
[3]Organismal and Evolutionary Biology Research Programme, University of Helsinki, 00014 Helsinki, Finland
[4]Ming Wai Lau Centre for Reparative Medicine, Karolinska Institutet, 171 77 Stockholm, Hong Kong
[5]Natural Resource Institute (LUKE), 00790 Helsinki, Finland
[6]Department of Evolution, Ecology and Behaviour, University of Liverpool, Liverpool CH64 7TE, UK
[7]Present address: School of Life Sciences, University of Warwick, UK
[8]Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803, USA
[9]Department of Biology, Lund University, 223 62 Lund, Sweden
[10]Department of Environmental and Biological Sciences, University of Eastern Finland, 70211 KUOPIO, Finland
[11]Present address: Biosafe–Biological Safety Solutions Ltd, Microkatu 1 M, 70210 Kuopio, Finland
[12]Department of Entomology and Nematology, University of Florida, Gainesville, FL 32611-0620, USA
[13]Department of Zoology, Loknete Vyankatrao Hiray Arts, Science & Commerce College, 422003, Maharashtra, India
[14]Department of Aquatic Ecology, Swiss Federal Institute of Aquatic Science and Technology, CH-8600 Dübendorf, Switzerland
[15]College of Plant Protection, China Agricultural University, Beijing 100083, China
[16]Department of Zoology, Stockholm University, 106 91 Stockholm, Sweden
[17]Viikki Plant Science Centre, Organismal and Evolutionary Biology Research Programme, University of Helsinki, 00014 Helsinki, Finland
[18]Present address: Luomus, Finnish Museum of Natural History, Zoology Unit, University of Helsinki, Helsinki, Finland
[19]Helsinki Institute of Life Science (HiLIFE), University of Helsinki, 00014 Helsinki, Finland
***Correspondence address**. Petri Auvinen. E-mail: petri.auvinen@helsinki.fi; Marjo Saastamoinen. E-mail: marjo.saastamoinen@helsinki.fi
†These authors contributed equally to the work.
*Corresponding authors contributed equally.

Abstract

Background: The Glanville fritillary (*Melitaea cinxia*) butterfly is a model system for metapopulation dynamics research in fragmented landscapes. Here, we provide a chromosome-level assembly of the butterfly's genome produced from Pacific Biosciences sequencing of a pool of males, combined with a linkage map from population crosses.

Results: The final assembly size of 484 Mb is an increase of 94 Mb on the previously published genome. Estimation of the completeness of the genome with BUSCO indicates that the genome contains 92–94% of the BUSCO genes in complete and single copies. We predicted 14,810 genes using the MAKER pipeline and manually curated 1,232 of these gene models.

Conclusions: The genome and its annotated gene models are a valuable resource for future comparative genomics, molecular biology, transcriptome, and genetics studies on this species.

Keywords: *Melitaea cinxia*, Glanville fritillary, genome, spatial ecology

Data Description

Context

Identifying and characterizing genes underlying ecologically and evolutionarily relevant phenotypes in natural populations has become possible with novel genomic tools that can also be used in "non-model" organisms. The Glanville fritillary (*Melitaea cinxia*, NCBI:txid113334) butterfly, and in particular its metapopulation in the Åland Islands (southwest Finland), is an ecological model system in spatial ecology [1, 2]. In Åland, the species inhabits a network of dry outcrop meadows and pastures and persists as a

classic metapopulation with high turnover in patch occupancy [1]. The network of 4,500 potential habitat patches has been systematically surveyed bi-annually for butterfly occupancy and abundance since 1993 [3], providing a vast amount of ecological data on population dynamics [2]. Experimental manipulations under more controlled conditions are also possible owing to the small size, high fecundity, and relatively short generation time of the species. Consequently, our understanding of the species includes knowledge of life history variation across development stages [4, 5], dispersal dynamics [6, 7], species interactions with host plants and parasitoids [8–12], and stress tolerance [13, 14]. During the past decade, the system has also been used to study genetic and evolutionary processes, such as identifying candidate genes underlying variation and evolution of dispersal in fragmented habitats [15] and host plant preference [16], and assessing allelic variation and its dynamics in space and time [17–19]. Several approaches have been used to explore the genetic underpinnings of phenotypic variation in the Glanville fritillary metapopulation, ranging from candidate gene approaches [13, 20] and quantitative genetics [21, 22] to whole-genome scans [23, 24], under both laboratory and natural environmental conditions.

The first *M. cinxia* genome assembly was released in 2014 [25]. This genome was produced from a combination of 454 sequencing for contig assembly, followed by scaffolding with Illumina paired-end (PE), SOLiD mate-pair reads and Pacific Biosciences (PacBio) data. The size of the final assembly was 390 Mb made up from 8,261 scaffolds, with a scaffold N50 of 119,328. Scaffolds were assigned to chromosomes on the basis of a linkage map produced from RAD sequencing [25]. We recently assessed the actual genome size using a *k*-mer–based approach on Illumina sequencing data and obtained estimates ranging from 488 to 494 Mb (Supplementary File S5). It was considered that a new genome, sequenced using PacBio long reads, would result in a more complete assembly and better represent the repetitive areas of the genome.

Here, a new sequencing and assembly of the *M. cinxia* genome has been carried out using a pool of 7 male butterflies from a single larval family collected from Sottunga, an island in an eastern part of the archipelago. Sequencing was conducted using the PacBio RSII sequencer. An initial assembly was created using FALCON [27, 28] followed by polishing performed with Quiver [27]. A new linkage map was created and used to assign the assembled scaffolds to their correct positions and orientations within the 31 chromosomes. The scaffolds were then gap-filled, producing a final assembly of 484 Mb with a scaffold N50 of 17,331,753 bp. The obtained genome size is well in line with the *k*-mer estimates. Gene prediction on the genome assembly was carried out using MAKER v 2.31.10 [29], which was run iteratively using several independent training sets. Manual annotation was performed for 1,232 of the gene models. The genome assembly increased greatly in contiguity and completeness compared to the first genome (Table 1), with chromosomal superscaffold N50 values of 17,331,753 bp in the new genome compared to 119,328 bp in the Version 1 genome.

The significant increase in assembly size warrants a further investigation of the composition of these added sequences. Initial observations of individual alignments from genome-to-genome alignment show many collapsed repeat regions in the Version 1 genome that are mapped to multiple chromosomes in Version 2.

Methods

An overview of the processing pipeline for the work is shown in Fig. 1.

Genomic samples and DNA extraction

Owing to the facultatively univoltine life cycle of the butterfly in Finland, experimental inbreeding of the species would have taken several years. Therefore, we chose to sample individuals from an island population, Sottunga, expected to harbour lower genetic diversity compared to less isolated populations. Sottunga is part of the Åland Islands archipelago in the northern Baltic Sea, and the population was introduced here in 1991 using individuals collected on the mainland of Åland Island [32]. This introduction was carried out with 71 larval families. The distance to the nearest *M. cinxia* population across the water is 5 km, and we therefore assume that the introduced population has remained (almost) completely isolated. Furthermore, the effective population size of *M. cinxia* in Sottunga has been very low during the past 24 years (on average 57 larval nests/year in 1993–2019), and it has experienced several strong bottlenecks [33]. Using genomic markers, Fountain et al. [17] demonstrated that samples from the Sottunga population separate clearly from samples collected on the mainland.

During the fall survey of 2014 (see Ojanen et al. for details of the survey [3]) we collected individuals from 1 larval group on the island of Sottunga (patch No. 1439, 60 8.1768 N 20 40.1214 E). The larvae were collected once they were in diapause and most likely comprise full siblings [18]. The larval group was kept in diapause (+5°C) until the following spring and then reared to adulthood under common garden conditions (28:8°C; 12L:12D) at the Lammi Biological Station, University of Helsinki. After eclosion, butterflies were sexed and stored at −80°C. High molecular weight DNA was isolated from 7 adult males using the caesium chloride method [25]. Several individuals were used to obtain enough starting material for constructing the Single Molecule, Real-Time (SMRT) sequencing library.

