PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Research

Cite this article: Kim K-W, De-Kayne R, Gordon IJ, Omufwoko KS, Martins DJ, ffrench-Constant R, Martin SH. 2022 Stepwise evolution of a butterfly supergene via duplication and inversion. *Phil. Trans. R. Soc. B* **377**: 20210207. https://doi.org/10.1098/rstb.2021.0207

Received: 27 November 2021
Accepted: 28 March 2022

One contribution of 15 to a theme issue 'Genomic architecture of supergenes: causes and evolutionary consequences'.

Subject Areas:
evolution, genomics

Keywords:
recombination suppression, structural variation, genome assembly, *Danaus chrysippus*, chromosomal rearrangement, repeats

Author for correspondence:
Simon H. Martin
e-mail: Simon.Martin@ed.ac.uk

Stepwise evolution of a butterfly supergene via duplication and inversion

Kang-Wook Kim[1,†], Rishi De-Kayne[1,†], Ian J. Gordon[2], Kennedy Saitoti Omufwoko[3], Dino J. Martins[3,4], Richard ffrench-Constant[5] and Simon H. Martin[1]

[1]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK
[2]Centre of Excellence in Biodiversity and Natural Resource Management, University of Rwanda, Huye Campus, Huye, Rwanda
[3]Department of Ecology and Evolutionary Biology, Princeton University, Princeton, USA
[4]Mpala Research Centre, Nanyuki, Kenya
[5]Centre for Ecology and Conservation, University of Exeter, Penryn Campus, Penryn, UK

K-WK, 0000-0003-1489-8264; RD-K, 0000-0001-5569-8061; KSO, 0000-0002-0321-5117; DJM, 0000-0003-1581-1642; Rf-C, 0000-0001-5385-9888; SHM, 0000-0002-0747-7456

Supergenes maintain adaptive clusters of alleles in the face of genetic mixing. Although usually attributed to inversions, supergenes can be complex, and reconstructing the precise processes that led to recombination suppression and their timing is challenging. We investigated the origin of the BC supergene, which controls variation in warning coloration in the African monarch butterfly, *Danaus chrysippus*. By generating chromosome-scale assemblies for all three alleles, we identified multiple structural differences. Most strikingly, we find that a region of more than 1 million bp underwent several segmental duplications at least 7.5 Ma. The resulting duplicated fragments appear to have triggered four inversions in surrounding parts of the chromosome, resulting in stepwise growth of the region of suppressed recombination. Phylogenies for the inversions are incongruent with the species tree and suggest that structural polymorphisms have persisted for at least 4.1 Myr. In addition to the role of duplications in triggering inversions, our results suggest a previously undescribed mechanism of recombination suppression through independent losses of divergent duplicated tracts. Overall, our findings add support for a stepwise model of supergene evolution involving a variety of structural changes.

This article is part of the theme issue 'Genomic architecture of supergenes: causes and evolutionary consequences'.

1. Introduction

Supergenes are clusters of loci at which adaptive combinations of alleles are co-inherited. This can facilitate the maintenance of complex phenotypes under balancing selection [1,2], or link locally adapted alleles in the face of migration [3] (Schaal *et al.* [4]). Supergenes are commonly assumed to be associated with a chromosomal inversion polymorphism [5], but detailed reconstruction of some supergenes has revealed more complex architectures, such as multiple adjacent inversions that arose in a stepwise manner [6], similar to the progressive spread of recombination suppression in sex chromosome evolution [7]. Unlike a single inversion, supergenes maintained by more complex mechanisms could protect more than two divergent alleles from recombination. Other forms of genomic rearrangement that disturb local synteny may also contribute to recombination suppression [8,9], but there are few well-characterized cases, and in some cases, the precise mechanism of recombination suppression is unknown [10] (Komata *et al.* [11]). One effective approach to explore the structural changes involved in supergene evolution and their relative timing is to compare whole-chromosomal assemblies for each of the supergene alleles [12].

[†]These authors contributed equally to this study.

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.5980556.

THE ROYAL SOCIETY
PUBLISHING

We investigated the evolution of a colour patterning supergene in *Danaus chrysippus*, a dispersive butterfly known as the African monarch, African Queen or plain tiger in different parts of its extensive range. It has bright warning coloration indicating its toxicity [13] and is divided into distinct geographical morphs with partly overlapping ranges [14]. Despite its vast range, genetic differentiation is almost non-existent between distant populations: F_{ST} is approximately zero across the genome, with the exception of a few broad peaks, including two that are associated with colour differences [15]. This implies selection for the maintenance of local morphs, possibly driven by learned predator avoidance favouring the most common local morph [16]. Colour variation in the forewing is controlled by the BC supergene on chromosome 15 (chr15) [15,17]. This broad region of suppressed recombination includes two colour patterning loci: B controls brown versus orange variation and localizes at the melanin pathway gene *yellow*, and C controls the presence/absence of a black forewing tip and white band. Three divergent BC alleles, which determine three distinct forewing phenotypes, have been described to date [15]. A previous genome assembly of one of the alleles (namely 'klugii') revealed that several genes on chr15 had increased in copy number in *D. chrysippus* relative to the outgroup *Danaus plexippus* (the monarch), suggesting that gene duplications may have been involved in the evolution of the BC supergene [15].

To reconstruct the steps in the evolution of the BC supergene, we generated haplotype-resolved chromosomal assemblies for the three known divergent alleles using long-read sequencing and trio-binning. We traced the evolution of the various structural changes across the phylogeny to determine the sequence of events. Our findings show that the supergene has evolved through multiple steps over the course of several million years and was preceded by extensive segmental duplications, which probably triggered recombination suppression.

2. Results

(a) Haploid assemblies of the three BC alleles

We generated four new haploid genome assemblies from two F_1 individuals by binning reads according to their parent of origin [18] (see the electronic supplementary material, table S1 for accession numbers and tables S2–S4 for assembly statistics). In all four assemblies, chr15 is represented by 1–4 contigs that could be ordered and oriented by eye (electronic supplementary material, figure S1). To identify the supergene allele represented by each assembly, we used an 'ancestry painting' approach based on sequence similarity to representative individuals of known genotype [15]. This confirmed that we had assembled all three alleles, as follows: MB18102MAT = 'chrysippus', MB18102PAT = 'klugii' and SB211PAT = 'orientis' (electronic supplementary material, figure S2). Unexpectedly, the fourth assembly (SB211MAT) showed mixed similarity to all three alleles, suggesting that it may be a rare recombinant that has not yet been observed in the homozygous state (electronic supplementary material, figures S2 and S3).

