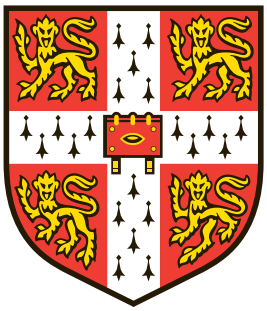
The genomics of adaptive colouration in *Hypolimnas* butterflies and the wood tiger moth, *Arctia plantaginis*



Anna Farre I Orteu

Supervisor: Prof Chris Jiggins

Department of Zoology, University of Cambridge

This thesis is submitted for the degree of

Doctor of Philosophy in Zoology

St John's College

September 2022

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text.

It is not substantially the same as any work that has already been submitted before for any degree or other qualification except as declared in the preface and specified in the text.

It does not exceed the prescribed word limit set by the Degree Committee of the School of Biology.

Anna Orteu

September 2022

Summary

The genomics of adaptive colouration in *Hypolimnas* butterflies and the wood tiger moth, *Arctia plantaginis*

Anna Orteu

Wing phenotypes in butterflies and moths are a striking example of adaptive evolution and are a tractable trait to dissect the genetic mechanisms underlaying adaptations. Studies of Lepidoptera, mainly mimetic species from the tropics, have led to two general patterns. First, a limited set of genes have been re-used multiple times in controlling the evolution of wing phenotypes, and second, structural variation often underlies such traits. Clarifying the generality of these two patterns requires the dissection of the genetic architecture of wing phenotypes in a wider variety of species. In this thesis, I explore the genetic basis of wing mimicry in *Hypolimnas* butterflies and of aposematism in the wood tiger moth using a range of genomics and transcriptomics methods.

First, I produced genome assemblies for two *Hypolimnas* species and investigate the evolution of W chromosomes in Lepidoptera. By comparing the *H. misippus* genome assembly to multiple Lepidoptera species, I provide evidence that suggests that the W chromosome has a shared origin across the Lepidoptera.

Second, I identify the genetic basis of forewing mimicry in *H. misippus* using a dataset of 335 individuals sequenced using haplotagging, a linked read sequencing technique. To analyse these data, I develop a method called Wrath for the visualisation and exploration of candidate structural variants. I find that transposable element insertions are associated with forewing phenotype and present evidence for the usefulness of Wrath to explore haplotagging data.

Third, I examine the evolution of mimicry in the *Hypolimnas* genus by identifying and comparing the genetic basis of wing phenotypes in *H. misippus* and *H. bolina.* Using a dataset of 214 whole genome sequences of *H. bolina* individuals together with my *H. misippus* data, I show that *cortex*, a gene involved in wing colour in many Lepidoptera, is the most likely candidate for the control of white colour elements in the two species. Furthermore, I present

evidence that the regulatory elements controlling the presence of hindwing white are likely not homologous between the two species. Additionally, I show that a region close to *optix,* another well-known colour gene, is associated with orange elements in *H. bolina.*

Finally, I explore the genetic basis of a complex phenotype involving aposematism and behavioural and physiological traits in the wood tiger moth, *Arctia plantaginis,* and show that this trait is associated with the duplication of a *yellow* family gene.

This work contributes to our understanding of the evolution of wing phenotypes in the Lepidoptera. Overall, my results highlight the importance of structural variation in the evolution of wing colouration while also emphasising the repeatability of the genetic basis of adaptive traits.

Acknowledgements

I feel very lucky to have had Chris Jiggins as my supervisor. I am very grateful for his guidance and support, as well as the freedom to explore my own scientific ideas. Chris has created a research environment that is engaging and challenging, but also fun and supportive, which is, in my opinion, the best characteristics that a research group can have—I have thoroughly enjoyed being part of it. I am also very thankful to Simon Martin for his supervision in the late stages of my PhD, his support has been crucial for the completion of this thesis.

I acknowledge the funding that has made possible this research. My PhD was primarily funded by the Natural Environment Research Council (NERC) and the Cambridge Trust, and partly funded by St John's College, the University of Oxford, the Lepidopterists' Society, and the American Society of Naturalists. And fieldwork was funded by the School of Biological Sciences. I am very grateful to all the academic and administrative staff in the Department of Zoology as well as in St John's College.

Being part of the Insect Evolution and Genomics Group has been a wonderful experience. I thank everyone in the group for contributing to create a supportive and fun atmosphere. I am very grateful for Ian's help in the lab, his dad jokes and caring advice. My PhD experience was made infinitely better by the other PhD students in the group, Rachel, Henry, Gabby, Kathy, Jonah, Eva, and Tom—I thank them for being wonderful friends and colleagues.

I am grateful to all the collaborators that have contributed to this PhD. To the collaborators in the *Hypolimnas misippus* project, Ian Gordon, Gerard Talavera, Ivy Ng'iru, Richard Durbin, Shane McCarthy, Dino Martins, Richard ffreanch-Constant, Steve Collins, David Smith, Frank Chan, and Marek Kucka. To the collaborators in the *Hypolimnas bolina* project, Emily Hornett, Louise Reynolds, and Greg Hurst. And the collaborators in the wood tiger moth project, Johanna Mappes, Melanie Brien, and Juan Galarza.

Being part of St John's College has been an experience that I will never forget. I am very grateful to the Children of Lady Margaret, Margherita, Rita, Jasmine, Greg, and Tim, for all the love and fun times, they have made this experience truly wonderful. And to my tutor, David Stuart, who has provided support in crucial times. I am also thankful to the Eagles, the

Cambridge University Association Football Club 2nds, for two great seasons of football and friendship, and to the football community at Pembroke college, who adopted me as one of their own.

I am very grateful to my friends and family back in Barcelona. To my mother, brother, grandmother and Carla for their unconditional support and encouragement. Without them I would have never accomplished this. I thank Berta, Elena, and Júlia, my chosen family, for never getting tired of listening to my stories about Evolution and encouraging me to follow my passion. And to Edu, for always believing in me.

Publications and collaborations

The general introduction chapter and the four data chapters have been written for publication, as individual pieces of work. These have been the result of fruitful collaborations.

For Chapter 1, Prof Chris Jiggins and I wrote the manuscript.

Chapter 1 is published:

- Appendix A: Orteu, A., Jiggins, C.D. The genomics of coloration provides insights into adaptive evolution. *Nat Rev Genet* **21**, 461–475 (2020).

For Chapter 3, *Hypolimnas misippus* sample processing and analysis and draft genome assembly was produced by Dr Shane McCarthy and Prof Richard Durbin at the University of Cambridge. I scaffolded the draft assembly using HiC data. The *H. bolina* samples for this chapter were reared by Dr Emily Hornett at the University of Liverpool. Dr Louise Reynolds (U. Liverpool) extracted the DNA and prepared it for Nanopore sequencing. Mathew Gemmell produced the draft Nanopore assembly. I extracted the DNA and prepared libraries for the *H. bolina* samples with Dr Ian Warren's help and performed the linkage map. I scaffolded and polished the draft *H. bolina* assembly. I dissected the samples for RNAseq, analysed them and annotated the two assemblies. I performed the genome assembly comparisons and sex chromosome analyses with support from Prof Chris Jiggins and Dr Simon Martin.

For Chapter 4, in 2019, I travelled to Kenya twice for a total of 3 months, but a severe drought made butterfly collection impossible. Thus most of the samples were obtained by local collectors Eunice Katili and Charo Ngumbao. Other samples were contributed and/or collected by me, Gerard Talavera, Ian Gordon, Ivy Ng'iru, Steve Collins and Jamie Ball. Dino Martins at the Mpala Research Centre provided the collection and export permits. Wing photographs were taken by Eva van Der Heijden. DNA extractions, haplotagging sequencing and demultiplexing was performed by Dr Frank Chan and Dr Marek Kucka at the Friedrich Miescher Laboratory at the Max Planck Campus in Tübingen, Germany. Dr Simon Martin, at the University of Edinburgh, contributed the demultiplexing code to Wrath, while I developed the structural variant visualisation pipeline. I performed the structural variant analyses of

Heliconius and *H. misippus* haplotagging data, the association study, and the analysis on the evolution of the alleles. The project was supervised by Dr Simon Martin and Prof Chris Jiggins.

For Chapter 5, Dr Emily Hornett, Dr Louise Reynolds, Prof Greg Hurst, Dr Anne Duplouy and Gabrielle Gloder collected and reared the *H. bolina* samples and sequenced dataset 1. I extracted the DNA and prepared libraries for the dataset 2 and performed all the data analyses with support from Dr Simon Martin and Prof Chris Jiggins.

For Chapter 6, samples were provided by Prof Johanna Mappes' research group at the University of Helsinky (formerly University of Jyväskylä), in Finland. Eugenie Yen processed the samples and performed preliminary analyses. Dr Melanie Brien performed the Quantitative Trait Mapping, GWAS and assembly analyses. I dissected the pupal samples, extracted the RNA, and performed the gene expression analyses with support from Prof Chris Jiggins.

Chapter 6 is part of an article available as a preprint:

- Brien, M.N., Orteu, A., Yen, E.C., Galarza, J.A., Kirvesoja, J., Pakkanen, H., Wakamatsu, K., Jiggins, C.D., Mappes, J., 2022. Colour polymorphism associated with a gene duplication in male wood tiger moths. *bioRxiv*.

Contents

‘The genetics of the Diadem [*Hypolimnas misippus*] is complex; it is not for the faint hearted and it is probably no accident that none of the big names in the field of butterfly genetics has ventured to tackle the problem. While it would be disingenuous to pretend that studying this beautiful creature has been a waste of time, it has been far from easy and many unresolved problems remain’

David A. S. Smith (2014)

1

Introduction

The genomics of colouration provides insights into adaptive evolution

1.1 Abstract

Colouration is an easily quantifiable visual trait that has proven to be a highly tractable system for genetic analysis and for studying adaptive evolution. The application of genomic approaches to evolutionary studies of colouration is providing new insight into the genetic architectures underlying colour traits, including the importance of large effect mutations and supergenes, the role of development in shaping genetic variation, and the origins of adaptive variation, which often involves adaptive introgression. An improved knowledge of the genetic basis of traits can facilitate field studies of natural selection and sexual selection, enabling strong selection and its influence on the genome to be demonstrated in wild populations.

1.2 Introduction

The study of colouration in animals and plants has long fascinated biologists (Cuthill et al., 2017). Colouration has a key role in modulating animal and plant fitness through its effects on many aspects of phenotype including courtship and mate preference, predator avoidance through camouflage, photoprotection, structural support, microbial resistance and thermoregulation (reviewed in (Protas and Patel, 2008)). Indeed any given visual cue is commonly under multiple selective pressures, which are often related to both interspecific and intraspecific communication (Cuthill et al., 2017; Quicke, 2017). For example, colouration can be adapted to reduce predation through camouflage, warning colouration or mimicry, while also functioning in pollinator attraction or signalling to conspecifics for mate choice or intrasexual competition (Caro and Allen, 2017; Hill et al., 2006).

Identifying and understanding the genetic basis of adaptations has become a key goal of evolutionary genetics. Determining the molecular mechanisms underlying fitness-related traits can reveal novel insights into the evolutionary forces that shape biodiversity and also provide the explicit link between genetic changes and natural selection. Colour traits have been well studied in ecological and traditional genetic contexts by early evolutionary biologists, so a great deal is known about both their adaptive value and the nature of their genetic control. However, until recently the precise genes and molecular mechanisms underlying colour variation remained largely unknown.

Now, ongoing advances in genomic methods are rapidly progressing our understanding of the genetic basis of colour-related traits (Box 1). In particular, assembly of reference genomes has become much more accessible, both technically and in terms of cost, which has enabled genome-scale analysis of genetic variation for association studies[1] and transcriptomic studies of gene expression. As an example, a recent study of wall lizards completed a reference genome, conducted a genome-scale analysis of a colour polymorphism[2] and identified causal

[1] ASSOCIATION STUDIES: Studies that correlate phenotypic variation with genetic variation. The most common methods are Genome-wide association studies (GWAS) and Quantitative trait locus (QTL) mapping studies.

[2] POLYMORPHISM: The occurrence of two or more distinct phenotypes or morphs of a species within a population.

genes, a process that until recently might have taken a decade of work (Andrade et al., 2019). Genomic methods are therefore permitting a new understanding of the molecular and developmental mechanisms underlying adaptive colouration in a wide variety of systems.

In this Review, we summarise the diverse ways in which organisms produce colours and their adaptive function, and then evaluate how our understanding of evolution has benefitted from studies of adaptive colouration, focusing on three major topics. First, we describe the genetic architecture[3] of adaptive colouration and how even complex adaptations can be controlled by genes of large effect. Second, we discuss the importance of developmental processes in constraining and directing the genetic changes that occur during evolution and perhaps even the overall direction of evolutionary change. Finally, we illustrate how genetic studies of colouration help us understand adaptive evolution more broadly.

1.3 Colouration and its adaptive value

Broadly, colours produced by organisms can be divided into pigmented and structural colours (Figure 1.1). Pigmented colours result from the accumulation of molecules (pigments) capable of selectively absorbing light, and are typically synthesized by the organism in a series of enzymatic steps (Figure 1.1a). By contrast, structural colours depend on physical structures that reflect light. In natural systems, colours often have both structural and pigment components (Shawkey et al., 2009; Shawkey and Hill, 2005). Regardless of type, colouration can be classified as having one of three main functions: signalling, in which visual cues affect interspecific and intraspecific communication; vision, in which pigments are an essential component of the animal's visual system; and physiological, in which pigments are required for a particular task within the organism, such as energy production through photosynthesis using chlorophyll.

[3] **GENETIC ARCHITECTURE:** The genetic basis underlying variation in a phenotypic trait. The main characteristics are the number of loci and their effect sizes, and their position relative to coding and regulatory sequences.

Box 1 | Genomic methods to study adaptive evolution.

Ongoing development of genomic technologies and the fall in sequencing prices has revolutionised the ways in which we study adaptive evolution, and species that in the past were challenging to study because of a lack of resources are now becoming accessible.

Commonly the first step in any study is to assemble a reference genome, often using a combination of long read and short read sequence data (Rice and Green, 2019). A reference genome permits a genome-wide search for loci underlying phenotype variation. For example, short read sequence data from individuals collected in wild populations that differ in colour phenotypes can be used to conduct a genome-wide association study (GWAS)(Pardo-Diaz et al., 2015). Alternatively, genotype-phenotype associations can be tested using laboratory populations in a Quantitative Trait Locus (QTL) mapping experiment (Liu, 1997) Typically, two populations or species are crossed and F2 hybrids analysed for both phenotype and genotype, to investigate associations across the genome. Often, GWAS and QTL mapping identify regions that contain multiple genes, and transcriptomic analyses can be used to identify genes that are differentially expressed between variants, thereby narrowing down an initial list of candidates. Transcriptomic analyse are also facilitated by good reference genomes, which ensure more accurate calling of transcript counts.

Once candidate loci have been identified, functional tests can be used to confirm they contribute to the phenotype. Methods that allow *in vivo* genetic manipulations are now a particularly powerful approach for studying colouration phenotypes. Genome editing technologies, such as those based on the CRISPR-cas9 system, are widely used (Pickar-Oliver and Gersbach, 2019), most commonly to introduce non-functional mutations into coding sequences (gene knock-outs). However, in the future it will be exciting to carry out more targeted and specific genetic manipulations that recapitulate evolutionary changes in natural populations, such as introducing or removing specific alleles from the genome, or even making individual mutations at putatively functional sites. RNAi is another molecular tool that can be used to lower the expression of a gene (knock-down), offering insights into gene function(Wilson and Doudna, 2013). Genomic data can also be combined with tests for

ecological function, such as translocation experiments, to confirm the presence of signatures of selection at the candidate loci (Figure 1.6).

1.3.1 Biological pigments

Pigments are highly diverse molecules that often have physiological functions: in addition to chlorophyll, they are found in haemoglobin in vertebrate blood cells and flavins in vitamin B_2, among others. Some pigments are taxonomically-restricted, such as the psittacofulvins responsible for the bright red, yellow and orange colours of parrot plumage (Cooke et al., 2017). Pterins, which can be white, yellow, orange or red, are commonly found in insects, including in the red eyes of the fruit fly and the orange in some *Pieris* butterfly wings, but are also responsible for the bright colours in some vertebrates, such as the wall lizard (Andrade et al., 2019). The ommochromes are red and yellow pigments that are restricted to invertebrates. They are important for vision in arthropods but have signalling functions in butterfly wings and squid chromatophores. Anthocyanins (red, purple and black) are restricted to plants and are responsible for the colouration of many types of flowers. By contrast, the melanins, eumelanin (brown-black) and pheomelanin (yellow-red), and the carotenoids (yellow, red and orange) are widespread across bacteria, fungi, plants and animals (Figure 1.1b). As well as playing a part in signalling, melanins are also important for thermoregulation and in protection against UV damage to tissues. Plants can synthesize carotenoids, which can be crucial in light harvesting during photosynthesis (Nisar et al., 2015). By contrast, animals were long thought to be unable to synthesize carotenoids, but recent work has shown that aphids can do so using genetic machinery acquired through horizontal transfer from fungi (Moran and Jarvik, 2010), and that birds can transform yellow carotenoids obtained from the diet to red carotenoids through ketolation. CYP2J19, a cytochrome P450 enzyme, has been recently identified as the enzyme responsible for this step (Lopes et al., 2016; Mundy et al., 2016). In birds, carotenoids have an important role in generating bright signals important in sexual selection and may play an important part in honest signalling of male quality.

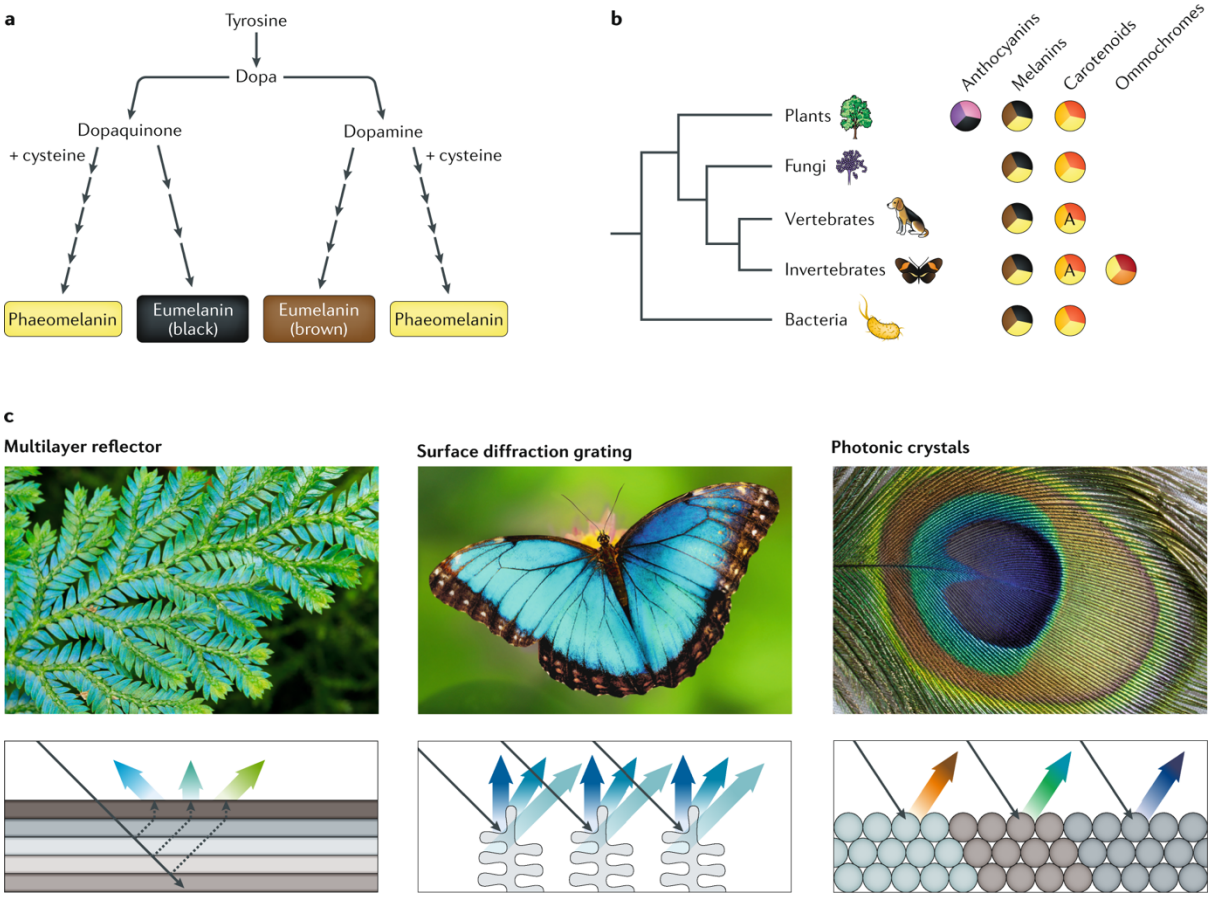


Figure 1.1 | **Different ways of making colour. a** | Pigments are compounds that selectively absorb light and therefore produce colour. In most cases pigments are synthesised biochemically by the organism. A simplified biosynthetic pathway for melanin pigments is shown, in which each arrow represents a biochemical step often catalysed by a different enzyme, providing multiple targets on which selection can act. **b** | The phylogenetic distribution of four common pigment types highlights that some are found in all groups (melanins and carotenoids), whereas others are more restricted (anthocyanins to plants, ommachromes to invertebrates). Colours in the pie charts represent the different colours of each type of pigment. 'A' indicates that the pigment is acquired through the diet, instead of being synthesised by the organism. **c** | Three types of physical structures underlying structural colouration are shown. Multilayer reflectors are found in the leaves of *Selaginella* plants(Hébant and Lee, 1984). Multiple stacked cuticular layers with different optical properties reflect the incident ray of light (solid black arrow) at each transition zone (dotted black lines). Interference of the reflected beams limits the reflected light to blue-green shades (coloured arrows). Surface diffraction gratings are found in Morpho butterflies(Vukusic et al., 1999). The periodical ridged structures found in the scales diffract light into multiple wavelengths. Interference of the diffracted rays from different ridges reduces the light reflected to shades of blue (coloured arrows). Photonic crystals are found in the multi-colour feathers of peacocks. These periodic structures affect the motion of photons such that certain wavelengths are reflected off (coloured arrows) rather than transmitted. Panel **a** is adapted

with permission from REF (Matsuoka and Monteiro, 2018). Panel c: Selaginella vogelii – Credit: Hervé Lenain/Alamy; Blue morpho – Credit: blickwinkel/H. Schmidbauer/Alamy; Peacock Feather – Credit: David Chapman/Alamy

1.3.2 Structural colouration

Structural colours can cover a huge variety of different forms (Figure 1.1c) and, because they are formed by reflected light, they are commonly iridescent; that is, the colour varies depending on the viewing angle. This variation often makes structural colours harder to quantify and study than pigments and, as a result, their developmental and genetic basis remains comparatively poorly understood. Some of the best-known classes of structural colours are multilayer reflectors, surface diffraction gratings, and photonic crystals. The shimmering blue iridescence of *Selaginella* plants is the result of a multilayer reflector (Figure 1.1c), and is caused by the presence of multiple cuticular layers in the leaf (Hébant and Lee, 1984). The distinct layers have different optical properties, which causes part of the light to be reflected in each layer transition. The reflected beams interfere with each other, reducing the spectrum of light reflected to shades of blue. An example of diffraction gratings is the wing scales of the *Morpho* butterfly, in which repetitive ridged structures split the incident ray of light into multiple wavelengths with different reflection angles (Figure 1.1c). The diffracted light from distinct ridges will interfere, limiting the spectrum of light reflected to blue. The combined effect of the vertical multilayer interference within each ridge and the diffraction among the different ridges produces a remarkably efficient wide-angled reflectance of blue light (Vukusic et al., 1999). Photonic crystals are periodic nanostructures that affect the motion of photons to produce coloured reflectance that can be found, for example, in the yellow and green colours of the Kaiser-i-Hind swallowtail butterfly, *Teinopalpus imperialis*, and in the multicoloured feathers of peacocks (Figure 1.1c).

1.4 Colour and adaptation

In this review, we focus on colours used in signalling, where the link between diversity in colouration and its adaptive value is well understood. Signalling through visual cues is essential for interspecific and intraspecific communication in animals and consequently colour attributes are under multiple selective pressures (Cuthill et al., 2017). For example, colouration commonly plays a role in sexual selection, with bright colours used as indicators of high-quality males (Figure 1.2a). Colouration can be involved in many other social functions, for example female primates can use colour signals to indicate sexual receptivity, such as in the Celebes crested macaque (*Macaca nigra*) (Figure 1.2b).

In addition, many predators rely on their vision to detect prey, thus camouflage can be a successful strategy for predator avoidance. Common forms of camouflage include background-matching through cryptic colouration (such as that seen in snowshoe hares and pygmy seahorses) (Figure 1.2c), disruptive colouration to break the body outline (such as zebra stripes and leopard spots), and masquerading as a common inedible object (as is seen in stick insects and leaf-mimicking butterflies)(Quicke, 2017). Escaping predation can also be achieved by predator deterrence through warning colouration or aposematism, in which unpalatable prey announce their toxicity to predators by displaying conspicuous colour patterns. Striking colours increase visibility and help predators remember the toxicity of the prey in future encounters. In some cases, toxic species share aposematic colouration and, by resembling one another, they reduce the cost of teaching predators, a phenomenon known as Müllerian mimicry (Vane-Wright, 1976). An example is the *Ranitomeya* genus of poison dart frogs, in which the four morphs of the toxic *R. imitator* resemble multiple toxic models from the same genus (Figure 1.2d)(Twomey et al., 2015). In other cases, palatable species resemble toxic ones to gain protection against predators, which is known as Batesian mimicry. For instance, the palatable African butterfly *Hypolimnas misippus* has very similar colour patterns to morphs of the unpalatable African Queen, *Danaus chrysippus* (Figure 1.2e)(Smith, 1976).

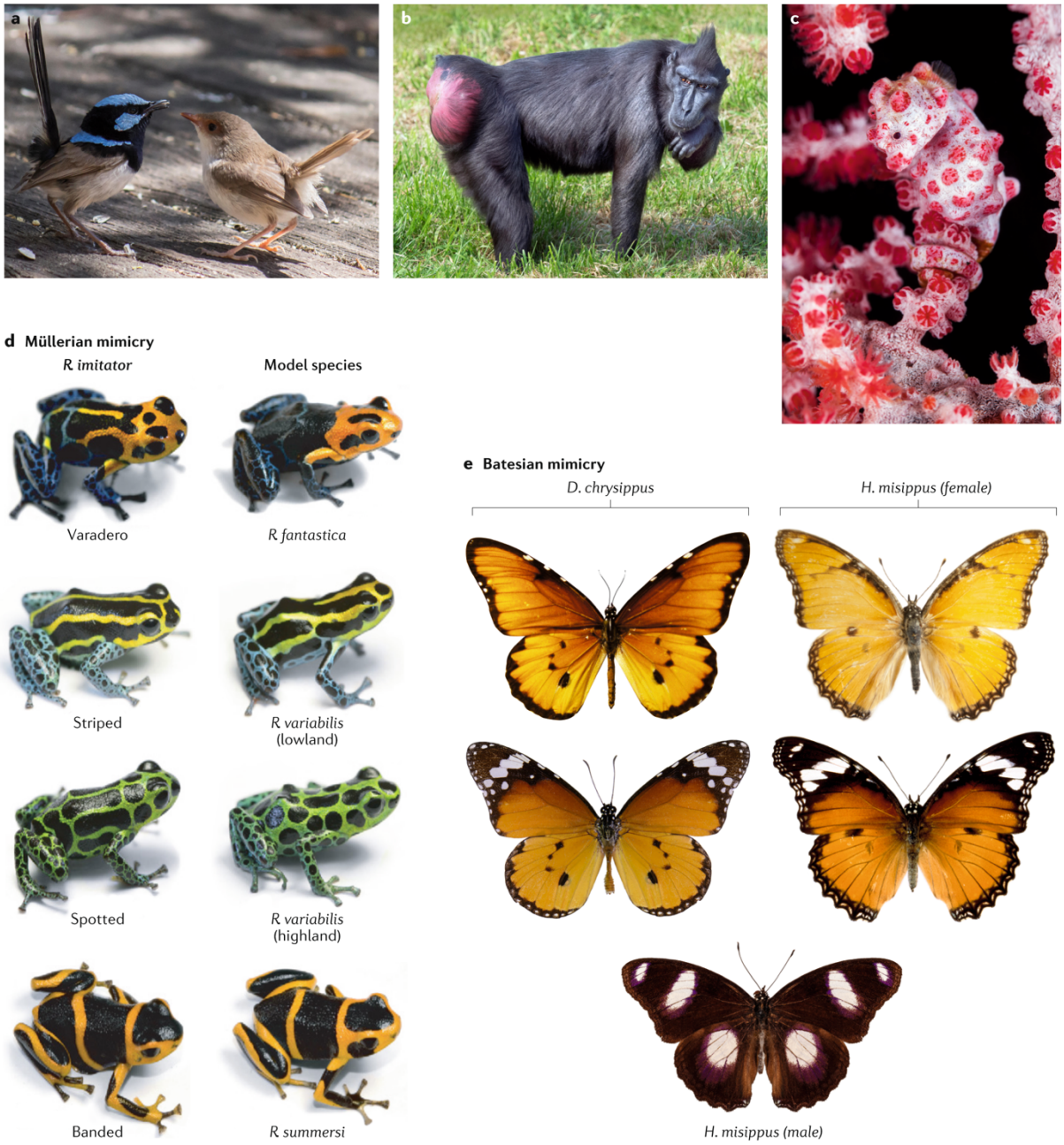


Figure 1.2 | **Signalling through colour**. **a, b** | An important function of colour is signalling to conspecifics, especially to possible future mating partners. During the mating season, male superb fairy-wrens (*Malurus cyaneus*) adopt conspicuous breeding plumage of contrasting shades of iridescent blue and black to attract females (panel **a**). The prominent red swelling of the female Celebes crested macaque (*Macaca nigra*) advertises their sexual receptivity (panel **b)**. **c** | Concealment through background matching allows animals to escape predation. The pygmy seahorse *Hippocampus bargibanti* has detailed resemblances to the gorgonian coral it inhabits. **d, e** | Warning colouration or aposematism is another effective way to avoid predation: toxic prey display conspicuous colour patterns that warn predators of their toxicity. Some aposematic species converge to the same colour pattern and therefore share the costs of teaching predators, which is known as Müllerian mimicry (Vane-Wright, 1976).

