A butterfly chromonome reveals selection dynamics during extensive and cryptic chromosomal reshuffling

Authors

Jason Hill[1,Ɨ], Ramprasad Neethiraj[1], Pasi Rastas[2], Nathan Clark[3], Nathan Morehouse[4], Maria de la Paz Celorio-Mancera[1], Jofre Carnicer Cols[5,6], Heinrich Dircksen[10], Camille Meslin[3], Kristin Sikkink[7], Maria Vives[5,6], Heiko Vogel[9], Christer Wiklund[1], Carol L. Boggs[8], Sören Nylin[1], Christopher Wheat[1,Ɨ]

Affiliations:

1 Population Genetics, Department of Zoology, Stockholm University, Stockholm, Sweden

2 Institute of Biotechnology, (DNA Sequencing and Genomics), University of Helsinki, Helsinki, Finland

3 Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, USA

4 Department of Biological Sciences, University of Cincinnati, Cincinnati, USA

5 Department of Evolutionary Biology, Ecology and Environmental Sciences, University of Barcelona, 08028 Barcelona, Spain

6 CREAF, Global Ecology Unit, Autonomous University of Barcelona, 08193 Cerdanyola del Vallès, Spain

7 Department of Ecology, Evolution and Behavior, University of Minnesota, St Paul, MN, USA

8 Department of Biological SciencesUniversity of South Carolina, Columbia, SC, USA

9 Department of Entomology, Max Planck Institute for Chemical Ecology, D-07745 Jena, Germany

10 Functional Morphology, Department of Zoology, Stockholm University, Stockholm, Sweden

[Ɨ] Authors for correspondence:

Jason Hill <jason.hill@zoologi.su.se>

Christopher Wheat <chris.wheat@zoologi.su.se>

Abstract

Taxonomic orders vary in their degree of chromosomal conservation with some having high rates of chromosome number turnover despite maintaining some core sets of ordered genes (e.g. Mammalia) and others exhibiting rapid rates of gene-order reshuffling without changing chromosomal count (e.g. Diptera). However few clades exhibit as much conservation as the Lepidoptera for which both chromosomal count and gene colinearity (synteny) are very high over the past 140 MY. In contrast, here we report extensive chromosomal rearrangements in the genome of the green-veined white butterfly (*Pieris napi*, Pieridae, Linnaeus, 1758). This unprecedented reshuffling is cryptic: microsynteny and chromosome number do not indicate the extensive rearrangement revealed by a chromosome level assembly and high-resolution linkage map. Furthermore, the rearrangement blocks themselves appear to be non-random, as they are significantly enriched for clustered groups of functionally annotated genes revealing that the evolutionary dynamics acting on Lepidopteran genome structure are more complex than previously envisioned.

Introduction

The role of chromosomal rearrangements in adaptation and speciation has long been appreciated and recent work has elevated the profile of supergenes in controlling complex adaptive phenotypes[1–4]. Chromosome number variation has also been cataloged for many species but analyses of the adaptive implications have mostly been confined to the consequences of polyploidy and whole genome duplication[5,6]. The identification of pervasive fission and fusion events throughout the genome is relatively unexplored since discovery of this pattern requires chromosome level assemblies. This leaves open the possibility of cryptic chromosomal dynamics taking place in many species for which this level of genome assembly has not been achieved. As chromosomal levels assemblies become more common, uncovering a relationship between such dynamics and adaptation or speciation can be assessed.

Here we focus upon the Lepidoptera, the second most diverse animal group with over 160,000 extant species in more than 160 families. Butterflies and moths exist in nearly all habitats and have equally varied life histories yet show striking similarity in genome architecture, with the vast majority having a haploid chromosome number of n=31[7–9]. While haploid chromosome number can vary from n = 5 to n = 223[10–12], gene order and content is remarkably similar within chromosomes (i.e. displays macrosynteny), regardless of haploid chromosome number. The degree of such synteny between species separated by up to 140 My is astounding as illustrated by recent chromosomal level genomic assemblies[7,13], as well as previous studies of the sequence and structure of lepidopteran genomes[14–17]. This ability of Lepidoptera to accommodate such chromosomal rearrangements, yet maintain high levels of macro and microsynteny (i.e. collinearity at the scale of 10s to 100’s of genes) is surprising. While a growing body of evidence indicates that gene order in eukaryotes is non-random along chromosomes, with upwards of 12% of genes organized into functional neighborhoods of shared function and expression patterns[18,19], to what extent this may play a role in the chromosomal evolution is an open question.

Variation in patterns of synteny across clades must arise due to an evolutionary interaction between selection and constraint[20], likely at the level of telomere and centromere performance. *Drosophila*, and likely all Diptera, differ from other eukaryotes studied to date in lacking the telomerase enzyme, and instead protect their chromosomal ends using retrotransposons[21]. This absence of telomerase is posited to make evolving novel telomeric ends more challenging, limiting the appearance of novel chromosomes and thereby resulting in high macrosynteny via constraint[22].

In contrast, Lepidoptera like most Metazoans use telomerase to protect their chromosomal ends which allows for previously internal chromosomal DNA to become subtelomeric in novel chromosomes[7,13]. Additionally all Lepidoptera have holocentric chromosomes in which the decentralized kinetochore allows for more rearrangements by fission, fusion, and translocation of chromosome fragments than is the case for monocentric chromosomes[23]. Thus, Lepidoptera should be able to avoid the deleterious consequences of large-scale chromosomal changes.

Here we present the chromosome level genome assembly of the green-veined white butterfly *P. napi*. Our analysis reveals large-scale fission and fusion events similar to known dynamics in other lepidopteran species but at an accelerated rate and without a change in haploid chromosome count. The resulting genome-wide breakdown of the chromosome level synteny is unique among Lepidoptera. While we are unable to identify any repeat elements associated with this cryptic reshuffling, we find the chromosomal ends reused and the collinearity of functionally related genes. These findings support a reinterpretation of the chromosomal fission dynamics in the Lepidoptera.