Using the same ancestry painting procedure, we also confirmed that a previous assembly (Dchry2.2) [19], represents a 'pure' klugii allele (electronic supplementary material, figure S2), despite having been assembled from diploid reads from a suspected heterozygous individual. We therefore examined the discarded haplotypic contigs from the Dchry2 assembly (Dchry2HAP) and identified the other chr15 haplotype, which represents the chrysippus allele (electronic supplementary material, figure S2). Therefore, in total, we have generated six assemblies of chr15: two representing the klugii allele, two representing the chrysippus allele, one representing the orientis allele and one representing a putative recombinant. The two independent assemblies of the klugii and chrysippus alleles allowed us to check for assembly errors. In both cases, the independent assemblies of the same chr15 allele were co-linear with no obvious errors detected (electronic supplementary material, figure S4).

(b) Multiple structural rearrangements in the evolution of the BC supergene

We predicted that each of the three BC alleles has undergone distinct structural changes resulting in recombination suppression. We therefore first analysed the alignments between each allele and the outgroup *D. plexippus* assembly, assumed to be structurally unaltered from the ancestral state. As predicted, each allele shows multiple structural differences from *D. plexippus* (figure 1; electronic supplementary material, figure S1). Most of these differences are unique to one of the three BC alleles, implying that they are derived changes. The structural events involve four adjacent chromosomal regions. The most strikingly rearranged is region 3 (1.3 megabases (Mb)), which has been duplicated and fragmented multiple times in all three resulting alleles. Many of the duplicated fragments have been inverted and translocated. These multiple duplications result in chr15 being considerably longer in *D. chrysippus* than in *D. plexippus*, with striking variation in length also seen among the BC supergene alleles owing to each carrying a different complement of duplicated fragments of region 3 (figure 1). Region 3 shows elevated levels of transposable elements (TEs), specifically LINEs, in both the *D. plexippus* and *D. chrysippus* (electronic supplementary material, figure S5), suggesting that TE activity may have triggered the segmental duplications.

Although duplications can cause assembly errors if their sequences are too similar, this issue does not appear to have affected our assemblies. Independent assemblies of the same allele show strong similarity (electronic supplementary material, figure S4) and read depth analyses show no evidence of large collapsed repeats (electronic supplementary material, figure S6). This suggests that the duplications and repeats are distinct enough to assemble separately, which is also confirmed by the lack of highly similar repeats showing up in whole-genome alignments (electronic supplementary material, figure S4).

The other three regions—1 (2 Mb), 2 (1.5 Mb) and 4 (0.4 Mb)—are each inverted in the orientis allele (figure 1). A portion of region 1 is also inverted in the chrysippus allele, but appears to share the same right-hand inversion breakpoint. In both the orientis and chrysippus alleles, region 2 is shifted to the right owing to upstream insertion of duplicated fragments of region 3. In all of these simpler rearrangements, either one or both edges are bordered by a duplicated fragment of region 3. We therefore hypothesized that the duplications of region 3 may have occurred first and triggered subsequent inversions.

(c) Multiple duplications deep in the history of the *Danaus* genus

To test the hypothesis that multiple duplications of region 3 initiated the formation of the supergene, we estimated the copy number of each gene on chr15 in six other *Danaus*

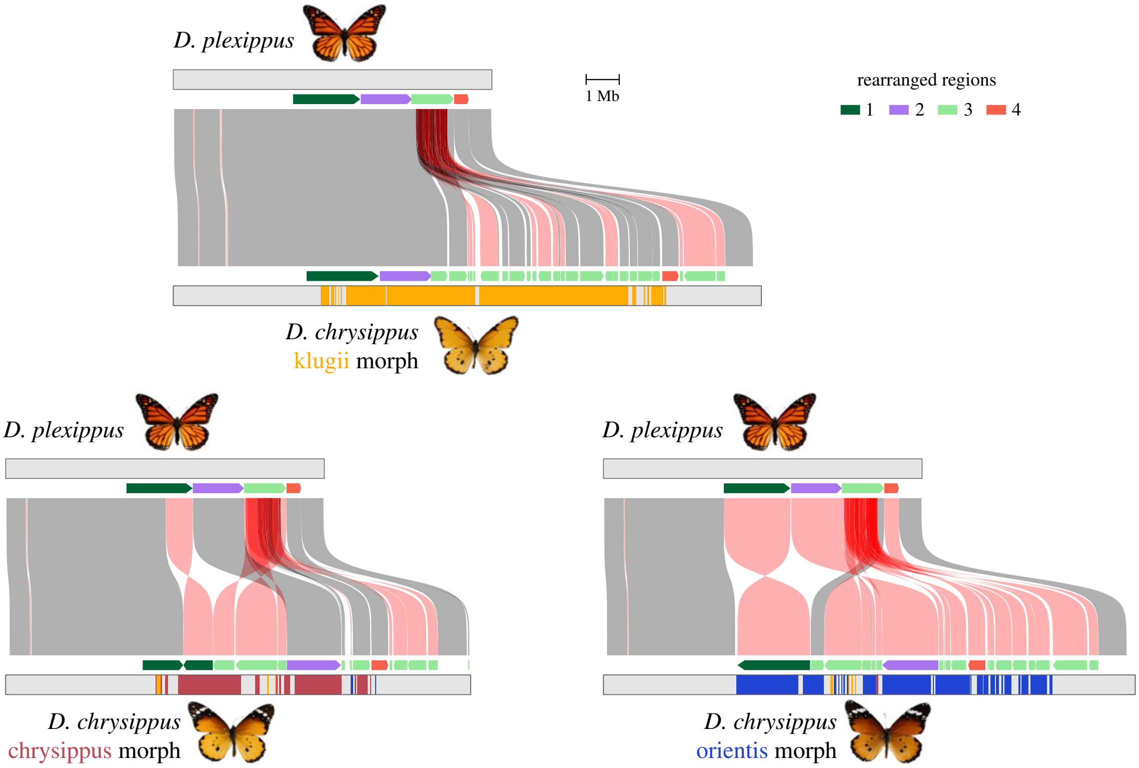

Figure 1. Structural changes in *D. chrysippus* supergene alleles relative to the outgroup *D. plexippus*. Connecting boxes indicate syntenic blocks (see Methods for details) between the outgroup (*D. plexippus*, MEX_DaPlex assembly) and each of the three BC supergene alleles in *D. chrysippus*. Blocks that are syntenic but in the reverse orientation are coloured red. Partial transparency is used to reveal duplicated syntenic blocks. Coloured arrows indicate the four regions that we identified as having experienced distinct rearrangements in *D. chrysippus*, with the direction of the arrow indicating their relative orientation. The coloured bars indicate 'ancestry painting' in 50 kb windows along the *D. chrysippus* chromosomes (see also the electronic supplementary material, figure S2). Regions lacking coloured bars could not be assigned ancestry because they showed similar sequence similarity to two or more different morphs. *D. plexippus* image attribution: Muséum de Toulouse, CC BY-SA 4.0, via Wikimedia Commons.

