Mimicry diversification in *Papilio dardanus* via a genomic inversion in the regulatory region of *engrailed-invected*

Martijn J.T.N. Timmermans¹,²,³*, Amrita Srivathsan⁴, Steve Collins⁵, Rudolf Meier⁴,⁶, Alfried P. Vogler¹,²

1) Department of Life Sciences, Natural History Museum, London, UK
2) Department of Life Sciences, Imperial College London, Silwood Park Campus, Ascot, UK
3) Department of Natural Sciences, Middlesex University, London, UK
4) Department of Biological Sciences, National University of Singapore, Singapore.
5) African Butterfly Research Institute, Nairobi, Kenya
6) Lee Kong Chian Natural History Museum, National University of Singapore, Singapore.

*) Corresponding author: m.timmermans@mdx.ac.uk

Email addresses: Meier, Rudolf (meier@nus.edu.sg); Srivathsan, Amrita (asrivathsan@gmail.com); Vogler, Alfried (a.vogler@imperial.ac.uk); Collins, Steve (collinsabri@gmail.com)
Abstract
Polymorphic Batesian mimics exhibit multiple protective morphs that each mimic different noxious models. Here we study the genomic transitions leading to the evolution of different mimetic wing patterns in the polymorphic Mocker Swallowtail *Papilio dardanus*. We generated a draft genome (231 Mb over 30 chromosomes) and re-sequenced individuals of three morphs. Genome-wide SNP analysis revealed elevated linkage disequilibrium and divergence between morphs in the regulatory region of *engrailed*, a developmental gene previously implicated in the mimicry switch. The diverged region exhibits a discrete chromosomal inversion (of 40 kb) relative to the ancestral orientation that is associated with the *cenea* morph, but not with the bottom-recessive *hippocoonides* morph or with non-mimetic allopatric populations. The functional role of this inversion in the expression of the novel phenotype is currently unknown, but by preventing recombination, it allows the stable inheritance of divergent alleles enabling geographic spread and local co-existence of multiple adaptive morphs.

Keywords: Supergene; Batesian mimicry; butterflies; genomic rearrangement; polymorphism
Background

Mimetic butterflies undergo profound evolutionary changes in wing patterns driven by selection for a common signal deterring visual predators [1]. In Batesian mimics, which imitate harmful models but are not chemically defended themselves, the fitness advantage of being mimetic is a function of the predator’s encounter frequency of palatable individuals among unpalatable ones. Thus, a rare phenotype has a better chance of survival than a frequent one and the lowered fitness with increasing abundance (negative frequency dependent selection) may favour the evolution of multiple forms that each resemble a different noxious model [2]. In various cases of Batesian mimics several such morphs co-exist as phenotypically discrete, genetically controlled variants within a single population [1, 2]. The African Mocker Swallowtail, *Papilio dardanus*, (Fig. 1) is a widely known example of a polymorphic Batesian mimic. The species has played a central role in the debate about the evolution of phenotypic diversity [3,4,5], starting with Trimen’s work in the 1860s [6]. Sometimes referred to as “the most interesting butterfly in the world” [3], well over 100 variants have been named, including geographic races (subspecies) and about a dozen genetically well-defined wing pattern morphs (forms) that may co-occur in populations [7–9]. Females only are mimetic and both sexual dimorphism and female polymorphisms presumably are driven by negative frequency dependent selection from predators [10–12].