Results

The *P. napi* genome was generated using DNA from inbred siblings from Sweden, a genome assembly using variable fragment size libraries (180 bp to 100 kb; N50-length of 4.2 Mb and a total length of 350 Mb), and a high density linkage map across 275 full-sib larva, which placed 122 scaffolds into 25 linkage groups, consistent with previous karyotyping of *P. napi*[24,25]. After assessment and correction of the assembly, the total chromosome level assembly was 299 Mb comprising 85% of the total assembly size and 114% of the k-mer estimated haploid genome size, with 2943 scaffolds left unplaced (**Supplementary Note 3**). Subsequent annotation predicted 13,622 gene models, 9,346 with functional predictions (**Supplementary Note 4**).

Single copy orthologs (SCOs) in common between *P. napi* and the first sequenced Lepidopteran genome, the silk moth *Bombyx mori* (Bombycidae), were identified. These revealed an unexpected deviation in gene order and chromosomal structure in *P. napi* relative to *B. mori* as well as another lepidopteran genome with a linkage map and known chromosomal structure, that of

Heliconius melpomene (Nymphalidae) (Fig 1a). Large-scale rearrangements that appeared to be the fission and subsequent fusion of fragments on the scale of megabases were present on every *P. napi* chromosome relative to *B. mori*, *H. melpomene*, and *Melitaea cinxia* (Nymphalidae) (fig 1b). We characterized the size and number of large scale rearrangements between *P. napi* and *B. mori* using shared SCOs to identify 99 clearly defined blocks of co-linear gene order (hereafter referred to as "syntenic blocks"), with each syntenic block having an average of 69 SCOs. Each *P. napi* chromosome contained an average of 3.96 (SD = 1.67) syntenic blocks, which derived on average from 3.5 different *B. mori* chromosomes. In *P. napi*, the average syntenic block length was 2.82 Mb (SD = 1.97 Mb) and contained 264 genes (SD = 219).

The indication that *P. napi* diverged radically from the thus far observed chromosomal structure of lepidopterans raised questions about how frequently a *P. napi* like chromosomal structure is observed *vs.* the structure reported in the highly syntenic *B. mori*, *H. melpomene*, and *M. cinxia* genomes. We accessed 22 publicly available lepidopteran genome assemblies and their gene annotations representing species that diverged up to 140 MYA in order to identify the genes corresponding to the SCO's used in the previous analyses. We used blastx (Diamond v0.9.10)[26] to place those genes on their native species scaffolds. With informations about each SCO's location on the *P. napi* and *B. mori* chromosomes we recorded how often a scaffold contained a cluster of genes whose orthologs resided on two *P. napi* chromosomes or two *B. mori* chromosomes. If two *P. napi* chromosomes were represented by only a single *B. mori* chromosome, then the scaffold was marked as containing an mori-like join. Conversely if two *B. mori* chromosomes were represented but only a single *P. napi* chromosome, then the scaffold was marked as containing a napi-like join. In total we found that 20 species have more mori-like joins, and two species of *Pieris* represented by 3 assemblies have more napi-like joins (Fig 2a). While this type of assessment is preliminary the indication is that the genome structure described here is novel to the genus *Pieris*.

We validated this novel chromosomal reorganization using four complementary but independent approaches to assess our scaffold joins. First, we generated a second linkage map for *P.*

napi, which confirmed the 25 linkage groups and the ordering of scaffold joins along chromosomes (Fig. 3; Supplementary Fig. 2). Second, the depth of the mate-pair (MP) reads spanning joins indicated by the first linkage map provides an independent assessment of the join validity. We therefore quantified MP reads spanning each base pair position along a chromosome (Fig. 3; Supplementary Fig. 2, Note 7), finding strong support for the scaffold joins. Third, we aligned the scaffolds of a recently constructed genome of *P. rapae*[27] to *P. napi*, looking for *P. rapae* scaffolds that spanned the chromosomal level scaffold joins within *P. napi*, finding support for 71 of the 97 joins (Supplementary Fig. 5). Fourth, we considered *B. mori* syntenic blocks that spanned a scaffold join within a *P. napi* chromosome as support for that *P. napi* chromosome assembly, and found that 62 of the 97 scaffold joins were supported by *B. mori* (Supplementary Fig. 2, Note 8,9).

To assess the novel chromosomal organization, we investigated the ordering and content of these syntenic blocks in *P. napi*. First, we tested whether telomeric ends of chromosomes were at all conserved between species despite the extensive chromosomal reshuffling (Fig. 4a). We found significantly more syntenic blocks sharing telomere facing orientations between species than expected ($P < 0.01$, two tailed t-test; Fig. 4b). We also identified a significant enrichment for SCOs in *B. mori* and *P. napi* located at roughly similar distance from the end of their respective chromosomes (Fig. 4c). Both of these findings are consistent with the ongoing use of telomeric ends, indicating that strong selection dynamics have favored their retention over evolutionary time. Second, we tested for gene set functional enrichment within the observed syntenic blocks by investigating the full set of annotated *P. napi* genes. We found that 57 of the 99 block regions in the *P. napi* genome contained at least three genes with a shared gene ontology (GO) term that was significantly less frequent in the rest of the genome ($P < 0.01$, fisher) (Supplementary fig. 3). We then tested whether the observed enrichment in the syntenic blocks of *P. napi* was greater than expected by randomly assigning the genome into similarly sized blocks. The mean number of GO enriched fragments in each of the simulated 10,000 genomes was 38.8 (variance of 46.6 and maximum of 52), which was significantly lower than the observed ($P < 0.0001$).

To assess the possible cause of the reshuffling, we surveyed the distribution of different repeat element classes across the genome, looking for enrichment of specific categories near the borders of syntenic blocks. While Class 1 transposons were found to be at higher density at near the ends of chromosomes relative to the distribution internally (Supplementary fig. 4), no repeat elements were enriched relative to the position of syntenic block regions. We therefore investigated whether any repeat element classes had expanded within *Pieris* compared to other sequenced genomes by assessing the distribution of repeat element classes and genome size among sequenced Lepidoptera genomes. In accordance with other taxa[28] we find an expected strong relationship between genome size and repetitive element content in *Pieris* species. Thus, while repetitive elements such as transposable elements are likely to have been involved in the reshuffling, our inability to find clear elements involved suggests these events may be old and their signal decayed.