The *Ranitomeya* genus of poison dart frogs provide one example. Multiple toxic *Ranitomeya* species act as models for the four morphs of the similarly toxic *R. imitator* (Twomey et al., 2015) (panel **d**). By contrast, in Batesian mimicry a palatable species mimics a toxic one to gain protection against predators. For example, the African butterfly *Hypolimnas misippus* resembles the unpalatable morphs of *Danaus chrysippus* (panel **e**). Permissions/credits: Panel a: Superb Fairywren – Credit: phototrip/Alamy. Panel b: Crested macaque – Credit: Ernie Janes/Alamy. Panel **c:** Pygmy Seahorse (Hippocampus Bargibanti) – Credit: Shahar Shabtai/Alamy. Panel **d** is adapted with permission from (Twomey et al., 2015). Panel e: African Monarch Butterfly (Danaus Chrysippus) – Credit: Vonkara1/iStock/Getty; Butterfly Danaid Eggfly (Hypolimnas misippus) – Credit: Paolo_Toffanin/iStock/Getty; Tiger Butterfly (Danaus chrysippus) – Credit: komkrit tonusin/Alamy; Hypolimnas misippus, diadem butterfly – Credit: The Natural History Museum/Alamy; Hypolimnas misippus male Danaid Eggfly – Credit: Domiciano Pablo Romero Franco/Alamy

1.5 Genetic architecture of colour traits

The genetic architecture of a trait describes the number and effect size of genes that influence variability within and between populations and its patterns of linkage and interaction with other traits. The genetic architecture can influence the potential for a trait to undergo evolutionary change and the degree to which different traits are genetically associated.

1.5.1 Mutational effect size

One question of general interest to evolutionary geneticists is the effect size of loci involved in adaptation. Mutations of large effect should make evolutionary change easier and faster, but conversely they are arguably more likely to be deleterious (Fisher, 1930). Furthermore, a body of theory predicts that a single bout of adaptation will involve loci with a distribution of effect sizes, with a few large effect and many small effect mutations expected (Orr, 2005, 1998). Both association studies and quantitative trait loci (QTL)[4] analyses can provide estimates of the effect sizes of individual loci (in terms of the amount of phenotypic variation explained by a particular locus), while genetic manipulation can investigate the importance of individual mutations towards a particular phenotypic change (Box 1). Such tools have confirmed that many recent adaptive changes in colouration do indeed involve a few loci of

[4] **QUANTITATIVE TRAIT LOCI (QTL):** Genomic regions at which there is a correlation between genetic variation and phenotypic variation in a trait of interest, among individuals from a laboratory-generated cross.

large effect, a few of which have been characterized at the level of individual mutations. In particular, genes encoding proteins involved in pigment synthesis pathways provide a large mutational target and simple genetic changes in these genes can have major effects on phenotype. Indeed, there are now many examples of large-effect coding sequence mutations in pigmentation genes, in both animals and plants, that influence adaptive colouration (Martin and Courtier-Orgogozo, 2017; Martin and Orgogozo, 2013). For example, the genetic basis of feather colour variation in budgerigars has been narrowed down to a single nucleotide polymorphism (SNP) in the coding region of a polyketide synthase gene, *MuPKS*, which has been shown to produce yellow pigment when heterologously expressed in yeast (Figure 1.3a)(Cooke et al., 2017). Similarly, island populations of flycatchers have independently evolved dark colouration through coding mutations in either the *MC1R* gene (which encodes the melanocortin 1 receptor) or the *Agouti* gene (which encodes the agouti signalling peptide)(Uy et al., 2016), and *in vitro* tests have confirmed that the observed coding changes in *MC1R* disrupt protein function (Hoekstra et al., 2006; Römpler et al., 2006). *MC1R* is a transmembrane receptor that, when activated, triggers a signalling cascade that results in the production of brown or black eumelanin. *MC1R* can be antagonized by *Agouti,* which reverts the cascade to the default production of yellow phaeomelanin. Similarly, a flower colour polymorphism in the Alpine orchid *Gymnadenia rhellicani* is regulated by a mutation that introduces a premature stop codon in a *R2R3-MYB* transcription factor. Transcription factors of this family have been shown to regulate expression of anthocyanin biosynthesis in multiple plant species (Albert et al., 2014; Yuan et al., 2016). In this alpine orchid, the loss of function of the transcription factor results in lower expression of an anthocyanin synthase, which controls the production of dark purple anthocyanidin pigments, and thus results in reduced accumulation of pigment and whiter colouration. This mutation is maintained in the population by heterozygote advantage[5], with both bee and fly pollinators attracted to the heterozygote red morph leading to higher fitness of that genotype (Kellenberger et al., 2019).

Large-effect mutations are not only found in coding sequences. In common canaries (*Serinus canaria*), a single regulatory mutation in the splice donor site of the SCARB1 gene causes a

[5] **HETEROZYGOTE ADVANTAGE:** A scenario in which the heterozygous genotype is fitter than the homozygous recessive or homozygous dominant genotypes.

change of plumage colouration from wild-type yellow to white (Toomey et al., 2017); SCARB1 encodes a membrane receptor for high density lipoprotein and is a mediator of carotenoid uptake. The SNP in the splice donor site causes abnormal splicing, leading to exon skipping and the loss of gene function. Similarly, compelling population genomic evidence suggests that a single transposable element (TE)[6] insertion in an intron of the *cortex* gene gave rise to the British melanic form of the peppered moth (*Biston betularia*), which, driven by natural selection, spread rapidly through industrial Britain; however, further studies are needed to prove the causal role of this mutation (van't Hof et al., 2016). Interestingly, variation at *cortex* (and two other loci, *WntA* and *optix*) has also been associated with the diverse colour patterns of *Heliconius* butterfly species and races (Mazo-Vargas et al., 2017; Nadeau et al., 2016; Reed et al., 2011; Westerman et al., 2018).

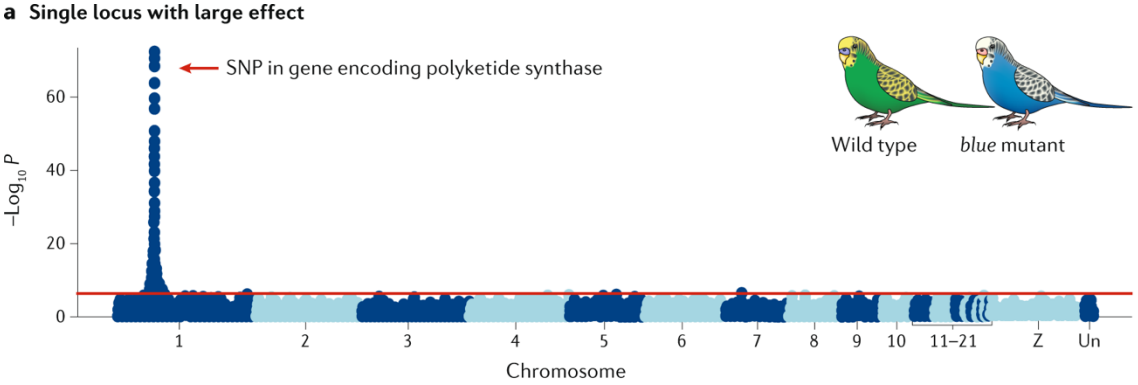
In other cases, alleles with major effects on phenotype can involve *cis*-regulatory alleles that encode complex spatial expression patterns. For example, differences in stripe patterning between two closely related species of the cichlid fish (*Pundamilia nyererei* and *Haplochromis sauvagei)* are controlled by a 1.1kb cis-regulatory element (CRE)[7] at the *agrp2* gene, which encodes agouti-related peptide 2, a protein that belongs to the Agouti family and is involved in melanosome aggregation (Kratochwil et al., 2018; Zhang et al., 2010). Similarly, *cis*-regulatory changes at the *LAR1* gene underlie spatial pattern variation in floral colours between two species of monkeyflowers, *Mimulus lewisii* and *Mimulus cardinalis*. The R2R3-MYB transcription factor encode by *LAR1* positively regulates the expression of the *Flavonol synthase* enzyme, which redirects metabolite flux from the production of pink anthocyanins to the biosynthesis of colourless flavonols, resulting in white petal in areas where *LAR1* is expressed (Yuan et al., 2016).

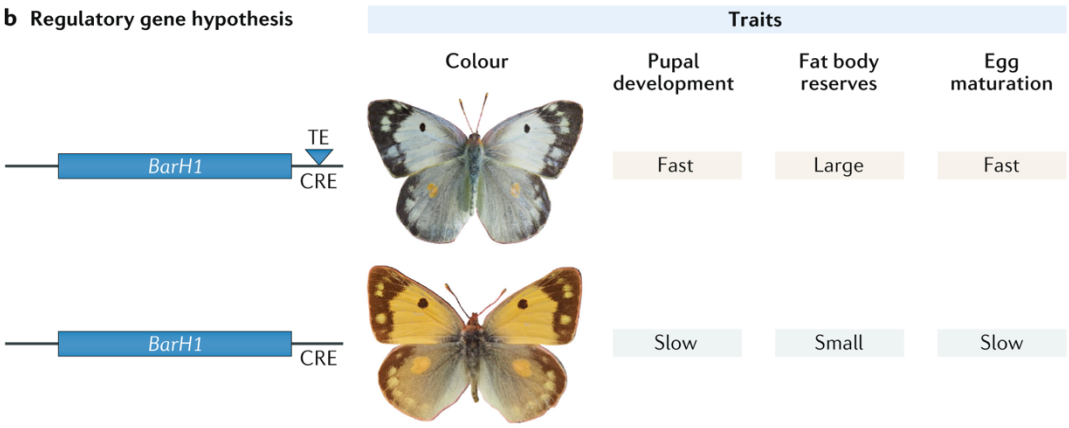
In summary, and somewhat contrary to the gradualist expectations of many evolutionary biologists, it is surprisingly common that evolutionary change is controlled by alleles, and even single mutations, with large effects.

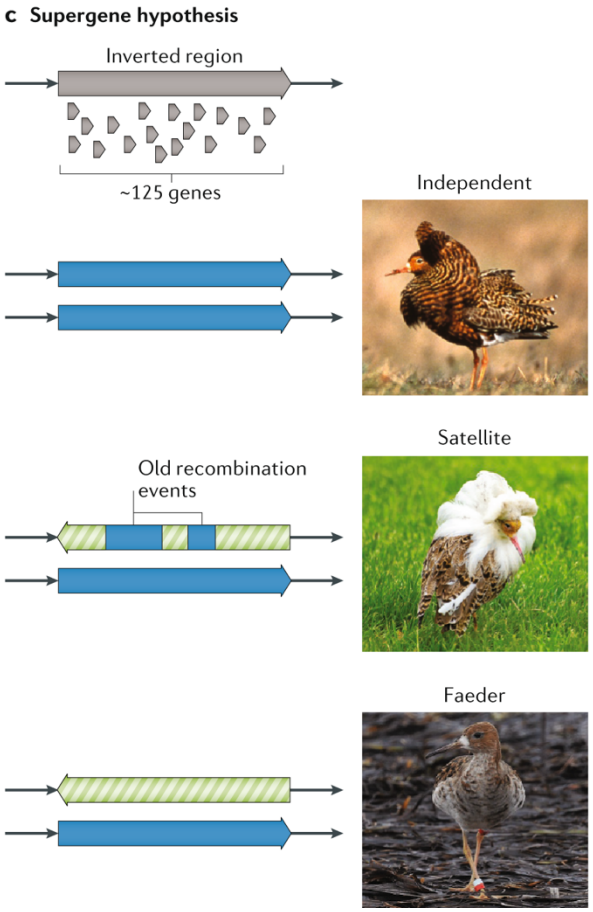
---

[6] TRANSPOSABLE ELEMENT: A genetic element that can move from one position in the genome to another.

[7] CIS REGULATORY ELEMENT (CRE): Genetic regions that regulate expression of a coding sequence on the same DNA strand.


a Single locus with large effect
SNP in gene encoding polyketide synthase
Wild type
blue mutant
–Log10 P
0
20
40
60
1
2
3
4
5
6
7
8
9
10
11–21
Z
Un
Chromosome


b Regulatory gene hypothesis
Traits
Colour
Pupal development
Fat body reserves
Egg maturation
BarH1
TE
CRE
Fast
Large
Fast
BarH1
CRE
Slow
Small
Slow


c Supergene hypothesis
Inverted region
~125 genes
Independent
Old recombination events
Satellite
Faeder

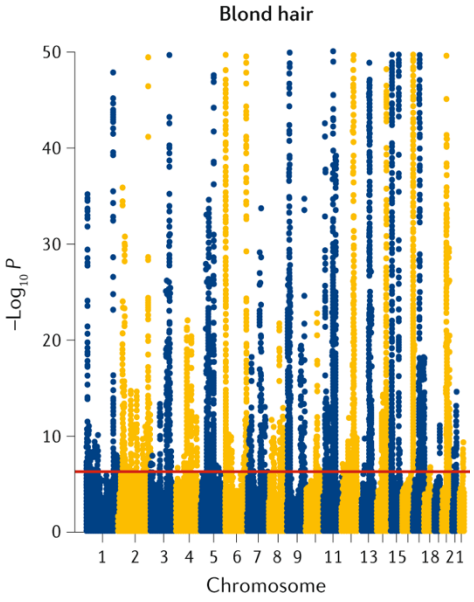

d Polygenic inheritance
Blond hair
–Log10 P
0
10
20
30
40
50
1
2
3
4
5
6
7
8
9
11
13
15
18
21
Chromosome

Figure 1.3 | **Genetic architecture of colour loci. a** | In many cases, the genetic basis of colour polymorphism has been narrowed down to single mutations of large effect. Shown is the results of a genome-wide association study (GWAS) to identify single nucleotide polymorphisms (SNPs) associated with colour morphs in budgerigars . Only one significant peak is detected, located on chromosome 1. The red line indicates the Bonferroni-corrected critical value (α = 0.05). Fine mapping and functional analysis confirmed that the causative variant associated with the SNP is a nonsynonymous mutation in *MuPKS*, a gene encoding a polyketide synthase, that abolishes yellow pigmentation and produces the blue morph (Cooke et al., 2017). **b** | The regulatory gene hypothesis for the inheritance of complex adaptive phenotypes states that a single gene is responsible for multiple co-inherited. An example that potentially supports this hypothesis is *Colias* butterflies, which present two morphs that differ in wing colour as well as numerous life-history related traits (Woronik et al., 2019). These forms are controlled by a single gene, *BarH-1*, which is over-expressed in the white morph during development because of a TE insertion in its *cis*-regulatory element (CRE). *BarH-1* supresses orange colouration in the wings and, either directly or as energetic side-effect, triggers the switch in multiple life-history traits. **c** | The supergene hypothesis states that a complex polymorphism is controlled by a single locus containing numerous genes that are inherited together owing to reduced levels of genetic recombination (often caused by an inversion). In the ruff, three morphs differing in body size, ornamentation and mating behaviour are controlled by a supergene (large grey arrow) that contains approximately 125 genes (small grey arrows). The supergene has 3 alleles (large blue and green arrows): an ancestral allele (blue) that is found in homozygosis produces the 'independent' morph; a derived allele (green) that can only be found in heterozygosis (because it is lethal in homozygotes) produces the 'faeder' morph; and a recombinant allele that is also found in heterozygosis produces 'satellites'. Hatched regions indicate reduced recombination due to the inversion. **d** | In other cases, colour variation is highly polygenic. The plot shows the results of a GWAS for loci associated with blonde hair colour in humans, which is influenced by many loci scattered across the genome (Morgan et al., 2018). The red line indicates the thereshold of significance ($\alpha = 5x10^{-8}$) Panel **a** is adapted with permission from (Cooke et al., 2017). Panel **b** is adapted with permission from (Woronik et al., 2019). Panel **c** is adapted with permission from (Jiggins, 2016). Panel **d** is reproduced with permission from (Morgan et al., 2018).

1.5.2 Complex adaptive polymorphisms

Some of the most compelling examples of adaptive colouration controlled by major effect loci are polymorphisms, particularly those involving complex phenotypes comprising multiple traits that are locally stable. Polymorphisms can be maintained if their selection coefficient[8] varies through time, across space or with local allele frequency or population density. For example, alternative mating strategies adopted by different phenotypic morphs can be subject to negative frequency dependent selection[9] such that no single allele becomes fixed and multiple distinct morphs are maintained, as illustrated by the ruff, a wading bird in which alternative male forms differ in physiology, body size, behaviour and plumage (Küpper et al., 2015; Lamichhaney et al., 2016; Widemo, 1998). However, the stability of such complex polymorphisms depends on a genetic architecture in which linkage between the multiple co-adapted traits ensures their joint inheritance. Early theoretical discussion focused on two possible genetic architectures that could stabilise such polymorphisms—the 'regulatory gene' hypothesis and the 'supergene' hypothesis (reviewed in (Thompson and Jiggins, 2014)).

The regulatory gene hypothesis states that changes in a single 'master' gene, acting at the top of regulatory gene networks and with multiple downstream targets, could lead to morph variation between individuals (Thompson and Jiggins, 2014). Several examples of complex polymorphisms associated with colour traits are now known to be controlled by a single master gene. In *Papilio* butterflies, the *doublesex* locus encodes a transcription factor that regulates both sex determination and a female-limited colour polymorphism (Iijima et al., 2018; Kunte et al., 2014; Nishikawa et al., 2015). *Colias* butterflies exhibit two alternative life history strategies differing in wing colour (and a suite of other traits) that are maintained by an energetical trade-off (Figure 1.3b)(Graham et al., 1980). White butterflies invest their resources in reproductive and somatic development instead of producing expensive colour pigments, which results in faster pupal development, larger fat body reserves, and faster egg maturation compared with orange females. This polymorphism has been mapped to a

[8] **SELECTION COEFFICIENT:** A measure of the difference in fitness between two genotypes, which are a necessary condition for the action of natural selection

[9] **NEGATIVE FREQUENCY DEPENDENT SELECTION:** An evolutionary process in which the fitness of a genotype or phenotype depends on its frequency in the population relative to other genotypes or phenotypes such that its fitness decreases as its frequency increases.

transposable element insertion in the white morph (also known as *alba*), downstream of *BarH-1*, a homeobox transcription factor that also affects eye colour in *Drosophila melanogaster* (Woronik et al., 2019). *BarH-1* is expressed in white but not orange forewing scales and CRISPR-Cas9 knock-out experiments have been shown that it supresses orange colouration in the wing. It is not known whether *BarH-1* interacts with other genes to control the developmental and physiological traits characteristic of the white form or whether the differences between morphs in life-history traits are a side-effect of the suppression of colouration, with the energy saved by reducing the production of pigment granules on the wings reallocated to reproductive and somatic development. In the Gouldian finch, two alternative colour morphs, black and red, differ in multiple traits including social dominance, stress hormone levels, sperm-length plasticity and sex allocation (Kim et al., 2019; Toomey et al., 2018). These morphs have been associated with a small intergenic region (~72-kbp) of the Z chromosome that likely regulates follistatin, a glycoprotein that antagonises TGF-*β* superfamily function and has been indirectly linked to variation in plumage colour in flycatchers and warblers (Lehtonen et al., 2012; Toews et al., 2016). Colour polymorphism in the wall-lizard described above also includes behavioural and ecological differences between morphs, and is controlled by two small regulatory regions near genes associated with pterin and carotenoid metabolism (Andrade et al., 2019). By contrast, in some plants a single transcription factor can regulate multiple pigmentation enzymes that, for example, control major differences in flower hue (Hopkins and Rausher, 2012, 2011); however, such transcription factors typically do not control balanced polymorphisms.

The supergene hypothesis proposes that a single locus (that is, the supergene) contains multiple functional elements - potentially multiple genes - that control a complex, adaptive phenotype (Schwander et al., 2014; Thompson and Jiggins, 2014). A key element of the supergene hypothesis is the maintenance of linkage between co-adapted loci. Although it was initially suggested that translocation of genes might have a role in establishing closer linkage (Ford, 1965), empirical evidence suggests that instead the occurrence of successive co-adapted mutations in close proximity is more important, a process known as Turner's sieve (Turner, 1967). Linkage is therefore assured through physical proximity in the genome. There may also be ongoing selection for further reduction in recombination, either through localisation to genomic regions with naturally low recombination rates, such as centromeres

(Schwander et al., 2014), or though accumulation of structural variation, such as TE insertions, tandem duplication and inversions (Iijima et al., 2018; Martin et al., 2020). However, reduced recombination itself could lead to an increase in structural mutations due to the reduced efficacy of purifying selection[10], so the direction of causality may be unclear. Inversions are the most common mechanism for reducing recombination and are associated with many supergenes underlying colour polymorphisms in birds (including the white-throated sparrow and the ruff) and in butterflies (such as *Heliconius numata*)(Joron et al., 2011; Küpper et al., 2015; Lamichhaney et al., 2016; Tuttle et al., 2016). Inversions can link multiple traits, such as the colouration, developmental, physiological and behavioural traits that differ between ruff alternative mating morphs (Figure 1.3c)(Küpper et al., 2015; Lamichhaney et al., 2016).

Understanding supergene architecture can shed light on their long-term maintenance. Polymorphism has generally been explained through ecological mechanisms, such as frequency dependent fitness and environmental heterogeneity, but an intriguing possibility is that intrinsic genetic mechanisms contribute to the maintenance of variation in ecologically important traits. Balanced lethal systems are genomic regions in which the accumulation of deleterious mutations on alternative haplotypes ensures heterozygote advantage and the long-term maintenance of variation (Muller, 1918). Supergene alleles associated with colour polymorphisms can similarly accumulate deleterious variation, owing to a small effective population size[11] and low recombination. For example, there is evidence that the supergene alleles in *H. numata* have higher TE activity and severe deleterious fitness effects in homozygous genotypes compared with heterozygous genotypes (Jay et al., 2019). Similarly, dominant supergene alleles in the ruff are lethal in homozygotes (Küpper et al., 2015; Lamichhaney et al., 2016). The reduced fitness of homozygotes effectively leads to heterozygote advantage, irrespective of the relative fitness of colour phenotypes, and could play an important part in maintaining dramatic colour polymorphisms. These examples demonstrate how a better understanding of the underlying genetics can inform our interpretation of the evolutionary processes maintaining variation in natural populations.

[10] **PURIFYING SELECTION:** The selective removal of deleterious alleles from the population.

[11] **EFFECTIVE POPULATION SIZE:** The number of individuals in an idealised population that would show the same degree of genetic drift as seen in the real population.

1.5.3 Polygenic adaptation

The ability to conduct genomic studies on larger populations through developments in high-throughput sequencing has led to an increasing interest in the genetic basis for highly polygenic inheritance[12], but it has generally not been well explored in relation to colouration. In traits where major effect loci have been identified as underlying adaptive change, there is often also a proportion of phenotypic variation controlled by loci of small effect, such as the distribution of effect sizes among colour forms of *Heliconius* butterflies (Papa et al., 2013). In other cases, variation in populations is largely polygenic. For example, a genome-wide association study (GWAS)[13] identified more than 150 genes influencing variation in abdominal pigmentation among lines derived from a single *D. melanogaster* population, including many previously uncharacterised genes that were confirmed to influence pigmentation using RNA interference (RNAi)-based targeted gene knockdown (Box 1)(Dembeck et al., 2015). Similarly, structural colours in butterflies have a polygenic basis (Brien Melanie N. et al., 2019), and in humans, analysis of traits such as hair colour have demonstrated the power of large-scale GWAS to identify hundreds of small effect loci influencing highly polygenic traits (Figure 1.3d)(Hysi et al., 2018; Morgan et al., 2018). Colouration traits can therefore offer a tractable system for the analysis of polygenic architecture and this is clearly an area with potential for future work. One specific future goal is to understand the underlying cause of different genetic architectures. It is possible that adaptive differences between populations and species in traits under strong directional selection will be more likely to be under the control of major effect loci compared with traits under stabilising selection[14], but this hypothesis will need to be tested by comparative analysis across many species.

[12] **POLYGENIC INHERITANCE:** Also known as polygenicity. The genetic control of a phenotype by multiple genes of small effect.

[13] **GENOME-WIDE ASSOCIATION STUDY (GWAS):** A study that correlates genetic variation between individuals across the genome with phenotypic variation among those same individuals, typically in a wild population. Associated regions are inferred to contain causal variants controlling phenotypic variation. More generally known as 'association studies'.

[14] **STABILISING SELECTION:** A selective force that maintains the population phenotypic mean and eliminates extreme values.

In summary, colouration traits demonstrate a wide range of genetic architectures, ranging from highly polygenic through to major effect loci and adaptive supergenes. While considerable progress has been made in describing these architectures in specific cases, major outstanding questions remain over the reasons for differences between species: How common are supergenes and why do they arise in some species but not others? When are major loci expected to regulate adaptive change? And, are there circumstances in which a polygenic architecture is more likely? One way to address these questions is to understand how genetic changes regulate development, and in turn how developmental processes might shape genetic architecture.

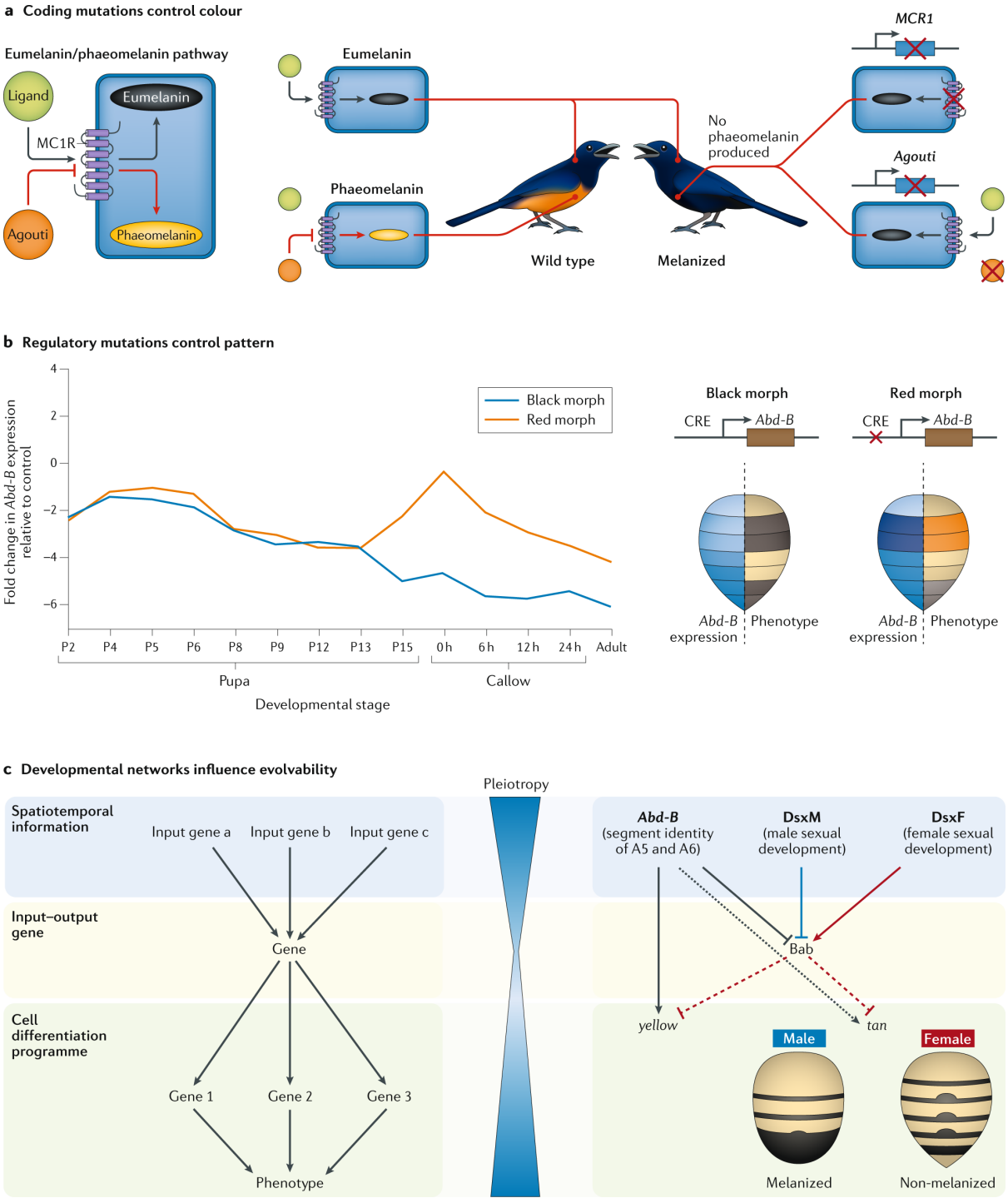


Figure 1.4 | Development can shape evolution. a | Colour is often controlled by coding mutations, particularly in genes found in the biochemical pathways that lead to the production of pigments. In the melanin synthesis pathway, ligand-bound *MC1R* activates the production of black and/or brown eumelanins, whereas if bound to its antagonist *Agouti,* the default production of pheomelanin is restored. The island flycatcher presents two morphs, one with chestnut belly plumage resulting from the production of pheomelanin, and one with only black plumage as a consequence of only eumelanin being produced. Two different mutations can produce the black morph: a mutation in *MC1R*, which activates the receptor

and leads to the production of eumelanin; and a mutation in *Agouti*, which abrogates its function as an *MC1R* antagonist and allows the ligand to bind to MC1R, activating the production of eumelanin (Uy et al., 2016). **b** | Colour pattern is often under the control of regulatory mutations. In bumble bees, a *cis*-regulatory element (CRE) mutation in the Abd-B gene changes its pattern of gene expression late in development (at the callow stage), which leads to a change in abdominal colour pattern from black to red (Tian et al., 2019). **c** | Input-output genes are found in the middle of developmental networks. They integrate complex spatiotemporal information from patterning genes and, as a result, trigger different cell differentiation programs by regulating downstream genes, which leads to distinct phenotypes (Stern and Orgogozo, 2009). Although patterning genes can trigger the development of phenotypes, they are likely to have pleiotropic effects. By contrast, some genes (such as input-output genes) may be able to initiate differentiation programs while showing low pleiotropy, which may make them more likely to be involved in evolutionary change. A potential example of such genes is *Bric à brac 1* and *2* (*Bab*) in *Drosophila melanogaster*, which integrate information about the abdominal segment identity from *Abd-B* and sex information from the female (F) and male (M) specific isoforms of *Doublesex* (*Dsx*), and in response regulate the pigment-related genes *tan* and *yellow* to control melanisation(Rogers et al., 2013). Segments A5 and A6 are melanised in males (*Bab* expression is repressed, allowing *yellow* and *tan* to be expressed (blue and black arrows)), whereas in females they are not (*Bab* is expressed and represses expression of yellow and tan (red arrows)). Solid lines indicate direct interactions, while dashed lines indicate indirect interactions. Panel **b** is adapted with permission from (Tian et al., 2019).