SMRT sequencing libraries and sequencing

Library construction for PacBio sequencing (PacBio RS II Sequencing System, RRID:SCR_017988) was carried out using the protocols recommended by the manufacturer (Pacific Biosciences, Menlo Park, CA, USA). Genomic DNA was sheared using a Megaruptor (Diagenode, Seraing, Belgium) followed by damage repair, end repair, hairpin ligation, and size selection using BluePippin (Sage Science, Beverly, MA, USA; RRID:SCR_020505). After primer annealing and polymerase binding, the DNA templates were sequenced on a PacBio RSII sequencer using P6/C4 chemistry and 360 min video time at the DNA Sequencing and Genomics Laboratory, Institute of Biotechnology, University of Helsinki, Finland [34].
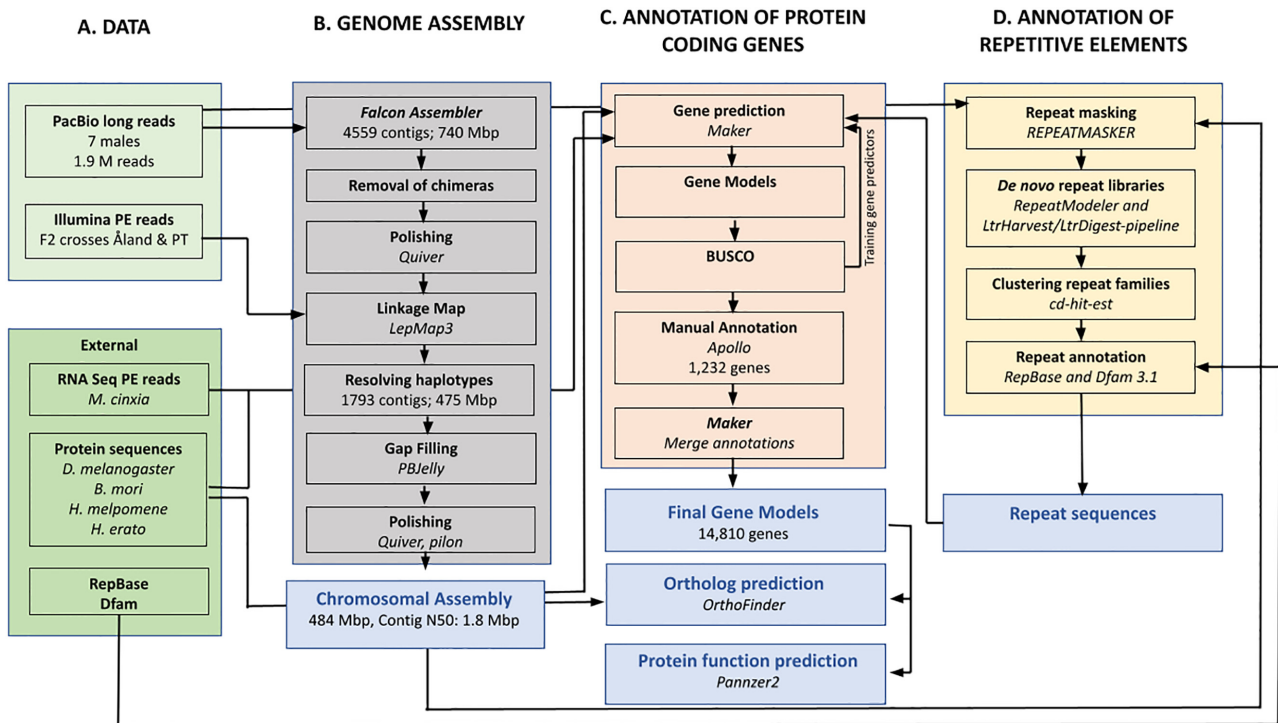
Genome Assembly

The genome was assembled using the FALCON assembler (FALCON-Integrate-1.8.6) [26, 27] with a read length cut-off of 18,000 bp. This cut-off was found to give the best contiguity for the assembly based on N50 value, while minimizing the percentage of possibly erroneous contigs. The erroneous contigs were detected by mapping markers of the linkage map from the previously published genome [25] to contigs and calculating the percentage of chimeric contigs. We tested 3 different read length cut-offs, 16,000, 18,000, and 20,000 bp, all of which included ~9% of chimeric contigs. The assembly was based on 1.9M PacBio reads, 24.4 Gb in total, with an N50 of 18,479 bp, which is ~50× coverage based on the final genome size. With the selected read cut-off the data produced 10.8 Gb of corrected reads that were further assembled using the FALCON software (Falcon, RRID:SCR_016089). The assembly yielded 4,559 primary contigs containing 739.9 Mb with an N50 of 340 kb and 1,661 alternative contigs containing 118.1

Table 1: Assembly statistics

Statistic	M. cinxia		Bombyx mori	Pieris napi v1.1
	Version 2	Version 1 Scaffolds		
Length (bp)	484,462,241	389,907,520	460,334,017	349,759,982
N (%)	<0.01	7.42	0.10	22.47
Scaffold count	31	8,261	696	2,969
Longest scaffold (bp)	22,190,643	668,473	21,465,692	15,427,984
Scaffold N50 length (bp)	17,331,753	119,328	16,796,068	12,597,868
Scaffold N50 count (L50)	13	970	13	13
Contig count	529	48,180	726	53,510
Contig N50 length (bp)	1,831,849	14,057	12,201,325	10,538
Contig N50 count (L50)	79	7,366	16	6,914

Assembly statistics were calculated for the *M. cinxia* Version 2 genome, *M. cinxia* Version 1 scaffolds, and *B. mori* using the assembly-stats program v 17.02 [30]. Statistics for *H. melpomene* v2.5 and *P. napi* v1.1 were obtained from LepBase [31].

Figure 1: An overview of the assembly and annotation process of the improved Glanville fritillary genome.

Mb with an N50 of 85,246 bp. The alternative contigs were automatically separated by the FALCON pipeline. The data were also assembled using miniasm software (0.2-r137-dirty) [35], which yielded similar results. The larger than expected initial assembly size, ~1.5 times the *k*-mer estimate, is due to the multiple haplotypes originating from the 7 individuals used in sequencing.

To evaluate the putative chimeric contigs and assembly errors suggested by the genetic map, the raw SMRT sequencing data were mapped to the assembly primary contigs using BWA (BWA-0.7.17, RRID:SCR_010910) with the MEM algorithm [36]. The alignments of the 425 regions discovered as possibly chimeric were visually inspected. Of these regions, 92 showed even read coverage and no evident signs of assembly errors, while 333 regions contained areas with low coverage and/or repeat regions indicated by high coverage that had led to erroneous overlaps and misassemblies. These errors were identified by positions where the majority of the reads did not fully align; i.e., the alignments ended mid-read. The assembly was

split in the positions where the coverage was at minimum. The resulting assembly was polished using the SMRT sequencing data and Quiver [26] software from the SMRT Tools-package (PacBio).

Linkage Map

Linkage mapping was constructed from whole-genome resequencing data of F2 crosses of *M. cinxia*. The grandparents of these F2 crosses are offspring of wild-collected *M. cinxia* originating from 2 distantly related *M. cinxia* populations around the Baltic Sea: the Åland Islands (ÅL) [1] and Pieni Tytärsaari (PT) populations [37]. Between-population crosses of type ÅL♂×PT♀ and ÅL♀×PT♂ were established to create the F1 population. Some of these F1 individuals were used to establish the F2 families, actively avoiding mating among siblings. A subset of the resulting full-sibling families were reared to adulthood, and 5 of these F2 families, together with their parents and grandparents, were selected for resequenc-

ing. In total, resequencing included 10 grandparental individuals, 10 F1 parents, and 165 F2 individuals (N: 185).

All the larvae from different generations completed development under common garden conditions (28:15°C; 12L:12D) utilizing fresh leaves of greenhouse-grown *Veronica spicata*. Diapausing larvae were kept in a growth chamber at +5°C and 80% relative humidity for ~7 months to mimic the normal wintertime conditions for these butterflies. Adults were kept in hanging cages (of 50 cm height and 40 cm diameter) at ~26:18°C; 9L:15D and fed *ad libitum* with 20% honey-water solution throughout the experiments.

Before DNA extraction the adult butterflies were stored at −80°C, and either thorax or abdomen tissue of these individuals was used for sequencing. Tissues were homogenized prior to extraction using TissueLyser (Qiagen, Venlo, The Netherlands) at 30/s for 1.5 mins with Tungsten Carbide Beads, 3 mm (Qiagen, Venlo, The Netherlands), and ATL buffer (Qiagen, Venlo, The Netherlands). DNA was extracted using the NucleoSpin 96 Tissue Core Kit (Macherey-Nage, Düren, Germany) according to the manufacturer's protocol with the exception that lysing time was extended to overnight. The samples were additionally treated with RNase A (Thermo Fisher Scientific, Waltham, MA. USA) before sequencing. Sequencing was performed using standard PE library preparation and Illumina HiSeq 2000 (Illumina HiSeq2000, RRID:SCR_020132) with 125 bp PE reads.

The mapping procedure followed the Lep-MAP3 [26] pipeline (biotools:lep-map3). First, individual fastq files were mapped to the contig assembly using BWA MEM (BWA-0.7.17) [36] and individual bam files were created using SAMtools (1.6) (SAMTOOLS, RRID:SCR_002105) [38, 39]. SAMtools mpileup and the scripts pileupParser2.awk and pileup2posterior.awk were used to obtain input data for Lep-MAP3. Then ParentCall2 (parameter: ZLimit: 2) and Filtering2 (parameters: dataTolerance: 0.0001; removeNonInformative: 1; familyInformativeLimit: 4) were run to obtain data with ≥4 informative families for each marker, resulting in a final input with almost 2.5M markers.