In *P. dardanus*, wing colours and patterns are controlled by a single Mendelian locus, *H*, whose various alleles segregate according to a well-defined hierarchy of dominance [3,13–15]. Phylogenetic analysis of subspecies and closely related species has led to the conclusion that mimicry has arisen fairly recently in *P. dardanus* and that the female mimetic forms are likely to have evolved from a ‘male-like’, presumed ancestral phenotype that is still found on Madagascar where the species is monomorphic and non-mimetic (Fig. 1) [16]. Segregation analysis in pedigree-broods using AFLP [17] and population genetics [18] have shown that the mimicry switch in *P. dardanus* is genetically linked to the *engrailed-invected* locus, a region that codes for two paralogous homeodomain transcription factors involved in anterior-posterior patterning [19].
Here, we study the genomic mechanisms that ultimately lead to the evolution of multiple mimetic phenotypes in *P. dardanus*. The simple Mendelian segregation of the wing colour and pattern traits led early geneticists to argue that a novel phenotype arises through a single macromutation [4,20]. However, the idea of achieving perfect mimics in a single step was generally dismissed by proponents of the Modern Synthesis [21,22] who argued that Mendelian inheritance alone was not sufficient to prove an origin through a single mutation. Instead, a two-step mechanism, first proposed by Nicholson [23], became the favoured hypothesis: a new mimetic phenotype originates via an initial large-effect mutation that provides at least moderate resemblance to a new mimicry model, after which genetically linked secondary mutations gradually improve the resemblance [24,25]. A gradual process of mimicry evolution was also favoured by computer simulations of varying recombination frequency and selection strength [26]. Under this hypothesis the initial mutation acts as a ‘genomic sieve’ [27] for closely linked mutations that improve the resemblance to the model; selection against non-mimetic intermediates then leads to the evolution of tighter linkage among genes determining colour and pattern [26,28], potentially producing a ‘supergene’ controlling multiple linked mutations, such that different polymorphic traits show Mendelian co-segregation [28–30].

A critical aspect of this process is that genetic recombination among functional sites is low, preventing the formation of intermediates with lower fitness. Molecular genetics studies in polymorphic butterflies, beetles and birds have detected associated genomic inversions as a mechanism that increases linkage of co-adapted mutations [31–36]. However, the importance of these inversions in the initial evolution and further diversification of polymorphic forms remains unclear. Mimetic polymorphism may exist with and without genomic inversions, as seen in the closely related Southeast Asian *Papilio polytes* and *P. memnon* whose mimicry locus (in the *dsx* genomic region) is contained in an inversion only in *P. polytes* [37].

To understand the genetic architecture underlying polymorphic mimicry in *P. dardanus* we use comparative genomics of three female ‘forms’ (Fig. 1). Specifically, among the numerous female-limited mimicry types the prevalent morph is the form *hippocoon* (f.
hippocoon), also referred to as f. hippocoonides in some parts of its range, which is a black-
and-white phenotype mimicking the danaiid Amauris niavius. This morph is widely
distributed on the African mainland and is recessive to all others. A further widespread
phenotype is the black-and-orange form cenea (f. cenea) present mostly in specific regions
of Kenya (subspecies P. d. polytrophus) and south-eastern Africa (subspecies P. d. cenea).
Numerous other mimetic morphs co-occur within populations of these two subspecies at
various frequencies throughout sub-Saharan Africa [7], but populations in Madagascar are
always monomorphic and have been recognised as a separate subspecies, P. d. meriones
[15]. Using a newly generated draft genome sequence we assess evidence for reduced
recombination and genetic divergence in P. dardanus, and search for local rearrangements
that might control the phenotypic switch. This first genome wide study of P. dardanus
allows greater insight into the evolution of multiple mimicry forms and their stable
inheritance in populations.

Results

Draft genome and linkage map

A draft genome sequence was constructed using a three-generation laboratory inbred male
of subspecies P. dardanus tibullus, which was homozygous for the bottom recessive f.
hippocoonides allele (Fig. 1). We obtained an assembly of 7,365 scaffolds (N50=596,599;
L50 = 99) with a total length of 231,123,043 bp, which was very similar to a genome size
estimate of 232 Mb obtained using k-mer counts (electronic supplementary
material, figure S1). We were able to annotate 12,795 potential protein coding sequences
(CDS) and obtained Gene Ontology annotations for 8,111 putative protein coding
sequences. The level of completeness was similar to published draft genomes of three
related Papilio species (electronic supplementary material, table S1 ). The entire mimicry
locus H [17,18] was contained in two scaffolds which were merged into a 2.5 Mb scaffold
using information from a publicly available BAC clone sequence from the same morph
[18].