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[29].

Illumina genome sequencing. Illumina sequencing was used for all data generation used in genome construction. A 180 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 [30]. Finally, a variable insert size library of 100 bp – 100,000 bp in length were generated using the Chicago and HiRise method[31]. 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[32]. 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 (bbtools, Brian Bushnell). 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 with AllpathsLG r50960[33] for initial contig generation and scaffolding (Supplementary Note 1). AllpathsLG was run with haploidify = true to compensate for the high degree of heterozygosity. The initial contig assembly's conserved single copy ortholog content was assessed at 78% for *P. napi* by CEGMA v2.5[34]. A further round of superscaffolding using the 40kb libraries alongside the 3kb and 7kb libraries was done using SSPACE v2[35]. Finally, both assemblies were 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).

Linkage Map. RAD-seq data of 5463 SNP markers from 275 full-sib individuals, without parents, was used as input into Lep-MAP2[36]. 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[37] and SAMtools[38] 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[39] 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[40] with sensitivity = slow. Reads that mapped uniquely were used to identify misassemblies in the Ultrascaffolds and arrange those fragments into chromosomal order. 54 misassemblies were identified and overall 115 fragments were joined together into 25 chromosomes using a series of custom R scripts (supplemental information) and the R package Biostrings[41]. 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 and a custom script was used to tabulate read spanning counts for each base pair in the assembly.

Synteny Comparisons Between *P. napi*, *B. mori*, and *H. melpomene*. A list of 3100 single copy orthologs (SCO) occurring in the Lepidoptera lineage curated by OrthoDB v9.1[42] was used to extract gene names and protein sequences of SCOs in *Bombyx mori* from KaikoBase[43] (Supplemental Note 5) using a custom script. Reciprocal best hits (RBH) between gene sets of *P. napi*, *P. rapae*, *H. melpomene*, *M. cinxia*, and *B. mori* SCOs were identified using BLASTP[44] and custom scripts. Gene sets of *H. melpomene* v2.5 and *M. cinxia* v1 were downloaded from LepBase v4 [45]. Coordinates were converted to chromosomal locations and visualized using Circos[46] and custom R scripts.

Synteny Comparison Within Lepidoptera. Genome assemblies and annotated protein sets were downloaded for 24 species of Lepidoptera from LepBase v4 [47] 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[26] 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 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 join’. Conversely if a target species scaffold contained 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 join'.

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[48,49]. We then placed the published chromosomal counts for 201 species[9,50] on this time calibrated phylogeny with ancestral chromosomal reconstructions for chromosome count, treated as a continuous character, using the contMap function of the phytools R package[51.]

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. Samtools was used to sort read mappings and merge them into an mpileup file (Supplemental Note 6). Variants were called with BCFtools[52] and filtered with VCFtools[53]. Linkage between SNPs was assessed with PLINK[54]. 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[55] 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 syntenic block position within chromosomes. Syntenic blocks (SBs) were identified as interior vs terminal and the ends of terminal blocks were marked as inward or outward facing. SBs were reshuffled into 25 random chromosomes of 4 SBs 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 SBs in *B. mori* that were also terminal and outward facing in *P. napi* was compared to this random distribution to derive the significance of deviation from the expected value. To test the randomness of gene location within chromosomes, orthologs were numbered by their position along each chromosome in both *B. mori* and *P. napi*. 10,000 random genomes were generated 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 syntenic blocks. Gene ontology set enrichment was initially tested within syntenic blocks of the *P. napi* genome using topGO[56] with all 13,622 gene models generated from the annotation. For each syntenic 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 syntenic 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 syntenic 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 selected from the *P. napi* genomes. The size distribution of the 99 syntenic 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 random compared to the syntenic block that defined its size. This random genome was tested for GO enrichment of the fragments in the same way as the syntenic 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* syntenic blocks.

Discussion

While massive chromosomal fission events are well documented in butterflies (e.g. *Leptidea* in Pieridae (n=28-103); *Agrodiaetus* in Lycaenidae (n=10-134)), their contribution to Lepidopteran diversity appears to be minimal as all these clades are very young[57–59]. However, our results challenge this interpretation. Rather, *P. napi* appears to represent a lineage that has undergone an impressive reconciliation of an earlier series of rampant fission events. Moreover, the subsequent fusion events exhibit a clear bias toward using ancient telomeric ends, as well as returning gene clusters to their relative ancestral position within chromosomes even when the other parts of the newly formed chromosome originated from other sources. Luckily these initial fission events have been frozen in time as reshuffled syntenic blocks, revealing the potential fitness advantage of maintaining certain functional categories as syntenic blocks.

Thus, despite the potential for holocentric species to have relaxed constraint upon their chromosomal evolution, we find evidence for selection actively maintaining ancient telomeric ends, as well as gene order within large chromosomal segments. Together these observations suggest that the low chromosome divergence in Lepidoptera over > 100 million generations is at least partially due to purifying selection maintaining an adaptive chromosomal structure.