SeparateChromosomes2 was run on the final data (parameters lodLimit: 20; samplePair: 0.2; numThreads: 48) to obtain 31 linkage groups with a total of 2.4M markers. OrderMarkers2 was run (parameter recombination2: 0) on each linkage group (chromosome). This map was used to anchor the contig assembly into chromosomes. To validate anchoring, the map construction was repeated in the same way except that OrderMarkers2 was run on the physical order of markers to reduce noise in the linkage map. Finally, the raw data were remapped to the gap-filled chromosome-level assembly and the linkage map was redone in the new physical order to infer final recombination rates.

Anchoring the genome and resolving haplotypes using the linkage map

The contigs were aligned against each other and lift-over chains were created by running the first 2 steps (batch A and B to calculate the alignment chain) of the HaploMerger2 [40] pipeline. By manually inspecting this chain (all.chain.gz), contigs fully contained in some longer contig were removed. Initial contig order and orientation within each chromosome was calculated by the median map position of each contig and the longest increasing subsequence of markers, respectively. For each chromosome, Marey map [41], a scatter plot of physical and linkage positions combining the genetic and physical maps, and contig-contig alignments from the chain were recorded. The contigs' orders and orientations were manually fixed when needed if the map had support for alternative orientation. If the contig-contig alignments linked contigs together, they were joined. Any assembly errors that were found were corrected by splitting the contigs accordingly. Also, partially haplotypic contigs were found and collapsed, i.e., alternative haplotype sequence removed, on the basis of the Marey maps and contig-contig alignments. This manual work facilitated the removal of additional haplotype contigs and regions and resulted in the haploid reference genome sequence including start and end positions of contigs in the correct order and orientation for each chromosome. Of 2,933 contigs in initial reference, 4 were chimeric and were split to 9 separate contigs. Of the resulting 2,938 contigs, 1,080 were included without any modification, 825 were trimmed on 1 or both ends, and 1,033 were completely contained and thus removed. Finally, the haplotype-corrected genome was gap-filled using PBJelly software (PBSuite_15.8.24; RRID:SCR_012091) [42] with the original SMRT sequencing data and polished with the Quiver tool [26] from the SMRT Tools-package 2.3.0 (PacBio) and with Pilon (1.21) (Pilon, RRID:SCR_014731) [43], which resulted in the final reference genome sequence of ~484 Mb.

The chromosomes were aligned against the *Heliconius melpomene* (2.5) [44, 45] and *Pieris napi* [46] genomes using the LAST aligner(938) [47] to check structural similarity between the species (Supplementary Figs S1–S13). An overview alignment for *H. melpomene* was created using D-GENIES (1.2.0) (D-GENIES, RRID:SCR_018967) [48] (Fig. 2). The data show a high level of collinearity between *M. cinxia* and *H. melpomene* chromosomes, as described before in Ahola et al. [25]. An interesting point is the lack of collinearity with sex chromosomes (*M. cinxia* chromosome 1 and *H. melpomene* chromosome 21). Furthermore, the visible vertical lines show the effect of long-read assembly on repeat resolution. With long reads spanning the repeats and allowing their accurate placement in the contigs, in *M. cinxia* the repeats are placed in single chromosomes whereas in *H. melpomene* they are present in all chromosomes.

Repeat masking and annotation

Genomic assemblies were masked with *de novo* repeat libraries by RepeatMasker v.4.0.9 (RepeatMasker, RRID:SCR_012954) [49]. *De novo* repeat libraries were constructed from original PacBio reads with lengths >30,000 bp and assembled scaffolds (pseudo chromosomes) using RepeatModeler v 1.0.10 (RepeatModeler, RRID:SCR_015027) [50] and the LtrHarvest/LtrDigest-pipeline [51, 52]. Repeat families were clustered using cd-hit-est applying the 80/80-rule (80% identity over 80% length) [53]. Repeat annotations were confirmed by RepBase Release 20,181,026 [54] and Dfam version 3.1 [55].

Transcriptome assembly

To aid construction of gene models, we capitalized on 2 transcriptome assemblies that were produced as part of separate projects in our laboratory to be presented in upcoming publications ([5], PRJNA670126). Importantly for gene model construction, they represent a wide range of transcriptional diversity, as the RNA-seq data are derived from various developmental stages (first instar larvae, fourth instar larvae, and adult thorax and abdomen). All individuals were lab-reared but originated from the same butterfly metapopulation. Transcriptome 1 was produced using a set of 78 individually sequenced female larvae (fourth developmental instar) [5], sequenced to an average depth of 17.3M reads (read lengths 85 and 65 bp for forward and reverse PE reads, respectively). Because the 2 sexes are practically indistinguishable in the
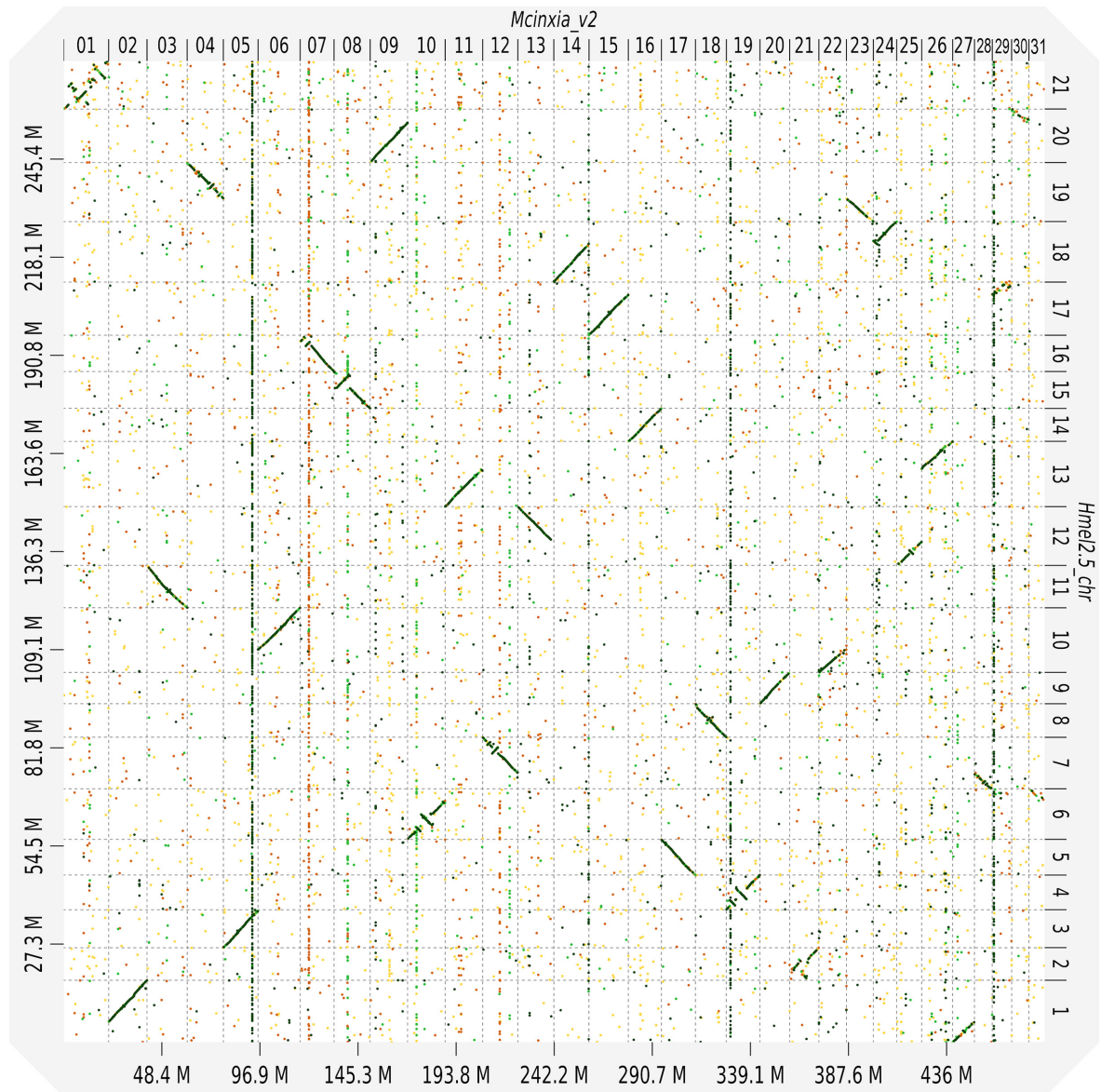
Figure 2: A dot-plot structural comparison of the *H. melpomene* genome against the *M. cinxia* v2 genome. The alignment was created using D-GENIES (1.2.0) [48]. The diagonal lines indicate the collinearity between the 2 species. The lack of collinearity in sex chromosomes is visible in the upper left corner between Mcnxia_v2 chr 01 and Hmel2.5 chr 21. The visible vertical lines show repeats that are resolved in Mcinxia_v2 but are present in all chromosomes in Hmel2.5_chr.