1.6 How does development shape evolution?

A greater understanding of evolution comes from considering the role of development in shaping and potentially channelling the evolutionary process. Colouration provides multiple examples in which mutations seem to be distributed non-randomly across the genome, with convergence[15] in phenotype often influenced by the same genes or the same types of mutations in different lineages. Studies of these examples provide insight into the relative importance of coding versus regulatory change, the extent to which the same genes are re-used in evolution, and the influence of the environment in the development of phenotypes.

[15] **CONVERGENCE:** The independent evolution of similar features in different lineages or species.

1.6.1 Coding changes control colour but regulatory changes control pattern

Morphological variation can be produced by changes in coding or *cis*-regulatory regions of genes. It has been proposed that coding sequence changes in genes that have multiple functions, especially those involved in signalling pathways that are deployed in multiple developmental contexts, will commonly be deleterious: even if a mutation is beneficial to one of its functions, it may be detrimental to another (Carroll et al., 2005). Nonetheless, coding mutations do commonly play a role in colouration, but they tend to have large phenotypic effects because any tissue or developmental stage expressing the gene will be affected. For example, flycatchers homozygous for either of the previously mentioned coding mutations in the *MC1R* and *Agouti* genes lack the ability to produce pheomelanin and therefore have only black plumage (Figure 1.4a)(Uy et al., 2016).

By contrast, subtle differences in local patterning are likely to be caused by regulatory changes, which can be more modular than coding changes(Martin and Courtier-Orgogozo, 2017); beneficial mutations that arise are therefore less likely to have negative pleiotropic effects on other functions (Carroll et al., 2005). Modularity in gene expression can be temporal and/or spatial in nature. For example, the evolution of dark *D. melanogaster* in montane habitats is the result of a tissue-specific regulatory change that reduces expression of *ebony*, a gene that supresses the production of dark melanin (Rebeiz et al., 2009). Studies using GFP reporter constructs showed that this dark regulatory allele reduces expression of *ebony* in the abdomen, but not in other tissues, such as the head and halteres, when compared to the light regulatory alleles. Indeed, regulatory mutations are responsible for every evolutionary pigmentation change in *Drosophila* spp. that has been mapped to date (Massey and Wittkopp, 2016), supporting the hypothesis that phenotypic variation is, on many occasions, controlled by *cis*-regulatory mutations. In another example, a GWAS with wild samples showed that a red-to-black colour switch in the bumble bee *Bombus melanopygus* is controlled by *cis*-regulatory changes in a ~4kb intergenic block downstream from *Abd-B. Abd-B* is a Hox gene that determines abdominal segment identity during development, which has been shown to positively regulate the pigmentation gene *yellow* and the transcription factors *bric à brac 1* and *bric à brac 2* (together referred to as *Bab).* Quantitative PCR analysis demonstrated that this regulatory change specifically expands the

highly conserved expression domain of the *AbdB* gene late in development; earlier functions of the AbdB transcription factor in determining abdominal fate are unaffected (Figure 1.4b)(Tian et al., 2019). Similarly, the gene *optix* encodes a transcription factor that is involved in brain and eye development in *D. melanogaster* (and presumably in all insects)(Seimiya and Gehring, 2000), but in nymphalid butterflies it is also expressed during wing development and acts to paint red colours onto the wings, with *cis*-regulatory alleles defining the precise patterns to be painted in any particular species. It seems likely that co-option of *optix* into wing patterning has occurred through gene regulatory changes that do not disrupt its existing functions in structures such as the eye and brain.

However, the simple model in which CREs are highly modular and tissue-specific is not always confirmed by experimental studies; indeed, many individual CREs control expression in multiple different tissues (Jory et al., 2012). For example, *optix* regulatory elements are associated with specific wing colour pattern elements (such as red rays or bands) in *Heliconius* populations, which is consistent with modularity. Yet, knocking out one of these CREs using CRISPR-Cas9 gene editing results in a broad phenotype that involves multiple pattern elements, which suggests pleiotropy[16] rather than modularity (Lewis et al., 2019). This result is consistent with earlier analysis of pigmentation in *D. melanogaster*, which used reporter transgenes to show that individual mutations within a CRE can have highly tissue-specific effects and avoid deleterious pleiotropy, even when the CRE is multi-functional (Rogers et al., 2013). This observation highlights the fact that it is the level of pleiotropy of a mutation that affects the likelihood of fixation of a novel variant, not the degree to which a particular gene or CRE has multiple functions. In other words, CREs may be developmentally multi-functional, affecting expression in multiple tissues, but evolutionarily modular in that specific mutations are tissue-specific. It is therefore misleading to label a gene or a CRE as being 'highly pleiotropic' - the term should be applied to mutations rather than functional elements themselves (this more restrictive definition of pleiotropy has recently been termed 'selectional pleiotropy' (Paaby and Rockman, 2013)).

[16] **PLEIOTROPY:** The effect of a single mutation on multiple aspects of the phenotype.

1.6.2 Developmental networks influence which genes are evolutionary hotspots

A pattern emerging from analysis of the genetic basis for adaptive change in colouration and other phenotypes, is the repeated co-option of the same genes into convergent evolutionary change, a phenomenon that implies constraint on which genes are likely to influence a particular phenotype. For example, linkage and association mapping has identified mutations in *Oca2* that underlie amelanic phenotypes in at least two populations of cave fish and in a domesticated corn snake population, and human albinism and skin pigmentation phenotypes in African populations (Crawford et al., 2017; Johanson et al., 2010, p. 2; Martin and Orgogozo, 2013; Saenko et al., 2015). *Oca2* encodes a membrane transporter protein thought to be involved in uptake of tyrosine, a melanin precursor molecule and, given it has a specific melanisation function, it is perhaps unsurprising that it is repeatedly co-opted to regulate adaptive colouration (Bilandžija et al., 2013). Similarly, MC1R and its antagonist Agouti control a signalling cascade leading to melanin production and have been associated with melanin-based colouration in many species, including deer mice (Linnen et al., 2013, 2009), white-sand lizards (Rosenblum et al., 2009) and island flycatchers (Uy et al., 2016) among others (Figure 1.4a). It is more surprising that genes with multiple roles in development, such as transcription factors and signalling molecules, can be repeatedly targeted during evolutionary change – yet there are now many such examples, including WntA, a Wnt-signalling pathway ligand that has been shown by both association studies and functional analysis to be repeatedly involved in establishing colour patterns among butterflies (Martin et al., 2012; Mazo-Vargas et al., 2017). One hypothesis is that the shape of developmental networks might constrain which genes are most free to evolve, with 'input-output' genes[17] lying at developmental switch points being most likely to regulate novelty (Figure 1.4c), but the generality of this hypothesis needs to be tested(Stern and Orgogozo, 2009). A likely example of input-output genes are the *Bab* transcription factors, which integrate segment and sex information to trigger sex and segment specific colouration in various *Drosophila* species (Figure 1.4c)(Dembeck et al., 2015; Massey and Wittkopp, 2016).

[17] **'INPUT-OUTPUT' GENES:** Genes that integrate complex spatiotemporal information and trigger alternative developmental outputs.

1.6.3 Phenotypic plasticity

Colour variation can also occur as a result of phenotypic plasticity[18], in which environmental cues lead to alternative colour morphs in genetically homogenous populations, often by inducing gene expression changes that alter developmental trajectories (Figure 1.5a). Background-dependent and temperature-dependent colour morphs in locusts, temperature-dependent pigmentation in *Drosophila* spp., seasonal polyphenism[19] in butterflies and seasonal camouflage in snowshoe hares are all examples of phenotypic plasticity (Järvi et al., 2019; Jones et al., 2018; Tanaka et al., 2016). Epigenetic and gene expression studies have shown that the darker abdominal pigmentation that occurs in *D. melanogaster* in response to lower temperatures involves changes in the transcriptional regulation of the pigmentation gene *tan* (Gibert et al., 2016); this differential gene expression is part of the phenotype of a plastic response to the environment. In this case dark colouration is likely an adaptive response to optimise heat absorption in colder environments.

A further challenge will be to understand how plasticity itself can evolve, by identifying genes that control differential responses of populations to environmental cues, as depicted by reaction norms[20]. Currently, there are two hypotheses: either plasticity is largely regulated by alterations in the mechanisms that sense environmental cues, which would implicate genes unrelated to colouration; or plasticity could be switched on by changes in downstream genes that respond to the environmental cues, and the genetic loci involved would be similar to those associated with the evolution of colour. Emerging results seem to support the second hypothesis. For example, genetic variation occurs at a CRE of the *Bab* locus in *D. melanogaster* strains with different reaction norms, and Bab protein is responsible for altering *tan* expression in response to temperature cues (Figure 1.4c)(Castro et al., 2018; Gibert et al., 2016); regulatory alleles at the *Agouti* locus control plasticity in the seasonal coat colour of snowshoe hares (Figure 1.5a)(Jones et al., 2018); and the loss of plasticity in side-blotched

[18] **PHENOTYPIC PLASTICITY:** The ability of a single genotype to produce a range of phenotypes depending on the environmental conditions.

[19] **POLYPHENISM:** A type of phenotypic plasticity in which a single genotype can produce two or more discrete alternative phenotypes depending on environmental conditions.

[20] **REACTION NORMS:** Patterns of phenotypic expression of a single genotype across differing environmental conditions.

lizards is associated with changes at two genes in the melanin pathway (*PREP* and *PRKAR1A*)(Corl et al., 2018). Although these early studies suggest that colour plasticity is often regulated by genes in pigmentation pathways, rather than by changes to the genetic mechanisms that underlie detection of environmental cues, more examples will be needed to test whether this pattern holds up as plasticity becomes better understood.

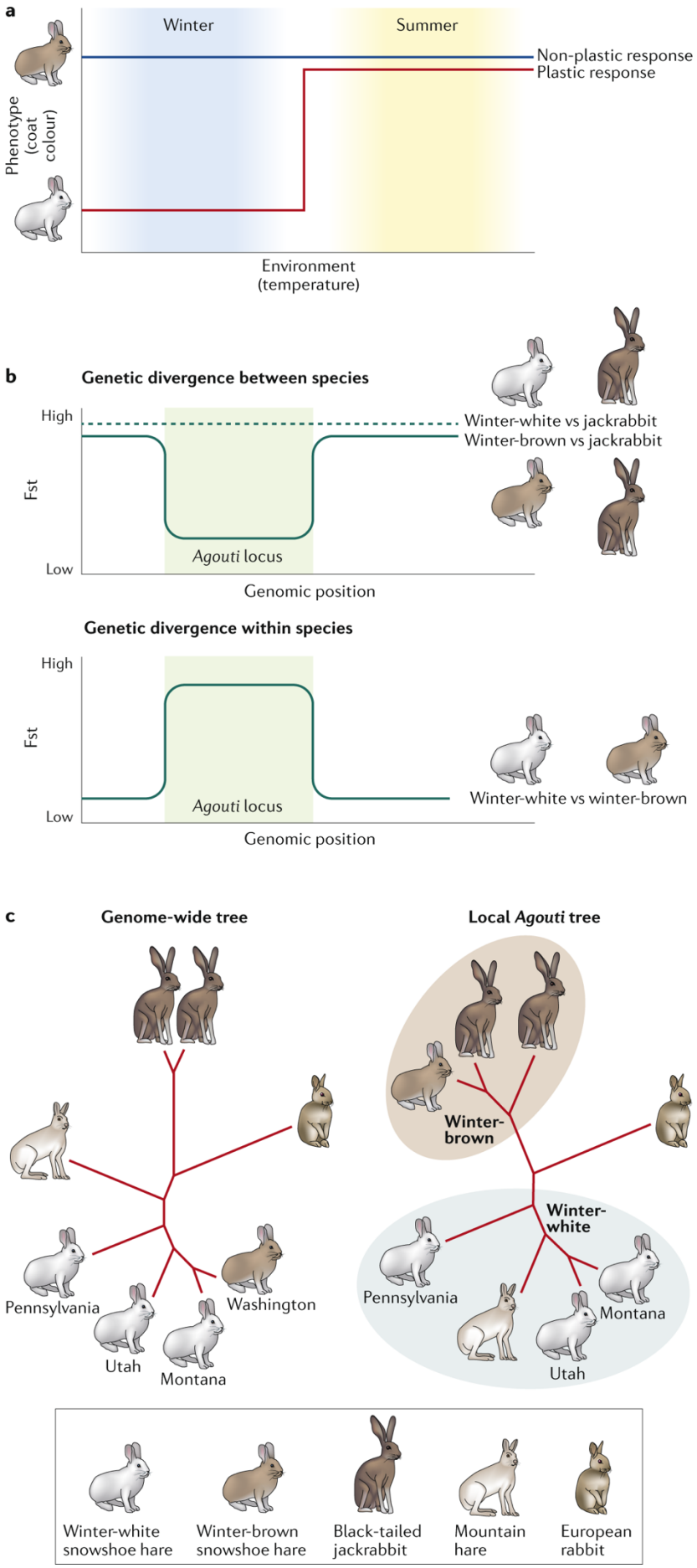

a
Winter
Summer
Non-plastic response
Plastic response
Phenotype (coat colour)
Environment (temperature)
b
Genetic divergence between species
High
Fst
Low
Agouti locus
Genomic position
Winter-white vs jackrabbit
Winter-brown vs jackrabbit
Genetic divergence within species
High
Fst
Low
Agouti locus
Genomic position
Winter-white vs winter-brown
c
Genome-wide tree
Local Agouti tree
Pennsylvania
Washington
Utah
Montana
Winter-brown
Winter-white
Pennsylvania
Montana
Utah
Winter-white snowshoe hare
Winter-brown snowshoe hare
Black-tailed jackrabbit
Mountain hare
European rabbit

Figure 1.5 | **An introgressed allele is responsible for the loss of phenotypic plasticity in snowshoe hares. a |** Snowshoe hares can show phenotypic plasticity in coat colour and maintain seasonal camouflage by moulting into a white coat in winter. However, some snowshoe hares have lost this phenotypic plasticity and retain the brown coat all year round. **b** | This winter-brown coat phenotype is controlled by an allele of the *Agouti* locus, and measures of genetic divergence indicate this allele has introgressed from black-tailed jackrabbits. Genetic divergence (Fst) measured between snowshoe hares and jackrabbits (top) shows that winter-brown snowshoe hares are less divergent than winter-whites from jackrabbits at the *Agouti* locus. Genetic divergence within snowshoe hares shows that winter-white and winter-brown snowshoe hares are highly divergent at the *Agouti* locus, but probably not at other loci. **c** | A tree based on genome-wide pairwise sequence similarity between snowshoe hares and related species recapitulates the species phylogeny. By contrast, a tree based on local sequence similarity at the *Agouti* locus indicates that the snowshoe hare winter-brown allele is more similar to the jackrabbit *Agouti* allele than to snowshoe hare winter-white alleles, providing evidence for introgression from jackrabbits into snowshoe hares. Drawings in panel a and b and panel **c** are adapted with permission from (Jones et al., 2018).

1.7 Understanding adaptation through colour

1.7.1 The origins of variation

Mutation and recombination have commonly been considered to be the primary sources of genetic variation required for adaptation to occur. However, an emerging pattern from genomic studies of colouration and other traits is that allelic variants underlying adaptive change are often more ancient than the populations in which they are found. For example, colour morphs of the invasive ladybird, *Harmonia axyridis,* are controlled by regulatory alleles at the locus encoding the transcription factor pannier (Ando et al., 2018; Gautier et al., 2018). Genomic comparisons of the two morphs have shown that genetic divergence is far greater at this locus than at surrounding genomic regions, which suggests an ancient origin for these colour alleles. Such patterns of high sequence divergence within species can result from either long-term balancing selection[21] at colour variants or more recent introgression[22] of divergent alleles between species. Phylogenetic reconstruction of *Agouti* alleles show that the loss of

[21] **BALANCING SELECTION:** Selective processes that maintain multiple alleles in a population, such as negative frequency dependent selection and heterozygote advantage.

[22] **INTROGRESSION:** The transfer of genetic material from one species to another through hybridization

plasticity in the winter coat colour of snowshoe hares is the result of introgression of an allele from the black-tailed jackrabbit that gives rise to brown winter coats (Figure 1.5b and c)(Jones et al., 2018), and in mountain hares, which present a winter-white/winter-grey polymorphism, the winter-grey variant introgressed through hybridisation with the Iberian hare (Giska et al., 2019). Similarly, a dark form of the wolf that is common in forest habitats has alleles very similar to those in domestic dogs, suggesting recent introgression (Anderson et al., 2009). These examples demonstrate how useful genetic variation can be acquired through introgression from close relatives. At a much finer scale, phylogenetic analysis of two adjacent CREs of *optix* that control red elements in multiple *Heliconius* species has shown a complex history involving multiple introgression events that generated novel combinations of phenotypes (Wallbank et al., 2016).

Introgression can also help to explain the evolution of supergene architectures that would otherwise require multiple genetic steps. In butterflies, introgression from *Heliconius pardalinus* has given rise to a novel allele at the *Heliconius numata* supergene locus (Jay et al., 2019). Levels of inter-species divergence associated with the introgressed allele were substantially lower than across the genome as a whole, which is consistent with introgression rather than long term balanced polymorphism. Similarly, in the white-throated sparrow the allele associated with the tan morph is more closely related to variants in a sister species, the harris sparrow, than to white morph alleles in the same species (Tuttle et al., 2016). Overall, it seems likely that introgression between species is a common and important evolutionary process (Marques et al., 2019).

1.7.2 Observing and testing evolution in the wild

An understanding of the genetic basis for adaptive traits can facilitate studies of natural selection in the wild. Direct tests for selection in natural environments can now take advantage of the ability to genotype the alleles under selection (Figure 1.6). For example, enclosure experiments with deer mice (*Peromyscus maniculatus)* on different soil substrates demonstrated rapid allele frequency changes at the *Agouti* locus that support the hypothesis of selection for crypsis against visual predators (Barrett et al., 2019). Similarly, experimental field studies that tracked genetic variation in natural populations of *Timema* stick insects over 25 years have provided support for negative frequency dependent selection on pattern

variants (striped and unstriped)(Nosil et al., 2018). Short-term changes in allele frequencies occurred substantially more rapidly at the patterning locus than at other similarly sized segments of the genome, indicating selection at these loci. In both cases, genomic data adds value to selection experiments by facilitating estimates of selection on genotypes rather than solely on phenotype, and by allowing changes in allele frequency at selected loci to be compared to the genomic average.

Natural selection also leaves a signature in the genome of wild populations. Therefore, an alternative approach to studying selection is to search for this genomic signature, in order to estimate the past action of natural selection. Colour variants are especially good for such studies, as there is often a well understood link between colour variation and genotype. For example, population genetic analysis of deer mice in the Nebraska Sand Hills demonstrated that coat colour in this species is composed of multiple traits and that each of these colour traits is associated with a distinct mutation within the *Agouti* locus (Linnen et al., 2013). These mutations have been independently selected, with larger effect variants showing evidence for stronger selection. A recent study explored the extensive radiation of *Heliconius* butterflies and provided evidence for strong natural selection at multiple colour loci across almost all sampled populations (Moest et al., 2020). These results support field evidence for consistently strong selection on colour forms in these butterflies (Kapan, 2001; Mallet and Barton, 1989), but additionally show that colour loci are some of the most strongly selected regions in the genome.

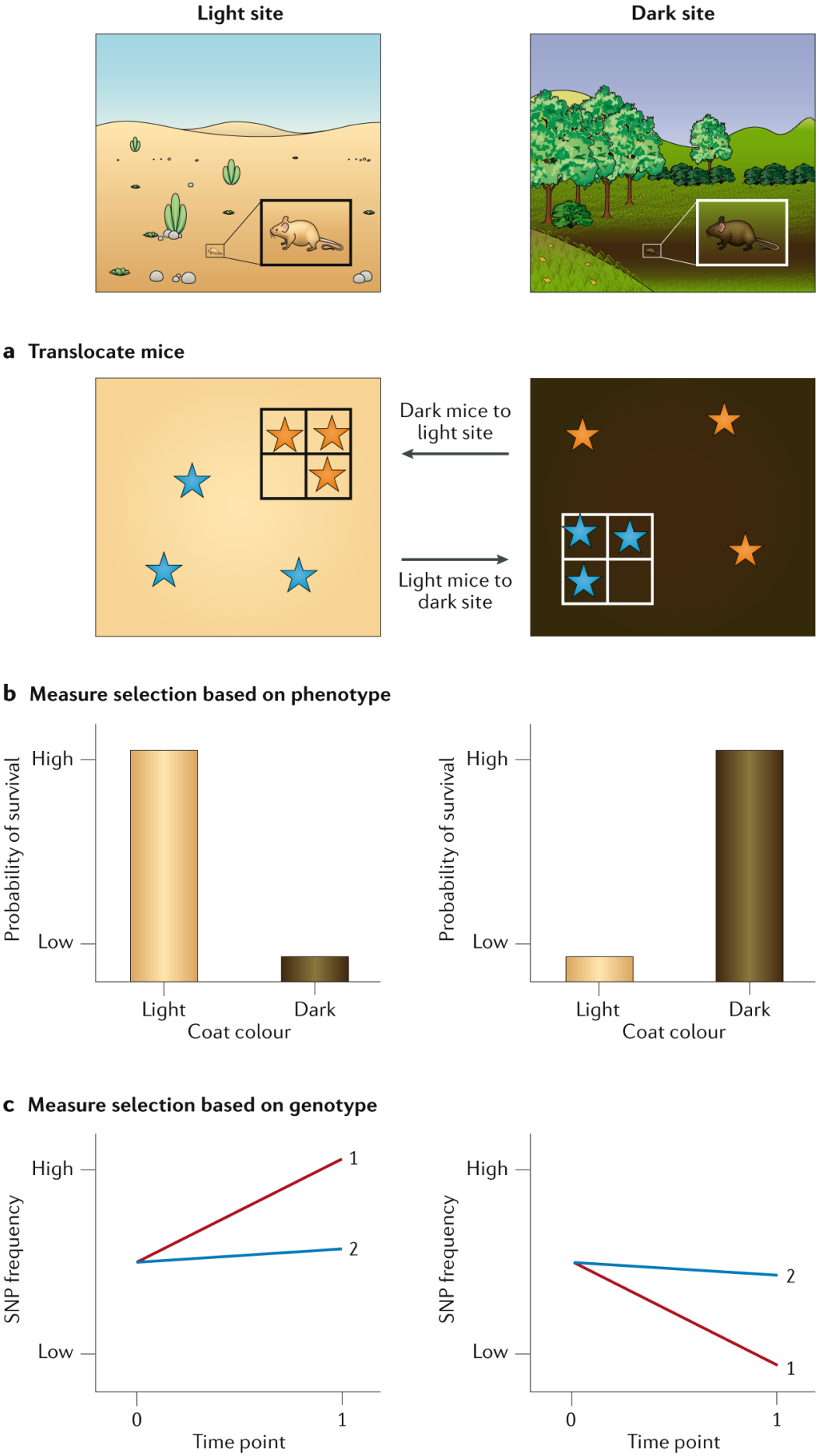

Light site
Dark site
a Translocate mice
Dark mice to light site
Light mice to dark site
b Measure selection based on phenotype
Probability of survival
High
Low
Light
Dark
Coat colour
c Measure selection based on genotype
SNP frequency
1
2
0
1
Time point

Figure 1.6 | **Estimating natural selection.** A mouse species presents two cryptic coat colour phenotypes (light and dark) that inhabit two types of soil colour (light and dark, respectively). **a** | A translocation experiment can be used to test whether the phenotypes are adaptive (such as in (Barrett et al., 2019)). Individuals from the light-coloured population on light-coloured soil (sampling sites indicated by blue stars) are translocated to enclosures in a dark-coloured soil area (blue stars in grid) and dark-coloured individuals (sampling sites indicated by orange stars) are transferred from the dark-coloured site to enclosures in the light-coloured site (orange stars in grid). **b** | Adult survival is measured and the probability of survival plotted against coat colour. The plots illustrate the hypothetical case in which coat colour is discrete and adaptive: light-coloured mice have a higher survival rate than dark-coloured mice in light-coloured soil sites and vice versa. If coat colour was non-adaptive, the probability of survival would be similar on either soil type regardless of coat colour. **c** | Selection at single nucleotide polymorphisms (SNPs) known to be associated with coat colour can be estimated by calculating the change in their frequency between different time-points: before (0) and after selection (1). The plotted lines indicate frequency changes of individual SNPs. SNP 1 shows positive selection in the light-coloured site and negative selection in the dark-coloured site, whereas SNP 2 is neutral in both sites. Thus, translocation experiments that incorporate genomic data provide the opportunity to track allele frequencies in the wild and measure the direction and strength of selection.

1.8 Conclusions and future perspectives

In many ways, studies of colouration have been at the forefront of evolutionary genetics and served as a model for investigating complex traits. Colouration represents some of the most striking and accessible adaptive traits in the natural world and were well understood before the genomics revolution, and colour-related traits are some of the best understood examples of genotype-phenotype links. Nonetheless, the extent to which colour traits are representative of other phenotypes remains unclear – for example, loci of large effect might be more common among colour traits if pigmentation has a relatively simple genetic basis compared with other morphological structures. Challenges remain, notably in understanding polygenic inheritance and the genetic control of phenotypic plasticity, but colour phenotypes offer excellent opportunities to study both of these phenomena. The long-term goal will be to synthesise what we learn from individual examples to infer general patterns regarding the evolution of genetic architectures. Finally, while genomic studies have informed our understanding of the action of natural selection in wild populations, there have been fewer attempts to study sexual selection in the wild using genomic approaches. Perhaps in the

future, genetic and genomic analyses will provide similar insights into the action of sexual selection on colour phenotypes in natural populations.

2

Introduction to the systems *Hypolimnas* butterflies and the wood tiger moth, *Arctia plantaginis*

Butterfly and moth (Lepidoptera) wing patterns have been an extensive source of knowledge on the evolution of adaptive phenotypes. The multiple examples of rapid speciation and polymorphism, together with the easiness of phenotyping, have made butterfly and moth wing phenotypes an ideal system for the study of evolution. Extensive knowledge on the ecological and genetic mechanisms driving wing pattern diversification and maintenance come from *Heliconius* butterflies (reviewed in (Jiggins, 2017)). The genus is best known for its multiple examples of Mullerian mimicry, in which unpalatable species converge to the same aposematic phenotypes in order to share the costs of teaching predators. Similarly, Batesian mimics of the genus *Papilio*, in which palatable species resemble toxic ones in order to gain protection from predators, have also provided key insights in the genetic bases of wing polymorphisms (Fukova et al., 2014; Kunte, 2009; Kunte et al., 2014; Nishikawa et al., 2015). These and other studies in multiple other Lepidoptera species have revealed that a relatively small set of genes are linked to adaptive phenotypes in multiple species (McMillan et al., 2020). Furthermore, another generality that has started to emerge from studies of wing phenotypes is that structural variation is often involved in the genetic architecture of these phenotypes (Joron et al., 2011; Livraghi et al., 2021; van't Hof et al., 2016; Woronik et al., 2019). Exploring diverse taxa across the Lepidoptera will provide crucial insight to clarify why some genes seem to be 'genetic hotspots' for adaptive evolution of wing patterns and the extent of the involvement of structural variation in the evolution of wing phenotypes. In this thesis, I have used two very distinct systems to explore the genetic basis of adaptive wing

phenotypes in Lepidoptera: *Hypolimnas* butterflies and the wood tiger moth, *Arctia plantaginis.*

1.9 *Hypolimnas* butterflies

Hypolimnas butterflies, commonly known as eggflies, are a tropical genus of butterflies containing 29 species with very diverse wing patterns (Sahoo et al., 2018). Two species, *Hypolimnas misippus* and *H. bolina*, present female limited Batesian mimicry of defended models (Vane-Wright et al., 1977). *H. misippus,* also known as the diadem butterfly, is a pantropical species, with polymorphic and mimetic females that resemble the four morphs of the toxic *Danaus chrysippus* and monomorphic non-mimetic males (Figure 2.1)(Swinhoe, 1896; Vane-Wright et al., 1977). Despite the striking detailed resemblances, there is a mismatch in frequencies of mimetic morphs and their models (Gordon et al., 2010; Pierre, 1980; Smith, 1976). In Africa, *D. chrysippus* morphs are largely allopatric with one morph being dominant in each region (Figure 2.2A), while *H. misippus* morphs are found distributed across, with best mimics not always being the most common form. For example, in Tanzania the most common morph of *D. chrysippus* has a frequency of 71% and presents orange forewing and hindwings (Smith, 1976). Contrastingly, in the same location, the matching *H. misippus* mimic is only present at a frequency of 23%, with the most common morph at frequency of almost 50% presenting black-and-white forewing apexes and orange hindwings. Furthermore, maladaptive intermediate morphs of *H. misippus* are often found. This has led to the suggestion that selection for mimicry is weak and raises the question of how the polymorphism is being maintained, if not driven by mimicry (Gordon et al., 2010).

One hypothesis explaining the maintenance of the polymorphism is that there is negative frequency dependent selection on wing phenotype through apostatic selection and that its strength varies over time following extreme changes in population size (Gordon, 1987; Gordon et al., 2010). In *H. misippus,* forewing and hindwing phenotype are determined by independent loci, selection for *H. misippus* erupt in large numbers suddenly, which has been suggested to produce a surge in predation on *H. misippus* and *D. chrysippus,* increasing the selective pressure for mimicry. Such a scenario would result in balancing selection and could maintain the polymorphism.

