Online Methods

Sample collection and DNA extraction. Pupal DNA was isolated from a 4th generation inbred cohort that originated from a wild caught female collected in Skåne, Sweden, using a standard salt extraction[1].

Illumina genome sequencing. Illumina sequencing was used for all data generation used in genome construction. One paired end (PE) and the two mate pair (MP) libraries were constructed at Science for Life Laboratory, the National Genomics Infrastructure, Sweden (SciLifeLab), using 1 PCR-free PE DNA library (180bp) and 2 Nextera MP libraries (3kb and 7kb) all from a single individual. All sequencing was done on Illumina HiSeq 2500 High Output mode, PE 2x100bp by SciLifeLab. An additional two 40kb MP fosmid jumping libraries were constructed from a sibling used in the previous library construction. Genomic DNA, isolated as above, was shipped to Lucigen Co. (Middleton, WI, USA) for the fosmid jumping library construction and sequencing was performed on an Illumina MiSeq using 2x250bp reads [2]. Finally, a variable insert size library of 100 bp – 100,000 bp in length were generated using the Chicago and HiRise method[3]. Genomic DNA was again isolated from a sibling of those used in previous library construction. The genomic DNA was isolated as above and shipped to Dovetail Co. (Santa Cruz, CA, USA) for library construction, sequencing and scaffolding. These library fragments were sequenced by Centrillion Biosciences Inc. (Palo Alto, CA, USA) using Illumina HiSeq 2500 High Output mode, PE 2x100bp.

Data Preparation and Genome assembly. Nearly 500 M read pairs of data were generated, providing ~ 285 X genomic coverage (Supplemental Table 1). The 3kb and 7kb MP pair libraries were filtered for high confidence true mate pairs using Nextclip v0.8[4]. All read sets were then quality filtered, the ends trimmed of adapters and low quality bases, and screened of common contaminants using bbduk v37.51[5]. Insert size distributions were plotted to assess library quality, which was high (Supplementary Fig. 1). The 180bp, 3kb, and 7kb, read data sets were used as input for AllpathsLG r50960[6] for initial contig generation and scaffolding (Supplementary Note 1). AllpathsLG was run with haploidify = true

to compensate for the high degree of heterozygosity. A further round of superscaffolding using the 40kb library alongside the 3kb and 7kb libraries was done using SSPACE v2[7]. Finally, the assembly was ultascaffolded using the Chicago read libraries and the HiRise software pipeline. These steps produced a final assembly of 3005 scaffolds with an N50-length of 4.2 Mb and a total length of 350 Mb (Supplementary Note 1). The final assembly's complete and single copy ortholog (SCO) content was 94% for *P. napi* as assessed by BUSCO v3.0.2[8] (for more details, see Supplementary Note 1).

Linkage Map. RAD-seq data of 5463 SNP markers from 275 full-sib individuals, without parents, was used as input into Lep-MAP2[9]. The RAD-seq data was generated from next-RAD technology by SNPsaurus (Oregon, USA)(Supplemental note 10). To obtain genotype data, the RAD-seq data was mapped to the reference genome using BWA mem[10] and SAMtools[11] was used to produce sorted bam files of the read mappings. Based on read coverage (samtools depth), Z chromosomal regions were identified from the genome and the sex of offspring was determined. Custom scripts[12] were used to produce genotype likelihoods (called posteriors in Lep-MAP) from the output of SAMtools mpileup.

The parental genotypes were inferred with Lep-MAP2 ParentCall module using parameters "ZLimit=2 and ignoreParentOrder=1", first calling Z markers and second calling the parental genotypes by ignoring which way the parents are informative (the parents were not genotyped so we could not separate maternal and paternal markers at this stage). Scripts provided with Lep-MAP2 were used to produce linkage file from the output of ParentCall and all single parent informative markers were converted to paternally informative markers by swapping parents, when necessary. Filtering by segregation distortion was performed using Filtering module.