The scaffolds were assessed for correct assembly using co-segregation of RADseq
polymorphisms generated for two pedigree broods (14 and 33 F1 individuals respectively).

5
For each brood, SNPs were selected that were heterozygous in the female parent and homozygous in the male parent. There is no crossing over in female Lepidoptera [38], and thus all heterozygous positions on a correctly assembled scaffold should show identical inheritance patterns in every offspring of a brood. Of the 7,365 scaffolds, 402 (total length: 193,743,404 bp) contained at least two polymorphic RADtags and could be included in this analysis. Using SNP markers within the RADtags that were the furthest apart in the physical maps of the scaffolds, 379 of these 402 scaffolds showed matching SNP patterns in all the progeny, while discrepancies were observed for the remaining 23 scaffolds, whose correct assembly could therefore not be confirmed (electronic supplementary material, figure S2). The RADseq data were further used to merge the scaffolds into 29 unordered bins to represent provisional groups of linked sequences. Of the 12,795 P. dardanus CDS, 9349 could be associated to one of these chromosome bins. Comparison with the well annotated *Heliconius melpomene* genome largely confirmed the groups (electronic supplementary material, figure S3). The 29 bins are not expected to include sex chromosomes as the analyses only used SNPs that are heterozygous in the female parent (female Lepidoptera are ZW, males ZZ). The data therefore suggests that *P. dardanus* exhibits 30 chromosomes (29 bins plus the sex chromosomes), in accordance with an AFLP study [17] and several related *Papilio* [39].

**Genomics of mimicry morphs**

Genomic differentiation of morphs was established by shotgun sequencing of specimens of *hippocoonides* (n=4), *cenea* (n=4), and an individual of the non-mimetic subspecies *P. d. meriones* (Fig. 1; Table 1; electronic supplementary material, table S2). Reads were mapped onto the genomic scaffolds that are longer than 100 kb (n=420). Genome-wide SNP analysis of 5-kb windows electronic supplementary material, figure S4) detected elevated F\textsubscript{st} values between the samples of *hippocoonides* and *cenea* individuals in various regions throughout the genome, including a region of ~75 kb covering the *engrailed*-invected locus. This latter ~75 kb region also showed elevated LD. No such pattern of joint elevated LD and F\textsubscript{st} was observed in any of the other 420 long contigs (electronic supplementary material, figure S4). These observations support the notion that within this region genetic subdivision is elevated and recombination is rarer than in other regions of
the *P. dardanus* genome (Fig. 2). The pinpointed region did not show evidence of elevated nucleotide diversity when analysing sequences from the *hippocoonides* and *cenea* morphs together (Fig. 2). However, sequence divergence (estimated as p-distance) between the *hippocoonides* individuals and the reference genome sequence (derived from a *hippocoonides* individual) was sharply lower in the pinpointed region than for the *cenea* individuals and the more divergent *P. d. meriones* (Fig. 2).

Closer inspection of the ~75 kb region revealed paired reads that were placed ~40 kb apart and in opposite orientation in all four f. *cenea* individuals (electronic supplementary material, figure S5). Such read-pairs were not observed in the four f. *hippocoonides* samples. This indicates that the genetically diverged region contains a ~40 kb inversion associated to the mimetic f. *cenea*. The inversion was not found in the non-mimetic *P. d. meriones* from Madagascar, which indicates that the bottom-recessive mimetic f. *hippocoonides* has the same arrangement as this male-like form, and therefore this specific arrangement is ancestral. The four f. *cenea* specimens represented two distinct subspecies from Kenya (*P. d. polytrophus* f. *cenea*) and South Africa (*P. d. cenea* f. *cenea*). The sequence data furthermore indicated that the Kenyan specimens carried a non-inverted allele too, suggesting they are heterozygous for f. *cenea* and f. *hippocoonides* (Hc/Hh) (Table 1) which is in agreement with breeding experiments (electronic supplementary material, figure S6). The South African f. *cenea* specimen was homozygous for the inversion; while homozygosity for the *cenea* allele has not been confirmed by breeding, it is likely because this morph is very common in this part of the species range [13].