species and an outgroup using short-read data (see the electronic supplementary material, table S1 for accession numbers). Most of the 476 genes for which we had sufficient data to quantify copy number were present as a single copy in all species. However, a cluster of 10 genes in region 3 shows high copy numbers in all three *D. chrysippus* morphs as well as three additional species: *Danaus petilia* (Australia), *Danaus gilippus* (Americas) and *Danaus melanippus* (Asia) (electronic supplementary material, figure S7 and table S5). Our inferred species tree, based on 5954 gene alignments confirms that the multiple duplications of region 3 originated at the base of subgenus *anosia* (figure 2*b*). The tree also resolves previous uncertainty regarding the placement of *Danaus eresimus*, which groups with *D. plexippus* and *Danaus erippus*, all of which lack the multiple duplications of region 3 genes. Four of the duplicated genes are homologous to hemicentin and nephrin (electronic supplementary material, table S5), members of the immunoglobulin family, but to our knowledge their functions in lepidoptera remain unknown.

We estimated a lower bound for the timing of origin of the copy number variation at the base of the *anosia* clade. d_a (absolute sequence divergence adjusted for estimated divergence in the ancestral species) at putatively neutral fourfold degenerate sites between *D. chrysippus* and *D. melanippus* is 8.5% (95% confidence interval (CI) = 8.4–8.6). Although we cannot convert this directly to a split time in years as the mutation rates and generation times for this clade are unknown, we can calibrate our estimate according to the estimated split between *Danaus* and *Tirumala* of 12.7 Ma [20]. d_a at fourfold degenerate sites between *D. chrysippus* and *T. formosa* is 14.5% (95% CI = 14.3, 14.6). This gives an estimate of 7.5 Ma as a lower bound for the origin of the region 3 duplications.

There is consistent variation in estimated copy number for each region 3 gene among species (figure 2; electronic supplementary material, table S5), with the highest copy numbers observed in the *D. chrysippus* morphs. This is most parsimoniously explained by further duplications in the *D. chrysippus* lineage, but there is also evidence for loss of copies (see below). Further work will be required to determine whether all of the duplicated copies are expressed and have phenotypic consequences.

(d) Structural polymorphisms persist through two speciation events

We next investigated the history of the three simpler rearrangements on chr15 by generating maximum-likelihood trees from concatenated coding sequences within each of the three regions involved (38, 18, 61 and 15 genes for regions 1.1, 1.2, 2 and 4, respectively; electronic supplementary material, figure S8). The trees for all of the inversions are incongruent with the species branching order (figure 2*c–e*), implying that the inversion polymorphisms are ancient and have persisted through speciation events. In region 1, the first event appears to have been an inversion of part of the region in the ancestor of the chrysippus allele, followed by a separate inversion of the whole of region 1 in the ancestor of the orientis allele (figure 2*c*). Both *D. gilippus* and *D. petilia* are most closely related to the chrysippus allele. This may reflect a scenario in which both