Following this, the SepareteChromosomes module was run on the linkage file and 25 chromosomes were identified using LOD score limit 39. Then JoinSingles module was run twice to add more markers on the chromosomes with LOD score limit of 20. Then SepareteChromosomes was run again but only on markers informative on single parent with LOD limit 10 to separate paternally and maternally

informative markers. 51 linkage groups were found and all were ordered using OrderMarkers module. Based on likelihood improvement of marker ordering, paternal and maternal linkage groups were determined. This was possible as there is no recombination in females (achiasmatic meiosis), and thus the order of the markers does not improve likelihood on the female map. The markers on the corresponding maternal linkage groups were converted to maternally informative and OrderMarkers was run on the resulting data twice for each of 25 chromosomes (without allowing recombination in female). The final marker order was obtained as the order with the higher likelihood from the two runs.

Chromosomal assembly. The 5463 markers that composed the linkage map were mapped to the *P. napi* ultrascaffolds using bbmap[5] with sensitivity = slow. Reads that mapped uniquely were used to identify misassemblies in the Ultrascaffolds and subsequently rearrange those fragments into the correct chromosomal order. 54 misassemblies were identified and a total of 115 fragments were joined together into 25 chromosomes using a series of custom R scripts (supplemental information) and the R package Biostrings[13]. Scaffold joins and misassembly corrections were validated by comparing the number of correctly mapped mate pairs spanning a join between two scaffolds. Mate pair reads from the 3kb, 7kb, and 40kb libraries were mapped to their respective assemblies with bbmap (po=t, ambig=toss, kbp=t). SAM output was filtered for quality (mapq > 20 and properly paired) and a custom script was used to tabulate read spanning counts for each base pair in the assembly.

Assembly and annotation of the *P. napi* mitochondrial genome. Contigs containing mitochondrial genome sequence were identified by a BLASTN search using published cytochrome oxidase I (COI) sequences against the *P. napi* assembled genome. The identified contigs were then imported into Geneious (version 5.6.6.) and assembled to form the whole mitochondrial genome. Sequencing and assembly errors were manually investigated and corrected by mapping sequencing reads to the assembled mitochondrial genome using CLC Genomics Workbench v.4.

In order to annotate the mtDNA, the protein coding genes (PCGs) were first predicted using Genewise[14] and then the Open Reading Frames (ORFs) were manually checked through alignment to the mitochondrial genome (NCBI Genbank acc. HM156697) of the closely related butterfly, *Pieris rapae*. Available *P. napi* partial mitochondrial sequences were also aligned to aid annotation of both protein coding and rRNA genes (Genbank accessions: AF170861; AM236011; GQ148917; DQ150035; DQ150071; LC090589). tRNA features were initially identified using tRNAscan-SE[15] and manually checked through alignment with the *P. rapae* mitochondrial genome.

The assembled mitochondrial genome of *P. napi* is 14945bp in length with a total AT content of 79.9%. These values fall well within the range of previously sequenced Lepidopteran mitogenomes. The mtDNA consists of the 37 genes typical of animals, 13 of which are protein coding, 22 are transfer RNAs (tRNAs), and the remaining two are ribosomal RNAs (rRNAs). In addition there is a non-coding AT-rich control region. The sequence of the control region typically does not assemble well and may therefore be incomplete. Few rearrangements have been observed among arthropod mitochondrial DNA, and usually these consist of translocations of tRNAs[16]. Here, when compared to three butterflies (*Pieris rapae* HM156697; *Pieris melete* NC_010568; *Melitaea cinxia* HM243592) and one moth (*Bombyx mori* AY048187), perfect synteny in gene order and orientation within the mitogenome is observed. For *P. napi* the start codon of 12/13 PCGs is the typical ATN (ATT: NAD2, ATP8, NAD3, NAD5, NAD6; ATG: COII, ATP6, COIII, NAD4, NAD4L, CYTB or ATA: NAD1). However the putative start codon for COI is CGA, which appears to be common across insects. The stop codons of *P. napi* PCGs are either the common TAA (NAD2, ATP8, ATP6, COIII, NAD5, NAD4, NAD4L, NAD6, CYTB, NAD1), TAG (NAD3), or a single T as an incomplete stop codon, which has been found in several other Lepidopteran mitochondrial genes (e.g. *Melitaea cinxia* HM243592).