We validated the inversion for several additional specimens of the two mimetic morphs by PCR amplification with boundary-defining primers (Fig. 3). PCR fragments confirmed the predicted inversion: all 4 additional f. *hippocoonides* individuals retained the arrangement of the draft genome physical map (based on primer pair A-B and C-D; electronic supplementary material, figure S7), consistent with the findings from the sequenced individuals. Four additional f. *cenea* individuals showed the ~40 kb inversion (primer pair A-C and B-D). These *cenea* females also showed the A-B and C-D fragments of the reference map, indicating they are heterozygous (Hc/Hh). Fisher exact tests for association
between phenotype and inversion were highly significant (P < 0.0001) (electronic supplementary material, table S3).

To test whether the genomic region surrounding * engrailed* and *invected* recombines freely, we used RAD data for two pedigree broods (homozygous f. *hippocoonides*), using 199 SNPs at sites with variants in the male but not the female parents in the ~2.5 Mb scaffold containing *engrailed-invected*. We detected seven recombination events in one brood and two in the other (Fig. 2; electronic supplementary material, table S4), for a relatively high recombination frequency of 7.8 cM/Mb (9 recombination events in 47 offspring, or 19.1 cM, over a distance of 2,458 Mb of the scaffold). These results were similar to those presented by Clark et al. [17], who analysed a cross between a heterozygous male (H*hb/H*hc*) and a homozygous female (H*hb/H*hb) and reported 5 recombination events between male-informative AFLP markers ACT and PD (highlighted on Fig. 2), which flank the *engrailed-invected* region, after scoring 35 F1 individuals.

**Discussion**

Our genomic analysis revealed a 40 kb inversion in *P. dardanus* at ~6,800 bp upstream of the *engrailed* start codon, which differentiates the haplotype associated with the *hippocoonides* and *cenea* morphs, and coincides with localized peaks in LD and F\textsubscript{st} between haplotypes of these morphs. Mimicry loci have been postulated to consist of several tightly linked, epistatically interacting loci that in concert determine adaptive phenotypes (i.e. acting as a supergene) [29]. Such interaction of multiple sites requires regions of reduced recombination preventing the segregation of co-adapted loci, which was broadly confirmed in recent work demonstrating inversions in mimicry-linked genomic regions of other mimetic butterflies [31,32,34,37,40]. We have not determined the sequence of the *cenea* (H\(_c\)) allele and do not know whether several independent mutations are required for the switch between f. *cenea* and f. *hippocoonides* to happen, but the fact that a recombination suppressing inversion exists suggests a genomic architecture consistent with the supergene hypothesis (although due to the linkage of mutations within the inversion, it will not be possible to uncover the functional sites without functional studies).
The inversion in *P. dardanus* is small, compared to those associated with the mimicry loci in the Batesian mimic *P. polytes* and the Müllerian mimic *H. numata*, which stretch over 130 kb and at least 400 kb, respectively, and in those species result in allelic divergence in several protein coding genes. The *P. dardanus* inversion also differs from those species by the fact that it is found in an extended regulatory region apparently devoid of protein coding sequences. The region contains various enhancer sequences [41,42] that in other species have been shown to exert cis-regulatory control of both *engrailed* and *invented* and therefore likely affect unlinked genes determining the colour pattern, as initially envisioned by the ‘regulatory hypothesis’ of Nijhout [8,43]. *Invented* also contains an intronic microRNA (miR-2768) conserved in Lepidoptera (Fig. 3; electronic supplementary material, figure S8), which has been shown to downregulate *cubitus interruptus* (*ci*), a gene that determines patterning of the wing primordia via the *hedgehog* signalling pathway in nymphalid butterflies [44].