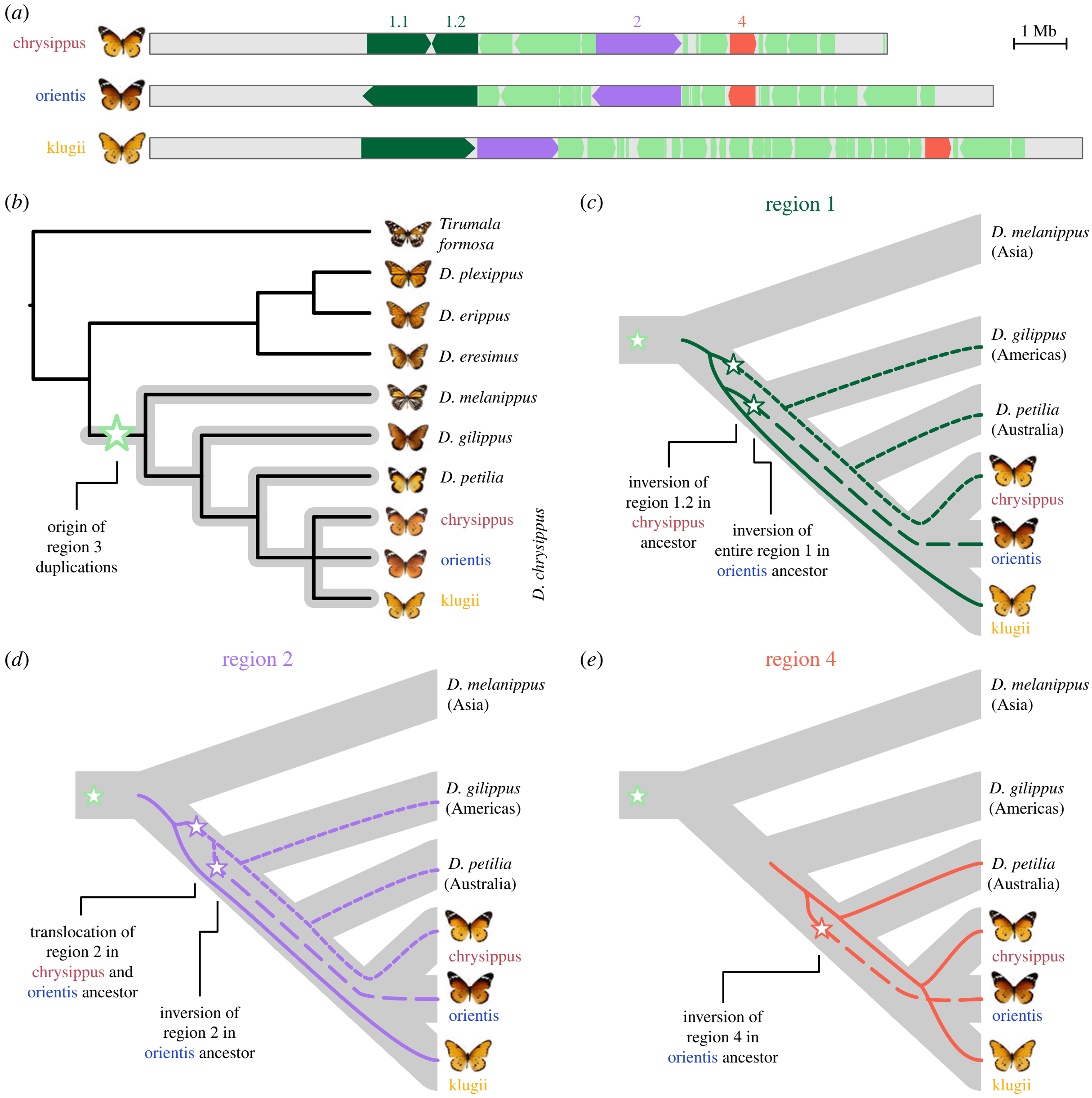

Figure 2. Tracing structural changes across the *Danaus* phylogeny. (*a*) The structure of the three supergene alleles, showing the position and orientation of regions 1, 2 and 4, which are interspersed by fragments of region 3. (*b*) Species tree for *Danaus* inferred using ASTRAL, based on 5954 gene alignments. The grey clade indicates all species with elevated gene copy number in region 3 (electronic supplementary material, figure S7 and table S5). The green star indicates the inferred origin of the region 3 duplications. (*c*–*e*) Inferred genealogies for regions 1 (including data from 18 genes—17 268 sites for region 1.2 and 38–36 576 sites for region 1.1), 2 (including data from 61 genes—45 214 sites) and 4 (including data from 15 genes—14 127 sites; see the electronic supplementary material, figure S8 for full maximum-likelihood trees). All three genealogies are incongruent with the species branching order. Node ages are therefore positioned to show the most likely origins of the structural variants (indicated by stars) to allow for the observed incongruence. For image attributions see the electronic supplementary material, figure S7.

D. gilippus (found in the Americas) and *D. petilia* (found in Australia) originated through dispersal from North Africa/Asia, where the chrysippus allele is currently found [15]. Using the dating procedure described above, we estimate the split giving rise to *D. gilippus* to 4.1 Ma, implying that the polymorphisms have persisted for at least this long.

Region 2 has also undergone two structural events: an inversion specific to the orientis allele and a shift (or translocation) caused by the insertion of duplicated fragments of region 3 (figure 2*a*). The tree for region 2 is similar to that for region 1, indicating that recombination suppression among all three supergene alleles predates the emergence of *D. gilippus* and *D. petilia*, who both inherited the chrysippus-like allele (figure 2*d*). The node separating the three variants could not be resolved (electronic supplementary material, figure S8C), but we infer that the translocation shared by both chrysippus and orientis occurred prior to the orientis-specific inversion (figure 2*d*). Finally, the tree for region 4 suggests that the inversion specific to the orientis allele occurred after the speciation event that gave rise to *D. gilippus*, but before the emergence of *D. petilia* (estimated at 1.7 Ma), which inherited the ancestral state (figure 2*e*).

(e) Sequence divergence reveals the landscape of recombination suppression

To understand how the different chr15 rearrangements contribute to recombination suppression in the BC supergene,

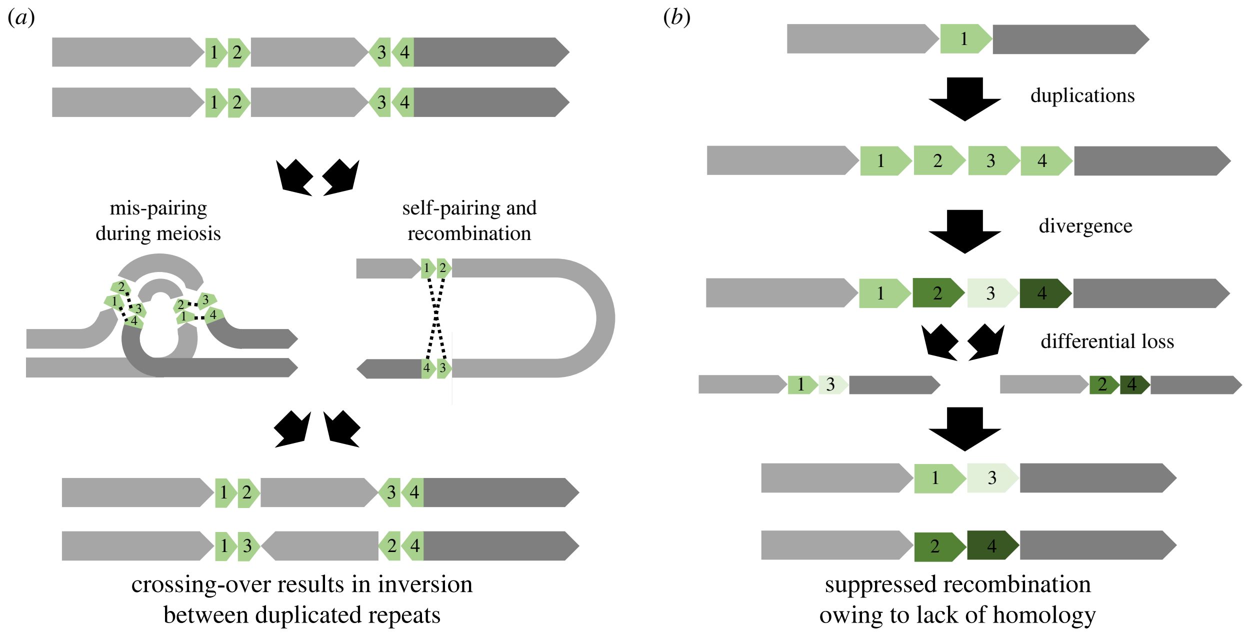
(a)
1 2 3 4
1 2 3 4
mis-pairing during meiosis
self-pairing and recombination
1 2 3 4
1 3 2 4
crossing-over results in inversion between duplicated repeats
(b)
1
duplications
1 2 3 4
divergence
1 2 3 4
differential loss
1 3
2 4
1 3
2 4
suppressed recombination owing to lack of homology

Figure 3. Two mechanisms by which segmental duplications can lead to recombination suppression. (*a*) Ectopic recombination between mis-paired paralogues in different parts of the chromosome can cause an inversion of the intervening region. This mechanism is most relevant when the paralogues are 'young' and still share strong sequence similarity. (*b*) Our proposed model of duplication, divergence and differential loss. Over long periods of time, old paralogues may become highly diverged in their sequences. If differential loss of paralogues occurs in different individuals, the remaining copies may be too divergent to pair up and crossover during meiosis. (Online version in colour.)

we examined d_{XY} (absolute sequence divergence between morphs) and d_a (divergence corrected for within-population diversity) across the chromosome. Elevated sequence divergence is a more reliable indicator of deep coalescence than relative measures such as F_{ST} [21] and is therefore a useful metric to identify regions of the genome at which recombination has been suppressed between segregating alleles for a long time. For each pair of alleles, clear regions of elevated sequence divergence are detectable, corresponding approximately with the regions that are rearranged between the pair (electronic supplementary material, figure S9). In regions 1 and 2, the level of sequence divergence is very high (d_{XY} approx. 6%, d_a approx. 4%), consistent with these rearrangements having occurred several million years ago, whereas region 4 shows notably lower divergence, consistent with a more recent inversion (electronic supplementary material, figure S9). In all three alleles—most notably klugii—some of the duplicated fragments of region 3 have little or no sequence homology with the other alleles (electronic supplementary material, figure S9). This implies that some of the paralogues are too divergent to allow cross-mapping of short reads, probably reflecting ancient duplications followed by independent losses of copies in each allele (see Discussion).