Bibliography

1. Schwander, T., Libbrecht, R. & Keller, L. Supergenes and complex phenotypes. *Curr. Biol.* **24,** R288–R294 (2014).
2. Kunte, K. *et al.* Doublesex Is a Mimicry Supergene. *Nature* **507,** 229–232 (2014).
3. Fishman, L., Stathos, A., Beardsley, P. M., Williams, C. F. & Hill, J. P. Chromosomal rearrangements and the genetics of reproductive barriers in mimulus (monkey flowers). *Evolution (N. Y).* **67,** 2547–2560 (2013).
4. Lamichhaney, S. *et al.* Structural genomic changes underlie alternative reproductive strategies in the ruff (Philomachus pugnax). *Nat. Genet.* **48,** 84–88 (2015).
5. Otto, S. P. & Whitton, J. Polyploid Incidence and Evolution. *Annu. Rev. Genet.* **34,** 401–437 (2000).
6. Van de Peer, Y., Mizrachi, E. & Marchal, K. The evolutionary significance of polyploidy. *Nat. Rev. Genet.* **18,** 411–424 (2017).
7. Ahola, V. *et al.* The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera. *Nat Commun* **5,** 1–9 (2014).
8. Lukhtanov, V. A. Sex chromatin and sex chromosome systems in nonditrysian Lepidoptera (Insecta). *J. Zool. Syst. Evol. Res.* **38,** 73–79 (2000).
9. ROBINSON, R. *Lepidoptera Genetics*. *Lepidoptera Genetics* (1971). doi:10.1016/B978-0-08-006659-2.50017-1
10. Brown, K. S., Von Schoultz, B. & Suomalainen, E. Chromosome evolution in Neotropical Danainae and Ithomiinae (Lepidoptera). *Hereditas* **141,** 216–236 (2004).
11. Kandul, N. P., Lukhtanov, V. A. & Pierce, N. E. Karyotypic diversity and speciation in Agrodiaetus butterflies. *Evolution (N. Y).* **61,** 546–559 (2007).
12. Saura, A., Schoultz, B. Von, Saura, A. O. & Brown, K. S. Chromosome evolution in Neotropical butterflies. *Hereditas* **150,** 26–37 (2013).

13. Davey, J. W. *et al.* Major Improvements to the Heliconius melpomene Genome Assembly Used to Confirm 10 Chromosome Fusion Events in 6 Million Years of Butterfly Evolution. *Genes|Genomes|Genetics* **6,** 695–708 (2016).

14. Yasukochi, Y. A Second-Generation Integrated Map of the Silkworm Reveals Synteny and Conserved Gene Order Between Lepidopteran Insects. *Genetics* **173,** 1319–1328 (2006).

15. Yasukochi, Y. *et al.* A FISH-based chromosome map for the European corn borer yields insights into ancient chromosomal fusions in the silkworm. *Heredity (Edinb).* **116,** 75–83 (2016).

16. Beldade, P., Saenko, S. V, Pul, N. & Long, A. D. A gene-based linkage map for Bicyclus anynana butterflies allows for a comprehensive analysis of synteny with the lepidopteran reference genome. *PLoS Genet.* **5,** e1000366 (2009).

17. Šíchová, J. *et al.* Fissions, fusions, and translocations shaped the karyotype and multiple sex chromosome constitution of the northeast-Asian wood white butterfly, Leptidea amurensis. *Biol. J. Linn. Soc.* **118,** 457–471 (2016).

18. Al-Shahrour, F. *et al.* Selection upon genome architecture: Conservation of functional neighborhoods with changing genes. *PLoS Comput. Biol.* **6,** (2010).

19. Gordon, J. L., Byrne, K. P. & Wolfe, K. H. Mechanisms of chromosome number evolution in yeast. *PLoS Genet.* **7,** 0–3 (2011).

20. Coghlan, A., Eichler, E. E., Oliver, S. G., Paterson, A. H. & Stein, L. Chromosome evolution in eukaryotes: A multi-kingdom perspective. *Trends Genet.* **21,** 673–682 (2005).

21. Levis, R. W., Ganesan, R., Houtchens, K., Tolar, L. A. & Sheen, F. miin. Transposons in place of telomeric repeats at a Drosophila telomere. *Cell* **75,** 1083–1093 (1993).

22. Muller, H. J. The remaking of chromosomes. *Collect. net* **13,** 181–198 (1938).

23. Maddox, P. S., Oegema, K., Desai, A. & Cheeseman, I. M. 'Holo'er than thou: Chromosome segregation and kinetochore function in C. elegans. *Chromosom. Res.* **12,** 641–653 (2004).

24. Lorkovic, Z. The genetics and reproductive isolating mechanisms of the Pieris napi - bryoniae group. *J. Lepid. Soc.* **16,** 5–19 (1955).

25. Maeki, K. & Kawazoe, A. On the Hybridization between Two Karyotype Lineages of Pieris napi Linnaeus from Japan. *Cytologia (Tokyo).* **59,** (1994).

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

27. Nallu, S. *et al.* The Molecular Genetic Basis of Herbivory between Butterflies and their Host-Plants. *bioRxiv* (2017). doi:10.1101/154799

28. Clark, A. G. *et al.* Evolution of genes and genomes on the Drosophila phylogeny. *Nature* **450,** 203–218 (2007).

29. 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).

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

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

32. 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).

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

34. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23,** 1061–1067 (2007).

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

36. 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).

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

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

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

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

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

42. 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).

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

44. 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).

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

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

47. 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

48. 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).

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

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

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

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

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

54. 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).

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

56. 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>

57. Dincă, V., Lukhtanov, V. A., Talavera, G. & Vila, R. Unexpected layers of cryptic diversity in wood white Leptidea butterflies. *Nat. Commun.* **2,** (2011).

58. Vila, R. *et al.* Phylogeny and palaeoecology of Polyommatus blue butterflies show Beringia was a climate-regulated gateway to the New World. *Proc. R. Soc. B Biol. Sci.* **278,** 2737–2744 (2011).

59. Lukhtanov, V. The blue butterfly Polyommatus (Plebicula) atlanticus (Lepidoptera, Lycaenidae) holds the record of the highest number of chromosomes in the non-polyploid eukaryotic organisms. *Comp. Cytogenet.* **9,** 683–690 (2015).