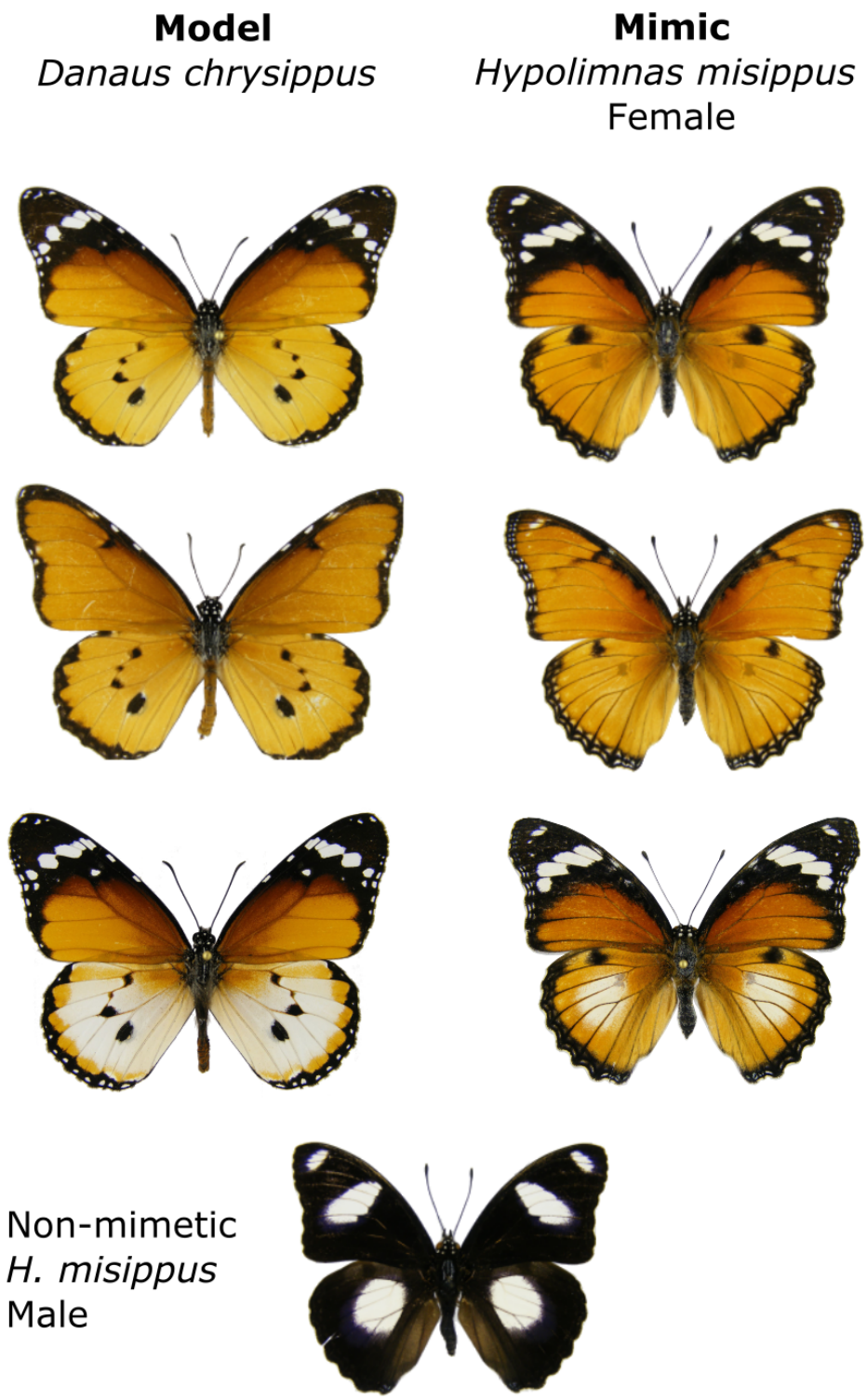


Figure 2.1 | Mimicry in *Hypolimnas misippus.* *H. misippus* is a mimic of the toxic *Danaus chrysippus.* Left column of images show three morphs of *D. chrysippus*, the right column of images are the corresponding *H. misippus* mimics, while the bottom image is a *H. misippus* male.

An alternative hypothesis is that associative overdominance at the loci controlling wing phenotype maintains the polymorphism, perhaps associated with structural variation. Structural variation generally produces a reduction in recombination that leads to decreased purifying selection and can result in the accumulation of deleterious mutations (Ohta and Kimura, 1970; Zhao and Charlesworth, 2016). In such cases, heterozygotes might have higher fitness, as the effect of recessive deleterious mutations is masked (Pamilo and Pálsson, 1998). An example of this has been seen in *Heliconius numata*, in which three inversions at a mimicry supergene show associative overdominance and, together with balancing selection, contribute to the maintenance of multiple morphs (Chouteau et al., 2017, 2016; Jay et al., 2021; Joron et al., 1999).

Understanding the genetic basis of wing phenotype in *H. misippus* is crucial to clarify the evolutionary forces maintaining the polymorphism. From large breeding experiments and association studies, we know that forewing polymorphism is determined by a Mendelian locus, the M locus, with two alleles (Figure 2.2B)(Smith and Gordon, 1987; VanKuren et al., 2019). The dominant allele *M* produces forewing with a black-and-white apex (*misippus* morph), while the recessive allele *m* produces all-orange forewings and intermediates in homozygotes, known as the *inaria* and *immima* morphs respectively. Furthermore, hindwing phenotype varies continuously from completely orange to having a large white spot in the wing and is determined by two loci of large effect, the A and the S loci, with two alleles each (Figure 2.2B)(Gordon and Smith, 1989). *AA* individuals have orange hindwings, while *aa* individuals present a large white spot and heterozygotes have an intermediate phenotype. There S locus also contributes to hindwing phenotype by supressing (hence the name, S) hindwing white. Identifying and dissecting these loci will help clarify the evolutionary history of wing mimicry in *H. misippus.*

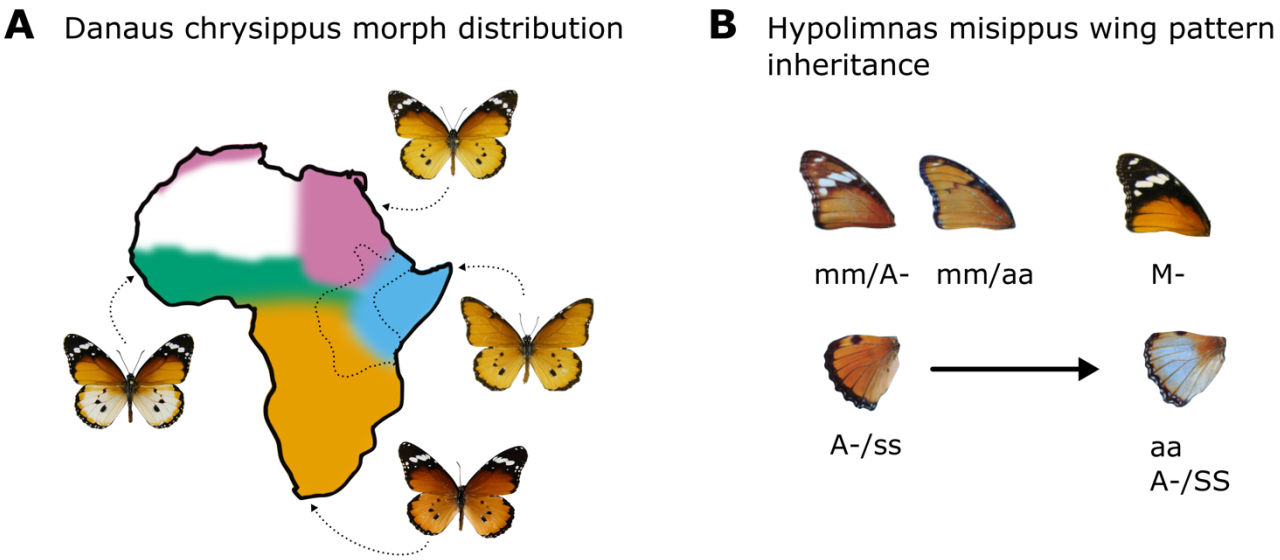


Figure 2.2 | A. Distribution of morphs of *Danaus chrysippus.* Dotted line area in East Africa indicates the contact zone between morphs. **B.** Model of wing pattern inheritance proposed by Gordon and Smith in (Gordon et al., 2010; Gordon and Smith, 1989; Smith and Gordon, 1987).

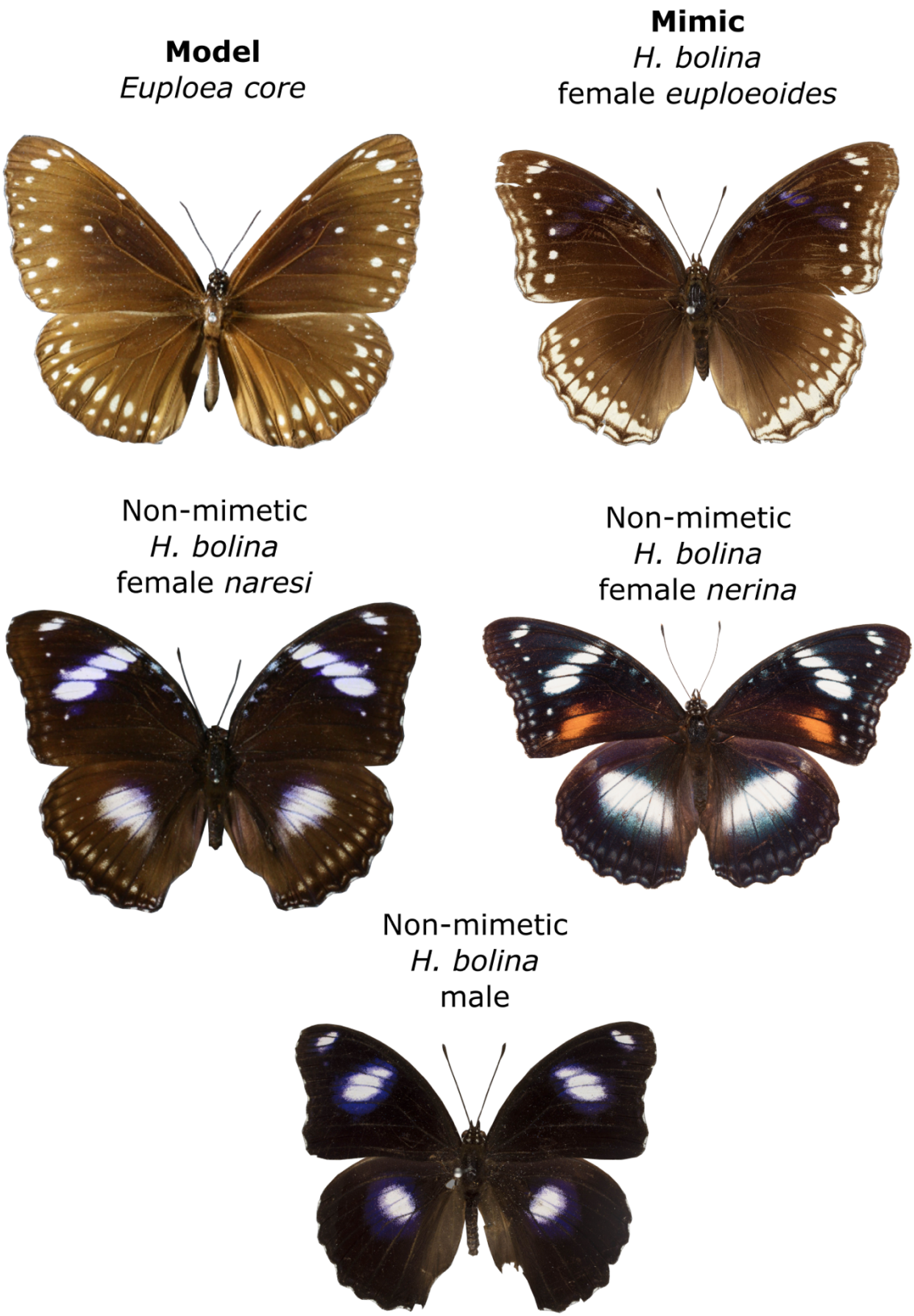


Figure 2.3 | Mimicry and polymorphism in *Hypolimnas bolina.* The *euploeoides* morph (top right) is the only mimetic morph of *H. bolina,* resembling species of the *Euploea* genus (top left).

Multiple similarities can be found between *H. misippus* and *H. bolina*. In southeast Asia, *H. bolina* is monomorphic with only the *euploeoides* morph found, which is a recognized Batesian mimic of multiple *Euploea* species (Figure 2.3). Contrastingly, in the south Pacific Islands and Australia, *H. bolina* presents multiple non-mimetic forms including the morphs *nerina* and *naresi* (Figure 2.3)(Clarke and Sheppard, 1975). The genetic basis of its wing pattern was extensively studied by Clarke and Sheppard, who hypothesised that the polymorphism was under the control of two autosomal loci (Clarke and Sheppard, 1975). The E locus controls the white pattern elements in the wings, determining the differences between the *euploeoides* morph and *nerina* and *naresi.* It determined the presence and absence of the hindwing white spot and the subapical white band in the forewing. The N locus controls the orange spot found in the forewings of *nerina* morphs. Intermediates are commonly found and mimicry seems to only apply to *euploeoides* morphs. Examining the genetic architecture of polymorphism in *H. bolina* will help clarify the selective forces behind its wing phenotype diversity.

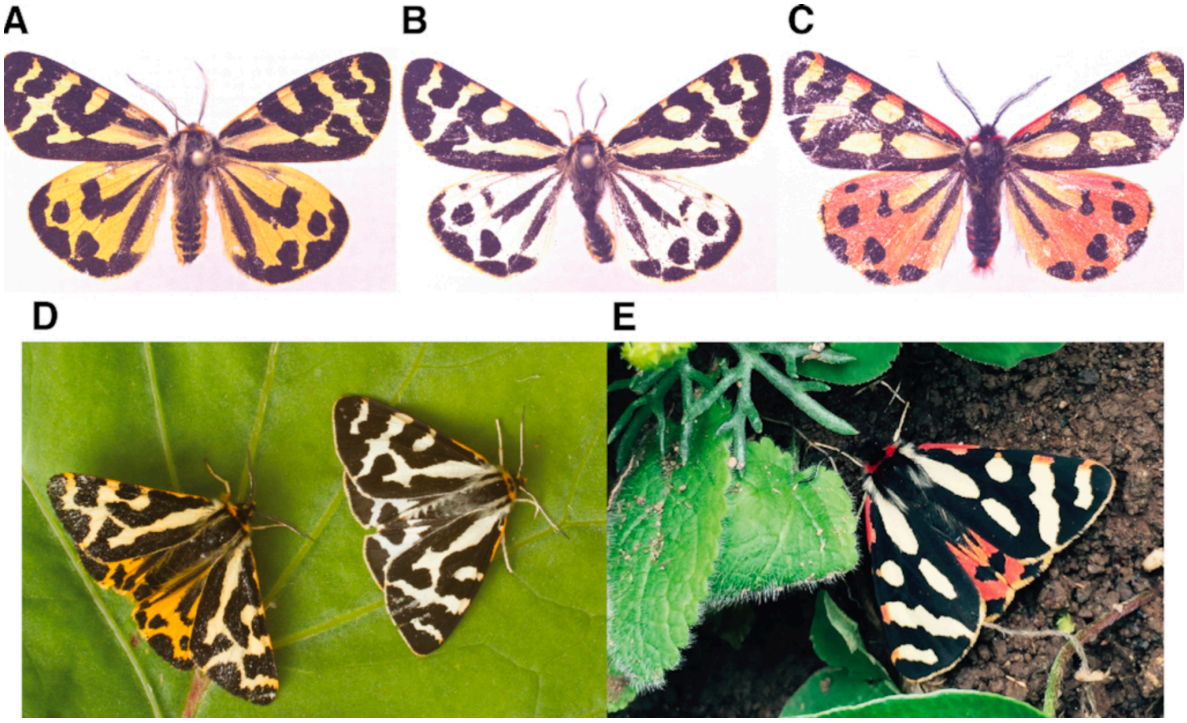


Figure 2.4 | Aposematism in male *A. plantaginis.* Male *A. plantaginis* present discrete morphs that vary in hindwing colouration. Images of the white, yellow and red morphs are shown. Figure reproduced from (Yen et al., 2020).

1.10 The wood tiger moth

The wood tiger moth, *Arctia plantaginis*, is an aposematic, protected species, presenting bright hindwing colouration to deter predators. Whilst female hindwing coloration varies continuously from red to yellow, males present discrete morphs that vary in frequency in different geographic locations (Figure 2.4). For example, some populations are monomorphic with only white individuals (e.g., Estonia), or yellow (e.g., Scotland), while in Finland, frequencies vary from 40-75%, and in Georgia there is a unique red form (Galarza et al., 2014; Hegna et al., 2015). These male morphs are determined by a single Mendelian locus with two alleles (Nokelainen et al., 2022; Suomalainen, 1938), the dominant 'white' and recessive 'yellow', and have been associated with distinct predator avoidance strategies that include differences in chemical defences (Rojas et al., 2019, 2017), mating success (Gordon et al., 2015; Nokelainen et al., 2012), flight ability (Rojas et al., 2015) and immune specialisations (Nokelainen et al., 2013). While yellow males have higher chemical defences and are subject to lower levels of predation, white males have a positive frequency dependant mating advantage. Also, white males invest more in flight activity for predator escape and finding mates, although yellow males might fly at more selective times, being active only during peak female calling periods. In summary, there is a trade-off between warning signal efficacy and mating success between yellow and white males that leads to the maintenance of this complex polymorphism (Rönkä et al., 2020). Despite the deep ecological understanding of this polymorphism, knowledge of its genetic basis is still lacking. Identifying and dissecting the genetic architecture of morph determination in *A. plantaginis* will shed light on the evolution of complex phenotypes.

3

The Hypolimnas misippus genome supports a common origin of the W chromosome in Lepidoptera

1.11 Abstract

Moths and butterflies (Lepidoptera) present heterogametic sex chromosomes with females carrying ZW chromosomes. The lack of W chromosomes in early diverging lepidopteran lineages, has led to the suggestion of an ancestral Z0 system in this clade and a B chromosome origin of the W. This contrasts with the canonical model of W chromosome evolution in which the W would have originated from the same homologous autosomal pair as the Z chromosome. Despite the distinct models proposed, the rapid evolution of the W chromosome has hindered the elucidation of its origin. Here, I present high-quality, chromosome-level genome assemblies of two *Hypolimnas* species (*Hypolimnas bolina and Hypolmnas misippus)* and use the *H. misippus* assembly to explore the evolution of W chromosomes in butterflies and moths. I show that in *H. misippus* the W chromosome presents higher similarity to the Z chromosome than any other chromosomes, suggesting a possible origin from the same homologous autosome pair as the Z chromosome. However, using other ditrysian genome assemblies containing assembled W chromosomes, I present contrasting evidence suggesting that the W chromosome might have evolved from a B chromosome instead. Crucially, by using a synteny analysis to infer homology, I show that W chromosomes are likely to share a common evolutionary origin in Lepidoptera. My study highlights the difficulty of studying the evolution of W chromosomes and contributes to better understanding its evolutionary origins.

1.12 Introduction

Sex chromosomes are highly variable in eukaryotes and have evolved independently multiple times (Bachtrog et al., 2014, 2011; Beukeboom and Perrin, 2014). In animals there are multiple types of chromosomal sex determination but two are predominant: male heterogametry such as seen in mammals, where males are XY and females XX, and female heterogametry as in birds, where females are ZW and males are ZZ (Beukeboom and Perrin, 2014). These heteromorphic sex chromosomes can potentially originate from different processes. One possibility is that they initially arise from a pair of autosomes that evolves genetic sex determination and, through a process of reduced recombination, sex specific mutations accumulate (Wright et al., 2016). This reduction in recombination can also lead to gene depletion and the accumulation of transposable and repetitive elements, which are common characteristics of sex specific (Y or W) chromosomes (Bachtrog et al., 2011; Wright et al., 2016). However, differentiating between cause and consequence can be difficult. Additionally, other autosomes may fuse to the sex chromosomes and become differentiated. Alternatively, the W/Y chromosomes can originate from the recruitment of a B chromosome, which are dispensable chromosomes that are found variably in populations and species (Yoshida et al., 2011).

Moths and butterflies have a ZW sex chromosome system in which females are heterogametic and show a lack of recombination (Turner and Sheppard, 1975). Whilst the Z chromosome has been shown to be highly conserved across the Lepidoptera (Fraïsse et al., 2017), the origin and evolution of the W chromosome remains unclear. The possible absence of W chromosomes in early-diverging ditrysian lineages and the deep conservation of the Z chromosome have been suggested to be evidence of an ancestral Z0 sex determination system for Lepidoptera and an origin for the W chromosome that is independent of the Z chromosome (Dalíková et al., 2017; Fraïsse et al., 2017; Lukhtanov, 2000; Sahara et al., 2012). Two main alternative hypotheses have been proposed for the origin of the W chromosome: evolution from a B chromosome (Dalíková et al., 2017; Lewis et al., 2021; Lukhtanov, 2000), or evolution from the homologous pair of an autosome that fused to the Z chromosome (Sahara et al., 2012).

Despite interest in understanding the evolution of the W chromosome in the Lepidoptera, the absence of high-quality reference genomes containing assembled W chromosomes has limited its study. The elevated repeat and TE content have made its assembly challenging until now. Long read sequencing technologies and the decrease in sequencing price have led to the production of high-quality Lepidopteran genome assemblies containing W chromosomes, which makes it increasingly possible to elucidate the enigmatic origin of the Lepidoptera W chromosome.

Hypolimas butterflies, commonly known as eggflies, are a phenotypically diverse genus that has served as a model for the study of ecology and evolutionary biology. Many *Hypolimnas* species are mimics of toxic species, which has shaped the diversification of wing colour pattern in the genus (Vane-Wright et al., 1977). Historically, two species, *Hypolimnas bolina* and *Hypolimnas misippus*, have received most of the attention. *H. bolina* and *H. misippus* diverged 8 million years ago and share many similarities, both have a pantropical distribution and the females of both species are polymorphic Batesian mimics of toxic models (Marsh et al., 1977; Sahoo et al., 2018; Smith, 1976). In contrast, males are monomorphic and have retained what is likely to be the ancestral phenotype of white-spotted black wings. Nonetheless, the two species also differ in many aspects. *H. misippus* females are mimics of the four morphs of the African Queen, *Danaus chrysippus*, while *H. bolina* females are mimics of several *Euploea* species. *H. bolina* has also received special interest for its coevolution with the endosymbiont *Wolbachia* (Charlat et al., 2009; Dyson et al., 2002). In *H. bolina*, *Wolbachia* has a male killing phenotype that promotes spread of the endosymbiont, while *H. bolina* has evolved a suppressor locus that counteracts it (Hornett et al., 2006).

The diversity in phenotype, precision of mimicry and intricate coevolution, make *Hypolimnas* a remarkable genus for evolutionary biology studies. However, there are few genomic resources to date. Here, I present chromosome-level assemblies for *H. misippus* (HypMis_v2) and *H. bolina* (HypBol_v1) and use our high-quality *H. misippus* assembly containing the Z and W chromosomes to explore the evolution of sex chromosomes in the Lepidoptera. First, I compare the synteny and TE content between the two *Hypolimnas* assemblies. Next, I evaluate the annotation completeness and gene content by comparison to the close relative the painted lady butterfly, *Vanessa cardui.* Then, I investigate the hypothesis of a B

chromosome origin of the W chromosome by comparing homology between the W chromosome, autosomes, and Z chromosome within *H. misippus*. Finally, I examine the origin of the W and Z chromosomes across the Lepidoptera by analysing synteny between *Hypolimnas misippus* and a diverse set of 10 Lepidoptera species in different ditrysian families.

1.13 Materials and methods

1.13.1 Hypolimnas *misippus* butterfly rearing and cross preparation

A trio-binning approach was used for sequencing of the *Hypolimnas misippus* assembly, which consists of the sequencing of the parents and offspring of a family. First, butterflies were obtained from Stratford-upon-Avon Butterfly Farm, UK, and reared in greenhouses in Madingley, Cambridge, UK. Larvae were fed *Portulaca oleracea* and *Portulaca quadrifida*. Adult butterflies were kept in a large cage (1.5 meters x 1.5 meters x 2 meters) and observed until mating when mating pairs were transferred to separate smaller cages during copulation. One mated pair was used for trio-based genome sequencing, their offspring reared until pupation and flash frozen as pupae by dropping them into liquid nitrogen.

1.13.2 *H. misippus* Trio binning genome assembly

One *H. misippus* family was used for sequencing and trio-binning genome assembly. The two parents were sequenced using Illumina short-read sequencing, which resulted in 368.67 and 346.56 million read pairs and a total yield of 55.67 and 52.33 Gbp of data from the trio mother and father respectively. One of the offspring was sequenced using PacBio long-read sequencing (total yield of 13.46 Gb of data and N50 of 13,493). A trio binning approach enabled the independent assembly of the two parental haplotypes. First, yak-r55 (Li, 2022) was used to create a kmer database from the parental Illumina data. Then, hifiasm-0.7-r256 (Cheng et al., 2020) was run in trio mode to assemble the HiFi PacBio long-read data, followed by the purging of haplotypes and overlaps using purgedups v1.2.3 (Guan, 2022). Haplotype 1 corresponds to the paternal haplotype, while haplotype 2 corresponds to the maternal one. All posterior analysis were done using the haplotype 2.

1.13.3 Curation of *HypMis_v2* with HiC

To place the assembled scaffolds into chromosomes, Hi-C sequencing was used, which is a chromosome conformation capture technology that provides information about the 3-D interactions between genomic loci. An offspring from an unrelated mating was flash frozen in liquid nitrogen and used for Hi-C sequencing and analysis. 386.99 million read pairs were produced, which represented 58,45 Gb of Hi-C data. Fastq files were converted to cram (two files, one for each read pair end), mapped to the haplotype 2 (the maternal haplotype) *H. misippus* assembly produced after the purgedups stage, and processed with Juicer v1.6 (Durand et al., 2016b). Juicer transforms raw Hi-C data into a list of contacts, which defined pairs of genomic positions that were in close physical contact during the experiment. Then, the main reference assembly was curated using the 3D-DNA pipeline, which corrects misassembles, anchors, orders and orients fragments of DNA based on the Hi-C data (Dudchenko et al., 2017; Durand et al., 2016). 3D-DNA generated assembly heatmaps as part of its workflow, which indicate the frequency of contact between pairs of genomic locations. Obvious errors in the genome assembly such as large genomic inversions were manually edited by examining the Hi-C heatmaps using the Juicebox tool (Dudchenko et al., 2018; Durand et al., 2016). Finally, the edited assembly was exported as an *.assembly* file and converted to a final fasta assembly file using the 'run-asm-pipeline-post-review.sh' script setting –editor-repeat-coverage to 6.

1.13.4 *HypBol_v1* genome curation with Ragout

The *H. bolina* reference assembly (*HypBol_v1*), was produced using a combination of Nanopore long-read sequencing and a linkage map. Pupae originating from the Philippines were ordered from Stratford-upon-Avon butterfly farm. One female pupa was used for sequencing. Hing-molecular DNA was extracted using a chloroform and glass capillary protocol. Sample was prepared for nanopore sequencing following the Long Read Club LSK109 library prep (bead free) and sequenced at the Centre for Genomic Research in Liverpool, UK. Adapters were removed using Porechop v0.2.4 (Wick, 2022) and reads assembled using redbean (Ruan and Li, 2020). Assembled contigs were split into bins using MaxBin2 (Wu et al., 2016) and Bacterial contigs and reads were removed from the data using

blobtools2 (Challis et al., 2020). This produced an assembly with 13,492 contigs and an N50 of 1.4 kbp.

Leveraging the high synteny expected from the two *Hypolimnas* assemblies, Ragout v2.3 (Kolmogorov et al., 2018, 2014) was used to improve the initial *HypBol_v1* assembly using *HypMis_v2* as a reference. First, the two genomes were soft-masked using RepeatMasker (Smit et al., 2015), creating the repeat library based on the genome being masked. Then, Cactus (Paten et al., 2011) was used to align both genomes using Python 3.8. The resulting HAL alignment was converted to MAF format using the hal2maf utility from the HAL program (Hickey et al., 2013). Finally, Ragout was run using the MAF alignment between *HypBol_v1* and *HypMis_v2* as input.

1.13.5 Rearing of *H. bolina* individuals

A linkage map was used to improve the *HypBol_v1* assembly and place the assembled scaffolds into chromosomes. Two families (178303XX and 182703XX) were reared and sequenced. First, female H. bolina purchased from Stratford Butterfly Farm (originating from the Philippines) were mated to wild-caught males from Moorea (French Polynesia) in French Polynesia (Gump research facility). Female Philippines-Moorea F1 hybrids were mated to pure Moorea F1 males. The F2 offspring of one of these crosses is family 178303XX. Male Philippines-Moorea F1 hybrids were mated to pure Philippine F1 females. The F2 offspring of one of these crosses Is family 182703XX. Butterflies were kept in a large outdoor cage for mating under observation. Any pairs were separated into a small cage. The female was then placed in an oviposition cup containing a small plant e.g. *Asystasia sp* and allowed to lay eggs. Hatched eggs are moved to a rearing box with suitable food plant e.g. Ipomoea sp. And caterpillars reared until pupation. Pupae are moved to a dry cup for emergence, adults left to dry for one day and then used for further matings or stored in -80˚C freezer.

1.13.6 Illumina library preparation of *H. bolina* family samples

The offspring of the *H. bolina* families (F1) were processed to extract the DNA and prepare the libraries for Illumina sequencing. DNA extractions were carried out using a custom protocol using PureLink buffers and homemade magnetic beads. Briefly, a small piece of thorax tissue (1/10) is placed in an 8-tube PCR strip. Then, 45 uL of PureLink Digestion buffer

and 10 uL of Proteinase K (20mg/mL) are added, and the mix is incubated at 58˚C with shaking (500 rpm) for 2-3 hours. Afterwards, 2uL of RNAseA are added (DNAse free, 10mg/mL) and incubated for 10min at room temperature. Then, 45uL of PureLink Lysis buffer is added to the mix and incubated at 58˚C for 30 minutes with shaking (500 rpm). Following that, a homemade magnetic bead mix is used to extract the DNA from the lysate. First, 37.5 uL of magnetic beads are added together with 75 uL of lysate to a 96-well plate. After mixing, the samples are incubated for 15 minutes at room temperature, then the plate is placed on a magnetic stand for 10 minutes, the supernatant removed, and the beads cleaned with 80% ethanol. After drying out, 50uL of 10mM Tris (pH=8) are added to elute the sample and incubated at 45˚C for 15 minutes without resuspending. Then, the beads are resuspended and incubated for 20 minutes at room temperature. Finally, the plate is placed on the magnetic stand and, after 10 minutes, the supernatant (the DNA) is transferred to a fresh tube.

The F1 were sequenced using a Nextera-based library preparation at intermediate coverage (~11X). A secondary purification using magnetic SpeedBeads™ (Sigma) was performed prior to Nextera-based library preparation. Libraries were prepared following a method based on Nextera DNA Library Prep (Illumina, Inc.) with purified Tn5 transposase (Picelli et al., 2014). PCR extension with an i7-index primer (N701–N783) and the N501 and N502 i5-index primers was performed to barcode the samples. Library purification and size selection was done using the same homemade beads as above. Pooled libraries were sequenced by Novogene Cambridge, UK. Libraries of the parental samples were prepared and sequenced to ~20X coverage by Novogene Cambridge, UK.

1.13.7 *H. bolina* linkage map construction and anchoring of the genome

A linkage map was produced with Lep-Map3 (Rastas, 2017) and then used to improve the *H. bolina* assembly and place the scaffolds onto chromosomes. First, sequences were mapped to the reference genome using bwa-mem (Li, 2013). PCR duplicates were marked using the MarkDuplicates from Picard tools. Sorted BAMs were then created using SAMtools (Li et al., 2009) and genotype likelihoods computed. The pedigree of individuals was checked and corrected using IBD (identity-by-descent) with a random subset of 10% of the markers (1,270,024 SNPs) following the IBD pipeline from Lep-Map3. These markers were also used to

construct the linkage map. Scaffolds were anchored into chromosomes based on the linkage map using LepAnchor (Rastas, 2020).