larval stages, the females were identified on the basis of homozygosity across a set of 22 Z-chromosome–specific single-nucleotide polymorphism loci [5]. To remove Illumina adapter sequences, we trimmed raw reads using Trimmomatic (Trimmomatic-0.35, RRID:SCR_011848) [56], and normalized using Trinity v2.6.5 (Trinity, RRID:SCR_013048) [57]. We then used 2 separate procedures to construct *de novo* transcriptome assemblies, Trinity (v2.6.5) and Velvet/Oases (1.2.10) [58]. Trinity was run with standard settings, whereas Velvet/Oases used a range of 7 *k*-mer sizes (21–71 bp), producing a separate assembly for each *k*-mer size. We then combined the resulting assemblies, filtered the combined assembly using the EvidentialGene (tr2aacds.pl VERSION 2017.12.21) [59] pipeline, and removed contigs smaller than 200 bp or expressed at a low level (<1 normalized counts per million), yielding the final assembly. Transcriptome 2 was constructed from a set of 12 adult females (thorax and abdomen, without ovaries) and 48 first instar larvae, as part of a separate gene expression

study (PRJNA670126). RNA from these 60 individual samples was sequenced to an average depth of 16.6M reads (86/74 bp PE). The stranded RNA-seq libraries were made using Ovation® Universal RNA-Seq System (Nugen) with custom ribosomal RNA removal. The libraries were PE sequenced on a NextSeq 500 using the 150 bp kit (Illumina) at the DNA sequencing and genomics laboratory Institute of Biotechnology University of Helsinki. We trimmed the reads using fastp (v0.20.0) [60], and used the HISAT2 2.0.4 (HISAT2, RRID:SCR_015530)/StringTie 1.3.5 (StringTie, RRID:SCR_016323) pipeline [61] to construct a genome-guided transcriptome assembly, mapping the RNA-seq reads to the new genome assembly. Transcriptome 1 yielded 69,182 putative transcripts with mean length of 727 bp (95% CI: 206–3,433), while Transcriptome 2 yielded 137,250 putative transcripts with mean length of 1,737 (95% CI: 203–9,106). These statistics should be interpreted with caution because the assemblies derive from different life stages, and different assembly and filtering approaches were used (re-

flecting differences in histories of the datasets as they were produced for different projects).

Gene model annotation

Initial gene predictions were obtained by running the MAKER v 2.31.10 [29] gene prediction program in an iterative procedure. In the first round of MAKER, transcriptome assembly 1, described above, was provided as evidence, and genes were predicted solely from the aligned transcripts. This resulted in 14,738 gene models. These gene models were then used for training the SNAP (2013–02-16) [62] and AUGUSTUS (3.3.2) (Augustus, RRID:SCR_008417) [63] gene predictors. A second round of MAKER was run providing the *de novo* transcripts from both transcriptomes (see previous paragraph), trained gene prediction models, repeat masking file, and protein data from other lepidopteran species. The MAKER settings were adjusted to allow prediction of gene models without requiring a corresponding transcript in the *de novo* transcriptome assembly. Following each round of MAKER gene prediction, the annotation completeness was assessed using BUSCO (Benchmarking Universal Single-Copy Orthologs, RRID:SCR_015008) [64, 65].

Manual annotation

Manual annotation was performed for 1,232 genes, using the Apollo collaborative annotation system Version 2.1.0 [66]. The collaborative annotation environment was set up in Ubuntu Linux 14.04 server with 250 GB RAM and 48 AMD Opteron 6 168 processing cores. This was later upgraded to a cloud server provided by the Finnish IT Center for Science (CSC) and run on Ubuntu Linux 18.04 with 200 GB RAM and 40 Intel Xeon model 85 processing cores. Evidence tracks were produced containing gene predictions from 3 rounds of MAKER, RNA-seq alignments of sequence reads, and protein alignments from other species (Table 2). RNA-seq alignments comprised a mixed-tissue pooled sample, an abdomen pooled sample, and 6 larval samples (from Transcriptome 1) selected to represent a diverse range and included, e.g., both sexes and different family backgrounds. A list of gene families that were considered of particular interest in butterfly research were identified for prioritization during the manual annotation. (Supplementary File S4). The gene annotators were able to select a family of genes for annotation, or a random selection from the prioritized families was given. Gene models were corrected by examining the evidence tracks in the browser, conducting blast searches, and examining multiple alignments of protein sequences. In total for the 1,232 genes, 1,455 messenger RNAs (mRNAs) were manually inspected, of which 814 genes and mRNAs were changed. Most changes were made to exon borders and mRNA exon structure, especially in the case of multiple isoforms.

Final gene models

Following the manual annotation, the SNAP [62] and AUGUSTUS [63] gene predictors were retrained using the manually annotated gene models. MAKER was run using the updated gene predictors, Transcriptome 1 and 2, and using a masking file for repeats. As a final step to incorporate the manually annotated gene models, MAKER was run, providing the previous MAKER file to pred_gff and the manually annotated models to model_gff. Gene functional prediction was performed using Pannzer v2 [68].

Ortholog identification

Predicted protein sequences from *Bombyx mori* [69] (January 2017 gene models), *P. napi* [46], and *H. melpomene* (Hmel2.5) [44, 45]

were downloaded from SilkBase [70], LepBase [31], and the Butterfly Genome Database [71]. OrthoFinder v2.3.3 (OrthoFinder, RRID: SCR_017118) [72] was run to identify orthologs between *M. cinxia*, *B. mori*, *P. napi*, and *H. melpomene* using blast as the search tool (Fig. 3 and Supplementary Fig. S14).

Data Validation and Quality Control

To assess the quality of the assembly, assembly statistics were generated using assembly-stats [30] and compared to the v1 genome, as well as the *H. melpomene*, *B. mori*, and *P. napi* genome assemblies (Table 1). The new genome contains 94 Mb more sequence than the previous scaffold assembly. On the basis of the observations of individual alignments in the full genome alignment between Version 1 and Version 2, there are many regions in Genome 1 that are aligned into multiple positions in Version 2. This points to collapsed repeat regions in Version 1 and more accurate repeat placement due to the long-read sequencing in Version 2. The N50 length and L50 value at scaffold or chromosome level improved greatly compared to the previous genome. To check for possible duplication or missing areas in the assembly, an assessment was made for the completeness of single-copy orthologs from BUSCO [64, 65] eukaryota, arthropoda, and metazoa gene sets (Table 3). In each of the gene sets, 93.0–94.9% of the expected single-copy orthologs were found in complete copies. The duplication rate was estimated to be between 1.4 and 1.5%. A total of 1,232 gene models were manually curated using the Apollo annotation system [66] to ensure the quality of the models. To test for contamination, the predicted protein sequences were checked with AAI-profiler [73] to identify sequences originating from different taxa (Supplementary Files S1–S3). Overall, 42% of the genome was composed of repeat sequences (Fig. 4 and Supplementary Figs S15–S20 [chromosome-specific repeat classes]). There were no clear differences in the repeat contents between chromosomes (Supplementary Table S1), which further supports the more accurate placement of repeats due to the long-read sequencing in Version 2. Long interspersed nuclear elements (LINEs) were the most prevalent.

Reuse potential

The substantial improvements in contiguity and gene annotation quality of the new genome will enable a range of important new studies and open up possibilities for future work. The results also demonstrate that with the use of proper computational tools and data, it is possible to obtain a high-quality, chromosome-scale reference genome even when a single individual organism will not provide enough high molecular weight DNA for long-read sequencing. Furthermore, we show the potential of the linkage mapping: it anchors contigs to actual chromosomes instead of just linking different contigs together as is done, for example, in the Hi-C approach. Moreover, the haplotype problem is not tackled by Hi-C. Our high-density linkage map allows us to put nearly all contigs into chromosomes. It is worth noting that the linkage map is not scaffolding directly but it puts contigs into map positions; scaffolding is possible if a contig spans 2 or more map positions. Otherwise, the contig can be placed only partially. In addition to the linkage map approach, we used extensive manual curation of the assembly to avoid chimeric parts and improve the assembly quality. Current research aims at identifying mechanisms underlying key life history adaptations, exploring the extent of natural variation and selection on these adaptations in wild populations, and integrating these insights with

Table 2: Evidence tracks that were used during the manual annotation of 1,232 *M. cinxia* genes

Evidence track	Type	Description
Maker 1	Gene prediction	Initial maker gene predictions based on EST alignments
Maker 2	Gene prediction	Second round of gene predictions from EST alignments, protein alignments, and gene predictors trained on maker 1.
RNA-seq abdomen pool	RNA-seq alignment	RNA-seq reads aligned to the genome with STAR [67]
RNA-seq mixed-tissue pool	RNA-seq alignment	
B. mori proteins	Protein alignment	Proteins sequences aligned to the genome with AAT
H. melpomene proteins	Protein alignment	
Drosophila melanogaster proteins	Protein alignment	
Heliconius erato proteins	Protein alignment	
RNA-seq female larvae family 80	RNA-seq alignment	RNA-seq reads aligned to the genome with STAR [67]
RNA-seq female larvae family 70	RNA-seq alignment	
RNA-seq female larvae family 119	RNA-seq alignment	
RNA-seq female larvae family 120	RNA-seq alignment	
RNA-seq male larvae family 80	RNA-seq alignment	
RNA-seq male larvae family 119	RNA-seq alignment	