3. Discussion

By generating chromosome-scale assemblies for all three known alleles of the *D. chrysippus* BC supergene, we have uncovered a stepwise process of supergene formation, with a pivotal early role of segmental duplications. Our results suggest two ways in which duplications may contribute to recombination suppression. The first has been described previously: duplicated regions can trigger inversions by promoting ectopic crossovers [22,23]. Meiotic pairing between paralogues can create inverted loops that lead to inversions if crossovers occur on both sides (figure 3*a*). In the BC supergene, two of the inverted regions (2 and 4) are bordered on both sides by duplicated fragments, consistent with their involvement in the inversions. Another two inversions involving region 1 are bordered only on one side by duplicated fragments. The re-use of breakpoints in multiple structural events has been seen elsewhere and is consistent with a role of duplications or repeats in driving recurrent inversion mutations [12,24–26].

The second mechanism by which duplications can contribute to recombination suppression has, to our knowledge, not been previously proposed. Our model has three steps (figure 3*b*). First, one or more tandem duplications occur and spread to fixation in the population. Second, over a long period, the paralogues become highly diverged through the accumulation of mutational differences. Third, independent losses of different paralogues in different individuals leads to the coexistence of haplotypes in a single species that may fail to pair up and form crossovers during meiosis owing to the lack of homology (electronic supplementary material, figure S9A). Our model might help to explain the existence of non-recombining regions in the absence of inversions, such as the *doublesex* polymorphism in *Papilio memnon* [10] (Komata *et al.* [11]). It remains to be seen whether duplications and copy number variation are a common feature of supergenes in other taxa.

Following the ancient duplications, four distinct inversions occurred in a stepwise manner to produce the three divergent BC supergene alleles we see today. This stepwise increase in the range of recombination suppression is similar to processes thought to underlie sex chromosome evolution [7,27]. It is possible that the expansion of the BC supergene has allowed the accumulation of additional co-adapted alleles and potentially extended its adaptive role to traits beyond colour pattern. The modular nature of the inversions also means that recombination can occur between them without creating unbalanced gametes and thereby generate recombinant supergene alleles. Indeed, wild individuals frequently show mosaic patterns of ancestry across the BC region [15], and one of the four new assemblies described here appears to be a recombinant (electronic supplementary material, figure S2). These findings, taken together with recent evidence for extensive gene conversion in inversions in *Drosophila* [28], suggest that the evolution of supergene alleles may be a more dynamic process than previously thought. As long-range sequencing and chromosomal assembly becomes feasible for an increasing number of study systems, we expect that the stepwise progression of recombination suppression described here will prove to be common.

4. Methods

(a) Samples, processing and sequencing

To produce haploid assemblies for the three BC alleles, we used a trio-binning approach in which long sequencing reads from a diploid sample are binned according to their parent of origin before assembling. We generated two broods (MB181 and SB211) using parents with distinct colour patterns to produce heterozygous F_1 progeny carrying two distinct supergene alleles. Specifically, brood MB181 comprised a father suspected to be homozygous for the klugii allele and a mother suspected to be heterozygous chrysippus/orientis. Brood SB211 comprised a father suspected to be homozygous for the orientis allele and a mother suspected to be heterozygous chrysippus/orientis. From brood MB181, DNA was extracted from a single pupa (MB18102) using the MagAttract HMW DNA Kit from Qiagen (Venlo, Netherlands). PacBio continuous long reads sequencing (Pacific Biosciences, Menlo Park, California, USA) was then carried out using four single molecule real-time (SMRT) cells on a PacBio Sequel to produce a total of 34.7 gigabases (Gb) of sequence data. From brood SB211, DNA was extracted from three pupae (SB21101, SB21102 and SB21104) and sequenced in a single SMRT cell of a PacBio Sequel II to produce HiFi reads totalling 12.5 Gb of sequence.

In addition to the long-read sequencing of F_1 progeny, we performed Illumina sequencing of parents of both broods for trio-binning (but used an aunt for MB181 because the mother was lost). We also generated Illumina reads for the F_1 pupa of brood MB181 for polishing the assembly, and from an individual of *Tirumala formosa*, to serve as an outgroup. DNA extraction was performed using the Qiagen DNEasy Blood and Tissue Kit. Sequencing was performed using the Illumina NovaSeq 600 platform with 300 cycles and an insert size of 350 bp (see the electronic supplementary material, table S1 for sample information and sequencing coverage).

(b) Genome assembly and annotation

Although trio-binning of long reads requires both parents, we reasoned that a sister of the mother of brood MB181 could be used to correctly bin reads from certain chromosomes. Lepidopteran females do not undergo crossing over, so daughters share identical chromosomal haplotypes with maternal aunts for some chromosomes. However, we needed to first confirm that the daughter and aunt shared a first-degree relationship for chr15, the chromosome of interest. To this end, we aligned the Illumina reads to a chromosomal-level *D. plexippus* assembly ('Dplex_v4', GCA_009731565, [29]) using BWA v. 0.7.15 mem [30], converted to BAM format using SAMtools v. 0.1.18 [31] and split by chromosome using BamTools split v. 2.5.1 [32]. We applied IBSrelate [33] to each chromosome separately and confirmed that the F_1 (MB18102) and aunt had parent–offspring level relationships for around half the chromosomes, including chr15 (electronic supplementary material, figure S10).

We used Canu v. 2.1 [34] to perform trio-binning and assembly for MB18102 (electronic supplementary material, table S2), with estimated genome size of 360 Mb and a minimum read length of 2 kb. For brood SB211, we had PacBio HiFi sequence data from three offspring (SB21101, SB21102 and SB21104). We therefore first binned reads using the *–haplotype* option of Canu and then pooled the binned reads representing each parent to obtain sufficient read depth for both assemblies. For assembly, both hifiasm [35] and HiCanu [36] were tested (electronic supplementary material, table S3). Based on contiguity of chr15-linked contigs (electronic supplementary material, figure S11), we chose to proceed with the hifiasm assembly for the paternal reads (SB211PAT), and the HiCanu assembly for the maternal reads (SB211MAT).