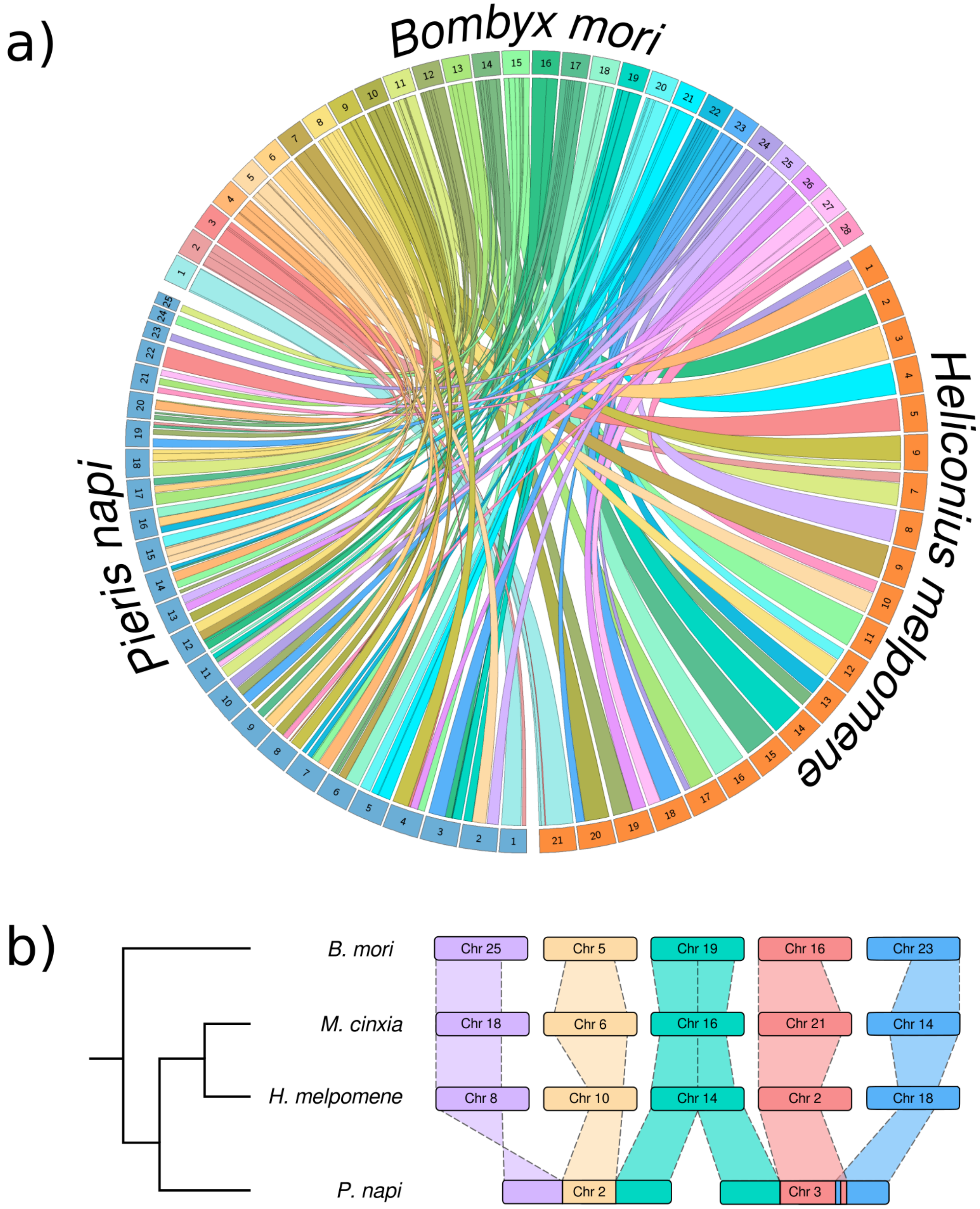
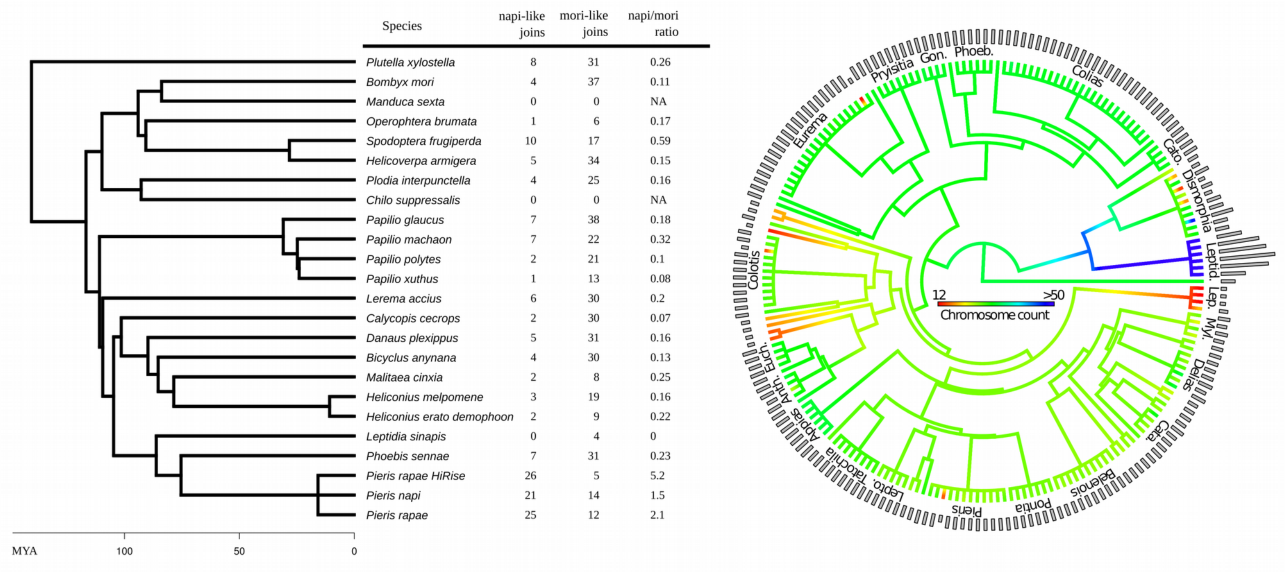