Table 3: BUSCO completeness estimates of the v2 genome based on the eukaryota, arthropoda, and metazoa gene sets

Lineage	BUSCO Category, No. (%)				
	Complete	Single-copy	Duplicated	Fragmented	Missing
Eukaryota	237 (93.0)	234 (91.8)	3 (1.2)	9 (3.5)	9 (3.5)
Arthropoda	960 (94.8)	946 (93.4)	14 (1.4)	16 (1.6)	37 (3.6)
Metazoa	905 (94.9)	891 (93.4)	14 (1.5)	16 (1.7)	33 (3.4)

the exceptional ecological, demographic, and climatic data available for this system. Future studies in this direction will help identify the mechanisms maintaining variation in life histories across spatial and temporal scales, and the extent to which phenotypic variation in these and other traits may contribute to a population's adaptive capacity under climate change. Several studies in different species illustrate how stress responses can be crucial for survival under variable environments, both within and between generations. The Glanville fritillary is being used to explore how environmental information is translated into adaptive phenotypic changes, and how these responses are transmitted to future generations, using transcriptomic and epigenetic approaches. Such studies will benefit from an improved annotation permitting exon-specific expression quantification, and identification of epigenetic marks and other functional variants outside coding regions. Exploiting current and past large-scale sampling efforts, these new studies apply population genomic approaches that are facilitated by the increased assembly contiguity, e.g., by permitting linkage disequilibrium and haplotype-based selection analyses. Other avenues of research enabled by the improved genome assembly include structural variation, regulatory evolution, recombination rate variation, and coalescent-based demographic analyses. The increasing availability of chromosome-level lepidopteran genomes such as ours permits exciting new comparative phylogenetic analyses, e.g., of chromosome and genome evolution.

Data Availability

The SMRT sequencing reads used for the genome assembly are available in the NCBI SRA and can be accessed with Bioproject PRJNA607899 accession No. SRR11184190.

The genome has been deposited to GenBank under Bioproject PRJNA607899.

The Illumina reads used for the linkage map are available in the NCBI SRA and can be accessed with Bioproject PRJNA608928 accession Nos. SRR11186917–SRR11187107.

Transcriptome 1 RNA-seq reads are available in NCBI GEO and can be accessed with accession No. GSE159376.

Transcriptome 2 RNA-seq reads are available in NCBI SRA and can be accessed with Bioproject PRJNA670126.

All supporting data and materials are available in the *GigaScience* GigaDB database [74].

Additional Files

Supplementary Figure S1: *M. cinxia* aligned against *H. melpomene* using the last aligner [47]. A: *M. cinxia* chromosome 1 (M01_B01_H21), B: chromosome 2 (M02_B04_H01a), C: chromosome 3 (M03_B15_H11), D: chromosome 4 (M04_B12_H19a), E: chromosome 5 (M05_B06_H03), and F: chromosome 6 (M06_B05_H10a).
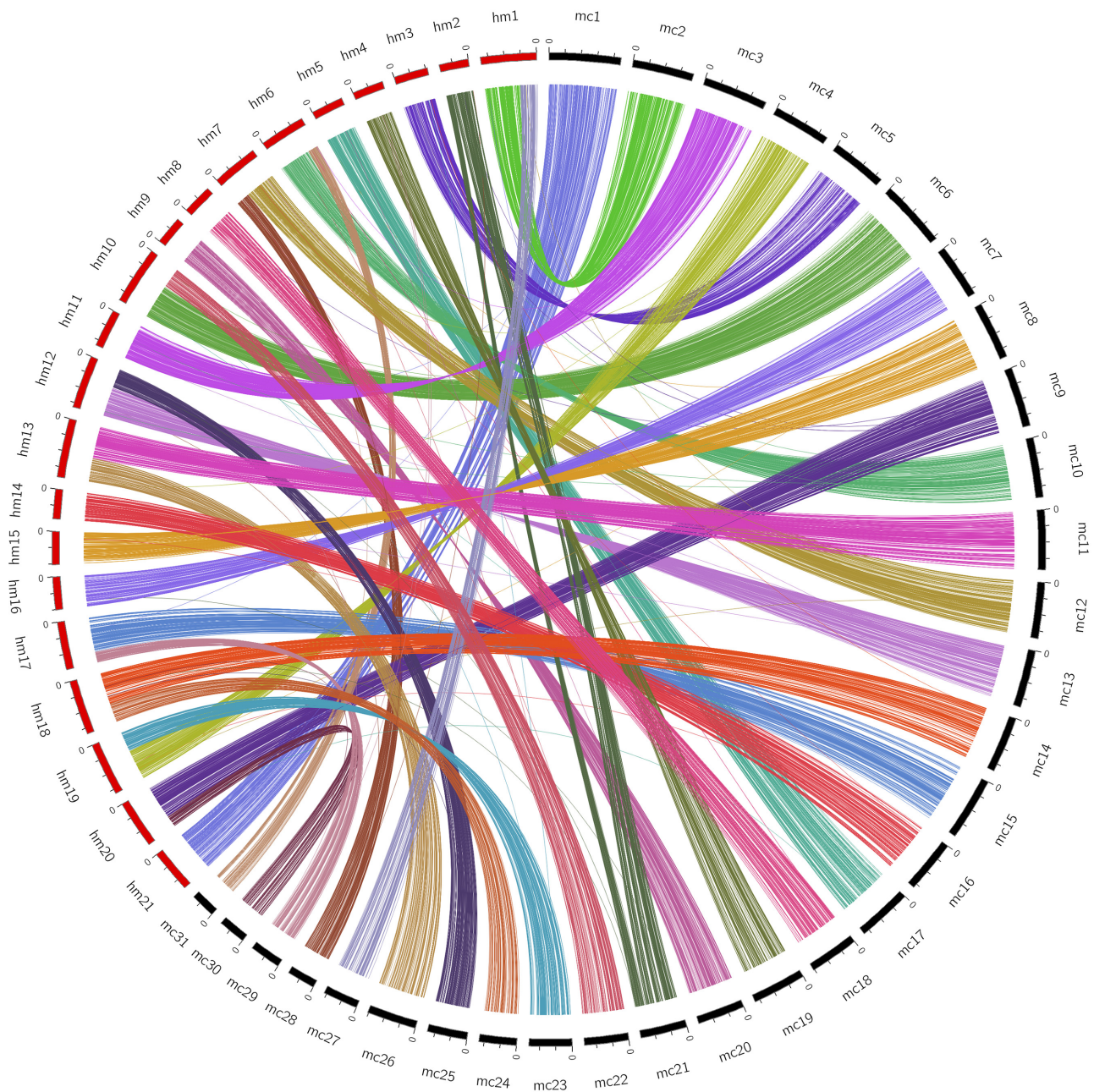
Figure 3: A circos plot showing the orthologs between *M. cinxia* and *H. melpomene*. Orthologs between *M. cinxia* and *H. melpomene* were identified using OrthoFinder and filtered for 1-to-1 orthologs. The internal links in the circos plot indicate the orthologs between *M. cinxia* and *H. melpomene*. The links are coloured according to the *M. cinxia* chromosome.

Supplementary Figure S2: *M. cinxia* aligned against *H. melpomene* using the last aligner [47]. A: *M. cinxia* chromosome 7 (M07_B18_H16), B: chromosome 8 (M08_B17_H15), C: chromosome 9 (M09_B10_H20a), D: chromosome 10 (M10_B09_H06a), E: chromosome 11 (M11_B22_H13a), and F: chromosome 12 (M12_B11a_H07a).

Supplementary Figure S3: *M. cinxia* aligned against *H. melpomene* using the last aligner [47]. A: *M. cinxia* chromosome 13 (M13_B08_H12a), B: chromosome 14 (M14_B23a_H18a), C: chromosome 15 (M15_B13_H17a), D: chromosome 16 (M16_B19_H14), E: chromosome 17 (M17_B03_H05), and F: chromosome 18 (M18_B25_H08).

Supplementary Figure S4: *M. cinxia* aligned against *H. melpomene* using the last aligner [47]. A: *M. cinxia* chromosome 19 (M19_B21_H04), B: chromosome 20 (M20_B07_H09), C: chromosome 21 (M21_B16_H02), D: chromosome 22 (M22_B28_H10b), E: chromosome 23 (M23_B26_H19b), and F: chromosome 24 (M24_B27_H18b).