1.13.8 *HypBol_v1* polishing with Pilon

After anchoring with the linkage map, three iterations of Pilon v1.24 (Walker et al., 2014) in diploid mode were run to correct the draft assembly by correcting bases, filling gaps and fixing mis-assemblies. Samples used for the Pilon correction were CAM035727, CAM035728, CAM035186, CAM035187, CAM035188, CAM035189.

1.13.9 Repeat annotation

Once the two final assemblies had been produced, they were each assessed for repeat content using a custom repeat library. First, a repeat database was built and the repeats of the two finished assemblies modelled using RepeatModeler v. 2.0.2a. Each custom library was then combined with the Lepidoptera library extracted from Dfam (Storer et al., 2021). This merged library was used to soft mask the genome using RepeatMasker v 4.1.0 with the cut-off score set to 250 and skipping the bacterial insertion element check. The resulting soft masked assemblies were used for gene annotation.

1.13.10 RNA-seq sample preparation

To annotate each genome, RNA-seq data was obtained from 4 *H. misippus* (2 adults and 2 pupae) individuals and 17 *H. bolina* (6 adults and 11 pupae). Butterflies were purchased from Stratford-upon-Avon Butterfly Farm and kept at room temperature until dissection. 4 tissues were dissected out from *H. misippus* pupae and placed in RNA-later (Sigma): wing discs, thorax, head, and abdomen; while *H. bolina* pupae 3 tissues were dissected: wing discs, thorax-head, and abdomen. Only abdomen, head and thorax samples were dissected from adults of either species. Two pooling strategies were followed: 1) 2 *H. bolina* pupae were pooled by individual, pooling head, thorax, abdomen, and wing discs together and sequenced at high coverage (50M reads). And 2) 4 *H. misippus* and 15 *H. bolina* adults and pupae were dissected into tissues and pooled by species and tissue. Each pooled sample, 4 for *H. misippus* and 3 for *H. bolina,* was sequenced to 20M reads. RNA was extracted using a modified Trizol protocol using the same protocol as in Chapter 6 (Brien et al., 2022).

1.13.11 RNA-seq mapping and gene annotation

First, low quality ends and adaptors from the RNA-seq data were trimmed using TrimGalore v 0.5.0 (Krueger, 2015). Then, the reads were mapped to the soft masked genomes using STAR v 2.5.0a (Dobin et al., 2013). Two rounds of mapping (2pass) were performed, including all the splice junction files in the second round. Then the resulting mapped reads and the soft masked genome assembly were used to generate a gene annotation using BRAKER v 2.1.5 (Brůna et al., 2021), running it a second time to add UTR annotations with options addUTR=on and skipAllTraining (Keilwagen et al., 2019). Completeness of the annotation was assessed using BUSCO (Seppey et al., 2019; Simão et al., 2015) using the Insecta_odb10 (n=1367) set of genes.

1.13.12 Homology with Merian units

A growing practice in Lepidoptera genetics is to name chromosomes of reference assemblies by their homology to *Melitaea cinxia*, which is the first assembled lepidopteran genome that retains the ancestral 31 chromosome karyotype (Ahola et al. 2014). The numbered chromosomes are known as Merian units in honour of the scientist Maria Sibylla Merian (Wulf, 2016) and allow for the tracking of conserved synteny blocks across the phylogeny of the Lepidoptera. *Melitaea cinxia* belongs to the *Nymphalinae* which is the same subfamily as *Hypolimnas*, and the two taxa diverged about 30 MYA (Espeland et al., 2018). To assign homology to the Merian units, lepidopteran_odb10 BUSCO assignment of the *Melitaea cinxia* genome was compared to the lepidopteran_odb10 BUSCO assignment of *HypBol_v2* and *HypBol_v1*. Merian units in *HypBol_v1* and *H. missipus* respectively. Were assigned by choosing the chromosome sharing the highest BUSCO hits with a *M. cinxia* chromosome.

Synteny of the two assemblies was assessed using two methods. First, the two final genome assemblies were aligned using D-GENIES (Cabanettes and Klopp, 2018), which produced a paf file that was used to detect candidate inversions between the two genome assemblies (Supplementary Figure 3.1). Second, synteny between the two assemblies was evaluated using Satsuma2 Synteny (Grabherr et al., 2010) and a circos plot generated using the circlize package v 0.4.14 (Gu et al., 2014) in R v 4.1.2 (Figure 3.1).

1.13.13 Synteny analyses

To shed light on the origin and evolution of the W chromosome in Lepidoptera, first, synteny-based homology between the *H. misippus* W chromosome, the Z chromosome and the autosomes was evaluated using Satsuma2. Satsuma2 is an aligner of whole genome assemblies intended to find homology based on sequence similarity. Satsuma2 first maps all genomic windows of the query genome to the target genome with a percentage of identity higher than 45%, and then filters those hits based on large scale synteny, that isit keeps only matches that are concordant with each other. Thus, Satsuma2 is not only aligning genomic windows, but also evaluating synteny between those blocks. Finally, Satsuma2 focusses on regions with high number of hits, to exhaustively evaluate the region around those hits, a strategy analogous to the battleship game. To evaluate synteny between the W and the remaining 31 chromosomes in *H. misippus,* the W chromosome was used as query and the 31 assembled chromosomes as target for Satstuma2. Resulting mapped regions were then filtered to keep only non-overlapping regions using the package GenomicRanges in R v4.1.2 (Lawrence et al., 2013). Satsuma2 does not require genome assemblies to be masked for their analysis. Because of its algorithm and filtering steps, repeat regions mapping to multiple places in the genome have decreased score and may be filtered out. However, to evaluate only the synteny of non-repeat regions, the RepeatMasker output, which details the coordinates of repeats in the genome, was used to filter out repeat regions of the W chromosome. Finally, the effect of sequence identity was evaluated by performing the analysis with no identity filter and applying a threshold of 70% of similarity.

The synteny between the W and the rest of the genome was also evaluated in 6 other Lepidoptera species: *Boloria selene* (GCA_905231865.2)*, Crocallis elinguaria* (GCA_907269065.1)*, Hemaris fuciformis* (GCA_907164795.1)*, Papilio machaon* (GCA_912999745.1)*, Watsonalla binaria* (GCA_929442735.1)*, Zygaena filipendulae* (GCA_907165275.2) and *Dryas iulia* (GCA_019049465.1). In all cases, the W chromosome was used as query, results were converted to non-overlapping regions and percentage of the W covered by matches calculated. No filter based on repeats was applied, as no repeat library was available for these species.

After that, Satsuma2 was used to compare the synteny of the *H. misippus* W (used as query) and Z chromosomes to 10 Lepidoptera species, the 6 from the previous analysis and also *Hemaris fuciformis* (GCA_907164795.1)*, Marasmarcha lunaedactyla* (GCA_923062675.1)*, Vanessa cardui* (GCF_905220365.1), and *Mythimna farrago* (GCA_910589285.2). Chromosomes of all the species analysed were re-named by their homology to *M. cinxia* as above. Again, results were converted to non-overlapping regions, repeat regions of the W chromosome filtered and percentage of the W chromosome covered by matches calculated. Using only the *H. misippus* W chromosome as query ensures that secondary and tertiary matches are also reported. A more specific analysis was produced by using the *H. misippus* assembly as query to search for synteny in the 10 Lepidoptera target species. Using whole assemblies as input limits the results to only primary matches, pairing all homologous autosomes.

1.13.14 BLAST of *H. misippus* W chromosome genes

To further explore the degree of homology between the *H. misippus* W chromosome and the autosomes and Z chromosome, protein sequence homology was assessed using BLASTp v2.4.0+ (BLAST, 2013; Kent, 2002). First a protein BLAST database was built using the protein sequences contained in the *H. misippus* autosomes and the Z chromosome in the HypMisi_v2 annotation. Then, protein sequences found in the W chromosome were extracted and blasted to the database setting the minimum e-value to 1e-10 and the maximum number of target sequences to be reported to 5,000. The best match for each protein sequence was selected based on e-value score, and if two matched had the same value percentage of identity was used. Finally, genome coordinates were extracted from the HypMisi_v2 annotation.

1.14 Results and Discussion

1.14.1 Genome assemblies and synteny between *HypMis_v2* and *HypBol_v1*

The size of the final assemblies were 438.07 MB for HypMis_v2 and 444.68 MB for HypBol_v1 , assembled into 218 and 59 scaffolds respectively. HypMis_v2 was sequenced using a trio-binning strategy, which, by using a combination of short and long-read sequencing of the parents and offspring of a cross, allows for the assembly of two parental haplotypes (Yen et

al., 2020). I assembled both parental haplotypes to a quasi-chromosome level and then scaffolded haplotype 2, which corresponds to the maternal haplotype, using Hi-C data (Suppl. table?). Hereafter, all mentions of the *H. misippus* assembly, HypMis_v2, refer to the Hi-C scaffolded haplotype 2.

HypMis_v2 was assembled into 32 chromosome-level scaffolds (>6 MB) and 186 unplaced scaffolds smaller than 1MB (Figure 3.1B) and has an N50 of 14.6 MB; while HypBol_v1 was assembled into 59 scaffolds all placed onto 31 chromosomes, with an N50 of 15.2 MB. Thus, both species have retained the ancestral karyotype of 31 chromosomes. This karyotype is present in other Nymphalinae species such as the painted lady *Vanessa cardui,* which has a comparable genome size (424.8 Mb) to the two *Hypolimnas* species and whose chromosomes are highly similar, without any fusion or fission observed, but with multiple large rearrangements (Supplementary Figure 3.1).

In general, all chromosomes were slightly larger in *HypBol_v1* compared to *HypMis_v2* with only 4 exceptions (chromosomes 13, 14, 28 and 31; Figure 3.1C). When aligning the two *Hypolimnas* assemblies, there were no fusions or fissions among chromosomes (Figure 3.1A). The only caveat to these conclusions is that *HypMis_v2* was used to scaffold *HypBol_v1* before using the linkage map, which might have altered the *H. bolina* chromosomal structure. Nonetheless the linkage map applied after the Ragout scaffolding should correct for any large-scale errors in scaffolding. Finally, 12 large inversions on multiple chromosomes are apparent in the alignment of the two assemblies (Suplementary Figure 3.1).

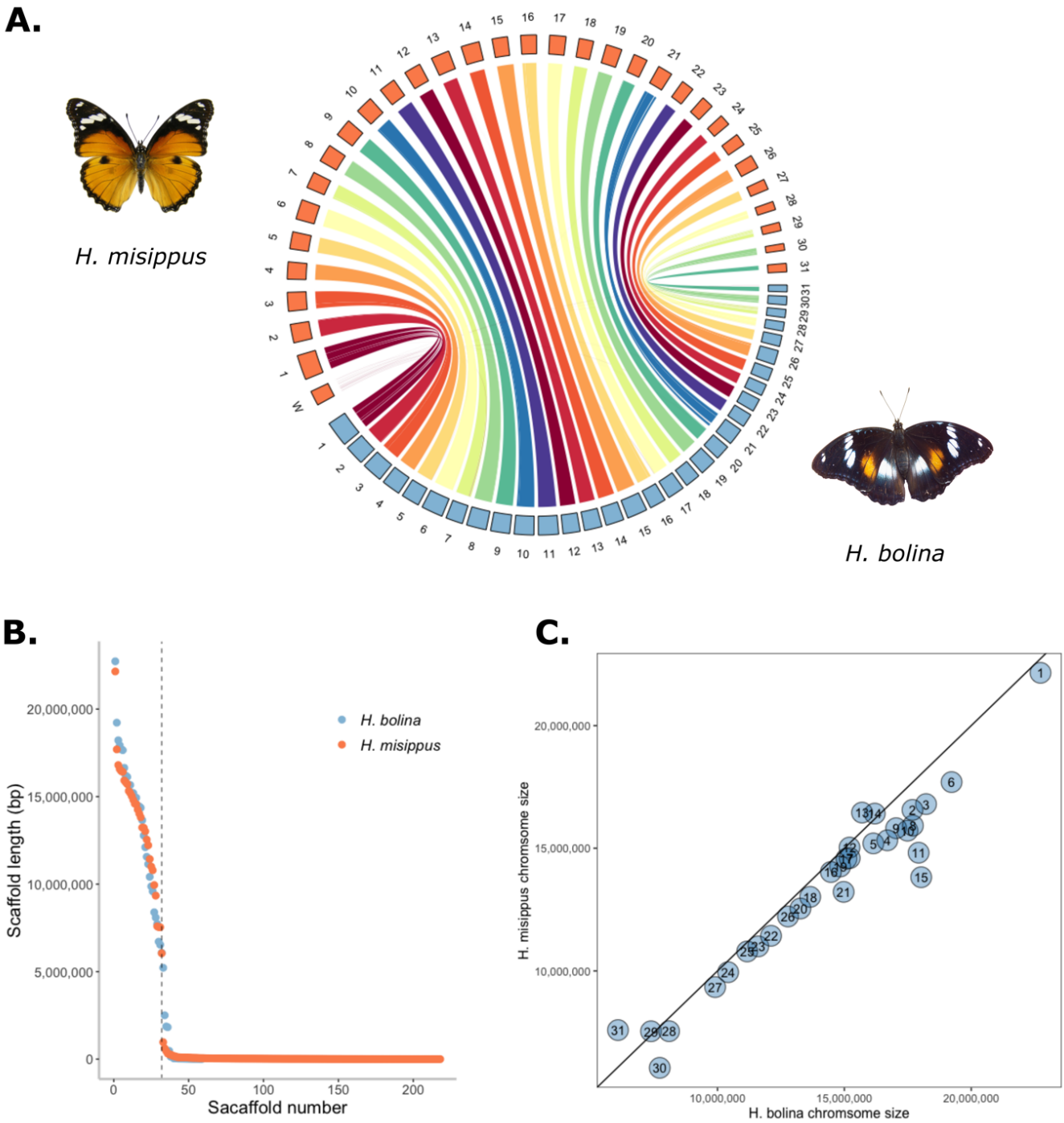


Figure 3.1 | Chromosome level assemblies for *Hypolimnas misippus* and *H. bolina*. **A.** Chromosomal synteny is conserved in both species. **B.** Chromosome level scaffolds have been assembled for both species. **C.** In general, *H. bolina* chromosomes are larger. Chromosome 1 refers to the Z chromosome.

1.14.2 Transposable elements and repeat content

In total, the two *Hypolimnas* genomes contain similar levels of repeats, 43.86% for HypBol_v1 and 42.34% for *H. misippus*, which are slightly higher than in the painted lady *Vanessa cardui* (37.27%), which has a comparable genome size (424.8 Mb). The composition of those repeats is similar in the two *Hypolimnas* genomes, while they differ substantially from the painted lady (Figure 3.2). The contribution of rolling-circles (also known as *Helitrons*) and DNA-transposons is more than double in *Hypolimnas,* while retroelements are >1.5 fold higher in the painted lady (Figure 3.2). This suggests a shift in TE activity, with *Helitrons* and DNA-transposons playing a more important role in *Hypolimnas* species. In both *Hypolimnas*, there has been both a recent and a more ancient expansion of *Helitron* family transposable elements (Figure 3.2B). Finally, the percentage of unclassified repeats and GC content is broadly the same in the three species.

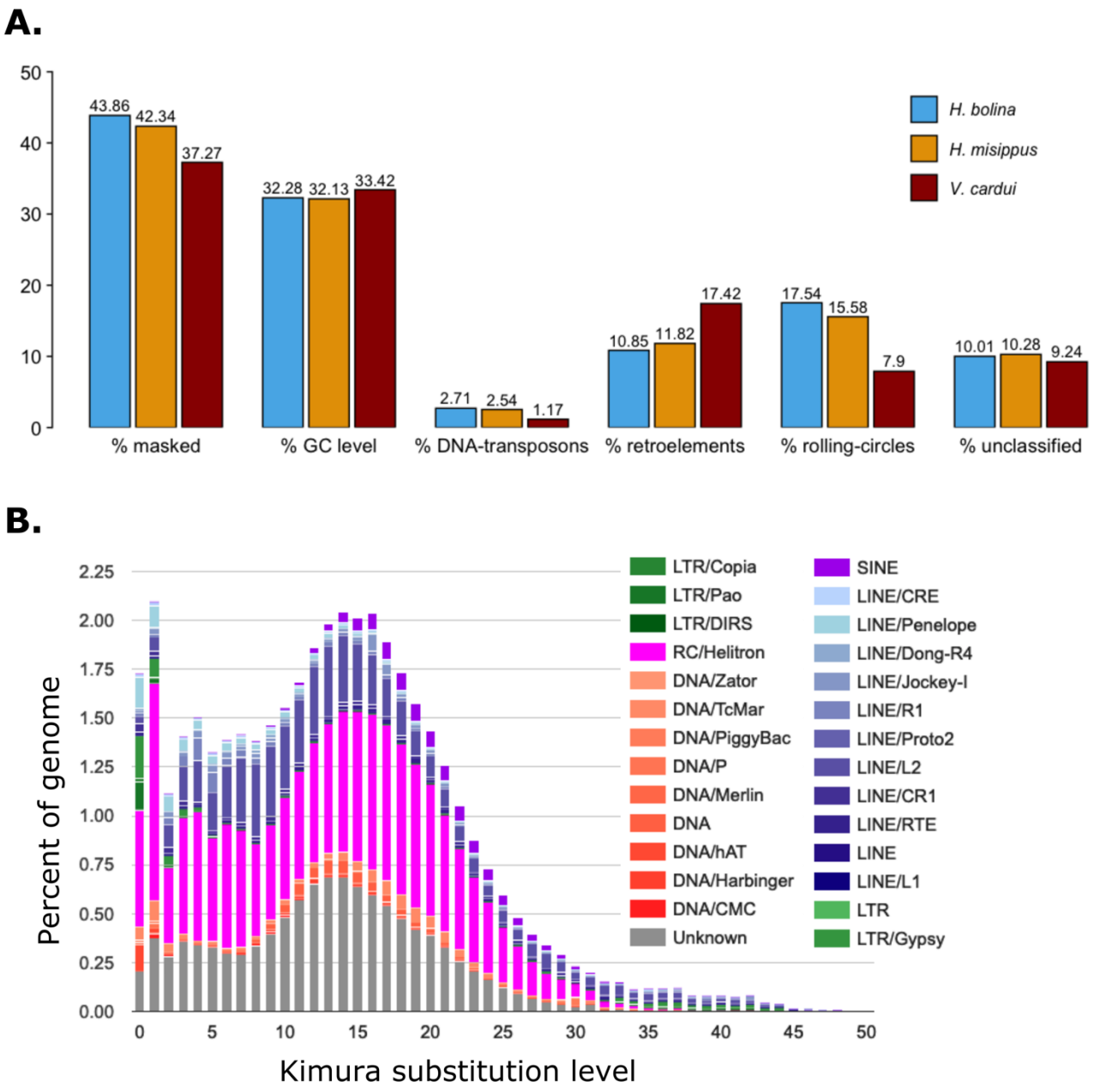


Figure 3.2 | A. Repeat content of the *Hypolimnas bolina, Hypolimnas misippus* and *Vanessa cardui* assemblies. **B**. Repeat landscape of the *H. misippus* assembly. Helitrons and LTR (Long Terminal Repeat) retrotransposons have undergone a recent expansion, as seen in the higher percentage of the genome covered by *Helitrons*.

1.14.3 Gene content and completeness of the annotation

In total 19,721 genes were annotated in HypBol_v1 using BRAKER and 21,784 coding mRNAs including all isoforms of the same gene; while for *H. misippus* there were 20,293 genes and 22,468 coding mRNAs. These numbers are higher than other Nymphalidae species such as the painted lady butterfly (*Vanessa cardui)*, whose latest annotation (v2.1) includes 13,223 genes and 19,836 mRNAs, with the number of genes being considerably larger in the two *Hypolimnas* (Figure 3.3). Analysis with BUSCO using the *insecta_odb10* benchmarking set showed that the completeness of the genome and annotation were 98.7% and 98.6% for *H. misippus,* and 94% and 92.4% for HypBol_v1 (Table 1). These scores are comparable to other published Nymphalidae assemblies and annotations such as that of *Danaus chrysippus* (Singh et al., 2022), with those for *HypBol_v1* slightly lower possibly due to errors introduced by Nanopore sequencing, which has a high per base error rate (although polishing with Illumina data was performed to minimise errors in the assembly).

Despite having a larger number of genes and mRNAs, the two *Hypolimnas* annotations show a smaller total number of exons and introns than *V. cardui* (Figure 3.3). This is because the two *Hypolimnas* annotations have more single exon genes and their mRNAs have, on average, a smaller number of exons, which might be a difference produced by the distinct annotation pipelines. The total length of mRNAs and genes in the two *Hypolimnas* is shorter then in *V. cardui*, which results in a smaller percentage of the genome covered by them. However, this trend is different for exons, which are on average the same length in the three species, have a comparable total length and cover a similar percentage of the genome. Thus, the longer total length of mRNAs and genes in *V. cardui* is due to a total longer length of introns, due to a higher number of introns per mRNA and an average longer length.

Table 3.1 | BUSCO scores for the genome assemblies and gene annotations of *H. misippus* and *HypBol_v1* calculated using the Insecta_odb10 (n=1367) set of genes. Scores for two parental haplotypes of *Hypolimnas misippus* are shown.

Type	Species	Assembly	Complete	Single	Duplicated	Fragmented	Missing
Genomes	*Hypolimnas misippus* haplotype 1 (parental)	Trio-HiFi	98.7	98.5	0.2	0.4	0.9
	Hypolimnas misippus haplotype 2 (maternal)	Trio-HiFi-HiC	98.7	98.5	0.2	0.4	0.9
	Hypolimnas bolina	Pilon polishing 3rd round	94.0	93.6	0.4	1.8	4.2
Annotations	*Hypolimnas misippus*	Trio-HiFi-HiC	98.6	89.8	8.8	0.6	0.8
	Hypolimnas bolina	Pilon polishing 3rd round	92.4	84.1	8.3	2.9	4.7

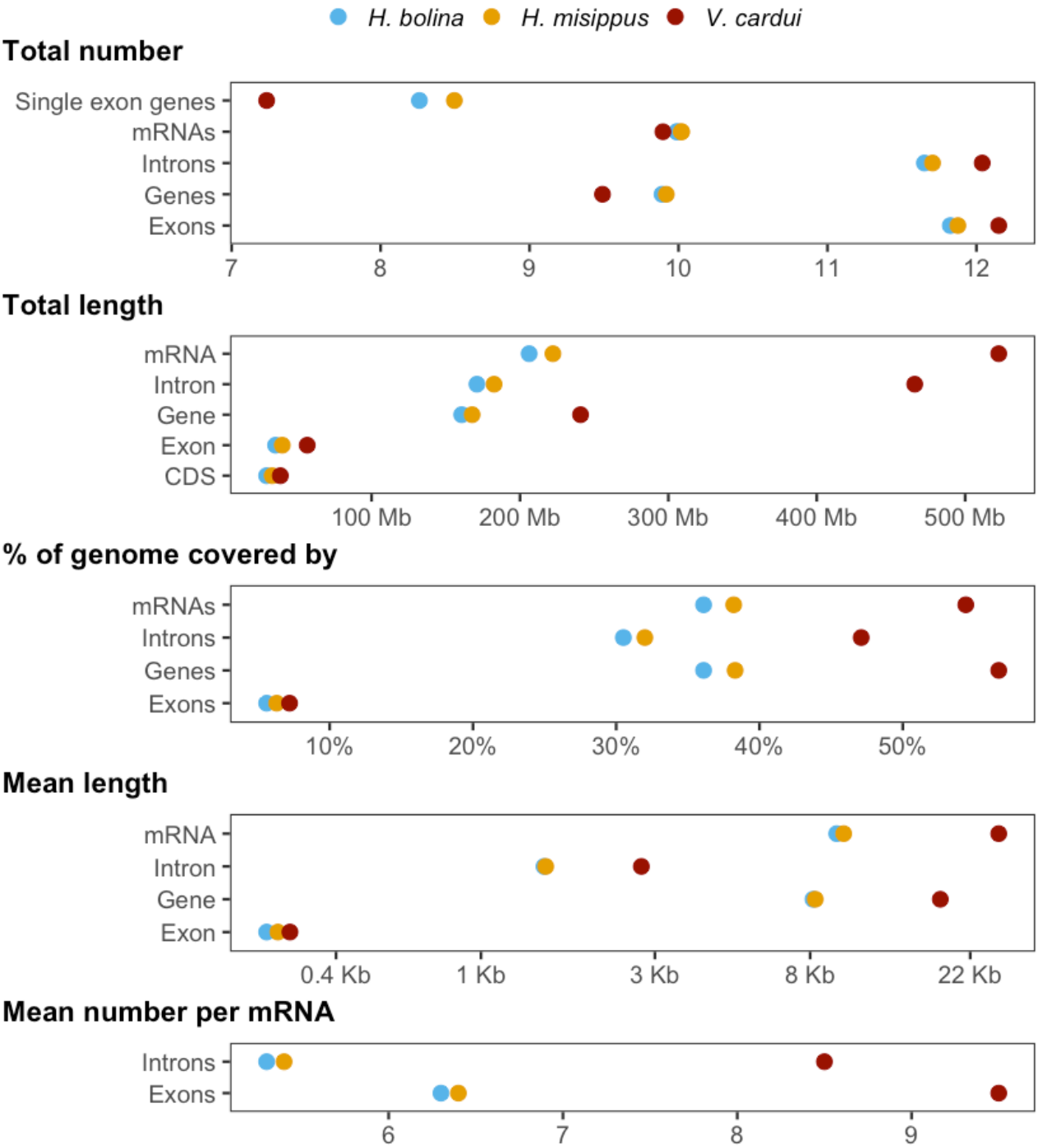


Figure 3.3 | Annotation comparison of *Hypolimnas misippus, H. bolina* and *Vanessa cardui.* For the statistics, pre-mRNA, that is mRNA including exons, have been used. Numbers are found in Supplementary Table 3.1.

1.14.4 Mixed evidence on the origin of W chromosomes from the same autosome pair as the Z

A few models have been proposed to explain the origin and evolution of the W chromosome in Lepidoptera (Fraïsse et al., 2017; Lukhtanov, 2000; Sahara et al., 2012). The deep conservation of the Z chromosome and its apparent lack of homology with the W have been suggested to be evidence of a B chromosome origin of the W chromosome (Fraïsse et al., 2017). Outside the Lepidoptera, it has been commonly found that heteromorphic sex chromosomes retain a degree of similarity, evidence of their shared autosomal origin (Wright et al., 2016). Thus, if the W chromosome originated from the same homologous autosome pair as the Z, we might expect to find some residual sequence similarity between the W and Z. Furthermore, the lack of similarity among W chromosomes of Lepidoptera species has led to the suggestion of multiple independent recruitments from B chromosomes (Lewis et al., 2021). Taking this into account, I set out to elucidate two questions. First, did the W chromosome evolve from the same homologous pair of autosomes as the Z or from a B chromosome? Second, do W chromosomes in Lepidoptera share a common origin, that is, did they evolve once or multiple times independently?

To specifically test the hypothesis that the Z and W have a common autosomal origin, I used a synteny analysis to infer sequence homology between the *H. misippus* W chromosome and the remaining 32 chromosomes. I found that the Z chromosome shows a higher degree of similarity to the W chromosome than any autosome, defined as the percentage of the W that is covered by non-overlapping matches to the target chromsome (Figure 3.4A). However, the W chromosome is the chromosome with the highest percentage of repeat regions (71% in *H. misippus*), which could be interfering with the analysis. Furthermore, the Z chromsome is the longest chromosome in *H. misippus* and has the highest absolute length of repeats (Supplementary Figure 2 and 3). To take this into account, I filtered out of the synteny analysis all the repeat regions in the *H. misippus* W chromosome. With this correction, the Z chromosome was the second most similar chromosome to the W after chromosome 2. These results differ from the homology tests performed in *Danaus plexipuss* (Lewis et al., 2021)*, Kallima inachus* (Yang et al., 2020) and *Dryas iulia* (Lewis et al., 2021), where no homology was found between the W and the Z chromosome.

To further explore the possible homology between the W and Z chromosomes, I extracted the predicted sequences of the proteins in the *H. misippus* W chromosome and searched for homology in the remaining chromosomes using BLASTp. Contrarily to the synteny results, I found that the Z chromosome placed 4th in the number of best BLAST protein hits with 94 proteins mapping to it, while 104 mapped to chromosome 2, 3 and 14. The average number of best BLAST hits was 66. These results suggests that the W chromosome did not evolve from the same homologous autosomal pair as the Z. However, as a result of the arrest of recombination and subsequent accumulation of mutations, sex specific chromosomes, that is W and Y, are often gene poor and highly degenerated compared to their autosomal counterparts (Wright et al., 2016). Additionally, lepidopteran W chromosomes have been shown to evolve rapidly (Vítková et al., 2007; Yoshido et al., 2013). Thus, the lack of homology between proteins of the W and other chromosomes is expected given the rapid turnover seen in the W chromosome and is in line with results shown in other Lepidoptera species (Lewis et al., 2021; Singh et al., 2022; Yang et al., 2020). Interestingly, chromosome 2 showed the highest sequence similarity in non-repeat regions and the highest number of best BLAST hits (Figure 3.4C).

In view of the differences between my results and previously published homology tests in other species, I decided to expand my analysis to include other Lepidoptera species. I performed synteny analysis between the W and other 31 chromosomes in 6 diverse taxa including *Dryas iulia, Boloria selene, Papilio machaon, Crocallis elinguaria, Watsonalla binaria,* and *Zygaena filipendulae*. These species were chosen from the high-quality assemblies containing a W chromosome publicly available and were specifically picked because they cover distinct ditrysian lineages (including 4 families: *Zyganoidea, Papilonoidea, Drepanoidea* and *Geometroidea*). No homology was observed between the W and Z in any species except for *P. machaon*, which contrasts with the results in *H. misippus* (Figure 3.4D). However, these analyses were performed using all regions as no repeat annotation was available for these species, which could be interfering with the results.