We assessed assembly completeness using BUSCO v. 3.1.0 [37] with arthropoda_odb9 and insecta_odb9 dataset. As expected, biassed binning occurred in brood MB181 and more PacBio reads were assigned to the father. The paternal assembly (MB18102PAT) therefore shows a high level of duplication whereas the maternal assembly (MB18102MAT) shows a high level of missing BUSCOs (electronic supplementary material, table S4). However, we assumed that these issues would only affect the chromosomes that do not show parent–offspring relationships with the aunt and therefore not affect chr15. For all assemblies, redundant (haplotypic) contigs were removed with Purge Haplotigs [38].

Polishing to correct sequencing errors was performed for the two MB18102 assemblies (the SB211 assemblies did not require polishing as they were generated using PacBio HiFi reads). Two rounds of polishing were carried out simultaneously on the MB18102MAT and MB18102PAT assemblies using Illumina data from the assembled F_1 individual in the program Hapo-G (allowing the correct mapping of reads to the correct haplotype). Following polishing, each of the MB18102 and SB211 assemblies was repeat masked and annotated using RepeatMasker [39] and BRAKER2 [40] pipeline, respectively (as in [19]; using the same Dchry2.2 repeat library and *D. plexippus* protein sets). The TE content of the genome was calculated in 50 kb windows and the RepeatMasker output using a custom script (Get.TE.Bed.sh; https://github.com/RishiDeKayne/GetTEBed).

(c) Assembly alignment and rearrangement detection

To identify contigs corresponding to chr15, we performed whole-genome alignment to two chromosomal-level *D. plexippus* assemblies: 'Dplex_v4' (GCA_009731565, [29]) and 'MEX_DaPlex' (GCA_018135715, [41]). Note that chromosome numbers in the D_plex4 assembly are different, and chr15 corresponds to chr7. For alignment, we used both minimap2 [42] with the 'asm20' parameter preset and the *nucmer* command in MUMmer v. 4.0.0 [43]. Alignments were explored visually using the online tool at https://dot.sandbox.bio. For final visualization, we used the minimap2 alignments to the outgroup *D. plexippus* 'MEX_DaPlex' assembly and discarded alignments less than 500 bp in length. For comparison among assemblies of *D. chrysippus* supergene alleles, we used minimap2 with preset 'asm10', and discarded alignments of less than 2 kb and with divergence (dv value output by minimap2) less than 0.2.

To identify syntenic blocks and delineate rearranged regions, we developed an algorithm that merges adjacent alignments. It is implemented in the script asynt.R available at https://github.com/simonhmartin/asynt. The algorithm has three steps. First, alignments are split into 'sub-blocks' that each corresponds to a unique tract of the reference assembly (each alignment corresponds to one sub-block, except overlapping intervals are defined as distinct sub-blocks). Second, sub-blocks below a minimum size in the reference are discarded. Third, adjacent sub-blocks that are in the same orientation and are below some threshold distance apart in the reference are merged to yield syntenic blocks. These three steps can be performed iteratively to first identify regions of fine-scale synteny and build these up into larger syntenic blocks. We used three iterations with minimum sub-block sizes of 200 bp, 2 kb and 20 kb, respectively, and a maximum distance between sub-blocks of 100 kb.

(d) Ancestry painting based on sequence divergence

We devised an approach to 'paint' each assembled haplotype according to its ancestry based on aligned short-read data from representative individuals known to carry one of the three known supergene alleles. We first selected several representative wild individuals identified as homozygous for one of the three supergene alleles in a previous study [15] (see the electronic supplementary material, table S1 for accession numbers). For each target assembly, short reads from the three sets of representative individuals were aligned using BWA mem v. 0.7.17 with default parameters. BAM files were generated using SAMTOOLS v. 1.9 and sorted using PICARDTOOLS (https://broadinstitute.github.io/picard/) SORTSAM v. 2.21.1. Duplicate reads were identified and removed using PICARDTOOLS MARKDUPLICATES. Genotypes were called using the repeat-masked references with bcftools v. 1.10.2 [31,44], using the mpileup, call and filter tools to retain only genotypes with an individual read depth (DP) ≥ 8 and genotype quality (GQ) ≥ 20. Diploid genotypes for each individual were exported, along with the haploid assembly genotype using the script parseVCF.py, and diversity (π) and divergence (d_{XY}) between each reference panel and the assembly were computed using the script popgenWindows.py (https://github.com/simonhmartin/genomics_general). Non-overlapping windows of 25 000 genotyped sites were used, and for each reference panel, at least 10 000 sites had to be genotyped in at least one of the individuals. We then calculated d_a by subtracting π for a given reference panel from d_{XY}. 'Ancestry painting' was performed by assigning, for each window, the reference panel that was most similar to the target assembly according to d_a, with the added requirement that the difference between the lowest and second lowest d_a value had to be greater than or equal to 0.01 (electronic supplementary material, figure S2).

(e) Copy number variation

We estimated gene copy number variation based on the read depth of Illumina reads. We used *D. plexippus* (Dplex_v4) as a reference as this species was inferred to represent the ancestral state in gene copy number. Reads were aligned and BAM files were processed as described above. To ensure that copy numbers were comparable among species and not biased by poor read mapping in highly divergent regions, we only considered read depth at exonic sites that had at least one read mapped in all individuals from all species. To this end, we generated a vcf file as described above, but without filtering for depth or GQ, and used bcftools view to retain only sites at which all individuals had non-zero depth. Finally, median depth was computed for each gene using the scripts parseVCF.py and windowStats.py (https://github.com/simonhmartin/genomics_general). Genes with fewer than 100 sites across all exons present in the dataset were excluded. Gene copy numbers were estimated by normalizing median depths by the median depth across all genes in the chromosome and rounding to the nearest whole number.

(f) Phylogenetic analyses

We generated alignments for the coding regions of each gene using available short-read Illumina sequence data for seven *Danaus* species and an outgroup from the sister genus *Tirumala* that was sequenced for this study (electronic supplementary material, table S1). The same VCF files described above (including invariant sites) for copy number analysis were used, with all genotypes with a GQ less than 20 and read depth (DP) < 5 set to missing data. We then removed all sites with any missing data, converted diploid genotypes to single bases with heterozygous genotypes expressed as IUPAC ambiguity characters and finally extracted alignments for each gene using the scripts parseVCF.py, filterGeno.py and extractCDSalignments.py available at https://github.com/simonhmartin/genomics_general. After generating these alignments for each gene, we discarded genes where alignments had greater than 500 Ns or comprised at least 75% Ns.