In *P. dardanus* the universally recessive *hippocoonides* form, despite being mimetic, apparently retains the presumed ancestral orientation found in the allopatric and genetically divergent (Fig. 2) Madagascan subspecies. This demonstrates that an inversion is not critical for the origin of mimetic forms, as also observed in *P. memnon* [37]. However, when multiple mimetic female forms are found in sympatry chromosomal inversions will assist stable segregation of divergent phenotypes, as has been shown for *P. polytes* [37]. Here we show that inversions are associated with multiple sympatric mimicry forms also in *P. dardanus* in mainland Africa. Balanced inversion polymorphisms may be maintained in populations by negative frequency-dependent selection (Type II polymorphisms of [45]). In addition, the spread of an advantageous phenotype is promoted when it is associated with an inversion (e.g. see [46]). The f. *cenea*-linked inversion has spread widely across the African continent and across subspecies boundaries, as evident from the presence of the *cenea* morph in *P. d. polytrophus* from Kenya and *P. d. cenea* from South Africa, geographically separated by at least 3000 km (Fig. 1). The fact that the same inversion is associated with the *cenea* morph in different subspecies, adds support for its role in defining the phenotype.
It still needs to be confirmed if the regulatory region of *engrailed-invected* plays any functional role in determining the pleiotropic changes of the wing. However, *P. dardanus* would not be unique in having regulatory changes underlying polymorphic mimicry. A recent study on the nymphalid *Hypolimnas misippus*, which displays sex-limited mimicry, revealed a 10-kb intergenic region upstream of the *Sox5/6* gene to be strongly associated to the wing phenotype, suggesting that a cis-regulatory element plays a role in pattern determination [47]. Inversions in an intron of the *pannier* locus determining colour polymorphism in a ladybird beetle have been shown to affect gene expression and to underlie phenotypic differences among colour morphs [48], also supporting *cis*-regulation of the phenotype through inversions of non-coding regions. If the 40-kb inversion in *P. dardanus* has *cis*-regulatory effects on the expression of one or more of *engrailed*, *invected* and miR-2768 (and possibly the adjacent gene *orange*), the genetic architecture of the region may be particularly conducive to the evolution of novel phenotypes. Thus, new inversions may provide the hypothesized major-effect shifts through their regulatory function that impacts the mosaic of pattern and colour elements of the wing.

Other morphs now need to be investigated for chromosomal rearrangements in this region, and may not exclusively involve inversions, given a previously reported duplication of *engrailed-invected* and a few neighbouring genes closely associated with one of the other *P. dardanus* female forms (f. *lamborni*) [18]. Preliminary results also suggest a genomic rearrangement in an individual of f. *planemoides*, which indicates that recombination-suppressing reordering of the *engrailed* region is an integral part of the evolution of new mimicry morphs. Determination of the phenotype likely works in concert with other changes in the *engrailed-invected* region, such as those in the first exon of *engrailed* found in the top-dominant f. *poultoni* and f. *planemoides* that exhibit a statistically significant overrepresentation of non-synonymous substitutions indicative of diversifying selection [49]. These divergent sites are outside of the newly detected inversions, perhaps suggesting that for some morphs a combination of the divergent *engrailed* coding region and the upstream inversion are required for correct specification of the phenotype. The presence of chromosomal rearrangements might suppress the recombination frequency even beyond
the inverted region, as already evident from the wider region of high LD and F_{st} extending
to ~75 kb (Fig. 3). Accordingly, recombinants producing maladaptive intermediate
phenotypes should exist but are rare, and such non-mimetic phenotypes may persist locally.

With each study of polymorphic systems, now including the prototypical *P. dardanus*, the
understanding of how discrete adaptive phenotypes evolve and are maintained in natural
populations improves: all currently described butterfly mimicry loci show the expected
signatures of allelic divergence, indicating that complex phenotypes indeed require
multiple sites and probably evolved in smaller steps. However, the mechanisms by which
tight linkage is achieved differ, as do the loci that determine the phenotypic switch.
Inversions are not necessary, but helpful to promote the capture of alleles under positive
selection, because they contribute to maintaining the alleles that would otherwise break up
genetically linked sites and lead to poor fitness. They might also contribute to the genetic
variation producing novel phenotypes, although for *P. dardanus*, the challenge remains to
determine any role of the inversion in gene expression or the regulation of downstream
pathways, in order to track the macro- and micro-mutations on the evolutionary trajectory
towards stable polymorphisms of mimicry forms.