Synteny Comparisons Between *P. napi, B. mori, M. cinxia,* and *H. melpomene*. A list of 3100 SCOs occurring in the Lepidoptera lineage curated by OrthoDB v9.1[17] was used to extract gene names and protein sequences of SCOs in *Bombyx mori* from KaikoBase[18] (Supplemental Note 5) using a custom script. Reciprocal best hits (RBH) between gene sets of *P. napi, B. mori, M. cinxia,* and *H. melpomene* SCOs were identified using BLASTP[19] and custom scripts. Gene sets of *H. melpomene* v2.5 and *M. cinxia* v1 were downloaded from LepBase v4 [20]. Coordinates from Blast tables were converted to chromosomal locations and visualized using Circos[21] and custom R scripts.

Synteny Comparison Within Lepidoptera. Genome assemblies and annotated protein sets were downloaded for 24 species of Lepidoptera from LepBase v4 [22] and other sources (Supplemental Table 4). Each target species protein set was aligned to its species genome as well as to the *Pieris napi* protein set using Diamond v0.9.10[23] with default options. The protein-genome comparison was used to assign each target species gene to one of it's assembled scaffolds, while the protein-protein comparison was used to identify RBHs between the protein of each species and its ortholog in *P. napi*, and *B. mori*. Using this information we used a custom R script to examine each assembly scaffold for evidence of synteny to either *P. napi* or *B. mori*. First, each scaffold of the target species genome was assigned genes based on the protein-genome blast results, using its own protein set and genome. A gene was assigned to a scaffold if at least 3 HSPs of less than 200bp from a gene aligned with >= 95% identity. Second, if any of these scaffolds then contained 5 genes whose orthologs resided on a single *B. mori* chromosome but two *P. napi* chromosomes, and those same two *P. napi* chromosome segments were also joined in the *B. mori* assembly, that was counted as a 'mori-like scaffold'. Conversely if a target species scaffold contained 5 genes whose orthologs resided on a single *P. napi* chromosome but two *B. mori* chromosomes, and those same two *B. mori* chromosome segments were also joined in the *P. napi* assembly, that was counted as a 'napi-like scaffold'.

Pieridae chromosomal evolution.

Chromosomal fusions and fissions were reconstructed across the family Pieridae by placing previously published karyotype studies of haploid chromosomal counts into their evolutionary context. There are approximately 1000 species in the 85 recognized genera of Pieridae and we recently reconstructed a robust fossil-calibrated chronogram for this family at the genus level[24,25]. We then placed the published chromosomal counts for 201 species[26,27] on this time calibrated phylogeny, using the maximal reported hapoid chromosomal count per species when more than one was recorded, with ancestral chromosomal reconstructions for chromosome count, treated as a continuous character, using the contMap function of the phytools R package[28].

Lepidopteran chronogram showing relationships among species with genomes.

The topology of the phylogenetic hypothesis is based on the consensus of relationships within Lepidoptera summarized by Mitter et al 2017[29], with the relationships between families being largely derived from Kawahara and Breinholt 2014[30], and within butterflies from Heikkilä et al 2012[31].

Second Linkage Map for *P. napi*. A second linkage map was constructed from a different family of *P. napi* in which a female from Abisko, Sweden was crossed with a male from Catalonia, Spain. Genomic DNA libraries were constructed for the mother, father, and four offspring (2 males, 2 females). RNA libraries were constructed for an additional 6 female and 6 male offspring. All sequencing was performed on a Illumina HiSeq 2500 platform using High Output mode, with PE 2x100bp reads at SciLifeLab (Stockholm, Sweden). Both DNA and RNA reads were mapped to the genome assembly with bbmap using default settings. Samtools was used to sort read mappings and merge them into an mpileup file (Supplemental Note 6). Variants were called with BCFtools[32] and filtered with VCFtools[33]. Linkage between SNPs was assessed with PLINK[34]. A custom script was used to assess marker density and determine sex-specific heterozygosity.