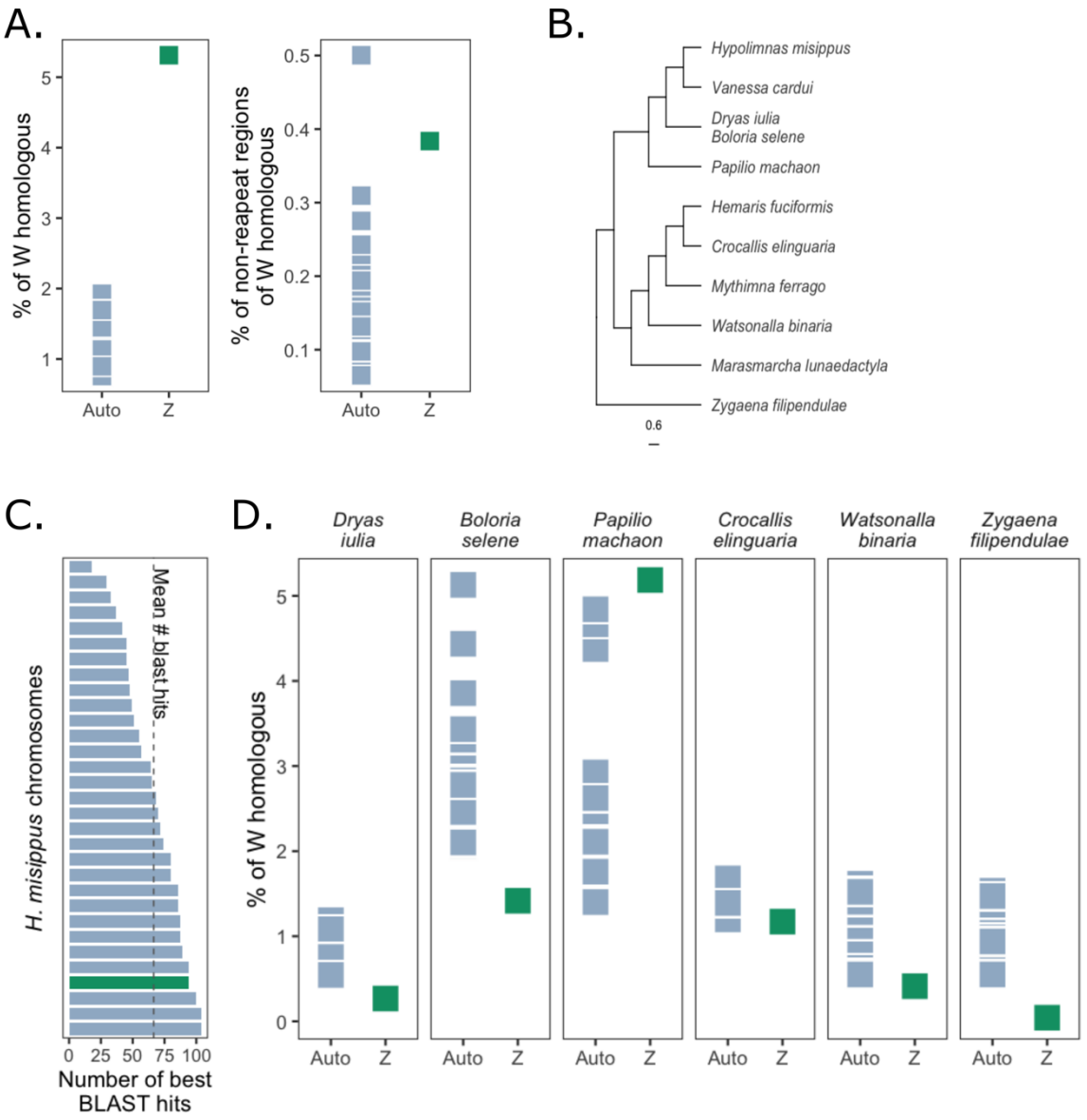


Figure 3.4 | Genomic synteny provides mixed evidence on the origin of the W chromosome. **A.** Synteny analysis of the *H. misippus* W chromosome and the autosomes and Z chromosome shows that the W and Z chromosome have the highest similarity. The analysis of non-repeat regions shows that the Z chromosome is the second most similar chromosome to the W after chromosome 19 (right). **B.** Phylogenetic tree modified from Kawahara, et al 2019 (Kawahara et al., 2019) including the species used in the synteny analysis. **C.** Number of best BLAST hits of the proteins found in the *H. misippus* W chromosome to the other chromosomes. Numbers summarised by chromosome with the Z shown in green. **D.** Synteny analysis of the *H. misippus* W chromosome compared to several Lepidoptera genome assemblies shows homology between W chromosomes across many species and suggests a common origin.

Taken together the homology tests between the W and remaining 31 chromosomes in *H. misippus* and other Lepidoptera provide mixed evidence for the origin of the W from the same homologous pair of autosomes as the Z. The lack of homology between the W and the Z in so many species could suggest a B chromosome origin or could be the result of the rapid evolution of the W.

1.14.5 Synteny analysis of the *H. misippus* W chromosome reveals a possible common origin of the W chromosome in ditrysian taxa

The second question I wanted to clarify was whether the W has a single origin in the Lepidoptera or evolved independently multiple times. Previously, the lack of similarity among W chromosomes of Lepidoptera species has been suggested to be the result of multiple independent recruitments from B chromosomes (Lewis et al., 2021). To explore the evolution of W chromosomes within ditrysian lineages, I performed synteny analysis of the *H. misippus* W chromosome compared to a diverse set of 10 Lepidoptera species including the 6 species from previous analysis and *Vanessa cardui, Hemaris fuciformis, Mythimna farrago, Watsonalla binaria* and *Marasmarcha lunaedactyla.* These species cover the breadth of ditrysian lineages, including 8 distinct families: *Zyganoidea, Papilonoidea, Pterophonoidea, Pyraloidea, Drepanoidea, Noctuoidea, Geometroidea* and *Bombycoidea*. When comparing the *H. misippus* W chromosome to these species, I found that there is a high level of similarity among W chromosomes, defined again as the percentage of the W chromosome covered by matches to target chromosomes (Figure 3.5A). If the W chromosomes of ditrysian lineages share a common origin, we would expect them to have a higher level of similarity than expected by chance and thus higher than the similarity to the autosomes or Z chromosome, which is what I observe. The highest degree of similarity is seen between W chromosomes of closely related species such as *H. misippus* and *Vanessa cardui* (45 MYA), but is also true even for highly divergent lineages such as *Zygaena filipendulae,* which diverged from *H. misippus* 156 MYA (Kawahara et al., 2019). Interestingly, there is no consistent pattern in the distribution of homologous regions along the *H. misippus* W chromosome, which means that different regions show homology to the *H. misippus* W chromosome across species and chromosome types (W, Z and autosomes;

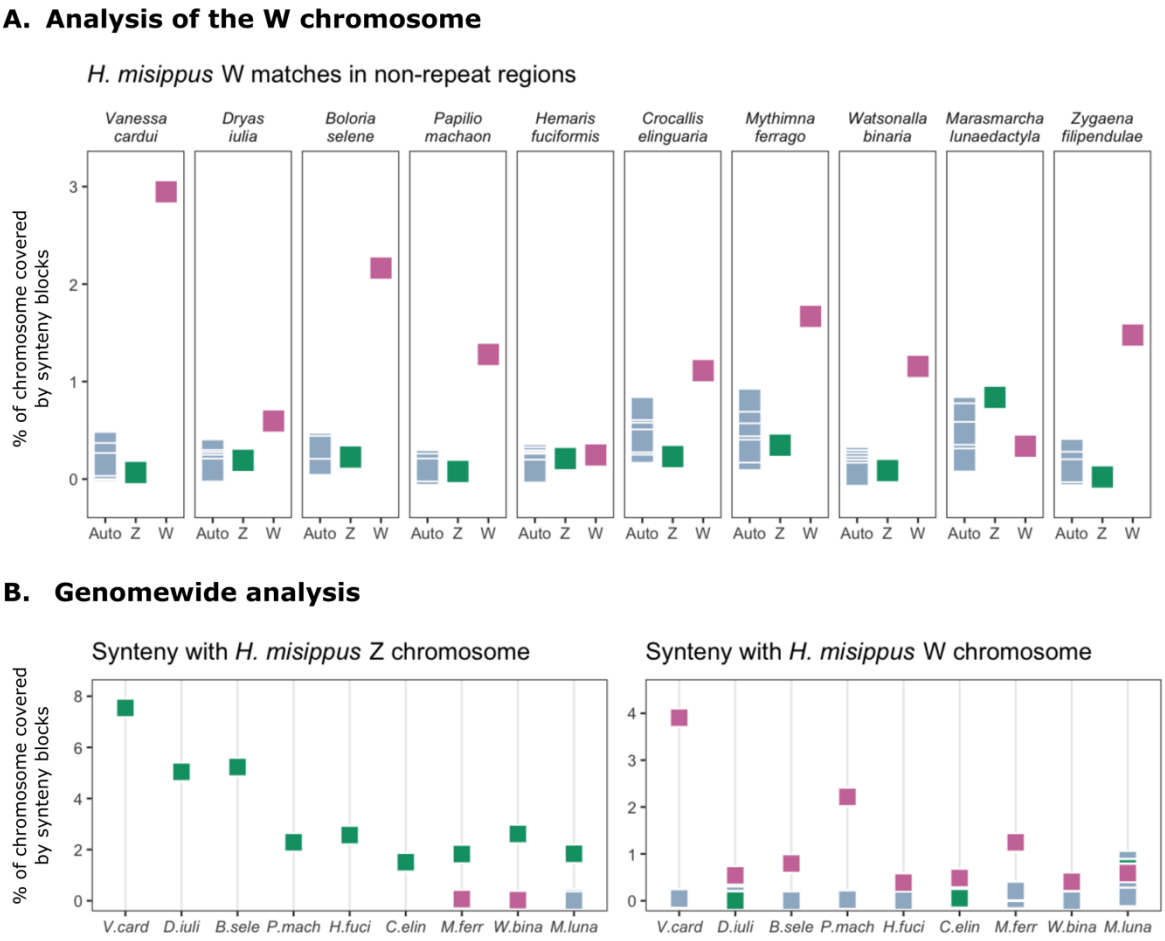


Figure 3.5 | Genome wide synteny analysis suggests a single origin of ditrysian W chromosomes and shows deep conservation of the Z chromosome. **A**. Percentage of query *H. misippus* W chromosome matched to the 32 chromosomes of 10 ditrysian species. W *H. misippus* chromosome used as query. **B.** Subset results of genome wide analysis. Percentage of Z (left) and W (right) chromosome covered by synteny blocks to other ditrysian species. W chromosome shown in pink, Z in green and autosomes in grey. Species names have been abbreviated: *Vanessa cardui, Dryas iulia, Boloria selene, Papilio machaon, Hemaris fuciformis, Crocallis elinguaria, Mythimna ferrago, Watsonalla binaria, Marasmarcha lunaedactyla* and *Zygaena filipendulae.*

Supplementary Figure 4). Finally, the Z chromosome does not show more similarity to the W than the autosomes.

Given these results showing conservation of the W chromosome across several species, I increased stringency by including all the *H. misippus* chromosomes and not only the W. This reduces the number of hits of the W chromosome on the target genome, because hits between homologous autosomes are also reported. Consistent with previous results, the W chromosomes have the highest degree of similarity in most species' comparisons (Figure 5B). Also, in line with results from other species, the Z chromosome shows deep conservation across the species studied (Fraïsse et al., 2017).

1.14.6 A multi-species synteny comparison reinforces the hypothesis of the single origin of the W chromosome and reveals a neo-Z chromosome

In light of the results of the *H. misippus* synteny analyses suggesting a single origin of the W chromosome in ditrysian lineages, I decided to explore this further by comparing the analysed ditrysian assemblies with each other. I performed pairwise synteny analyses between all species pairs and observed consistent results. First, the Z chromosome is deeply conserved in multiple species pairs and the homology between Z chromosomes of different species decays at the same rate as autosomal homology with species divergence (Supplementary Figure 5). Second, the W chromosome shows a more variable pattern of homology between species but is conserved even with distant species pairs (Supplementary Figure 5). Nonetheless, several of the comparisons found no homology between W chromosomes. Finally, the cross-species comparison revealed that the Z chromosome of *Marasmarcha lunaedactyla* (chromosome name OV181339.1; identity as the Z chromosome assigned in the public assembly) presented high levels of synteny with chromosome 19 of most other species (Supplementary figures 6 and 7). This could suggest that a fusion took place between the ancestral chromosome 19 and the Z chromosome of *M. lunaedactyla* creating a neo-Z chromosome. The pattern of homologous regions seen in the OV181339.1 chromosome was similar for all species compared, with approximately half of the chromosome showing homology to chromosome 19 and half to the Z chromosome (Supplementary figures 7). Furthermore, the pattern of homology with other species showed that recombination events between the parts belonging to the ancestral Z and ancestral chromosome 19 have resulted in an interleaved pattern of

homology blocks. Consistent with these results, BUSCO matches of chromosome OV181339.1 are shared with chromosome 19 and chromosome 1 (the Z chromosome) of *Melitaea cinxia* (Supplementary Figures 6 and 7).

1.15 Conclusions

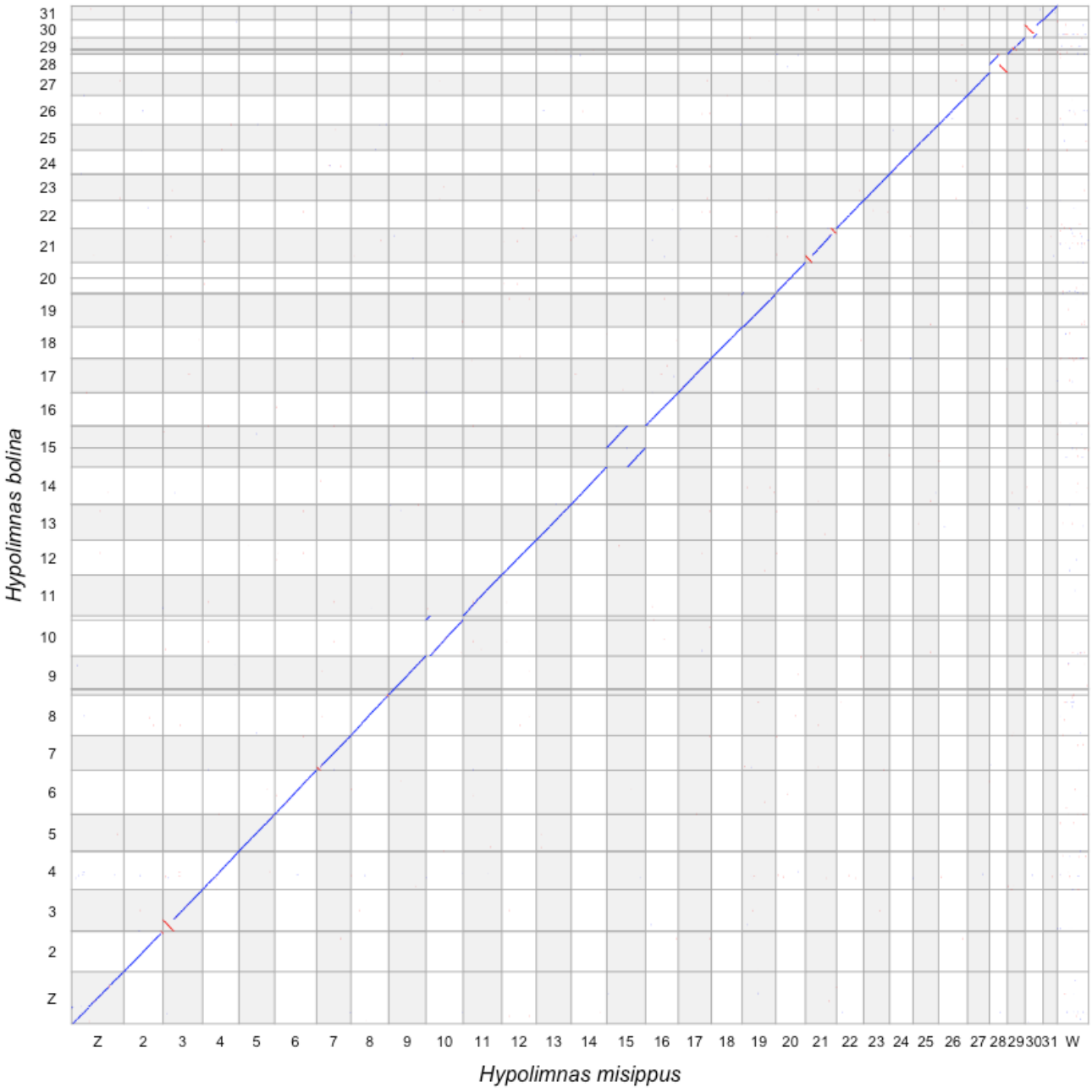
Hypolimnas species have become a focus for studies of evolutionary genetics, due to their mimetic colouration and co-evolution with *Wolbachia* parasites. Here, I have assembled chromosome level reference genomes for *Hypolimnas misippus* and *bolina* and RNA-informed annotations for both genomes. HypMis_v2 is a significant improvement on a previously published genome, with higher contiguity and higher N50 and BUSCO scores. I also used RNA-seq data from *H. misippus* to inform our annotation, as opposed to only homology methods.

Using the *H. misippus* assembly, I have provided evidence of a possible shared origin of W chromosomes in the ditrysian lineages of the Lepidoptera. My results show that there is a degree of synteny between the *H. misippus* W and the Z, which could suggest a common autosomal origin of the heterogametic pair. However, the rapid evolution of the W chromosome may be limiting the inference of homology (Vítková et al., 2007; Yoshido et al., 2013). Further studies in other Lepidoptera species are necessary to clarify whether the W chromosome evolved from an autosome or from a B chromosome, particularly including non-ditrysian lineages. Finally, my results suggest that W chromosomes of ditrysian lineages may share a common origin. These contrasts with evidence of rapid turnover of W chromosomes in Lepidoptera. However, it is possible that some species have retained the ancestral W, while some other have recruited other B chromosomes as suggested. Including more species across the Lepidoptera will shed light on the possible shared origin of W chromosomes.

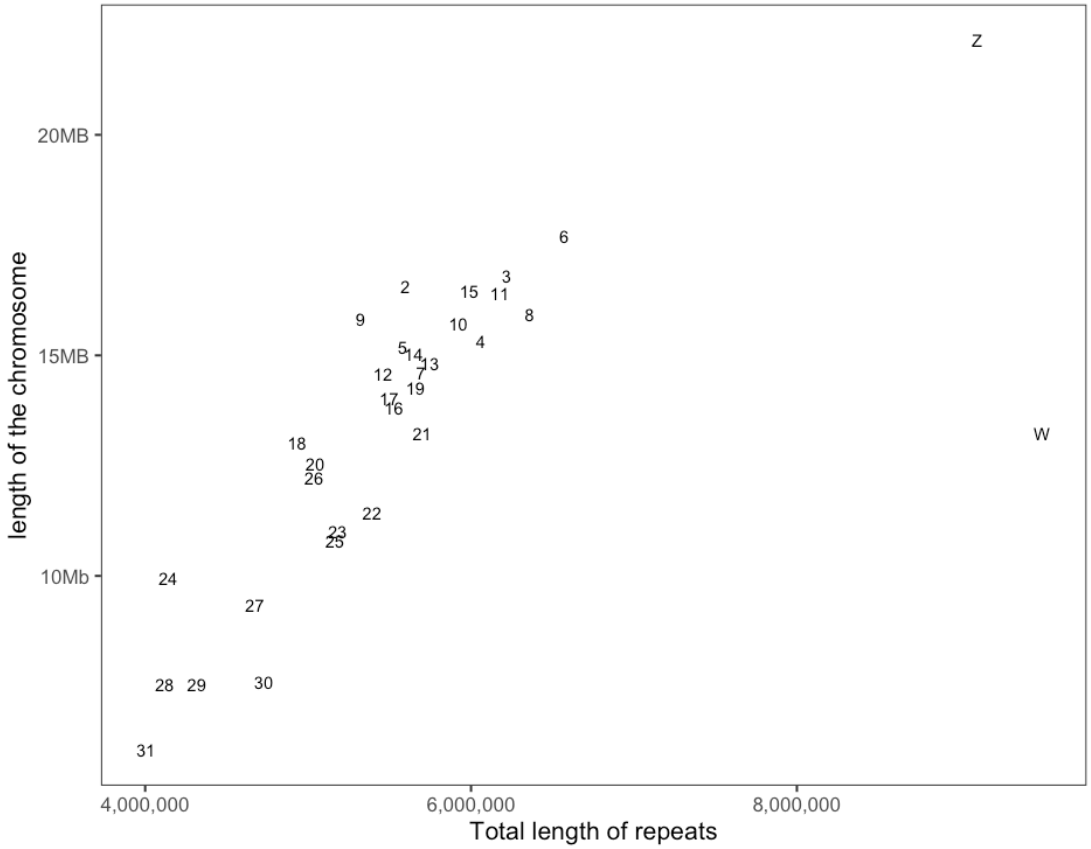
1.16 Supplementary Information

Feature	*HypBol_v1*	*HypMisi_v2*	*Vanessa cardui*
Total length: Gene	160709320	167739620	240814091
Total length: mRNA	206176689	222209628	522618316
Total length: CDS	29265705	32999245	38521385
Total length: Exon	35299426	39816515	56622152
Total length: Intron	170992076	182514332	465996164
Mean number per mRNA: Exons	6.3	6.4	9.5
Mean number per mRNA: Introns	5.3	5.4	8.5
Mean length: Gene	8149	8265	18211
Mean length: mRNA	9464	9890	26346
Mean length: Exon	258	277	299
Mean length: Intron	1489	1505	2754
Total number: Genes	19721	20293	13223
Total number: mRNAs	21784	22468	19836
Total number: Exons	136597	143687	189042
Total number: Introns	114813	121219	169206
Total number: Single exon genes	3868	4891	1387
% of genome covered by: Genes	36.1	38.3	56.7
% of genome covered by: mRNAs	36.1	38.2	54.4
% of genome covered by: Exons	5.6	6.3	7.2
% of genome covered by: Introns	30.5	32	47.1

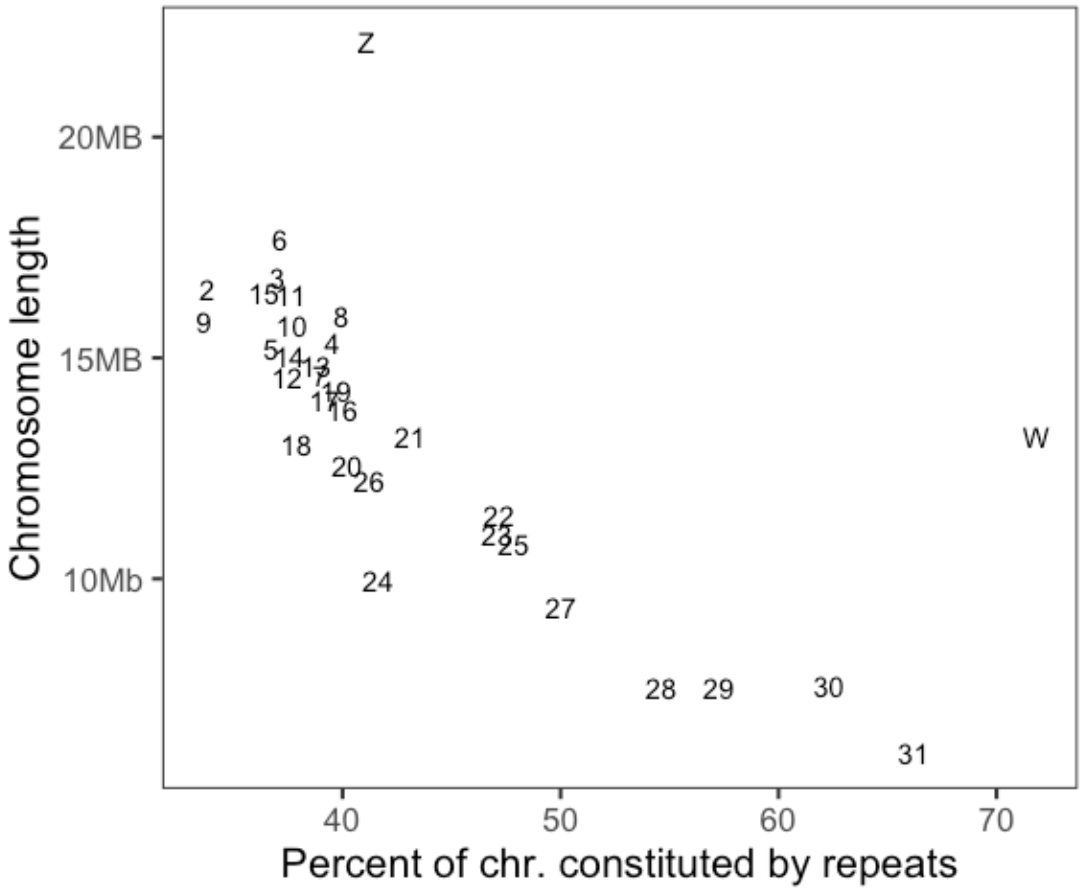
Supplementary Table 3.1 | Annotation statistics of *Hypolimnas misippus (HypMisi_v2)* and *H. bolina (HypBol_v1)* compared to *Vanessa cardui.*



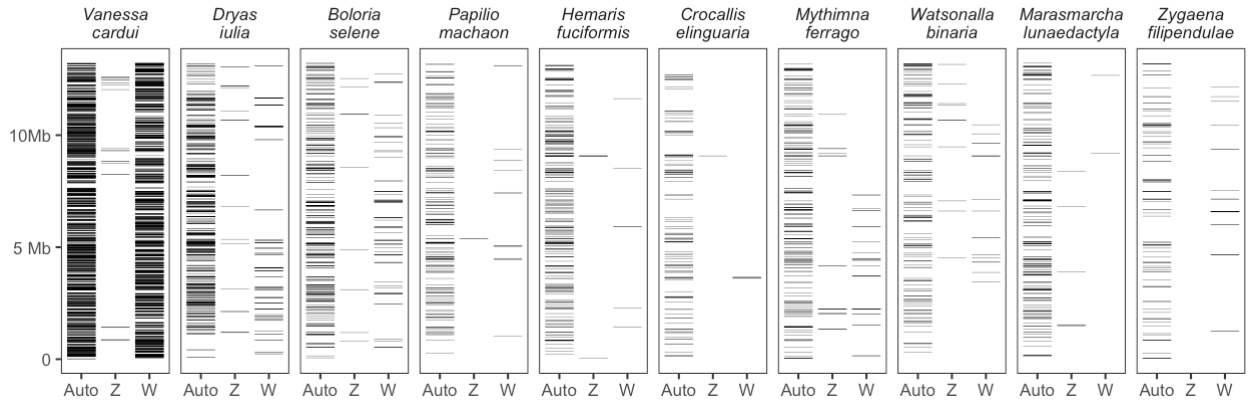
Supplementary Figure 3.1 | The D-Genies whole assembly alignment of *Hypolimnas misippus* and *H. bolina* reveals 12 inversions between the two genomes (shown in red).



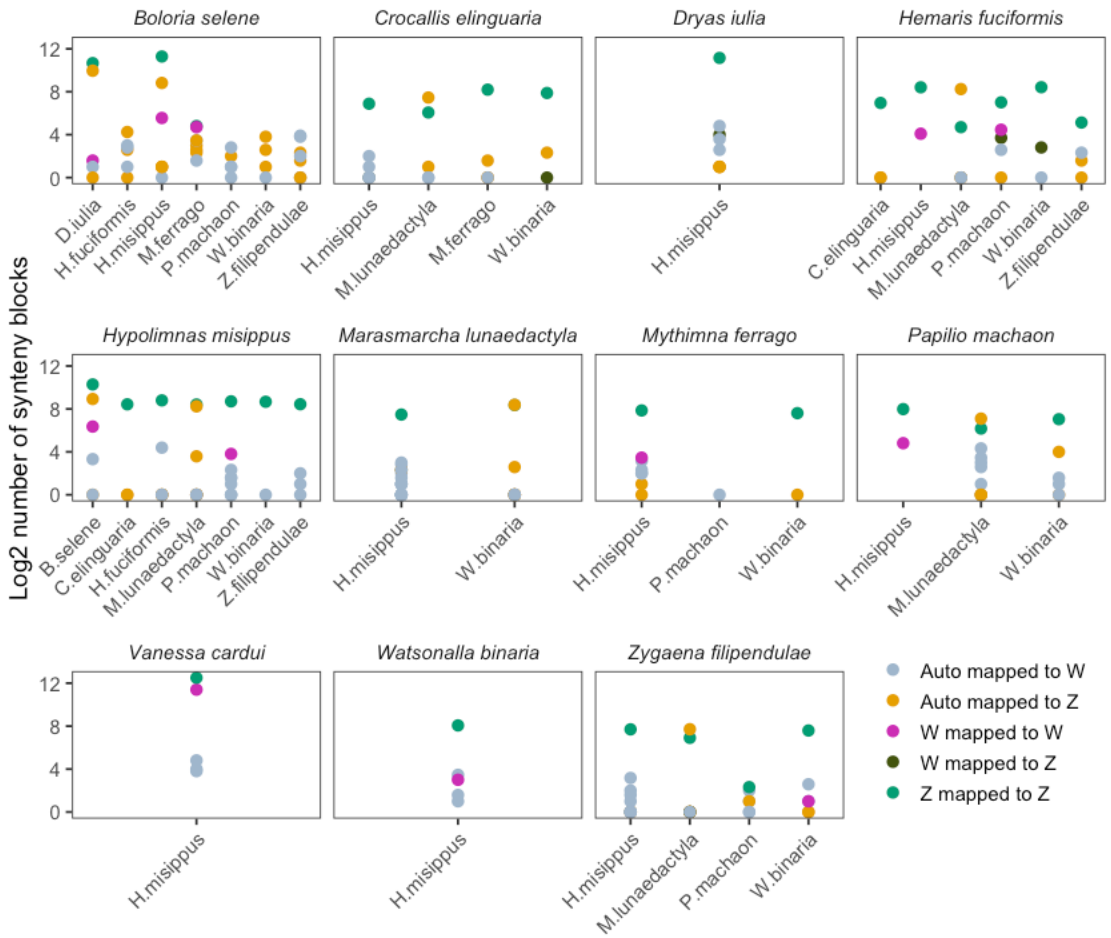
Supplementary Figure 3.2 | Repeat content of the *HypMisi_v2* by chromosome. Larger chromosomes tend to have more repeats. The W chromosome deviates from the correlation and shows a higher repeat content for its length (see also Supplementary Figure 3).