Figure 1 a) Chromosomal mapping between the moth *Bombyx mori* (Bombycoidea) and the butterflies *Pieris napi* (Pieridae) and *Heliconius melpomene* (Nymphalidae). These species last shared a common ancestor > 100 million generations ago[49]. Depicted are the reciprocal best hit orthologs identified between *B. mori* and *P. napi* (n=2354) and between *B. mori* and *H. melpomene* (n=2771). Chromosome 1 is the Z chromosome in *B. mori* and *P. napi* and 21 is the Z chromosome in *H. melpomene*. Chromosomes 2-25 in *P. napi* are ordered in size from largest to smallest. Links between orthologs originate from the *B. mori* chromosome and are colored by their chromosome of origin, while *P. napi* chromosomes are colored blue and *H. melpomene* chromosomes are colored orange. Links are clustered into blocks of synteny and each ribbon represents a contiguous block of genes spanning a region in both species. **b)** Two largest autosomes of *P. napi* and their synteny to other Lepidoptera and their phylogenetic relationship. The sister taxa and the more distant *B. mori* share a high degree of macro synteny while the *P. napi* genome required multiple chromosomal fusion and fission events to be patterned in the way that is observed. Band width for each species is proportional to the length of the inferred chromosomal region of orthology, although the individual chromosomes are not to scale.



Species	napi-like joins	mori-like joins	napi/mori ratio
Plutella xylostella	8	31	0.26
Bombyx mori	4	37	0.11
Manduca sexta	0	0	NA
Operophtera brumata	1	6	0.17
Spodoptera frugiperda	10	17	0.59
Helicoverpa armigera	5	34	0.15
Plodia interpunctella	4	25	0.16
Chilo suppressalis	0	0	NA
Papilio glaucus	7	38	0.18
Papilio machaon	7	22	0.32
Papilio polytes	2	21	0.1
Papilio xuthus	1	13	0.08
Lerema accius	6	30	0.2
Calycopis cecrops	2	30	0.07
Danaus plexippus	5	31	0.16
Bicyclus anynana	4	30	0.13
Malitaea cinxia	2	8	0.25
Heliconius melpomene	3	19	0.16
Heliconius erato demophoon	2	9	0.22
Leptidia sinapis	0	4	0
Phoebis sennae	7	31	0.23
Pieris rapae HiRise	26	5	5.2
Pieris napi	21	14	1.5
Pieris rapae	25	12	2.1

Figure 2 a) A time calibrated phylogeny of currently available Lepidopteran genomes (n=24) and estimates of their macrosynteny with *B. mori* and *P. napi*, with time in million years ago (MYA). Macrosynteny was estimated by quantifying the number of times a scaffold of a given species contained *B. mori* orthologs from two separate chromosomes and *P. napi* orthologs from a single chromosome (napi-like join), or vice versa (mori-like joins)(see Supplemental Note for more details). **b)** A time calibrated ancestral state reconstruction of the chromosomal fusion and fission events across Pieridae (n=201 species). As only a time calibrated genus level phylogeny exists for Pieridae, all genera with > 1 species are set to an arbitrary polytomy at 5 MYA, while deeper branches reflect fossil calibrated nodes. The haploid chromosomal count of tips (histogram) and interior branches (color coding) are indicated, with the outgroup set to n=31 reflecting the butterfly chromosomal mode. Genus names are indicated for the larger clades (all tips labels in Supplemental Material).