Supplementary Figure S5: *M. cinxia* aligned against *H. melpomene* using the last aligner [47]. A: *M. cinxia* chromosome 25 (M25_B20_H12b), B: chromosome 26 (M26_B14_H13b), C: chromosome 27 (M27_B24a_H01b), D: chromosome 28 (M28_B02_H07b), E: chromosome 29 (M29_B24b_H17b), and F: chromosome 30 (M30_B23b_H20b).
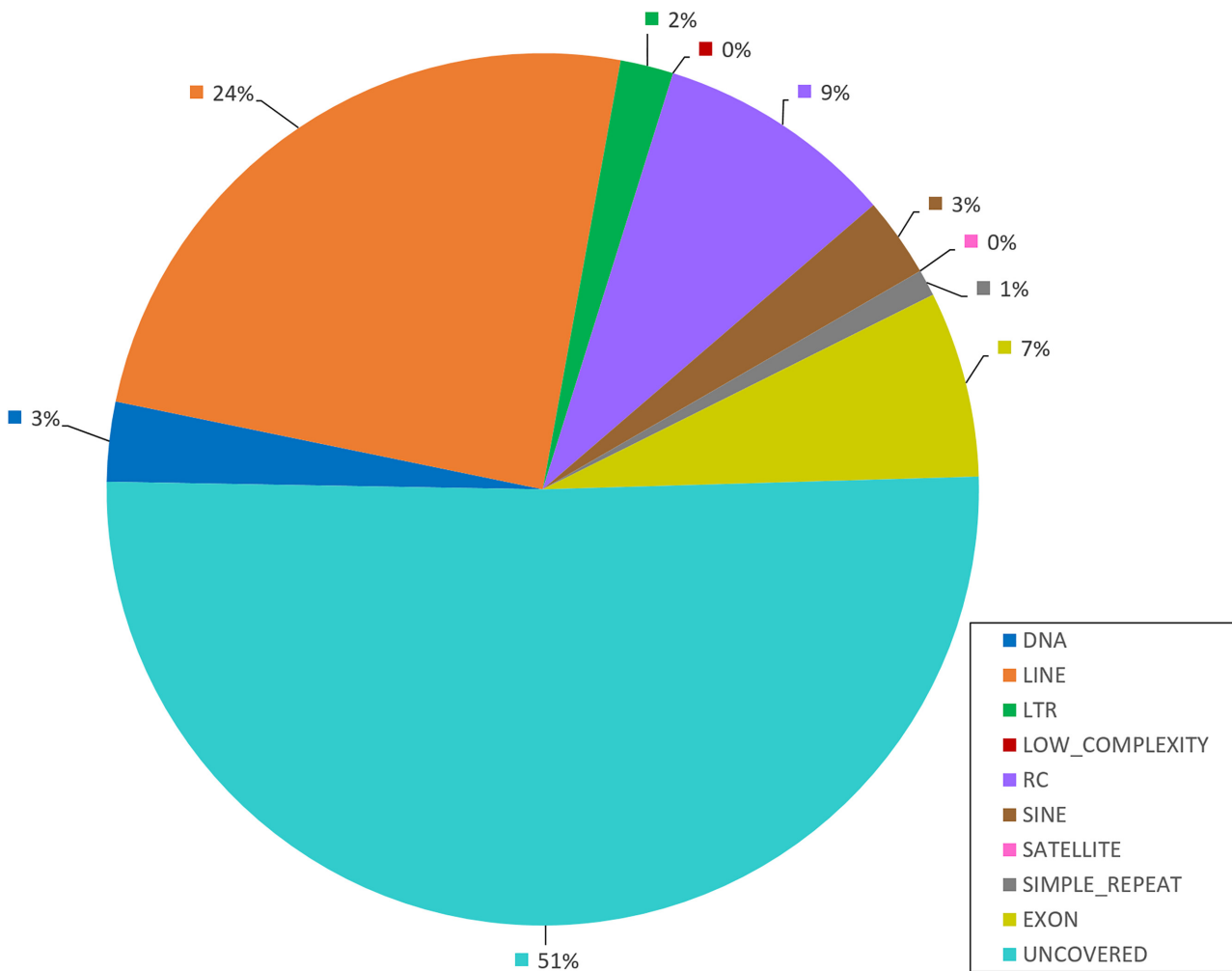
Figure 4: Relative amounts of different repeat classes in *M. cinxia* genome. Repeat classes and coverage of the *M. cinxia* genome v2: DNA: Class II; LINE: long interspersed nuclear elements; LTR: long terminal repeats; LOW_COMPLEXITY: low-complexity repeated DNA; RC: rolling circle elements (e.g., helitrons); SINE: short interspersed nuclear elements; Satellite: satellite DNA; SIMPLE_REPEAT: simple repeated motifs; EXON: exonic regions; UNCOVERED: rest of the chromosomes.

Supplementary Figure S6: *M. cinxia* aligned against *H. melpomene* using the last aligner [47]. *M. cinxia* chromosome 31 (M31_B11b_H06b).

Supplementary Figure S7: A dot-plot showing the structure of *P. napi* genome against *M. cinxia* genome v.2. The diagonal lines indicate the collinearity between the 2 species.

Supplementary Figure S8: *M. cinxia* aligned against *P. napi* using the last aligner [47]. A: *M. cinxia* chromosome 1 (M01_B01_H21), B: chromosome 2 (M02_B04_H01a), C: chromosome 3 (M03_B15_H11), D: chromosome 4 (M04_B12_H19a), E: chromosome 5 (M05_B06_H03), and F: chromosome 6 (M06_B05_H10a).

Supplementary Figure S9: *M. cinxia* aligned against *P. napi* using the last aligner [47]. A: *M. cinxia* chromosome 7 (M07_B18_H16), B: chromosome 8 (M08_B17_H15), C: chromosome 9 (M09_B10_H20a), D: chromosome 10 (M10_B09_H06a), E: chromosome 11 (M11_B22_H13a), and F: chromosome 12 (M12_B11a_H07a).

Supplementary Figure S10: *M. cinxia* aligned against *P. napi* using the last aligner [47]. A: *M. cinxia* chromosome 13 (M13_B08_H12a), B: chromosome 14 (M14_B23a_H18a), C: chromosome 15 (M15_B13_H17a), D: chromosome 16 (M16_B19_H14),

E: chromosome 17 (M17_B03_H05), and F: chromosome 18 (M18_B25_H08).

Supplementary Figure S11: *M. cinxia* aligned against *P. napi* using the last aligner [47]. A: *M. cinxia* chromosome 19 (M19_B21_H04), B: chromosome 20 (M20_B07_H09), C: chromosome 21 (M21_B16_H02), D: chromosome 22 (M22_B28_H10b), E: chromosome 23 (M23_B26_H19b), and F: chromosome 24 (M24_B27_H18b).

Supplementary Figure S12: *M. cinxia* aligned against *P. napi* using the last aligner [47]. A: *M. cinxia* chromosome 25 (M25_B20_H12b), B: chromosome 26 (M26_B14_H13b), C: chromosome 27 (M27_B24a_H01b), D: chromosome 28 (M28_B02_H07b), E: chromosome 29 (M29_B24b_H17b), and F: chromosome 30 (M30_B23b_H20b).

Supplementary Figure S13: *M. cinxia* aligned against *P. napi* using the last aligner [47]. *M. cinxia* chromosome 31 (M31_B11b_H06b).

Supplementary Figure S14: Orthologs between *M. cinxia* and *P. napi* were identified using OrthoFinder and filtered for 1-to-1 orthologs. The internal links in the circos plot indicate the orthologs between *M. cinxia* and *P. napi*. The links are coloured according to the *M. cinxia* chromosome.

Supplementary Figure S15: Repeat classes and coverage of the *M. cinxia* genome v.2. A: M. cinxia chromosome 1

(M01_B01_H21), B: chromosome 2 (M02_B04_H01a), C: chromosome 3 (M03_B15_H11), D: chromosome 4 (M04_B12_H19a), E: chromosome 5 (M05_B06_H03), and F: chromosome 6 (M06_B05_H10a). (DNA: Class II; LINE: long interspersed elements; LTR: long terminal repeats; Low_complexity: low-complexity repeated DNA; RC: rolling circle elements [e.g., Helitrons]; SINE: short interspersed nuclear elements; Satellite: satellite DNA; Simple_repeat: simple repeated motifs; Exon: exonic regions; Uncovered: rest of the chromosome).

Supplementary Figure S16: Repeat classes and coverage of the *M. cinxia* genome v.2. A: *M. cinxia* chromosome 7 (M07_B18_H16), B: chromosome 8 (M08_B17_H15), C: chromosome 9 (M09_B10_H20a), D: chromosome 10 (M10_B09_H06a), E: chromosome 11 (M11_B22_H13a), and F: chromosome 12 (M12_B11a_H07a). (DNA: Class II; LINE: long interspersed elements; LTR: long terminal repeats; Low_complexity: low-complexity repeated DNA; RC: rolling circle elements [e.g., Helitrons]; SINE: short interspersed nuclear elements; Satellite: satellite DNA; Simple_repeat: simple repeated motifs; Exon: exonic regions; Uncovered: rest of the chromosome).

Supplementary Figure S17: Repeat classes and coverage of the *M. cinxia* genome v.2. A: *M. cinxia* chromosome 13 (M13_B08_H12a), B: chromosome 14 (M14_B23a_H18a), C: chromosome 15 (M15_B13_H17a), D: chromosome 16 (M16_B19_H14), E: chromosome 17 (M17_B03_H05), and F: chromosome 18 (M18_B25_H08). (DNA: Class II; LINE: long interspersed elements; LTR: long terminal repeats; Low_complexity: low-complexity repeated DNA; RC: rolling circle elements [e.g., Helitrons]; SINE: short interspersed nuclear elements; Satellite: satellite DNA; Simple_repeat: simple repeated motifs; Exon: exonic regions; Uncovered: rest of the chromosome).