For species tree inference, we further excluded all genes on chromosomes thought to carry rearrangements [15]: chr4 (chr11 in Dplex_v4), chr7 (chr15), chr15 (chr7), chr17 (chr16), chr22 (chr20) and chr30 (chr28). The resulting 5954 gene alignments were analysed using the multispecies coalescent tool ASTRAL [45], implemented in the package PARGENES v.1.1.2 ([46]; which produces individual gene trees with RAxML [47] and then initiates an ASTRAL run on this full set of trees). Default parameters for PARGENES were used.

To explore the history of individual rearrangements, we subsetted our filtered set of gene alignments to only the genes found within the bounds of the rearranged region (region 1 was divided into two to account for the separate inversions in orientis and chrysippus). We then ran RAxML v. 8.2.12 [47] on a concatenated alignment for each subset using the GTRGAMMA model with 100 rapid bootstrap replicates (-f a -N 100).

(g) Sequence divergence in sliding windows

We calculated sequence divergence between reference panels representing each supergene allele using the same genotype data described in the ancestry painting section above. Non-overlapping windows of 25 000 genotyped sites were used, and at least 10 000 sites had to be genotyped in at least one of the individuals from each panel. Nucleotide diversity (π) and divergence (d_{XY}) were computed using the script popgenWindows.py (https://github.com/simonhmartin/genomics_general), and we then calculated d_a by subtracting the average π for a given pair from d_{XY}.

Ethics. Butterfly collection was performed with permission of the land owners and with permits where relevant NACOSTI/P15/3290/3607, NACOSTI/P15/2403/3602 (National Commission for Science and Technology, Kenya).

Data accessibility. Sequencing reads and assemblies are available at the European Nucleotide Archive project (see the electronic supplementary material, table S1 for accession numbers [48]). Assemblies are available in the European Nucleotide Archive project accession PRJEB52180. Additional data files are available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.xwdbrv1g0 [49], including genome assemblies and annotations, repeat library and windowed repeat content, whole-genome alignments, VCF and genotype files, window-based diversity and divergence measures and read depth, sequence alignments for genes and phylogenetic trees. Scripts for assembly polishing, the analysis of repeat content, genome annotation and phylogenetic tree construction are available at https://github.com/RishiDeKayne/Danaus_supergene_structure. Scripts for genome alignment and synteny block inference, ancestry painting and divergence analyses, and read depth and copy number analyses are available at https://github.com/simonhmartin/Danaus_supergene_structure.

Authors' contributions. K.-W.K.: formal analysis, investigation, visualization, writing—original draft, writing—review and editing; R.D.-K.: formal analysis, investigation, visualization, writing—original draft, writing—review and editing; I.J.G.: conceptualization, funding acquisition, project administration and resources; K.S.O.: investigation; D.J.M.: project administration, resources and supervision; R.f.-C.:

conceptualization, funding acquisition, writing—review and editing; S.H.M.: conceptualization, formal analysis, funding acquisition, investigation, methodology, project administration, supervision, visualization, writing—original draft, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Competing interests. We declare we have no competing interests.

Funding. This work was supported by a Royal Society University Research Fellowship (grant no. URF\R1\180682) and Enhancement Award (grant no. RGF\EA\181071) awarded to S.H.M., a Swiss National Science Foundation Early Postdoc Mobility Fellowship (grant no. P2BEP3_195567) awarded to R.D.-K. and a National Geographic Society Research Grant (grant no. WW-138R-17) awarded to I.J.G.

Acknowledgements. We thank Alex Mackintosh for providing advice on genome assembly and annotation and valuable feedback on our interpretations of the results. We thank Chay Graham for helpful comments on the manuscript.