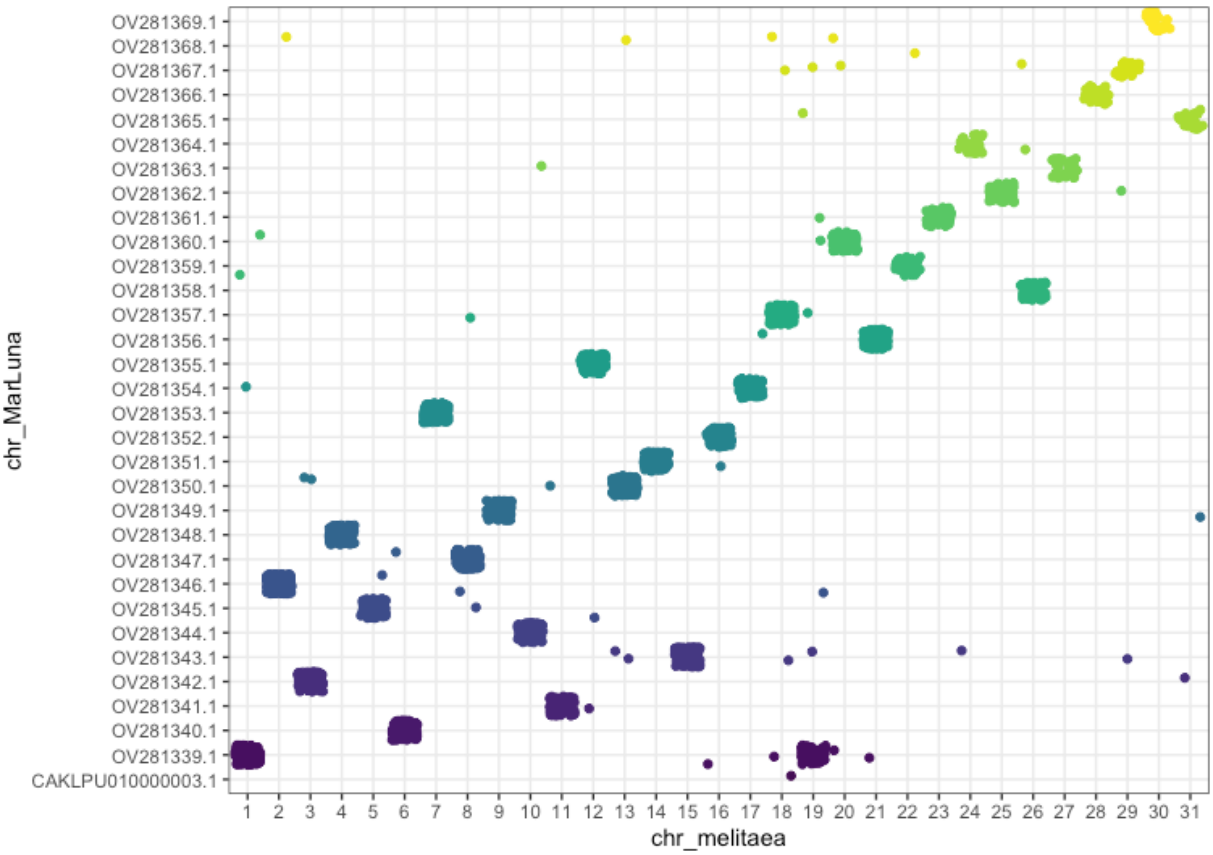
Supplementary Figure 3.3 | Percentage of each chromosome of the *HypMisi_v2* assembly constituted by repeats. The W chromosome has the highest percentage of repeats.



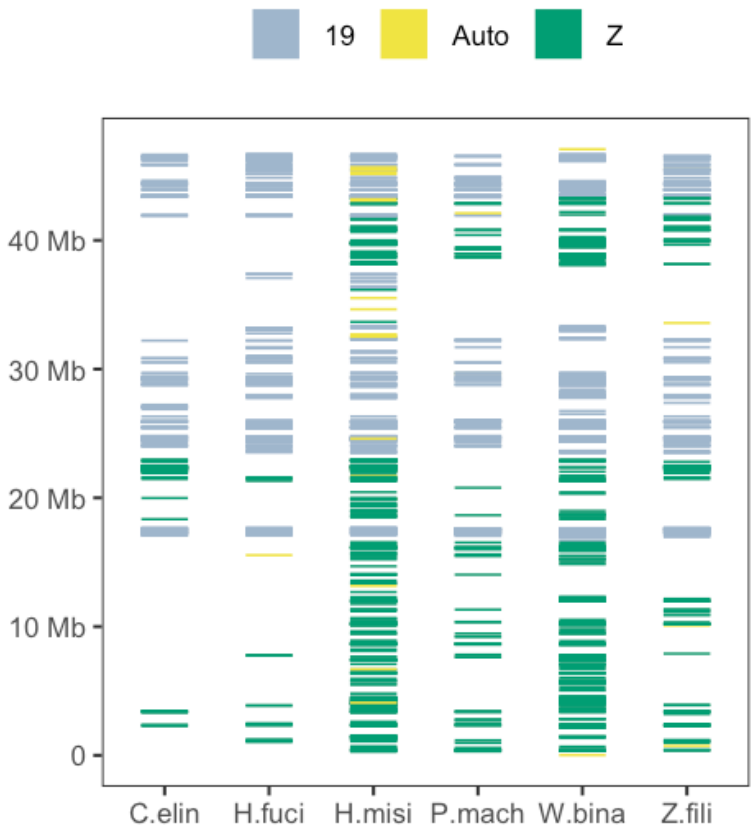
Supplementary Figure 3.4 | The positions of synteny blocks between the *H. misippus* W chromosome and the chromosomes of other species do not reveal a conserved section of the W chromosome across species. Left-most column shows all synteny blocks with autosomes plotted.



Supplementary Figure 3.5 | Synteny analysis using Satsuma2 between 10 ditrysian species reveals deep conservation of the Z and some conservation of the W chromosome across species. Auto refers to any autosomes.



Supplementary Figure 3.6 | Homology of *M. lunaedactyla's* chromosomes with *Melitaea cinxia* reveals a neo-Z chromosome. Chromosome OV281339.1 of *M. lunaedactyla* shares BUSCOs with chromosome 1 (Z) and chromosome 19 of *M. cinxia*, suggesting a fusion of these two.



Supplementary Figure 3.7 | Synteny analysis using Satsuma2 of the neo-Z chromosome of *M. lunaedactyla* compared to other ditrysian assemblies reveals the patterns of homology with the ancestral Z and chromosome 19.

2

Transposable element insertions are associated with Batesian mimicry in the pantropical butterfly Hypolimnas misippus

1.17 Abstract

Structural variants can play a crucial role in adaptation and speciation. *Hypolimnas misippus* is a pantropical species of butterfly that has female limited Batesian mimicry. *H. misippus* females are polymorphic showing detailed resemblances to the four morphs of *Danaus chrysippus* and are maintained as a stable polymorphism, with little correlation to local model frequencies. Using haplotagging, a linked-read sequencing technology, I show that two large insertions are found at the intergenic locus associated with mimicry in *H. misippus* and validate these using molecular methods. I demonstrate that the insertions spanning 2.4 kb and 4.8 kb are formed by several transposable elements (TE) and are found in the dominant allele which produces a mimetic phenotype. Through comparative analysis with *Hypolimnas* species, I explore the evolution of the locus and show that the dominant allele containing the insertions is derived, suggesting that the TEs disrupt a cis-regulatory element that leads to the reversion of the phenotype. Our results add to the growing evidence of the importance of TE insertions in the evolution of colour phenotypes in Lepidoptera. Additionally, I develop a program, Wrath, for the visualisation and exploration of candidate structural variants from haplotagging data.

1.18 Introduction

Structural variation (SV) forms a large part of the genetic variation observed in wild populations and can play a key role in adaptation and speciation (Auton et al., 2015; Wellenreuther et al., 2019). SVs are typically defined as events larger than 50 bp and include various combinations of gains, losses and rearrangement of genetic material, which can have extensive effects on gene content, as well as genetic contiguity (reviewed in (Ho et al., 2020)). These effects have major roles in adaptation and speciation in many species (reviewed in (Faria et al., 2019; Hoffmann and Rieseberg, 2008; Kondrashov, 2012)) as well as human disease (Weischenfeldt et al., 2013; Zeevi et al., 2019). For example, inversions have often been associated with complex phenotypes, as reduced recombination at the inversion ensures the joint inheritance of co-adapted alleles. Examples of this are seen in elytra colouration in ladybirds and reproductive morph switches in the ruff ((Ando et al., 2018; Gautier et al., 2018; Küpper et al., 2015; Lamichhaney et al., 2016); reviewed in (Küpper et al., 2015; Thompson and Jiggins, 2014) and Chapter 1). In other cases, gene duplications might give rise to adaptive loci through neo-functionalisation as seen in heterostyly in *Primula* plants and in the complex phenotypes of the wood tiger moth (see Chapter 5)(Brien et al., 2022; Li et al., 2016). Structural variants are crucial in the evolution of adaptive traits and improvements in sequencing techniques are accelerating our ability to detect and study them.

Butterfly wing patterns are a classic example of adaptive evolution. Evolutionary genetic studies have dissected the loci controlling wing pattern in several species of butterflies from a wide range of ecotypes and families (Beldade and Brakefield, 2018; Jiggins et al., 2017). Crucially, a few examples have found that SVs are involved in the control of wing phenotypes. For example, three large inversions found in *Heliconius numata* lock in multiple co-adapted loci that control several wing pattern elements (Joron et al., 2011). Similarly, transposable element insertions in some *Heliconius* species and in the peppered moth *Biston betularia* have been shown to cause the switch between adaptive wing phenotypes by altering the expression of the gene *cortex* (Livraghi et al., 2021; van't Hof et al., 2016). The diversity of phenotypes and importance in adaptive evolution make butterfly wing phenotypes a suitable model for the study of structural variation in adaptations.

Hypolimnas misippus is a pantropical species of butterfly that presents a complex case of Batesian mimicry and offers an ideal opportunity to study the involvement of structural variation in adaptation and in the maintenance of polymorphism (Marshall and Poulton, 1902; Trimen and Bowker, 1889). Females are mimetic and polymorphic, with detailed resemblances to the four morphs of the toxic *Danaus chrysippus* (Figure 4.1)(Smith, 1973). Despite the striking mimicry, a puzzling mismatch exists in the geographical distribution of *H. misippus* and *D. chrysippus* morphs across Africa, where the most abundant models are not necessarily reflected in the frequency of mimics at a given location (Gordon et al., 2010). This, together with the fact that maladaptive intermediate morphs of *H. misippus* are commonly found, suggests current selection for mimicry might be weak and raises the question of how the polymorphism is maintained (Gordon et al., 2010; Gordon and Smith, 1998). One hypothesis is that the strength of the negative frequency dependent selection fluctuates over time, following extreme changes in population density observed in *H. misippus* (Smith, 1976). Another hypothesis is that SV at the locus could be maintaining the phenotypic morphs in the population through a process known as associative overdominance. In these cases, reduced recombination stemming from a SV leads to the accumulation of deleterious mutations in one of the alleles and to decreased fitness in heterozygotes. Such a case is seen in *Heliconius numata* (Jay et al., 2021). Exploring the structure of the loci controlling wing mimicry in *H. misippus* will shed light to the forces maintaining the phenotypes in the population.

Wing colouration in *H. misippus* is determined by two loci of large effect, the M and A loci, determining forewing and hindwing pattern respectively (Gordon and Smith, 1989; Smith and Gordon, 1987; VanKuren et al., 2019). The existence of a third locus, the hindwing white suppressor S, has also been hypothesised (Gordon and Smith, 1989). The M locus is a Mendelian locus with two alleles, individuals bearing a copy of the M dominant allele (diploid genotype *M-*) present the mimetic black and white forewing tips which is known as the *misippus* morph; whereas recessive homozygotes (*mm*) have mimetic orange or maladaptive brown forewings, known as the *inaria* and *immima* morphs respectively. Previous work has identified the M locus to an intergenic region of 10 kb near genes of interest such as *pink* and *Sox 5/6 (VanKuren et al., 2019)*. However, not much is known about the structure of the locus itself.

Despite the importance of SVs in phenotypic variation, their study is limited by the difficulty of detecting them using high throughput DNA sequencing (Mahmoud et al., 2019). The identification of SVs using short-read data is harder than the identification of SNPs for two reasons. First, many SVs are longer than the average short-read length (300-500 bp) and thus the entirety of the SV is not usually contained in one read. This means that SVs are identified by a combination of mapping signatures, such as unexpectedly low or high coverage, discordant read pairs and split reads at breakpoints, for which many programs have been developed (for example (Iakovishina et al., 2016; Layer et al., 2014; Rausch et al., 2012; Sindi et al., 2012); reviewed in (Mahmoud et al., 2019)). And second, SVs are often mediated by repetitive elements and thus their breakpoints are found in repetitive regions (Carvalho and Lupski, 2016; Payer et al., 2017; Sharp et al., 2005). This results in low mapping qualities and high levels of multi-mapping, which hampers the identification of SVs through their mapping patterns at breakpoints. Long-read sequencing, in contrast, provides long range linkage information with reads spanning the repetitive and problematic regions, but is limited by cost (Ho et al., 2020; Sedlazeck et al., 2018).

Linked-read sequencing has emerged as an alternative, utilising short-read scalability while retaining linkage information (Marks et al., 2019). The newly developed 'Haplotagging' is a simple, linked-read technique that can be applied to hundreds of individuals at low-cost and has the potential to improve the study of structural variants in wild populations (Meier et al., 2021). In this approach, large DNA molecules are individually barcoded while they are being broken up for short-read sequencing. This approach retains haplotype phase information, as well as long range linkage information that can be used to detect SVs. Importantly, haplotagging can be easily scaled up to population level by multiplexing, which makes it an ideal tool for the study of adaptation and speciation in non-model organisms (Meier et al., 2021).

There are multiple software solutions for the identification of SVs from linked-read data, including LongRanger (Sudmant et al., 2015), Leviathan (Morisse et al., 2021), NAIBR (Elyanow et al., 2018) and GROC-SV (Spies et al., 2017). LongRanger and GROC-SV are very well curated tools for the analysis of linked-reads produced by 10X-Genomics, while NAIBR uses the BAM files produced by LongRanger for its SV detection pipeline. Whilst these programs could be

used for haplotagging data, the data would need to be converted to the 10X-Genomics format for input to these programs. Finally, Leviathan can take haplotagging data as input and produces a list of detected candidate SVs and their predicted breakpoints, however, unlike the above tools, Leviathan does not produce graphic visualisations of barcode sharing.

Here, I study the involvement of SVs in adaptive evolution in *H. misippus* using haplotagging data. To explore the candidate SVs in *H. misippus* haplotagging data, I develop Wrath (WRapped Analysis of Tagged Haplotypes), a tool for the visualisation and exploration of SVs tailored for haplotagging data. First, I validate Wrath by using it on published haplotagging data with known SVs. Then, I apply it to haplotagging data from *Hypolimnas misippus* and identify TE insertions at the locus associated with wing mimicry in this species. Finally, I perform a cross-species comparison within the genus *Hypolimnas* to investigate the evolutionary history of the adaptive alleles.

1.19 Results

1.19.1 Visualisation and exploration of SVs from haplotagging data using Wrath

To analyse haplotagging data, I developed Wrath (WRapped Analysis of Tagged Haplotypes), a program for the exploration and visualisation of SVs consisting of three steps.

1.19.2 Barcode parsing

Haplotagging reads are produced using magnetic beads that present a modified Tn5 enzyme on their surface carrying sequencing adapters, each with a unique barcode. During library preparation, DNA molecules wrap around the beads and are cut into smaller fragments and barcodes attached to them. Thus, reads belonging to the same DNA molecule present the same unique barcode, and the small size of the beads ensures that each barcode combination is unique to one or a small number of molecules. In the sequencing files, barcode information is included as four nucleotide sequences of 6 bp each (two per index read). To analyse haplotagging data, first, molecule information needs to be included as a BX tag in the information fields of the fastq files, a process known as molecule demultiplexing. Once the reads include information on their molecule-of-origin in their BX tag, they are ready to be mapped.

1.19.3 SV visualisation

Using mapped reads and a reference genome, Wrath plots heatmaps of barcode sharing with a single command. Wrath can also produce a list of candidate SVs. Haplotagging reads belonging to the same DNA molecule present the same unique barcode and are expected to map in close proximity in the genome in the absence of rearrangements. Thus, I can use patterns of barcode sharing between more distant genomic windows that exceed the background expectation to identify SVs.

Wrath divides a given chromosome into *n* windows of size *m* (m needs to be specified, by default 10 kb) and identifies the barcodes attached to the reads mapping in each of the genomic windows. Window size is chosen based on two factors. First, computational overhead, as Wrath builds a matrix of *nxn* dimensions which can require a large amount of

memory for large values of *n*. And second, molecule size, which depends on several factors such as sample preservation, DNA extraction and library preparation. By default, molecules are assumed to be centred around 50kb in length, although they can be much larger (Meier et al., 2021). Window size needs to be smaller than molecule size (e.g., 10kpb window size and 50 kb molecule size) as the identification of SVs is only possible if molecules span more than one window.

Once the chromosome has been split into windows, Wrath determines the barcodes that are present in each of those windows and calculates the jaccard index for each pair of windows along the chromosome and stores the value in a matrix of *nxn* dimensions. The jaccard index is an index of similarity that quantifies barcode sharing between windows:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Where J is the jaccard value between window A and B from a given chromosome.

The highest values are expected around the diagonal, which then decay exponentially with distance from it. This is because windows that are closer to each other are expected to share more barcodes than windows that are further apart, as DNA molecules span more than one window. Structural variants such as moderate to large inversions (>50 kb), intrachromosomal translocations and long duplicated regions are expected to deviate from the background distribution of barcode sharing. For example, inversions show up as bowtie patterns of excessive barcode sharing between windows that are far apart (FIGURE 1). I define excessive sharing as barcode sharing that is statistically higher than that expected by the distance between the windows given that barcode sharing decays exponentially from the diagonal. Conversely, SV of a much smaller size than the average molecule length cannot be detected with linked-reads. In those cases, short-read methods are more appropriate.

Finally, the construction of the matrix and calculation of jaccard indices for each genomic window can be a computationally expensive task, for that Wrath can be run in parallel, which minimizes computational time.

1.19.4 Exploration of candidate SVs

Wrath has the additional functionality of producing a list of candidate SVs. Wrath detects SVs where there is an excess of barcode sharing, such as in inversions, interchromosomal translocations and duplications. The distance of each entry of the matrix to the diagonal is calculated and a double exponential decay model fitted to the data such that:

$$y \sim e^{\left(a+b*e^{(x*(-c))}\right)}$$

Where *x* is the distance of each entry to the diagonal, *y* is the value of the entry (barcode sharing between a pair of windows, and *a*, *b* and *c* are parameters of the function calculated from the data.

The prediction bands fitted by the model includes the background distribution of barcode sharing and any windows whose values are outside the prediction bands (α=0.05) are then classified as putative SVs. Once the model has been fitted, Wrath outputs a list of putative SVs with their genomic coordinates and produces plots of the fitted model and identified outliers. The putative SVs are not classified into SV types and are intended to be used for prioritisation processes before further exploration.

Wrath can be applied to single populations to visualise and explore putative SVs in each chromosome of the genome. It can also be applied to detect SVs in different populations separately, which can then easily be compared and scanned for overlaps using bedtools (see case-studies below).

1.19.5 Visualisation of inversions, deletions, translocations, and duplications in a dataset of wild *Heliconius* butterflies using Wrath

To test Wrath's performance, I applied it to an existing haplotagging dataset of the two tropical butterfly species, *Heliconius erato* and *melpomene* (Meier et al., 2021). The dataset published in Meier et al., 2021 contains sequenced individuals from, *H. melpomene* and *H. erato.* Each species presents two morphs or races with mimetic wing patterns which hybridise: *H. melpomene plesseni* and *malleti* and *H. erato notabilis* and *lativitta*. I used this dataset to explore which SVs are found in these populations. First, I searched for any SVs present in the dataset genome-wide and identified 3,072 large (>50 kb) putative SVs in *H.*

melpomene and 2,885 in *H. erato*. I then explored these using the heatmaps produced by Wrath (Figure 4.1). Patterns of barcode sharing observed in the heatmaps can be used to identify the type of SV present in the samples. For example, inversions result in a bowtie pattern in the heatmap, as more barcodes are shared than expected between loci that are far apart in the reference genome (Figure 4.1). I produced heatmap plots for all large scaffolds of *H. melpomene* and *H. erato* and explored the SVs present in the dataset (Figure 4.1C, Supplementary Data). For example, a known SV in chromosome 2 in *H. erato* was clearly visible in the heatmap (Supplementary Figure 4.1). With this, I show that Wrath can visualise patterns of barcode sharing and help prioritise the order of exploration of SVs, as visual examination of the haplotagging data helps explore the SV content in the samples. However, it needs to be considered that some SVs can have similar or matching signals. For example, patterns of interchromosomal translocations like that shown in Figure 4.1A can also be produced by TE insertions.

Finally, I assessed the time required to run on the dataset. In a small subset of the data and running with 20 threads, the parallel implementation ran 10x faster than a single-threaded implementation of the same algorithm (Supplementary Figure 4.2).

1.19.6 Multiple deletions are found at the colour associated locus near *optix* in *H. melpomene* and *H. erato*

I was interested to know whether structural variation was associated with wing pattern in *Heliconius*. Using Wrath, I found multiple deletions of 1-10 kb at the locus associated with red pattern elements near the gene *optix* are associated with wing pattern (Figure 4.1D). Deletions leave an area depleted of barcode sharing, as reads do not map to the area, which can be visually identified using the heatmaps (Figure 4.1A-C). However, only polymorphic deletions can be identified, as deletions that are fixed in the population compared to the reference cannot be distinguished from assembly artefacts or poor mapping (e.g. repetitive regions for which mapping reads are filtered of due to low mapping quality). Larger deletions (>10 kb) leave an additional signature in the heatmap, showing increased barcode sharing between the breakpoints in the shape of a triangle (Figure 4.1A-C). This is because more molecules than expected span through the breakpoints in individuals presenting the deletion.

This signature is harder to detect in smaller deletions, as barcode sharing becomes higher between loci that are close together.

I identified 7 polymorphic deletions of 1-10 kb in *H. erato* and 1 of 1 kb in *H. melpomene* at the *optix* locus by visual inspection of the heatmaps (Figure 4.1D). This locus contains cis-regulatory elements that control the expression of optix and influence development of red colour elements, which have been functionally tested with CRISPR (Lewis et al., 2019). Thus, one possibility is that these deletions are disrupting the function of an *optix* CRE and affecting *optix* expression, although they will need to be functionally tested. The deletion detected in *H. melpomene* was detected previously using short-read sequencing of two different sub-species (Wallbank et al., 2016).

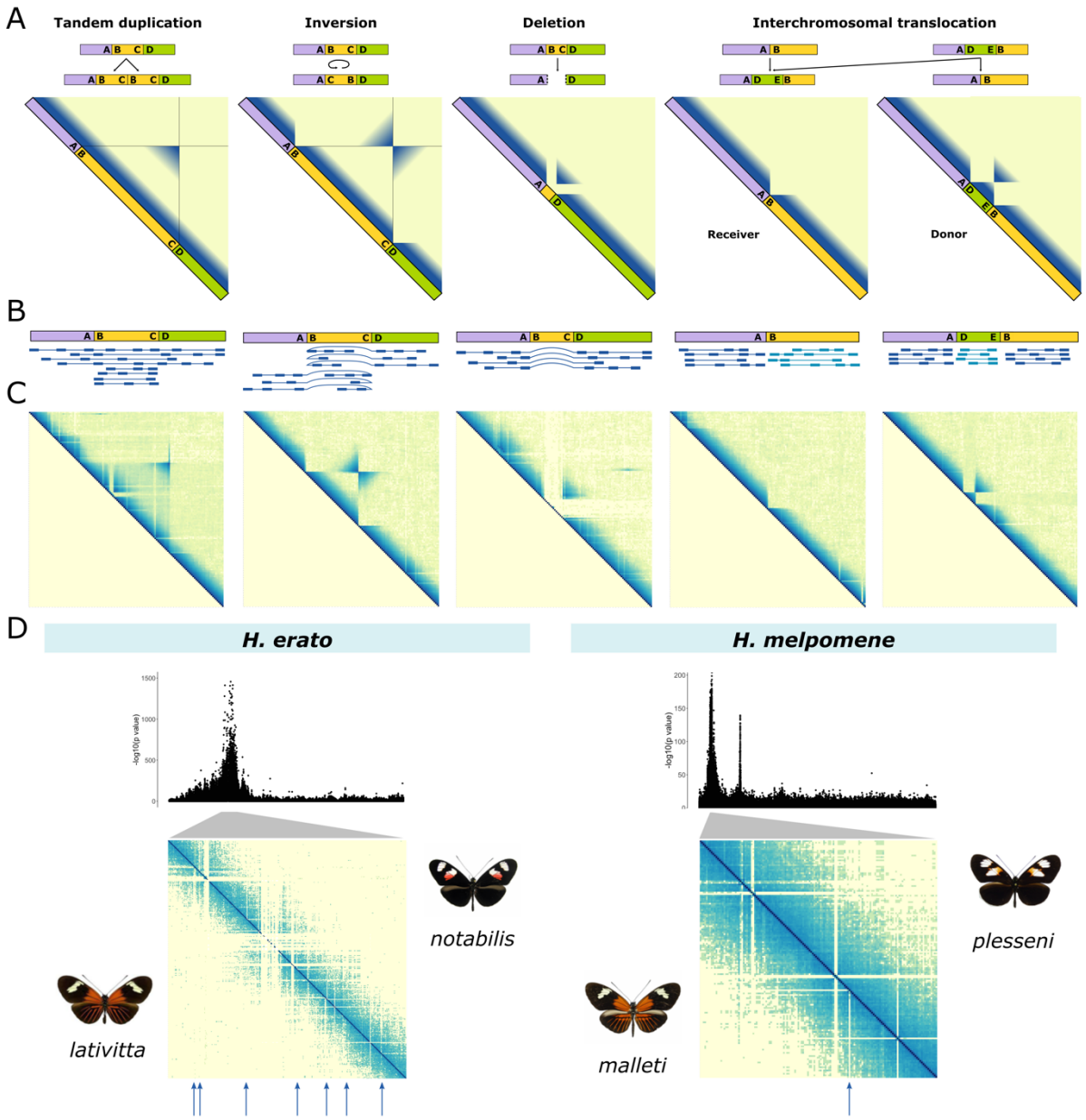


Figure 4.1 | A. Hypothetical Wrath outputs for some SV types. On top of each heatmap are depicted the reference genome and its order of loci and below a depiction of the rearranged genome containing an SV. Points A, B, C and D depict different loci around the breakpoints of the SV that I use as a guide through the diagrams. **B.** Linked-read mapping pattern on the reference genome for each of the hypothetical SVs. **C.** Wrath output from the *Heliconius* dataset depicting possible SVs that match the hypothetical predictions. **D.** Wrath output heatmaps of the region around one of the loci (the *optix* locus) associated with colour pattern in *H. erato* (left) and *H. melpomene* (right). Each triangle half of the matrix depicts barcode sharing for one of the races—in each case depicted by the side. On top of the heapmap is plotted the Manhattan plot of the GWAS association of colour pattern between each pair of races. These have been zoomed in to show only the region around the associated *optix* locus. Grey triangles depict the correspondence of regions between the GWAS and heatmap.

1.19.7 Forewing mimicry in *Hypolimnas misippus* is controlled by the M locus

To explore the genetic underpinnings of this complex case of Batesian mimicry, I first confirmed the previously described identity of the M locus. I sequenced 335 *Hypolimnas misippus* females collected in Kenya and other locations in Africa using haplotagging. The dataset contains 277 *misippus* individuals and 54 *inaria/immima,* sequenced to 0.81 coverage on average (Supplementary Table 1 and Supplementary Figure 4.3). By using haplotagging with a large dataset, I could sequence to low coverage per individual without compromising statistical power to detect loci associated with mimicry. This is because, although read coverage is low, molecular coverage (i.e. coverage of DNA molecules) is higher in linked read data, as SNP information of reads belonging to the same DNA molecule can be used for imputation and phasing (Marks et al., 2019; Meier et al., 2021). Also, using a large population sample (>200 individuals) ensures the detection of regions associated with the trait of interest (Lou et al., 2021).

First, I parsed and demultiplexed the data using Wrath and then imputed SNPs and phased haplotypes, which resulted in the identification of 46.1 M SNPs. Mean haplotype length after phasing was 109.02 kb (Supplementary Figure 4.5). I used the phased haplotypes to perform a GWAS to identify the M locus. The highest peak of association with variation in forewing phenotype was found on chromosome 26 (6,731,000 - 6,743,400 bp; Figure 4.2A). I confirmed identity with the region previously identified as the M locus (VanKuren et al., 2019) by aligning the identified locus using to the *HypMisi_v2* assembly using BLASTn. Furthermore, genes surrounding the peak of association are the same as in VanKuren et al., 2019 VanKuren et al., 2019). The gene g1737, identified as Sox 5/6 in (VanKuren et al., 2019) has the closest similarity (BLASTp) Sox102F in *D. melanogaster* but with *Sox 5/6* with Principal Component Analysis of the whole of chromosome 26 showed no evidence for population structure in the data (Supplementary Figure 4.6). In contrast, when using just the associated region for PCA, samples of the same phenotype were found closer together (Supplementary Figure 4.7). Closer examination of the associated region in the GWAS result revealed 3 clearly separated peaks of association (Figure 4.2D), a pattern that could be caused by structural variation at the locus; to examine this more closely, I visualised the barcode sharing at the region using Wrath.