Supplementary Figure S18: Repeat classes and coverage of the *M. cinxia* genome v.2. A: *M. cinxia* chromosome 19 (M19_B21_H04), B: chromosome 20 (M20_B07_H09), C: chromosome 21 (M21_B16_H02), D: chromosome 22 (M22_B28_H10b), E: chromosome 23 (M23_B26_H19b), and F: chromosome 24 (M24_B27_H18b). (DNA: Class II; LINE: long interspersed elements; LTR: long terminal repeats; Low_complexity: low-complexity repeated DNA; RC: rolling circle elements [e.g., Helitrons]; SINE: short interspersed nuclear elements; Satellite: satellite DNA; Simple_repeat: simple repeated motifs; Exon: exonic regions; Uncovered: rest of the chromosome).

Supplementary Figure S19: Repeat classes and coverage of the *M. cinxia* genome v.2. A: *M. cinxia* chromosome 25 (M25_B20_H12b), B: chromosome 26 (M26_B14_H13b), C: chromosome 27 (M27_B24a_H01b), D: chromosome 28 (M28_B02_H07b), E: chromosome 29 (M29_B24b_H17b), and F: chromosome 30 (M30_B23b_H20b). (DNA: Class II; LINE: long interspersed elements; LTR: long terminal repeats; Low_complexity: low-complexity repeated DNA; RC: rolling circle elements [e.g., Helitrons]; SINE: short interspersed nuclear elements; Satellite: satellite DNA; Simple_repeat: simple repeated motifs; Exon: exonic regions; Uncovered: rest of the chromosome).

Supplementary Figure S20: Repeat classes and coverage of the *M. cinxia* genome v.2. *M. cinxia* chromosome 31 (M31_B11b_H06b). (DNA: Class II; LINE: long interspersed elements; LTR: long terminal repeats; Low_complexity: low-complexity repeated DNA; RC: rolling circle elements [e.g., Helitrons]; SINE: short interspersed nuclear elements; Satellite: satellite DNA; Simple_repeat: simple repeated motifs; Exon: exonic regions; Uncovered: rest of the chromosome).

Supplementary File S1: Report from AAI-profiler on predicted protein sequences.

Supplementary File S2: Results from AAI-profiler on matrix-format on predicted protein sequences.
Supplementary File S3: Results from AAI-profiler on krona plot format on predicted protein sequences.
Supplementary File S4: List of prioritized gene families selected based on particular interest in butterfly research
Supplementary File S5: Kmer analysis for genome size estimation
Supplementary Table S1: Repeat contents of chromosomes

Abbreviations

BLAST: Basic Local Alignment Search Tool; bp: base pairs; BUSCO: Benchmarking Universal Single-Copy Orthologs; BWA: Burrows-Wheeler Aligner; CI: confidence interval; Gb: gigabase pairs; LINE: long interspersed nuclear element; LTR: long terminal repeat; Mb: megabase pair; mRNA: messenger RNA; NCBI: National Center for Biotechnology Information; PacBio: Pacific Biosciences; PE: paired-end; RAD sequencing: restriction-site associated DNA sequencing; RAM: random access memory; SMRT: Single Molecule, Real-Time; SRA: Sequence Read Archive.

Ethics Approval and Consent to Participate

There are no ethical policies related to working with insect data. The Glanville fritillary is not considered endangered in the Åland Islands and no permits are required for sampling. However, we note that within this project the larval sampling for genetic analyses is done non-invasively in the field, ensuring insignificant demographic impact. In addition, because the sampling takes place prior to diapause (Åland) when mortality is generally the highest, the collection has negligible effect on family survival or the demographic characteristics of populations.

Competing Interests

The authors declare that they have no competing interests.

Funding

Authors' Contributions

O.-P.S. assembled the genome, processed the chimeric contigs, performed the gap filling and the polishing of the assembly, and participated in the genome analysis.

V.A. was responsible for the initial idea of the approach for the genome-related activities, coordinated the first part of the project, designed and produced data for the linkage map, and worked on solving the haplotypes from the initial assembly.

D.B. performed gene prediction, functional annotation, ortholog prediction, and managed the manual annotation.

J.I.K. installed and managed the Apollo annotation server.

S.I. was responsible for larval rearing and preparation of butterfly crosses.

P.R. performed the linkage mapping and anchored the genome onto chromosomes.

V.O. assembled the transcriptomes used for gene prediction.

L. Pellegrini manually inspected the chimeric contigs.

A.R. performed DNA extraction.

D.B., J.I.K., V.O., T.D., M.F.D., A.D., I.C.D., P.H., A.K., S.S.K., S.O.K., E.L., S.L., J.M., A.N., M.C.-M., V.P., T.S., A.I.T., V.T., E.v.B., J.Ö.-U., and M.S. participated in manual annotation.

J.T. performed the annotation of transposable elements and repeat classes.

L. Paulin was responsible for the management of the DNA sequencing.

M.J.F. was responsible for the management of the genome analysis.

P.A. was responsible for the initial idea of the approach for the genome-related activities, and the management of the genome analysis.

M.S. was responsible for the management of the *M. cinxia* database and genome analysis.

O.-P.S., D.B., V.A., P.R., J.T., J.I.K., V.O., L.P., M.J.F., P.A., and M.S. wrote the manuscript.

References

1. Hanski, I. Habitat connectivity, habitat continuity, and metapopulations in dynamic landscapes. *Oikos* 1999;**87**:209–19.
2. Ovaskainen, O, Saastamoinen, M. Frontiers in metapopulation biology: the legacy of Ilkka Hanski. *Annu Rev Ecol Evol Syst* 2018;**49**:231–52.
3. Ojanen, SP, Nieminen, M, Meyke, E, et al. Long-term metapopulation study of the Glanville fritillary butterfly (*Melitaea cinxia*): survey methods, data management, and long-term population trends. *Ecol Evol* 2013;**3**(11):3713–37.
4. Saastamoinen, M, Hirai, N, Van Nouhuys, S. Direct and transgenerational responses to food deprivation during development in the Glanville fritillary butterfly. *Oecologia* 2013;**171**(1):93–104.
5. Kahilainen, A, Oostra, V, Somervuo, P, et al. Alternative developmental and transcriptomic responses to host plant water limitation in a butterfly metapopulation. *Mol Ecol* 2021:doi:10.1111/mec.16178.
6. Hanski, I, Saastamoinen, M, Ovaskainen, O. Dispersal-related life-history trade-offs in a butterfly metapopulation. *J Anim Ecol* 2006;**75**(1):91–100.
7. Niitepõld, K, Smith, AD, Osborne, JL, et al. Flight metabolic rate and Pgi genotype influence butterfly dispersal rate in the field. *Ecology* 2009;**90**(8):2223–32.
8. Hanski, I, Singer, MC. Extinction-colonization dynamics and host-plant choice in butterfly metapopulations. *Am Nat* 2001;**158**(4):341–53.
9. Rosa, E, Woestmann, L, Biere, A, et al. A plant pathogen modulates the effects of secondary metabolites on the performance and immune function of an insect herbivore. *Oikos* 2018;**127**:1539–49.
10. Rosa, E, Minard, G, Lindholm, J, et al. Moderate plant water stress improves larval development, and impacts immunity and gut microbiota of a specialist herbivore. *PLoS One* 2019;**14**(2):e0204292.
11. Salgado, AL, Saastamoinen, M. Developmental stage-dependent response and preference for host plant quality in an insect herbivore. *Anim Behav* 2019;**150**:27–38.
12. Van Nouhuys, S, Lei, G. Parasitoid-host metapopulation dynamics: the causes and consequences of phenological asynchrony. *J Anim Ecol* 2004;**73**:526–35.
13. De Jong, MA, Saastamoinen, M. Environmental and genetic control of cold tolerance in the Glanville fritillary butterfly. *J Evol Biol* 2018;**31**(5):636–45.
14. Saastamoinen, M, Ikonen, S, Wong, SC, et al. Plastic larval development in a butterfly has complex environmental and genetic causes and consequences for population dynamics. *J Anim Ecol* 2013;**82**:529–39.
15. Niitepõld, K, Saastamoinen, M. A candidate gene in an ecological model species: phosphoglucose isomerase (Pgi) in the Glanville fritillary butterfly (*Melitaea cinxia*). *Ann Zool Fennici* 2017;**54**:259–73.
16. De Jong, MA, Wong, SC, Lehtonen, R, et al. Cytochrome P450 gene CYP337 and heritability of fitness traits in the Glanville fritillary butterfly. *Mol Ecol* 2014;**23**(8):1994–2005.
17. Correction for Fountain et al., Predictable allele frequency changes due to habitat fragmentation in the Glanville fritillary butterfly. *Proc Natl Acad Sci U S A* 2016;**113**(36):E5363.
18. Fountain, T, Husby, A, Nonaka, E, et al. Inferring dispersal across a fragmented landscape using reconstructed families in the Glanville fritillary butterfly. *Evol Appl* 2018;**11**:287–97.
19. Dileo, MF, Husby, A, Saastamoinen, M. Landscape permeability and individual variation in a dispersal-linked gene jointly determine genetic structure in the Glanville fritillary butterfly. *Evol Lett* 2018;**2**(6):544–56.
20. Haag, CR, Saastamoinen, M, Marden, JH, et al. A candidate locus for variation in dispersal rate in a butterfly metapopulation. *Proc Biol Sci* 2005;**272**(1580):2449–56.
21. Mattila, ALK, Hanski, I. Heritability of flight and resting metabolic rates in the Glanville fritillary butterfly. *J Evol Biol* 2014;**27**(8):1733–43.
22. Klemme, I, Hanski, I. Heritability of and strong single gene (Pgi) effects on life-history traits in the Glanville fritillary butterfly. *J Evol Biol* 2009;**22**(9):1944–53.
23. Kvist, J, Mattila, ALK, Somervuo, P, et al. Flight-induced changes in gene expression in the Glanville fritillary butterfly. *Mol Ecol* 2015;**24**(19):4886–900.
24. Kvist, J, Wheat, CW, Kallioniemi, E, et al. Temperature treatments during larval development reveal extensive heritable and plastic variation in gene expression and life history traits. *Mol Ecol* 2013;**22**(3):602–19.
25. Ahola, V, Lehtonen, R, Somervuo, P, et al. The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera. *Nat Commun* 2014;**5**: 4737.
26. Rastas, P, Paulin, L, Hanski, I, et al. Lep-MAP: fast and accurate linkage map construction for large SNP datasets. *Bioinformatics* 2013;**29**(24):3128–34.
27. Chin, C-S, Alexander, DH, Marks, P, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* 2013;**10**(6):563–9.
28. Chin, C-S, Peluso, P, Sedlazeck, FJ, et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods* 2016;**13**(12):1050–4.