References

1. Küpper C *et al.* 2016 A supergene determines highly divergent male reproductive morphs in the ruff. *Nat. Genet.* **48**, 79–83. (doi:10.1038/ng.3443)
2. Merot C, Llaurens V, Normandeau E, Bernatchez L, Wellenreuther M. 2020 Balancing selection via life-history trade-offs maintains an inversion polymorphism in a seaweed fly. *Nat. Commun.* **11**, 670. (doi:10.1038/s41467-020-14479-7)
3. Kirkpatrick M, Barton N. 2006 Chromosome inversions, local adaptation and speciation. *Genetics* **173**, 419–434. (doi:10.1534/genetics.105.047985)
4. Schaal SM, Haller BC, Lotterhos KE. 2022 Inversion invasions: when the genetic basis of local adaptation is concentrated within inversions in the face of gene flow. *Phil. Trans. R. Soc. B* **377**, 20210200. (doi:10.1098/rstb.2021.0200)
5. Wellenreuther M, Bernatchez L. 2018 Eco-evolutionary genomics of chromosomal inversions. *Trends Ecol. Evol.* **33**, 427–440. (doi:10.1016/j.tree.2018.04.002)
6. Joron M *et al.* 2011 Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature* **477**, 203–206. (doi:10.1038/nature10341)
7. Charlesworth D, Charlesworth B, Marais G. 2005 Steps in the evolution of heteromorphic sex chromosomes. *Heredity* **95**, 118–128. (doi:10.1038/sj.hdy.6800697)
8. Ozias-Akins P, Akiyama Y, Hanna WW. 2003 Molecular characterization of the genomic region linked with apomixis in *Pennisetum/Cenchrus*. *Funct. Integr. Genomics* **3**, 94–104. (doi:10.1007/s10142-003-0084-8)
9. Li J *et al.* 2016 Genetic architecture and evolution of the *S* locus supergene in *Primula vulgaris*. *Nat. Plants* **2**, 16188. (doi:10.1038/nplants.2016.188)
10. Iijima T, Kajitani R, Komata S, Lin CP, Sota T, Itoh T, Fujiwara H. 2018 Parallel evolution of Batesian mimicry supergene in two *Papilio* butterflies, *P. polytes* and *P. memnon*. *Sci. Adv.* **4**, eaao5416. (doi:10.1126/sciadv.aao5416)
11. Komata S, Kajitani R, Itoh T, Fujiwara H. 2022 Genomic architecture and functional unit of mimicry supergene in female limited Batesian mimic *Papilio* butterflies. *Phil. Trans. R. Soc. B* **377**, 20210198. (doi:10.1098/rstb.2021.0198)
12. Yan Z *et al.* 2020 Evolution of a supergene that regulates a trans-species social polymorphism. *Nat. Ecol. Evol.* **4**, 240–249. (doi:10.1038/s41559-019-1081-1)
13. Brower LP, Edmunds M, Moffitt CM. 1975 Cardenolide content and palatability of a population of *Danaus chrysippus* butterflies from West Africa. *J. Entomol. Ser. A* **49**, 183–196. (doi:10.1111/j.1365-3032.1975.tb00084.x)
14. Smith DAS, Owen DF, Gordon IJ, Lowis NK. 1997 The butterfly *Danaus chrysippus* (L.) in East Africa: polymorphism and morph-ratio clines within a complex, extensive and dynamic hybrid zone. *Zool. J. Linn. Soc.* **120**, 51–78. (doi:10.1111/j.1096-3642.1997.tb01272.x)
15. Martin SH *et al.* 2020 Whole-chromosome hitchhiking driven by a male-killing endosymbiont. *PLoS Biol.* **18**, e3000610. (doi:10.1371/journal.pbio.3000610)
16. Smith DAS. 1979 The significance of beak marks on the wings of an aposematic, distasteful and polymorphic butterfly. *Nature* **281**, 215–216. (doi:10.1038/281215a0)
17. Smith DAS, Gordon IJ, Allen JA. 2010 Reinforcement in hybrids among once isolated semispecies of *Danaus chrysippus* (L.) and evidence for sex chromosome evolution. *Ecol. Entomol.* **35**, 77–89. (doi:10.1111/j.1365-2311.2009.01143.x)
18. Koren S *et al.* 2018 *De novo* assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.* **36**, 1174–1182. (doi:10.1038/nbt.4277)
19. Singh KS, De-Kayne R, Omufwoko KS, Martins D, Bass C, ffrench-Constant R, Martin SH. 2022 Genome assembly of *Danaus chrysippus* and comparison with the Monarch *Danaus plexippus*. *G3* **12**, jkab449. (doi:10.1093/g3journal/jkab449)
20. Chazot N *et al.* 2019 Priors and posteriors in Bayesian timing of divergence analyses: the age of butterflies revisited. *Syst. Biol.* **68**, 797–813. (doi:10.1093/sysbio/syz002)
21. Cruickshank TE, Hahn MW. 2014 Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Mol. Ecol.* **23**, 3133–3157. (doi:10.1111/mec.12796)
22. Montgomery EA, Huang SM, Langley CH, Judd BH. 1991 Chromosome rearrangement by ectopic recombination in *Drosophila melanogaster*: genome structure and evolution. *Genetics* **129**, 1085–1098. (doi:10.1093/genetics/129.4.1085)
23. Arguello JR, Connallon T. 2011 Gene duplication and ectopic gene conversion in *Drosophila*. *Genes* **2**, 131–151. (doi:10.3390/genes2010131)
24. Puerma E, Orengo DJ, Aguade M. 2016 Multiple and diverse structural changes affect the breakpoint regions of polymorphic inversions across the *Drosophila* genus. *Sci. Rep.* **6**, 36248. (doi:10.1038/srep36248)
25. Maggiolini FAM *et al.* 2020 Single-cell strand sequencing of a macaque genome reveals multiple nested inversions and breakpoint reuse during primate evolution. *Genome Res.* **30**, 1680–1693. (doi:10.1101/gr.265322.120)
26. Corbett-Detig RB, Said I, Calzetta M, Genetti M, McBroome J, Maurer NW, Petrarca V, Della Torre A, Besansky NJ. 2019 Fine-mapping complex inversion breakpoints and investigating somatic pairing in the *Anopheles gambiae* species complex using proximity-ligation sequencing. *Genetics* **213**, 1495–1511. (doi:10.1534/genetics.119.302385)
27. Otto SP *et al.* 2011 About PAR: the distinct evolutionary dynamics of the pseudoautosomal region. *Trends Genet.* **27**, 358–367. (doi:10.1016/j.tig.2011.05.001)
28. Korunes KL, Noor MAF. 2019 Pervasive gene conversion in chromosomal inversion heterozygotes. *Mol. Ecol.* **28**, 1302–1315. (doi:10.1111/mec.14921)
29. Gu L, Reilly PF, Lewis JJ, Reed RD, Andolfatto P, Walters JR. 2019 Dichotomy of dosage compensation along the Neo Z chromosome of the Monarch butterfly. *Curr. Biol.* **29**, 4071–4077. (doi:10.1016/j.cub.2019.09.056)
30. Li H, Durbin R. 2010 Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595. (doi:10.1093/bioinformatics/btp698)
31. Li H *et al.* 2009 The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079. (doi:10.1093/bioinformatics/btp352)
32. Barnett DW, Garrison EK, Quinlan AR, Strömberg MP, Marth GT. 2011 BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* **27**, 1691–1692. (doi:10.1093/bioinformatics/btr174)
33. Waples RK, Albrechtsen A, Moltke I. 2019 Allele frequency-free inference of close familial relationships from genotypes or low-depth sequencing data. *Mol. Ecol.* **28**, 35–48. (doi:10.1111/mec.14954)
34. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017 Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736. (doi:10.1101/gr.215087.116)

35. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021 Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175. (doi:10.1038/s41592-020-01056-5)
36. Nurk S *et al.* 2020 HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* **30**, 1291–1305. (doi:10.1101/gr.263566.120)
37. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015 BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212. (doi:10.1093/bioinformatics/btv351)
38. Roach MJ, Schmidt SA, Borneman AR. 2018 Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinf.* **19**, 460. (doi:10.1186/s12859-018-2485-7)
39. Smit AFA, Hubley R, Green P. 2015 RepeatMasker Open-4.0. 2013–2015. See http://www.repeatmasker.org.
40. Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. 2021 BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR genomics and bioinformatics* **3**, lqaa108.
41. Ranz JM *et al.* 2021 A *de novo* transcriptional atlas in *Danaus plexippus* reveals variability in dosage compensation across tissues. *Commun. Biol.* **4**, 791. (doi:10.1038/s42003-021-02335-3)
42. Li H. 2018 Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100. (doi:10.1093/bioinformatics/bty191)
43. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. 2018 MUMmer4: a fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944. (doi:10.1371/journal.pcbi.1005944)
44. Danecek P *et al.* 2021 Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008. (doi:10.1093/gigascience/giab008)
45. Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, Warnow T. 2014 ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* **30**, i541–i548. (doi:10.1093/bioinformatics/btu462)
46. Morel B, Kozlov AM, Stamatakis A. 2019 ParGenes: a tool for massively parallel model selection and phylogenetic tree inference on thousands of genes. *Bioinformatics* **35**, 1771–1773. (doi:10.1093/bioinformatics/bty839)
47. Stamatakis A. 2014 RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313. (doi:10.1093/bioinformatics/btu033)
48. Kim K-W, De-Kayne R, Gordon IJ, Omufwoko KS, Martins DJ, ffrench-Constant R, Martin SH. 2022 Stepwise evolution of a butterfly supergene via duplication and inversion. FigShare. (doi:10.6084/m9.figshare.c.5980556)
49. Martin SH. 2022 Data from: Stepwise evolution of a butterfly supergene via duplication and inversion. Dryad Digital Repository. (doi:10.5061/dryad.xwdbrv1g0)