Annotation of *P. napi* genome. Genome annotation was carried out by the Bioinformatics Short-term Support and Infrastructure (BILS, Sweden). BILS was provided with the chromosomal assembly of *P.*

napi and 45 RNAseq read sets representing 3 different tissues (head, fat body, and gut) of 7 male and 8 female larva from lab lines were separate from the one used for the initial sequencing. Sequence evidence for the annotation was collected in two complementary ways. First, we queried the Uniprot database[35] for protein sequences belonging to the taxonomic group of Papilionoidea (2,516 proteins). In order to be included, proteins gathered in this way had to be supported on the level of either proteomics or transcriptomics and could not be fragments. In addition, we downloaded the Uniprot-Swissprot reference data set (downloaded on 2014-05-15) (545,388 proteins) for a wider taxonomic coverage with high-confidence proteins. In addition, 493 proteins were used that derived from a *P. rapae* expressed sequence tag library that was Sanger sequenced.

Permutation test of collinear block position within chromosomes. Collinear blocks (CBs) were identified as interior vs terminal and the ends of terminal blocks were marked as inward or outward facing (i.e. teleomere facing). CBs were reshuffled into 25 random chromosomes of 4 CBs in a random orientation and the number of times that a terminal block occurred in a random chromosome with the outward end facing outward was counted. This was repeated 10,000 times to generate a random distribution expectation. The number of terminal outward-facing CBs in *B. mori* that were also terminal and outward facing in *P. napi* was compared to this random distribution to derive the significance of our observation. To test the randomness of gene location within chromosomes, the previously identified SCOs were numbered by their position along each chromosome in both *B. mori* and *P. napi*. We then generated 10,000 random genomes as above. Distance from the end of the new chromosome and distance from the end of *B. mori* chromosome were calculated for each ortholog and the results were binned. P-values were determined by comparing the number of orthologs in a bin to the expected distribution of genes in a bin from the random genomes. All test were done using a custom R script.

Gene set enrichment analysis of collinear blocks. Gene ontology set enrichment was initially tested within collinear blocks of the *P. napi* genome using topGO[36] with all 13,622 gene models generated

from the annotation. For each collinear block within the genome, each GO term of any level within the hierarchy that had at least 3 genes belonging to it was analyzed for enrichment. If a GO term was overrepresented in a collinear block compared to the rest of the genome at a p-value of < 0.01 by a Fisher exact test, that block was counted as enriched. 57 of the 99 collinear blocks in the *P. napi* genome were enriched in this way. Because arbitrarily breaking up a genome and testing for GO enrichment can yield results that are dependent on the distribution of the sizes used, we compared the results of the previous analysis to the enrichment found using the same size genomic regions, randomly drawn from the *P. napi* genome without replacement. The size distribution of the 99 collinear blocks were used to generate fragment sizes into which the genome was randomly assigned. This resulted in a random genome of 99 fragments which in total contained the entire genome, but the content of a given fragment was a random genomic region.. This random genome was tested for GO enrichment of the fragments in the same way as the collinear blocks in the original genome, and the number of enriched blocks counted. This was then repeated 10,000 times to generate a distribution of expected enrichment in genome fragments of the same size as the *P. napi* collinear blocks.

Data availablility

Illumina reads were submitted to the European Nucleotide Archive (ENA) under accession PRJEB24862. The assembly and annotation are available at Lepbase http://ensembl.lepbase.org/Pieris_napi_pnv1x1/. Pierid phylogenetic tree is under accession xxx at Dryad.

Bibliography

1. Aljanabi, S. M. & Martinez, I. Universal and rapid salt-extraction of hight quality genomic DNA for PCR-based techniques. *Nucleic Acids Res.* **25,** 4692–4693 (1997).

2. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456,** 53–59 (2008).

3. Putnam, N. H. *et al.* Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.* **26,** 342–350 (2016).

4. Leggett, R. M., Clavijo, B. J., Clissold, L., Clark, M. D. & Caccamo, M. NextClip: an analysis and read preparation tool for Nextera Long Mate Pair libraries. *Bioinformatics* **30,** 566–8 (2014).

5. Bushnell, B. BBTools. (2017). at <https://jgi.doe.gov/data-and-tools/bbtools/>

6. Gnerre, S. *et al.* High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci.* **108,** 1513–1518 (2011).

7. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27,** 578–579 (2011).

8. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31,** 3210–3212 (2015).

9. Rastas, P., Calboli, F. C. F., Guo, B., Shikano, T. & Merilä, J. Construction of Ultradense Linkage Maps with Lep-MAP2: Stickleback F2 Recombinant Crosses as an Example. *Genome Biol. Evol.* **8,** 78–93 (2015).

10. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. **0,** 1–3 (2013).

11. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

12. Kvist, J. *et al.* Flight-induced changes in gene expression in the Glanville fritillary butterfly. *Mol. Ecol.* **24,** 4886–4900 (2015).

13. Pages H, Gentleman R, Aboyoun P, et al. Biostrings: String objects representing biological sequences, and matching algorithms. *R Packag. version* **2,** 2008 (2008).

14. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14,** 988–995 (2004).

15. Lowe, T. M. & Eddy, S. R. TRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25,** 955–964 (1996).

16. Boore, J. L., Macey, J. R. & Medina, M. Sequencing and comparing whole mitochondrial genomes of animals. *Methods Enzymol.* **395,** 311–348 (2005).

17. Zdobnov, E. M. *et al.* OrthoDB v9.1: Cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Res.* **45,** D744–D749 (2017).

18. Shimomura, M. *et al.* KAIKObase: an integrated silkworm genome database and data mining tool. *BMC Genomics* **10,** 486 (2009).

19. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST:a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).

20. Kersey, P. J. *et al.* Ensembl Genomes 2016: More genomes, more complexity. *Nucleic Acids Res.* **44,** D574–D580 (2016).

21. Krzywinski, M. *et al.* Circos: An information aesthetic for comparative genomics. *Genome Res.* **19,** 1639–1645 (2009).

22. Challis, R. J., Kumar, S., Dasmahapatra, K. K., Jiggins, C. D. & Blaxter, M. Lepbase: The Lepidopteran genome database. *bioRxiv* 56994 (2016). doi:10.1101/056994

23. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12,** 59–60 (2014).

24. Wahlberg, N., Rota, J., Braby, M. F., Pierce, N. E. & Wheat, C. W. Revised systematics and higher classification of pierid butterflies (Lepidoptera: Pieridae) based on molecular data. *Zool. Scr.* **43,** 641–650 (2014).

25. Edger, P. P. *et al.* The butterfly plant arms-race escalated by gene and genome duplications. *Proc. Natl. Acad. Sci.* **112,** 8362–8366 (2015).

26. ROBINSON, R. *Lepidoptera Genetics*. *Lepidoptera Genetics* (1971). doi:10.1016/B978-0-08-006659-2.50017-1

27. Lukhtanov, V. A. Karyotype evolution and systematics of higher taxa of Pieridae (Lepidoptera) of the World. *Ent. Obozr. 70 619-636, 3 figs* (1991).

28. Revell, L. J. phytools: An R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* **3,** 217–223 (2012).

29. Mitter, C., Davis, D. R. & Cummings, M. P. Phylogeny and Evolution of Lepidoptera. *Annu. Rev. Entomol.* **62,** 265–283 (2017).

30. Kawahara, A. Y. & Breinholt, J. W. Phylogenomics provides strong evidence for relationships of butterflies and moths. *Proc. R. Soc. B Biol. Sci.* **281,** 20140970–20140970 (2014).

31. Heikkila, M., Kaila, L., Mutanen, M., Pena, C. & Wahlberg, N. Cretaceous origin and repeated tertiary diversification of the redefined butterflies. *Proc. R. Soc. B Biol. Sci.* **279,** 1093–1099 (2012).

32. Narasimhan, V. *et al.* BCFtools/RoH: A hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics* **32,** 1749–1751 (2016).

33. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27,** 2156–2158 (2011).

34. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).

35. The UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.* **43,** D204-12 (2015).

36. Alexa A and Rahnenfuhrer J. topGO: Enrichment Analysis for Gene Ontology. *R package version 2.26.0.* (2016). at <http://bioconductor.org/packages/release/bioc/html/topGO.html>