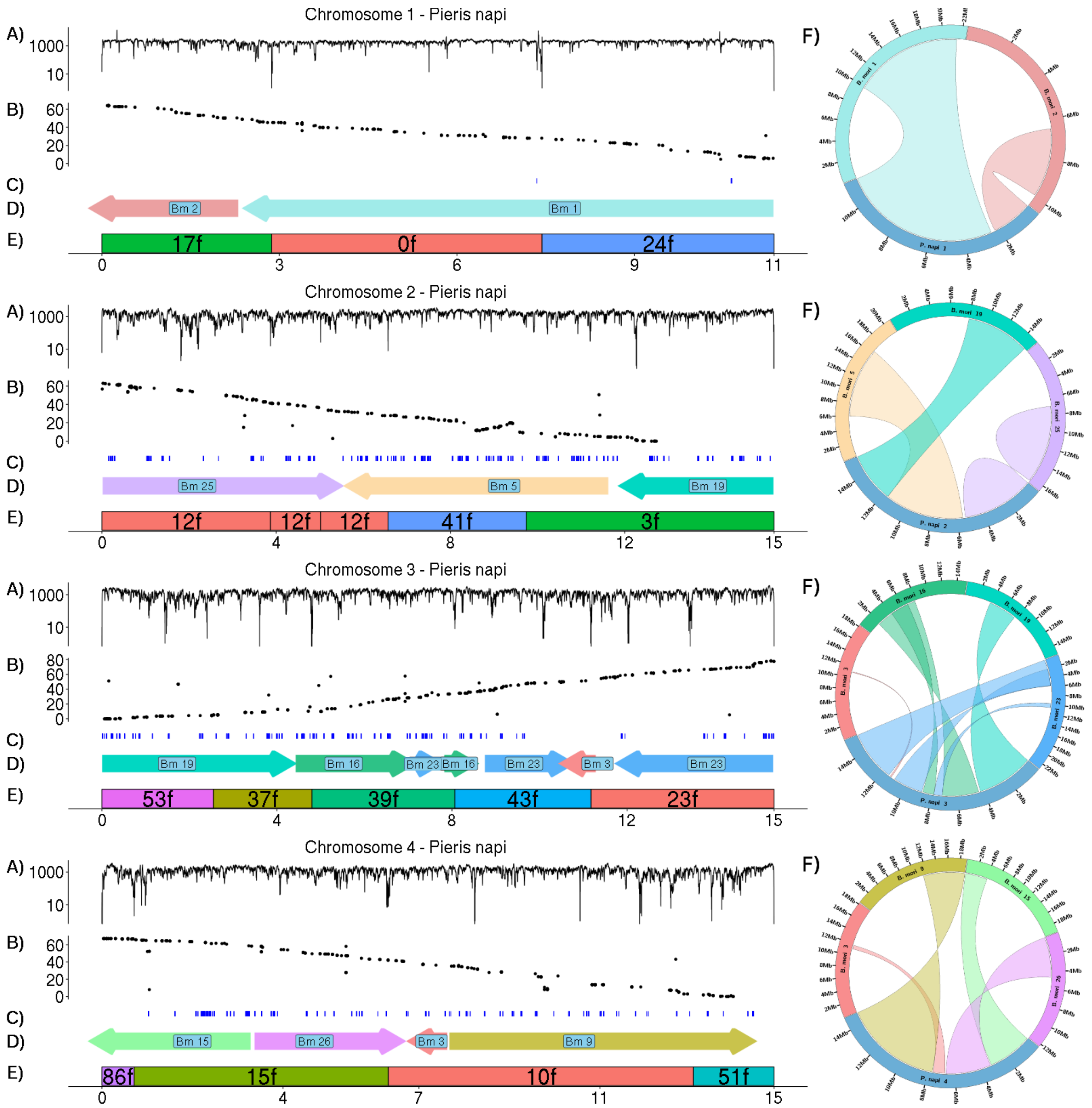


Figure 3 Validation of syntenic relationship between *B. mori* and first four *P. napi* chromosomes. (a) Mate pair spanning depth across each chromosome summed for the 3kb, 7kb, and 40kb libraries. Spanning depths averaged 1356 across the whole genome. Of the scaffold join positions 74 of 97 were spanned by > 50 properly paired reads (mean = 117.8, S.D. = 298.7) which we considered good evidence for correct assembly at scaffold boundaries while the remaining 23 scaffold joins had 0 mate pair spans. (b) RAD-seq linkage markers and recombination distance along chromosomes from the first linkage map that was used for genome assembly. (c) Results from the second linkage map of maternally inherited markers, using RNA-Seq and whole genome sequencing. All markers within a chromosome are completely linked due to suppressed recombination in females (i.e. recombination distance is not shown on Y axis). (d) Syntenic block origin and orientation colored and labeled by the *B. mori* chromosome containing the orthologs, as in Fig. 1 (e) Component scaffolds of each chromosome labeled to indicate scaffold number and orientation. (f) To the right of each P. napi chromosome is a circos plot showing the location and orientation of syntenic blocks within each *B. mori* chromosome that comprise a given *P. napi* chromosome. Ribbons representing the blocks of synteny are colored by their orthologs location in the *B. mori* genome. Relative orientation of a block is shown by whether the ribbon contains a twist. Remaining chromosomes shown in Supplementary Fig. 2.