**Methods**

**Genome sequencing, assembly and annotation**

The draft genome sequence was generated from an inbred male specimen of subspecies *P.
dardanus tibullus* (electronic supplementary material, table S2). Genomic DNA was used
for construction of Illumina TruSeq libraries (insert sizes of 300 bp and 800 bp) and a
Nextera mate-pair (MP) library prior to sequencing on Illumina platforms, followed by
standard procedures for adapter removal and quality trimming. GenomeScope [50] was
used to estimate genome size by obtaining the mean of the k-mer count distribution.
Sequencing errors were corrected using QUAKE v0.3.5 [51] using JELLYFISH v1.1.11
for k-mer counting [52]. Using an estimated genome size of 200 Mb we used k=17 for error
correction and Quake was run using default parameters. Genome assembly was conducted
using Platanus v. 1.2.4 [53], using only paired-end data for generating initial contigs, while
using mate-pair data for subsequent steps as recommended by the developers (number of
links for scaffolding = 10). For improving accuracy of the assembly, removing redundancy
and further scaffolding we used HaploMerger2 (Release 20151124) [54]. WindowMasker
v1.0.0. was first used to mask repetitive regions and all-against-all whole genome
alignments were then obtained using LASTZ and reciprocally-best whole-genome
alignments using chainNET to generate an improved haploid assembly.

The haploid assembly was further scaffolded using SSPACE v3.0 (number of links = 10),
using both paired-end and mate-pair libraries. Insert sizes were estimated by using the
library *_insFreq.tsv file generated by Platanus. This assembly was further refined by the
removal of tandem assembly errors and gaps in the assembly were closed using GapCloser.
Lastly, to remove scaffolds that could be from contaminations, we built a custom database
consisting of representative bacterial genomes from NCBI RefSeq 6, four reference
genomes for Papilio sp. (P. machaon: GCA_001298355.1; P. polytes: GCF_000836215.1;
P. xuthus: GCF_000836235.1; and P. glaucus: GCA_000931545.1) and a reference human
genome (GRCh38.p7). All scaffolds were searched against the reference database using
BLASTN with e-value of 1E-5. Genome completeness of this draft genome and other
Papilio genomes was assessed using BUSCO version 3 [55]. The assembly was annotated
using MAKER2 [56] with gene predictors trained by AUGUSTUS [57] using the BUSCO
ortholog set. Predicted protein and RNA sequences from genome assemblies of other
Papilio species were used as evidence. For functional annotations, protein sequences were
matched to SWISS-PROT [58] using BLASTP (E-value 1e-5) and subject to InterProScan
[59] for detection of protein signatures.

Scaffold clustering and mimicry locus genetic recombination
Sets of unordered linked scaffolds (“chromosome bins”) were obtained by SNP segregation
in RADseq data generated for two P. dardanus broods of 14 and 35 offspring. RAD library
construction was performed using PstI restriction digestion and barcoded libraries were
sequenced (100 bp single-end reads). Reads were de-multiplexed using the process_radtags
script of the package Stacks and subsequently mapped onto the genomic scaffolds using
bbmap (sourceforge.net/projects/bbmap/) (setting: ambiguous=toss local=t). The resulting
SAM files were sorted and converted to BAM files using SAMtools. Picard-tools-1.117
(http://broadinstitute.github.io/picard) was used to add read group information and merge the individual files of each brood into a single BAM file (i.e. one merged file per brood). These files were then converted to VCF format using the HaplotypeCaller program of GATK. Positions with 18x coverage or less for at least one of the samples within a brood were removed using SNPsift and the file converted to OneMap format using the vcf_to_onemap_input version 1.0 python script and positions heterozygous in the female parent (Onemap notation: ‘a,b’) and homozygous in the male (Onemap notation: ‘a,a’) parent (OneMap crosstype: D1.10) were extracted. For each scaffold with at least two segregating RADtags we tested co-segregation of the most distant SNPs to detect inconsistencies in segregation pattern, indicating incorrect assemblies. Co-segregation of SNPs was subsequently used to group scaffolds into linkage groups. CDS from linkage groups were compared to the Heliconius melpomene genome (version 2) and the positions of sequence matches on 21 H. melpomene chromosomes were recorded. The Perl GD::SVG library was used to visualise the positions of sequence matches. The RAD data was also used to investigate recombination within the scaffold containing engrailed-invected. SNPs homozygous in the female parent and heterozygous in the male parent were extracted and inspected manually for evidence of genomic recombination.