29. Campbell, MS, Holt, C, Moore, B, *et al.* Genome annotation and curation using MAKER and MAKER-P. *Curr Protoc Bioinformatics* 2014;**48**:4.11.11–39.

30. Challis, R. Assembly-stats 2017. Zenodo 2017. http://doi.org/10.5281/zenodo.594927.

31. Challis, RJ, Kumar, S, Dasmahapatra, KK, *et al.* Lepbase: the Lepidopteran genome database. bioRxiv 2016:doi:10.1101/056994.

32. Thomas, CD, Hanski, I. Metapopulation dynamics in changing environments: butterfly responses to habitat and climate change. In: I Hanski, OE Gaggiotti, eds. *Ecology, Genetics and Evolution of Metapopulations.* Elsevier. 2004:489–514.

33. van Bergen, E, Dallas, T, Di Leo, MF, *et al.* The effect of summer drought on the predictability of local extinctions in a butterfly metapopulation. *Conserv Biol* 2020;**34**(6):1503–11.

34. Salojärvi, J, Smolander, O-P, Nieminen, K, *et al.* Genome sequencing and population genomic analyses provide insights into the adaptive landscape of silver birch. *Nat Genet* 2017;**49**:904–12.

35. Li, H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* 2016;**32**(14):2103–10.

36. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv 2013:1303.3997.

37. Mattila, ALK, Duplouy, A, Kirjokangas, M, *et al.* High genetic load in an old isolated butterfly population. *Proc Natl Acad Sci U S A* 2012;**109**(37):E2496–505.

38. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 2011;**27**(21):2987–93.

39. Danecek, P, Bonfield, JK, Liddle, J, *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* 2021;**10**(2): doi:10.1093/gigascience/giab008.

40. Huang, S, Kang, M, Xu, A. HaploMerger2: rebuilding both haploid sub-assemblies from high-heterozygosity diploid genome assembly. *Bioinformatics* 2017;**33**(16):2577–9.

41. Chakravarti, A. A graphical representation of genetic and physical maps: the Marey map. *Genomics* 1991;**11**(1):219–22.

42. English, AC, Richards, S, Han, Yi, *et al.* Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS One* 2012;**7**(11):e47768.

43. Walker, BJ, Abeel, T, Shea, T, *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 2014;**9**(11): e112963.

44. Davey, JW, Barker, SL, Rastas, PM, *et al.* No evidence for maintenance of a sympatric *Heliconius* species barrier by chromosomal inversions. *Evol Lett* 2017;**1**(3):138–54.

45. Davey, JW, Chouteau, M, Barker, SL, *et al.* Major improvements to the *Heliconius melpomene* genome assembly used to confirm 10 chromosome fusion events in 6 million years of butterfly evolution. *G3 (Bethesda)* 2016;**6**(3):695–708.

46. Hill, J, Rastas, P, Hornett, EA, *et al.* Unprecedented reorganization of holocentric chromosomes provides insights into the enigma of lepidopteran chromosome evolution. *Sci Adv* 2019;**5**(6):eaau3648.

47. Kiełbasa, SM, Wan, R, Sato, K, *et al.* Adaptive seeds tame genomic sequence comparison. *Genome Res* 2011;**21**(3): 487–93.

48. Cabanettes, F, Klopp, C. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ* 2018;**6**:e4958.

49. Smit, AFA, Hubley, R, Green, P. RepeatMasker Open-4.0. 2013-2015. http://www.repeatmasker.org. Accessed 13 Dec 2021.

50. Smit, AFA, Hubley, R. RepeatModeler Open-1.0. 2008-2015. http://www.repeatmasker.org. Accessed 13 Dec 2021.

51. Ellinghaus, D, Kurtz, S, Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* 2008;**9**:18.

52. Steinbiss, S, Willhoeft, U, Gremme, G, *et al.* Fine-grained annotation and classification of de novo predicted LTR retrotransposons. *Nucleic Acids Res* 2009;**37**(21):7002–13.

53. Fu, L, Niu, B, Zhu, Z, *et al.* CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012;**28**(23):3150–2.

54. Jurka, J. Repbase Update - a database and an electronic journal of repetitive elements. *Trends Genet* 2000;**16**(9):418–20.

55. Hubley, R, Finn, RD, Clements, J, *et al.* The Dfam database of repetitive DNA families. *Nucleic Acids Res* 2016;**44**(D1):D81–9.

56. Bolger, AM, Lohse, M, Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;**30**(15):2114–20.

57. Grabherr, MG, Haas, BJ, Yassour, M, *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 2011;**29**(7):644–52.

58. Schulz, MH, Zerbino, DR, Vingron, M, *et al.* Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* 2012;**28**(8):1086–92.

59. Gilbert, D. Gene-omes built from mRNA-seq not genome DNA. *F1000Res* 2016;**5**:1695 (poster). https://doi.org/10.7490/f1000research.1112594.1.

60. Chen, S, Zhou, Y, Chen, Y, *et al.* fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 2018;**34**(17):i884–90.

61. Pertea, M, Kim, D, Pertea, GM, *et al.* Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc* 2016;**11**(9):1650–67.

62. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* 2004;**5**:59.

63. Lomsadze, A. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res* 2005;**33**(20):6494–506.

64. Simão, FA, Waterhouse, RM, Ioannidis, P, *et al.* BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 2015;**31**(19):3210–2.

65. Waterhouse, RM, Seppey, M, Simão, FA, *et al.* BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol Biol Evol* 2018;**35**(3):543–8.

66. Dunn, NA, Unni, DR, Diesh, C, *et al.* Apollo: democratizing genome annotation. *PLoS Comput Biol* 2019;**15**(2):e1006790.

67. Dobin, A, Davis, CA, Schlesinger, F, *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;**29**(1):15–21.

68. Törönen, P, Medlar, A, Holm, L. PANNZER2: a rapid functional annotation web server. *Nucleic Acids Res* 2018;**46**(W1):W84–8.

69. Kawamoto, M, Jouraku, A, Toyoda, A, *et al.* High-quality genome assembly of the silkworm, *Bombyx mori. Insect Biochem Mol Biol* 2019;**107**:53–62.

70. Kawamoto, M. SilkBase 2017. http://silkbase.ab.a.u-tokyo.ac.jp/cgi-bin/download.cgi. Accessed 13 Dec 2021.

71. Reed, B, Pillardy, J. Butterfly Genome Database 2016. http://butterflygenome.org. Accessed 13 Dec 2021.

72. Emms, DM, Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* 2019;**20**(1):238.

73. Medlar, AJ, Törönen, P, Holm, L. AAI-profiler: fast proteome-wide exploratory analysis reveals taxonomic identity, misclassification and contamination. *Nucleic Acids Res* 2018;**46**(W1):W479–85.

74. Smolander, O, Blande, D, Ahola, V, *et al.* Supporting data for "Improved chromosome level genome assembly of the Glanville fritillary butterfly (*Melitaea cinxia*) based on SMRT sequencing and linkage map." *GigaScience Database* 2021. http://dx.doi.org/10.5524/100915.