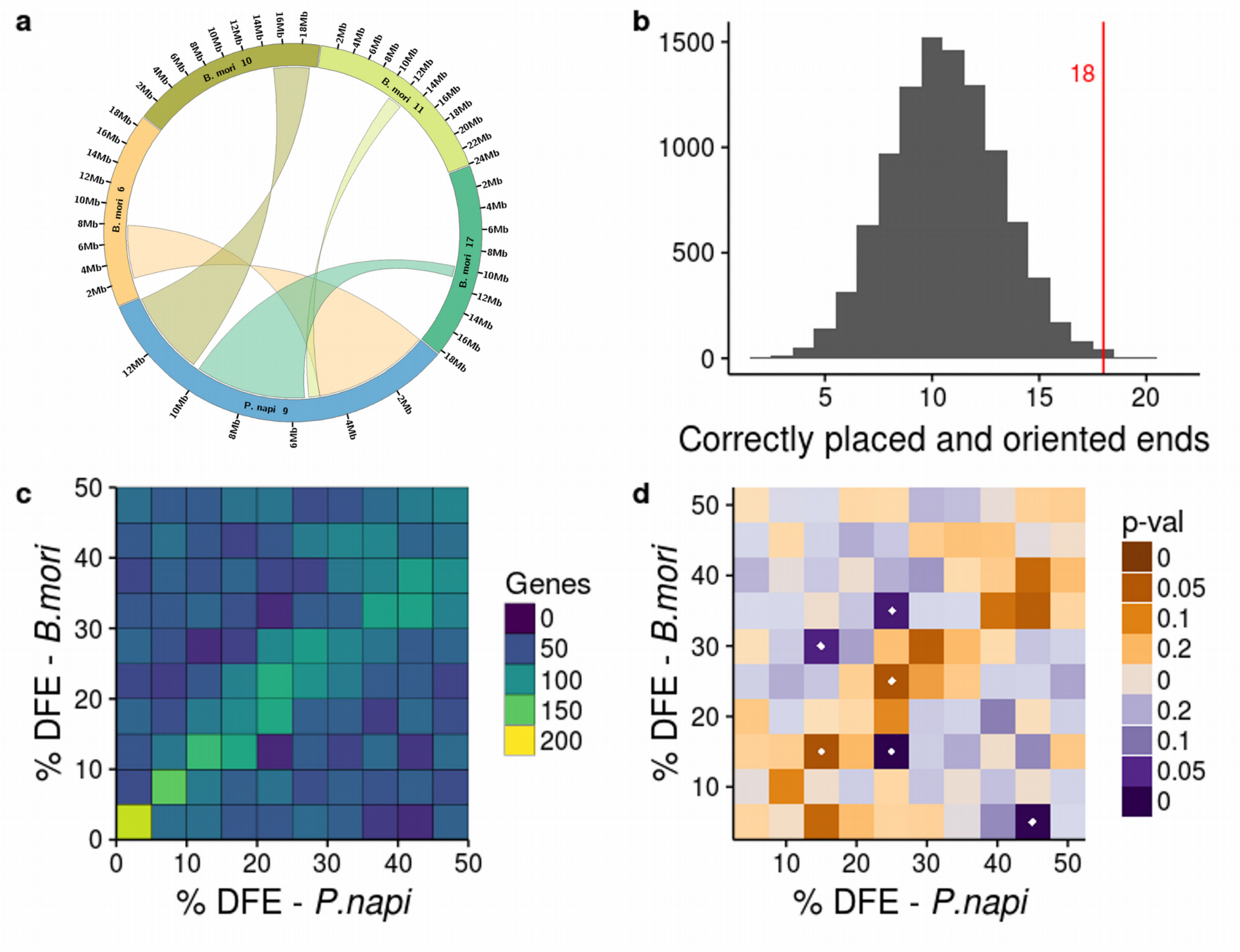


Figure 4. Comparison of gene content of and chromosomal location of syntenic blocks between *Pieris napi* and *Bombyx mori* in observed and randomly generated expectation genomes. (a) Observed pattern of conserved syntenic block location within *P. napi* Chromosome 9, wherein telomere facing and interior syntenic blocks are conserved between species despite shuffling. (b) Histogram of the number of syntenic blocks that are terminal on the *B. mori* genome and also occur in the terminal position on chromosomes in a simulated genome, from 10,000 simulated genomes (average 10.7, std dev= 6.8). (c) Percentage distance from the end (DFE) of a chromosome of a single copy gene in *P. napi vs.* DFE of that gene's single copy ortholog (SCO) in *B. mori*. Counts binned on the color axis. (d) Comparison between the observed DFE distribution and the expected distribution generated from 10,000 genomes of 25 chromosomes constructed from the random fusion of syntenic blocks. Bins in which more genes occur in the observed genomes than the expected distribution are in orange, less genes in blue, P < 0.05 in either direction are denoted by a white dot. SCO spatial distribution was significantly higher than expected along the diagonal (two bins with p < 0.05), while significantly lower than expected off the diagonal (four bins with p < 0.05).

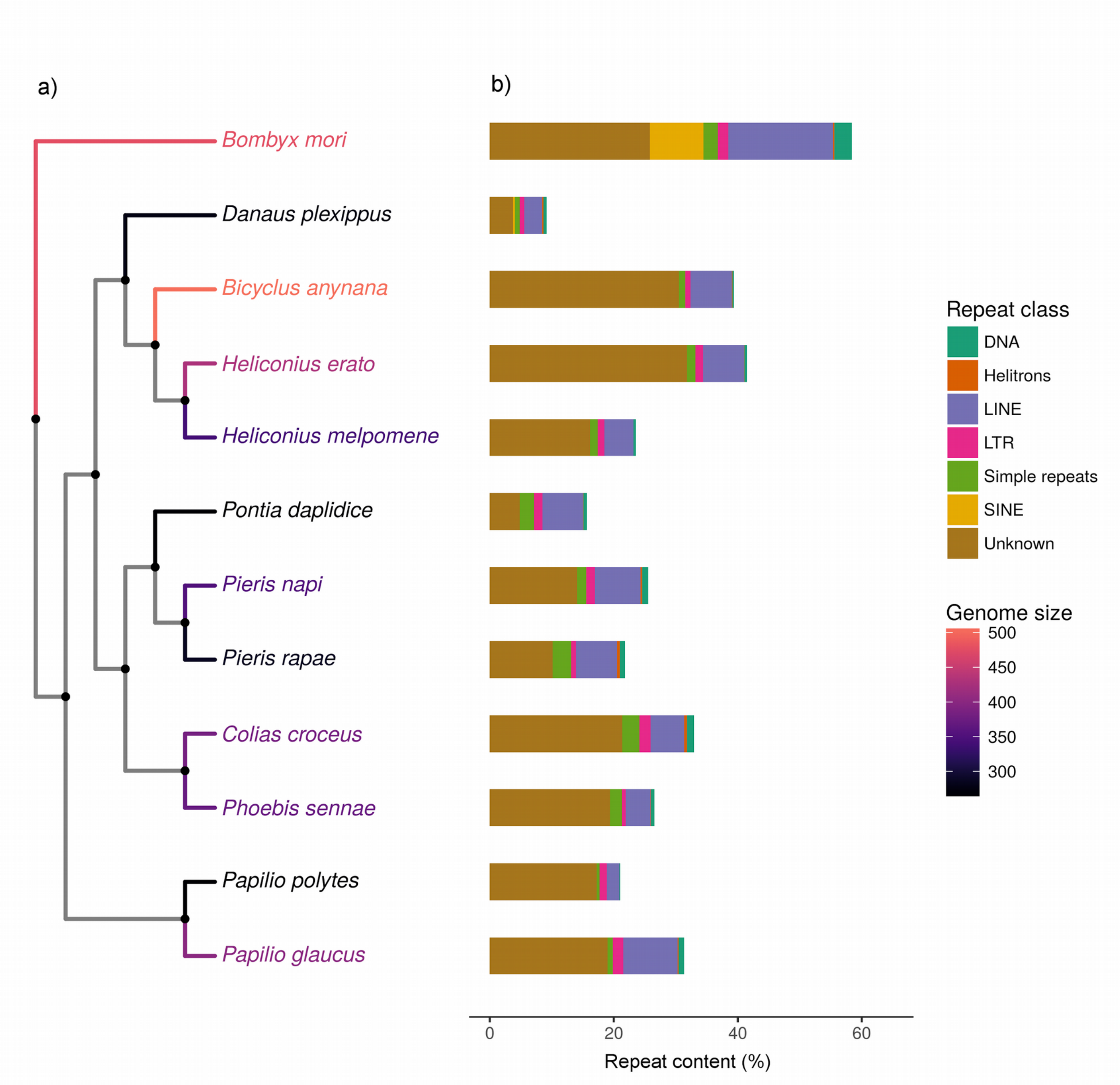


Figure 5. The genomic size and repeat content of Lepidopteran genomes placed in a phylogenetic context. (a) Phylogenetic relationships represented as a cladogram, with terminal branches and species names colored by genome size estimates from k-mer distributions of read data. (b) The fraction of repeat content of each genome, color coded by repeat class.