Population genomics of the P. dardanus supergene
Genomic data for eight specimens (Table 1) were mapped onto all scaffolds >100 kb using the BWA-MEM algorithm [60], merging the data for hippocoonides and cenea specimens into two separate files. Mean coverage was calculated for both for 5 kb sliding windows using SAMtools depth function and a custom perl script. To remove repetitive regions, sites with >400x coverage were masked for this analysis. The two files were merged and Kelly’s ZnS statistic (the average of the LD measure r² calculated between all pairs of SNPs) [61], nucleotide diversity (pi), and mean p-distance to the reference genome sequence were calculated using PopBam (sliding window 5 kb) [62]. Fst values were calculated using VCFtools 0.1.12 [63] contrasting the hippocoonides and cenea morphs (window size 5 kb). PCR was used to validate a genomic inversions (Fig. 2) using additional hippocoonides and cenea specimens (electronic supplementary material, figure S7) and the following primers: A) 5’-gktgtcggatttttgccgcta-3’, B) 5’-
aactaaaactrtyagacacgcaa-3’, C) 5’-tyaaccggtcagacaagttt-3’ and D) 5’-
amatggeatgractgmcga-3’. Fisher exact tests (two-tailed) were performed to test for
association between phenotype and presence of an inversion (taking the dominance
hierarchy into account) (electronic supplementary material, table S3).

Acknowledgements

The authors thank Martin Thompson for access to DNA from butterfly samples from South
Africa. Rebecca Clark kindly provided laboratory cross details on the sequenced Papilio
dardanus f. cenea from Kenya. We also would like to thank all reviewers who have
commented on various submissions of the manuscript. Sequencing libraries were
constructed and sequenced at the NHM London, the Department of Biochemistry
(University of Cambridge), and Genepool (University of Edinburgh).

Data accessibility

Sequence data that support the findings of this study have been deposited in GenBank with
the accession codes PRJNA451133, PRJNA600400, PRJNA600373 and SAMN05819004.

Author’s contributions

MJTNT participated in the design of the study, carried out the molecular lab work, analysed
data, and drafted the manuscript; AS carried out bioinformatics analyses and drafted the
manuscript; SC participated in the design of the study and provided specimens; RM
provided bioinformatics resources and critically revised the manuscript; APV participated
in the design of the study and drafted the manuscript. All authors gave final approval for
publication and agree to be held accountable for the work performed therein.

Competing interests

There are no competing interests.

Funding
This study was funded by NERC Postdoctoral Fellowship NE/I021578/1 (to MJTNT) and NERC NE/F006225/1 (to APV). AS was supported by SEABIG (R-154-000-648-646 and R-154-000-648-733).

References


Table 1: Samples used for sequencing.

<table>
<thead>
<tr>
<th>Voucher number</th>
<th>Geographic origin</th>
<th>Subspecies</th>
<th>Phenotype</th>
<th>H genotype</th>
<th># Paired-End Reads</th>
<th>Estimated coverage</th>
<th>Reference orientation</th>
<th>40 kb inversion</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMNH746848</td>
<td>Kenya</td>
<td>polytrophus</td>
<td>hippocoonides</td>
<td>H_{h}/H_{h}</td>
<td>38039853</td>
<td>41</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>BMNH746826</td>
<td>Kenya</td>
<td>polytrophus</td>
<td>hippocoonides</td>
<td>H_{h}/H_{h}</td>
<td>55548066</td>
<td>60</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>BMNH847389</td>
<td>Kenya</td>
<td>polytrophus</td>
<td>hippocoonides</td>
<td>H_{h}/H_{h}</td>
<td>35666867</td>
<td>39</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>BMNH746846</td>
<td>South Africa</td>
<td>cenea</td>
<td>hippocoonides</td>
<td>H_{h}/H_{h}</td>
<td>40600315</td>
<td>44</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>BMNH746453</td>
<td>Kenya</td>
<td>polytrophus</td>
<td>cenea</td>
<td>H_{c}/H_{h}</td>
<td>49512794</td>
<td>54</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>BMNH746764</td>
<td>Kenya</td>
<td>polytrophus</td>
<td>cenea</td>
<td>H_{c}/H_{h}</td>
<td>56085254</td>
<td>61</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Troph-c-02-46</td>
<td>Kenya</td>
<td>polytrophus</td>
<td>cenea</td>
<td>H_{c}/?</td>
<td>42184635</td>
<td>46</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>BMNH847353</td>
<td>South Africa</td>
<td>cenea</td>
<td>cenea</td>
<td>H_{c}/?</td>
<td>39434400</td>
<td>43</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>BMNH740167</td>
<td>Madagascar</td>
<td>meriones</td>
<td>meriones</td>
<td>H_{c}/?</td>
<td>31242775</td>
<td>34</td>
<td>x</td>
<td></td>
</tr>
</tbody>
</table>

H genotype: H_{c} = H_{cenea}, H_{h} = H_{hippoconides}. #Paired End Reads: number of raw reads generated for each specimen. Estimated coverage is calculated via: (number of raw reads * read length) / length of genome assembly. Read length was 125 bp. The actual coverage is expected to be lower due to not all reads passing quality control and the presence of contamination. “x” in the last three columns indicates whether the specimen carried an allele with the reference orientation and the 40 kb inversion.
Figure 1: Phenotypic variation in *Papilio dardanus* and samples used. Top: Seven female forms and a male. Bottom: Origin of samples for sequencing and population genetic analyses, from four subspecies: *P. dardanus polytrophus* (Kenya), *P. dardanus tibullus* (Kenya), *P. dardanus cenea* (South-Africa), and *P. dardanus meriones* (Madagascar). The specimen of subspecies *P. dardanus tibullus* was used for the construction of the draft genome sequence. The tree depicts the relationships among these four subspecies and is based on a tree presented in [16]. Three female forms were analysed: *hippocoonides*, *cenea*, and ‘male-like’.
Figure 2: Population genomic analysis of the full *engrailed-invected* containing scaffold. Thin vertical red lines: exons of various genes in the region. Note the exons of * engrailed* shown by large arrows and the upstream region marked in grey. Brood 59 and Brood 48: Recombination events in pedigree broods. SNPs from RADseq data for two broods are mapped on the *engrailed-invected* scaffold, shown by black triangles. Red triangles mark the intervals with confirmed recombination events, and the number of recombination events within these intervals are circled. The central band in the figure shows the map of the scaffold with exons (red arrows) and the upstream region of engrailed (grey). ACT and PD indicate the position of the AFLP markers of [17]. Linkage disequilibrium (LD; Kelly’s
ZnS statistic), $F_{st}$, nucleotide diversity ($\pi$), coverage and p-distance (to the reference genome) for the scaffold, calculated for the f. cenea and f. hippocoonides samples in 5 kb windows. Coverage and p-distance was calculated separately for the four cenea and for four hippocoonides specimens. The p-distance to the reference genome is also given for the P. dardanus meriones sample. Scales are in million base pairs.

**Figure 3:** Length and relative position of the inversions in the upstream regulatory region of *engrailed*. At the top, the map of the *engrailed-invected* region is shown, with short arrows indicating exons and the miR-2768 [44] shown in blue. Below the map is the direction of boundary-defining primers. The grey shading indicates the extent of the 40 kb inversion associated with f. cenea. For each of the forms, dark grey – light grey shading is used to indicate directionality of the 40 kb region. Scale is in base pair.