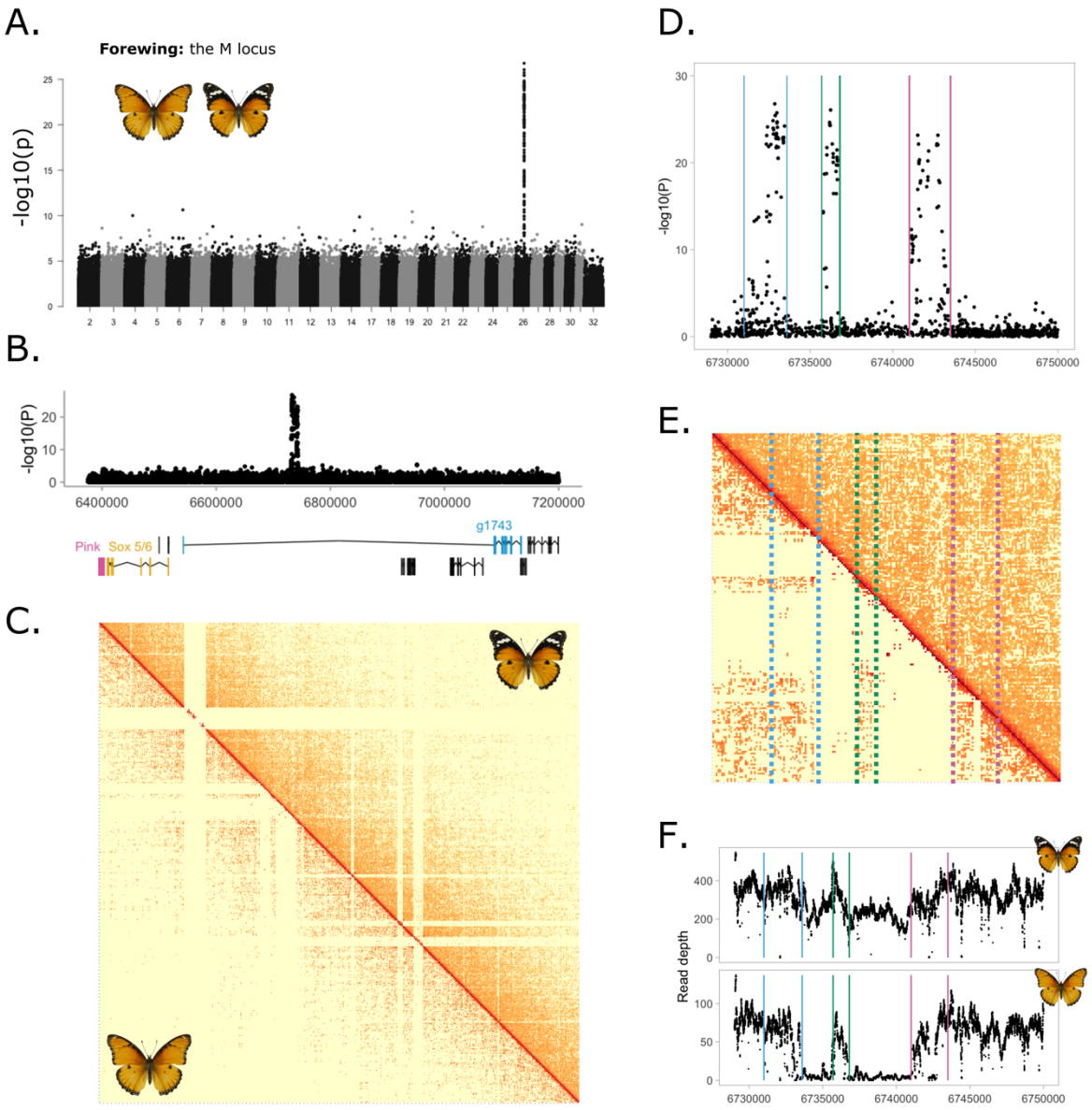


Figure 4.2 | Two deletions are found at the locus associated with forewing mimicry in *mm* individuals. A. GWAS of forewing phenotype shows a unique peak at chromosome 26. **B.** The GWAS peak showing local annotation track with the three main candidate genes coloured. *Pink* shown in pink, *Sox 5/6* in yellow and *g1743* in blue. **C.** Heatmap of barcode sharing of the region around the association peak. Upper triangle shows barcode sharing for *M-* individuals and lower triangle *mm* individuals. A different pattern of barcode sharing between *mm* and *M-* individuals is seen at the associated region. **D.** At a finer scale the association peak reveals a three-peak structure. **E.** A zoom in of the barcode sharing heatmap reveals signal depletion between the peaks of association in *mm* individuals, a signature of deletions. **F.** Read coverage around the associated region supports the hypothesis of deletions, as *mm* individuals present almost 0 coverage between the association peaks, while *M-* individuals have more constant coverage throughout the region.

1.19.8 *misippus* individuals show multiple TE insertions at the M locus

The region associated with differences in forewing pattern is 10 kb in length, thus a very small window size is necessary to elucidate whether there is any structural variation at the locus. I therefore visualised barcode sharing using a window size of 100 bp around the M locus and identified two putative deletions at the M locus relative to the reference genome in recessive homozygotes (Figure 4.2C-D). Calling them deletions is only relative to the comparison reference genome and makes no reference to their ancestral or derived state. These two indels perfectly match the locations of the troughs of association seen in the GWAS analysis where read coverage is almost zero in *mm* individuals (Figure 4.2E). This explains the decline in association in the two regions, as SNPs cannot be confidently called in *mm* individuals.

To verify the presence of these indels, I designed PCR primers flanking each indel, and at the breakpoints, and amplified them in the two phenotypes (3 individuals per phenotype; Supplementary Figure 4.8). This confirmed that two insertions of 2.4kb and 4.3kb are present in the dominant *misippus* phenotype relative to the recessive *inaria/immima.* A set of transposable elements (TE) insertions detected with RepeatMasker compose the entirety of the two insertions, which are situated in a 3'-UTR intron of the gene g16415, an ankyrin repeat and sterile alpha motif domain containing gene of unknown function (Figure 4.2B and 3A and Supplementary Table 2 and 3). Insertion A (most downstream) is composed of a tandem duplication of *Helitron* family transposable element and an unknown TE, while insertion B is composed of 3 *Helictrons,* 4 LINEs and 2 unknown TEs (Figure 4.3A and Supplementary Table 4). Given that the insertions are found in the dominant allele, the most plausible explanation is that the insertion is modifying the expression of a nearby gene, either *g16415* or others such as *Sox 5/6* and *pink* by acting as or affecting existing cis-regulatory elements.

Finally, I applied Wrath genome-wide to explore candidate SVs across the genome of *H. misippus*. I detected 6,335 large SVs (50 kb) in homozygous recessive individuals and 6,504 in heterozygous and homozygous dominant individuals. Of these, 3,184 and 3,097 are shared between the two groups with a minimum overlap of 80%. The differences in number of overlaps between the two groups happen because a large SV from one group might overlap with more than one SV in the other.

1.19.9 Comparisons with other *Hypolimnas* species suggests that the *m* allele is ancestral

I next explored the evolutionary history of the M locus. The presence of the TE insertions could either be ancestral or derived and to test these hypotheses I explored this region in the H. bolina genome. First, I aligned the *H. misippus* and *H. bolina* reference genomes using Satsuma2, an aligner intended for inferring homology from sequence similarity (Grabherr et al., 2010). All alignments of the region around the M locus mapped on chromosome 26 in the *H. bolina* genome, suggesting that there have not been any interchromosomal rearrangement in the region in *H. misippus* since the most recent common ancestor (Figure 4.3B). Also, the alignment shows that, while the peaks of association had matching sequences in chromosome 26 in *H. bolina,* the two indels between the peaks had no matches in the *H. bolina* genome (Figure 4.3B). This evidence also suggest that the *m* allele could be the ancestral allele shared with *H. bolina.* However, the high TE content of the insertions in the *M* allele could potentially be interfering with genome alignment and short read mapping.

To further explore the origin of the alleles, I analysed whole genome resequencing data from 4 other *Hypolimnas* species produced in Chapter 5, including 214 *H. bolina,* 4 *H. anthedon,* 4 *H. deceptor* and 2 *H. usambara* (Figure 4.3D), sequenced to an average coverage of ~6.5X. First, I mapped all resequenced *Hypolimnas* to the *H. misippus* reference genome and quantified read coverage at the M locus (Figure 4.3C). Given the differences in numbers of individuals per species, I grouped all *H. bolina* individuals together and placed the other *Hypolimnas* in another group. Read coverage followed the same pattern in both groups, with a pronounced decrease in coverage matching the regions of TE insertions. This suggests that the TE insertions are not present in these species.

To investigate in more detail the evolution of the *M* and *m* alleles, I evaluated the variation at the locus using Principal Component Analysis (PCA) using the SNPs of the associated region for the *H. misippus* and *H. bolina* samples (Figure 4.3E) and generated PCAs in genomic windows. Read coverage is very low per individual (~1X), which makes the use of phylogenetic

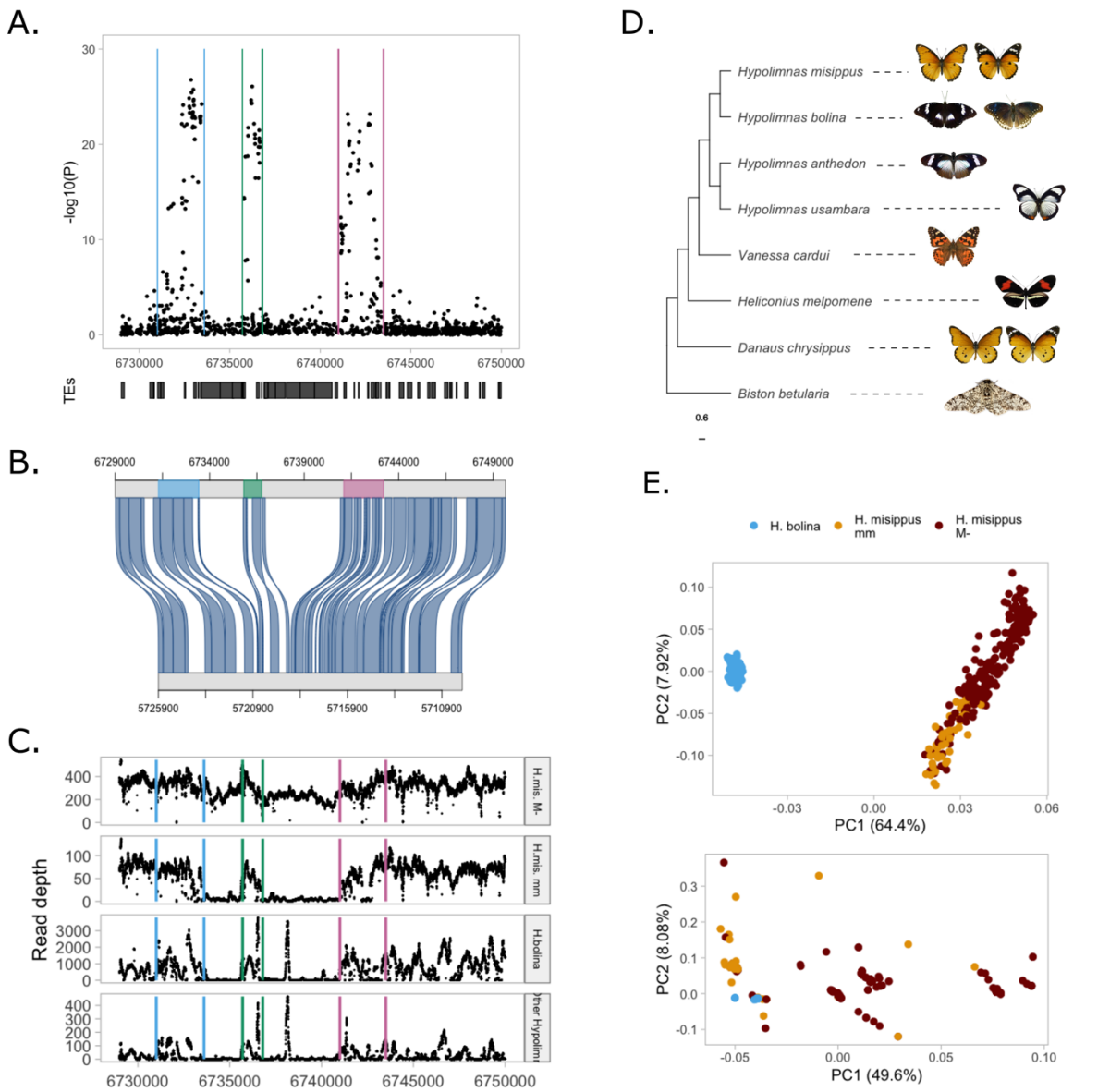


Figure 4.3 | TE insertions are found in the recessive allele *m*. A. Zoom in of the association peak showing a TE annotation track. The regions between the association peaks are composed by TE insertions. **B.** The alignment of the *H. misipppus* and *H. bolina* refrence genomes shows that *H. bolina* does not present the TE insertions. **C.** Read coverage of *H. bolina* and other *Hypolimnas* species suggests that other *Hypolimnas* do not carry the insertions and that those are thus derived. **D.** A phylogeny of the *Hypolimnas* species used and other Lepidoptera. *H. deceptor* is missing and would be in the same clade as *H. anthedon* and *H. usambara.* Phylogeny extracted from (Kumar et al., 2022). **E.** Principal Component Analysis (PCA) of the locus associated with forewing phenotype reveals structure by species, while PCA of the top associated SNP region of the first peak (Chromosome 26: 6732649-6732923) reveals that the *m* allele is more similar to the *H. bolina* samples, supporting the hypothesis of the recessive allele being ancestral.

trees to explore the evolutionary history of the alleles unreliable. If enough SNPs are in linkage disequilibrium with the causal mutation, I would expect that the relationship between the alleles in the PCA would reflect colour phenotype to a greater degree than the rest of the genome. Particularly, I would expect two clusters, one with only haplotypes of the dominant allele of the *M*- individuals and another with all phased haplotypes from *mm* individuals and some phased haplotypes from M- individuals, as many of them are expected to be heterozygotes. And second, I would expect the ancestral allele to be more similar to the *H. bolina* alleles.

Across the whole associated region that contains the M locus, the PCA reflects the species relationships, with one cluster for each species (Figure 4.3E top). After that, I generated PCAs of the associated region in genomic windows of 50 SNPs and examined them. I identified a region containing the top associated SNPs of the most upstream peak of association (6732649-6732923) for which the relationship between the samples did not reflect species' relationship (Figure 4.3E bottom). One of the clusters is formed almost entirely by phased haplotypes from *M*- individuals, except for 4 haplotypes belonging to *mm* individuals. The other cluster contains all the remaining *m* phased haplotypes, the *H. bolina* haplotypes and haplotypes from *M*- individuals. The closer relationship of *m* phased haplotypes and the *H. bolina* haplotypes suggests that the *m* allele is ancestral. Taken all together, the read-coverage, reference genome alignment and local PCA at the associated region suggest that the recessive m allele that produces orange wings in homozygosis is ancestral to the dominant *M* allele that produces black-and-white forewing tips.

1.20 Discussion

Here, I identify two large insertions associated with wing pattern mimicry in the diadem butterfly, *Hypolimnas misippus,* a complex case of Batesian mimicry. *H. misippus* show detailed mimicry of the four morphs of the toxic *Danaus chrysippus* (Smith, 1976). Although the accuracy of the mimicry would suggest strong selection, intermediates are often found (Gordon et al., 2010) and a mismatch exists between the model present in a region and the most common mimic, which suggests that other evolutionary forces might be at play (Gordon et al., 2010). Despite the complexity of the mimicry, a single Mendelian locus (M locus) controls forewing mimicry in this species (Smith and Gordon, 1987; VanKuren et al., 2019). Negative-frequency dependant selection has been invoked as the force maintaining the phenotypes in the population (Gordon, 1987; Gordon et al., 2010). Another alternative is that structural variation maintains associative overdominance and prevents their loss even in scenarios of weak mimicry selection. An example of this has been shown in *Heliconius numata* in which a supergene containing three inversions controls wing phenotype (Jay et al., 2021). The inversions result in a region of low recombination, which in turn lead to the accumulation of deleterious mutations. This has resulted in the decrease in fitness of homozygous individuals and the maintenance of the polymorphism a process known as associative overdominance. Here, I present evidence of structural variation at the locus controlling forewing phenotype in the mimetic *H. misippus.* I detected two insertions of 2.4kb and 4.3kb in the dominant allele of the M-locus formed of several TE insertions. However, compared to the reduced recombination produced by a large inversion that can lead to associative overdominance, it is unclear if such small-scale SV can have a similar influence on the maintenance of the polymorphism.

The TE insertions identified are found in the intron of the gene g16415, an ankyrin repeat and sterile alpha motif domain containing of unknown function. This represents a similar case to the peppered moth, where a 10-kb TE insertion increases the expression of *cortex*, resulting in the production of melanic morphs sterile alpha motif domain containing gene of unknown function (van't Hof et al., 2016). Given that the insertion is found in an intergenic region, the most plausible explanation is that the insertion is modifying the expression of a nearby gene, either g16415 or others such as *Sox 5/6* and *pink* by acting as or affecting existing cis-

regulatory elements. Similarly, variation in wing colour in the *Heliconius melpomene/timareta* lineage is associated with a TE insertion in the cis-regulatory region of *cortex,* suggesting that cis-regulatory structural variation controls these mimetic phenotypes (Livraghi et al., 2021). One hypothesis is that the g16415 gene is involved in pattern determination in *H. misippus* and the TE insertion influences its expression leading to a change in phenotype. Given that the insertion is found in the dominant allele, an increase in expression of the candidate gene or a certain isoform are possible explanations, which could be achieved by the disruption of a repressor or the generation of a novel enhancing function. Alternatively, the M-locus could be modifying the expression of some other gene nearby, such as *pink* and *Sox 5/6*. Overall, this case adds more evidence that cis-regulatory mutations are associated with pattern variation, while coding mutations are more likely to be associated with colour, and sheds light on the adaptive importance of TEs (Casacuberta and González, 2013; Orteu and Jiggins, 2020; van't Hof et al., 2016). Finally, by comparison to other *Hypolimnas* species, I show that the insertions are the derived allele. This suggests that the orange phenotype evolved first and that insertions at the M locus subsequently reverted the phenotype to a more *Nymphalidae*-like phenotype. Melanised apexes in the forewing with subapical white bands are a common wing phenotype in Nymphalids such as Danaids or some *Nymphalinae,* including *Antanartia* and *Vanessa species* (e.g. *Vanessa cardui,* Figure 4.4.3D) and have been suggested to be ancestral to the clade.

Additionally, here I present Wrath, a user-friendly, flexible, and fast tool for the visualisation of haplotagging data and exploration candidate SVs. Wrath produces heatmap plots of barcode sharing that can be used to visually inspect the data and identify candidate SVs. Wrath can be run with any chosen window size, which gives flexibility to the user and allows for the detection of SVs of different sizes. By default, Wrath targets the detection of large SVs, defined as >50 kb, as such large SVs are probably of biological importance and can be easily detected using large window sizes, which minimises the computational resources necessary. Also, to minimize computational time, Wrath can be run in parallel. When comparing population or groups, Wrath can be run on each separately to automatically detect SVs, which can then be compared. Finally, Wrath can be run at chromosomal level or on only a specific region of the genome, which allows for flexibility depending on the research question. Thus, a candidate region can be explored at a finer scale using Wrath and detect small SVs.

I tested the effectiveness of Wrath using our dataset of *Hypolimnas misippus* and a published dataset from *Heliconius* butterflies. Using the *Heliconius* dataset, I demonstrate the success of Wrath at visualising all kinds of SVs such as inversions, inter- and intrachromosomal translocations, indels and duplications at the genome wide level. Furthermore, I show that using Wrath small structural variants can also be explored, which I demonstrate by detecting 1-kb deletions at the *optix* locus in *H. melpomene* and *H. erato* through visual inspection of barcode sharing matrices. However, Wrath is limited in automatically detecting SVs, for which other programs exist. Furthermore, although linked-reads can provide a solution on the detection of structural variants, they are limited by SV size, as they do not provide additional information than short reads for their detection, and the ambiguity of some signals (e.g. translocations and transposable element insertions might have the same signal).

Altogether, my study highlights the importance of structural variation in the evolution of adaptive phenotypes. Our results in *Hypolimnas misippus* add to the mounting evidence that TEs have an important role in the evolution of colour and mimetic phenotypes in Lepidoptera. Furthermore, our results in the *Heliconius* dataset contribute evidence of the possible involvement of deletions in wing pattern differences between races. Finally, with both analyses I have shown that Wrath is an easy and flexible way of visualising haplotagging data and exploring candidate SVs.

1.21 Methods

1.21.1 Analyses of *Heliconius* data

Data from (Meier et al., 2021), was used. First, reads were pre-processed, adapters and low-quality ends were trimmed using TRIMMOMATIC (Bolger et al., 2014). Then, I mapped the reads to the respective reference genome, H. erato v1.0 (Nadeau et al., 2014) and H. mepomene v2.5 (Davey et al., 2017, 2016), using BWA-mem (Li, 2013) and marked PCR duplicates using the MarkDuplicates utility from Picard tools (broadinstitute.github.io/picard). Alignment (BAM) files were used as input for Wrath, which I ran separately for each scaffold in the reference genome with a window size of 10 kb. Reads with a mapping quality below 10 were filtered out by Wrath. To examine the colour loci specifically, I ran Wrath with a window size of 1 kb and specifying the desired coordinates.

1.21.2 Hypolimnas misippus sampling

Samples were collected in multiple locations in Africa. Most sample bodies were preserved in 100% ethanol immediately after collection, with some exceptions that were air dried. Samples in ethanol were kept at room temperature for 2 moths and then stored at -80C. A small (1/8) piece of the thorax was used for sequencing.

1.21.3 *H. misippus* sequencing library construction

Haplotag libraries were prepared essentially as described in Meier et al., 2021, with the following modifications. Broadly, haplotag libraries were prepared in batches of 96 samples. Briefly, genomic DNA sample was diluted to 0.15 ng/µl with 10 mM Tris, pH8 and quantified with Quant-iT PicoGreen dsDNA Assay Kit (Thermo Fisher Scientific). We used only 1.2 µl of haplotagging beads (~0.88 million beads, each carrying one of 885K well-specific barcodes) per sample; 30 µl of WASH buffer (20 mM Tris pH8, 50 mM NaCl, 0.1% Triton X-100); 10 µl of 5x tagmentation buffer (50 mM TAPS pH 8.5 with NaOH, 25 mM MgCl2, 50% N,N-dimethylformamide); and 25 µl of 0.6% SDS for Tn5-stripping following tagmentation. For sub-sampling, 1/10th of the beads+DNA (0.15 ng DNA per sample) from each of the 96 samples was pooled into a single 8-tube-PCR-strip, and then again from every 8 pools into 4 final samples pools. With only 4 pooled samples on the magnetic stand, the buffer was

removed, and 20 µl of 1x Lambda Exonuclease buffer, supplemented with 10 units of Exonuclease I and 5 units of Lambda Exonuclease (New England BioLabs), was added to each sample. Samples were incubated at 37 °C for 30 minutes, and then washed twice for 5 minutes with 150 µl of WASH buffer. DNA library was then amplified using Q5 High-Fidelity DNA Polymerase (New England BioLabs) in four 25 µl PCR reaction according to manufacturer's instructions, using 4 µl of 10 µM TruSeq-F AATGATACGGCGACCACCGAGATCTACAC and TruSeq-R CAAGCAGAAGACGGCATACGAGAT primers, with the following cycling conditions: 10 min at 72°C followed by 30 sec 98°C and 10 cycles of: 98°C for 15 sec, 65°C for 30 sec and 72°C for 60 sec. Libraries were pooled after PCR into a single library pool, size selected using Ampure magnetic beads (Beckman Coulter), Qubit quantified, and adjusted with 10 mM Tris, pH8, 0.1 mM EDTA to 2.5 nM concentration for sequencing.

Sequencing and demultiplexing. Pooled libraries were sequenced by a HiSeq 3000 (Illumina) instrument at the Genome Core Facility at the MPI Tübingen Campus with a 150+13+12+150 cycle run setting, such that the run produced 13 and 12nt in the i7 and i5 index reads, respectively. Sequence data were first converted into fastq format using bcl2fastq v2.17.1.14 with the following parameters --use-bases-mask=Y150,I13,I12,Y150 --minimum-trimmed-read-length=1 --mask-short-adapter-reads=1 --create-fastq-for-index-reads (Illumina).

1.21.4 SNP calling and imputation

First, molecules were de-multiplexed. When using haplotagging, the molecule of origin information is embedded in the read name in the fastq file as a string of four barcodes of six nucleotides each. This was done as described in described in Meier et al., 2021 to generate the modified fastq files. The molecule ID is then included in the BX tag of each read. Barcode mismatches caused by sequencing errors are allowed as long as there is an unambiguous closest match. Once the BX tag was created, I pre-processed the reads, cutting adapters and low-quality ends using TRIMMOMATIC, and marking PCR duplicates using the MarkDuplicates utility from Picard tools (broadinstitute.github.io/picard) with two specific options CREATE_INDEX=TRUE and READ_ONE_BARCODE_TAG=BX. I then de-multiplexed the individuals using their barcodes and included their individual ID information in the read group field.

SNPs were identified using the mpileup utility of bcftools v1.11 (Danecek et al., 2021), running each chromosome separately including the INFO/AD,AD,DP,DV,DPR,INFO/DPR,DP4,SP tags in the output (-a option), setting the minimum mapping quality to 10 (-q) and the minimum base quality to 20 (-Q), ignoring Read Group tags (--ignore-RG) and removing duplicates (-F 1024), and the optput directly piped to bcftools call using the alternative model for multiallelic and rare-variant calling (--multiallelic-caller), including only variants in the output (--variants-only) and the fields GQ and GP (-f GQ,GP). Then, using bcftools query (-f), I selected the generated a file containing the chromosome, position, reference and alternative alleles for each SNP and with that produced a file of SNP positions that I could use as one of the inputs for the SNP imputation program STITCH (Davies et al., 2016). Following that I generated genomic windows of 500 kb using bedtools over which I could iterate to run the remainder of the pipeline.

I ran STITCH separately for each of the genomic intervals using all our bam files as input. STITCH imputes SNPs from read and linked-read information but requires fine tuning of the input parameters. To optimise the values, I tested multiple values and compared the results, evaluating their performance using the M locus (Supplementary Methods). SNPs at the M locus are expected to be 0/0 or 0/1 for *misippus* individuals and 1/1 for *inaria* individuals. Options that optimised the results were K=30, method=diploid, nGen=500, readAware=TRUE, keepInterimFiles=FALSE, shuffle_bin_radius=500, expRate=5, iSizeUpperLimit=500000, keepSampleReadsInRAM=TRUE, use_bx_tag=FALSE. I concatenated the resulting imputed variant calls (vcf files) using bcftools concat.

1.21.5 SNP phasing in *H. misippus*

To phase the SNPs into haplotypes, I used HapCut2(Edge et al., 2017), which I ran separately for each of the 500 kb intervals used for imputation and each individual separately. First, I filtered SNPs based on their informativeness (INFO_SCORE >= 0.2) and selected all heterozygous SNPs. I used this as input for the --extractHAIRS utility of HapCut2 together with the BAM files with marked duplicates and the option --10X turned on, which indicates that the input contains linked reads. This produced a filed with unliked fragments, which I then used as input for the LinkFragments.py script of HapCut2, which integrates the information of the linked reads. I specified a maximum distance of 50kb. Then, I used the linked fragment

file and vcf as input for the HAPCUT2 utility with the option --nf 1 --threshold 30 --error_analysis_mode 1 --call_homozygous 1. Finally, I integrated the resulting vcf to out vcf of homozygous sites.

1.21.6 Phenotyping and GWAS

I photographed forewings and hindwings of each individual in a standardised set-up, using a green background and a colour checker. Phenotypes were scored by hand following the phenotype categorisations of (Gordon et al., 2010), coding *misippus* morphs as 1 and *inaria* as 0. All sample phenotypes are found in Supplementary Table 5.4 (Chapter 5). Using the merged HAPCUT2 output vcf file as input and the phenotype scores, I performed a GWAS with Plink v1.9(Purcell et al., 2007) using the option –assoc and –adjust to correct for multiple testing.

1.21.7 Detection of SVs in *H. misippus*

Genome wide SVs were identified using Wrath using the same method as for the *Heliconius* data. I used the intersect utility from bedtools v2.30.0 to assess overlap between SVs identified in homozygous recessive, and heterozygous and homozygous dominant individuals, setting the minimum fraction of overlap to 0.8 for both sets and extracting only one match per SV (intersectBed with options -f 0.8 -F 0.8 -u).

1.21.8 DNA extractions and amplification of the M locus in *H. misippus*

DNA extractions were carried out using a custom protocol using PureLink buffers and homemade magnetic beads. Briefly, a small piece of thorax tissue (1/10) is placed in a 8-tube PCR trip. Then, 45 uL of PureLink Digestion buffer and 10 uL of Proteinase K (20mg/mL) are added, and the mix is incubated at 58˚C with shaking (500 rpm). Thereafter, I added 2uL of RNAseA (DNAse free, 10mg/mL) and incubated it 10min at room temperature. Then, I added 45uL of PureLink Lysis buffer and incubated at 58˚C for 30 minutes with shaking (500 rpm). I then used a homemade magnetic bead mix to extract the DNA from the lysate. First, I added 37.5 uL of magnetic beads and 75 uL of lysate to a 96-well plate. After mixing, I incubated 15 minutes at room temperature, placed the plate in a magnetic stand for 10 minutes, removed the supernatant and cleaned the beads with 80% ethanol. After drying out, I added 50uL of

10mM Tris (pH=8) to elute and incubated at 45˚C for 15min without resuspending. Then, I resuspended the beads and incubated for 20 minutes at room temperature. Finally, I placed the plate on the magnetic stand and, after 10 minutes, transferred the supernatant (the DNA) to a fresh tube.

To amplify the regions of interest, I designed primers at each side of the deletions and at the breakpoints (SUP FIGURE XX). I used a Q5 High-Fidelity 2X Master Mix from New England BioLabs and with 35 cycles. I used 8 individuals, 3 *inaria/immima* (CAM035230, CAM035232, CAM035239, CAM035240, CAM035244, CAM035245, CAM035249, CAM035250).

1.21.9 Reference genome alignments

To identify putative homologous regions of the reference genomes of *H. bolina* and *H misippus*, I aligned the two references to each using Satsuma2 (Grabherr et al., 2010) with default parameters. I visualised the resulting alignments using the asynt R functions (Kim et al., 2022).

1.21.10 Sample preparation and genome wide analysis of *H. bolina* samples and other *Hypolimnas* species

214 wild and reared *H. bolina* samples (dataset2) from Chapter 5 were used. Briefly, samples DNA was extracted and DNA Nextera libraries prepared using custom protocols described in Chapter 5. Short-read data from whole genomes were sequenced to ~6.5X in coverage. Reads were trimmed using fastp (Chen et al., 2018) and mapped to the reference genome *HypMisi_v2* and *HypMisi_v1* using BWA-MEM2 (Vasimuddin et al., 2019). PCR duplicated were marked using MarkDuplicatesSpark from GATK (Auwera and O'Connor, 2020) and SNPs called bcftools v1.11 (Danecek et al., 2021) mpileup like described above. Imputation was carried out with STITCH (Davies et al., 2016) using the same settings as for *H. misippus.*

1.21.11 Read depth analysis and identification of indels

I calculated read depth from the bam files using the depth utility from samtools (Danecek et al., 2021) with the --a option to output depth for all sites, including those with no reads mapping to them. I visualised the optput in R v 4.1.2 using the ggplot2 package (Wickham, 2016).

Individual BAM files with marked duplicates were subset for the region of interest using tabix and merged using samtools merge. This merged BAM files were then visualised using IGV. The candidate indels were identified through visual inspection. SNPs associated with the indel were as proxies to identify the individuals carrying the deletion and insertion. WHICH SNPS. 100% of the reads carrying

1.21.12 Principal Component Analysis

VCFs of phased haplotypes were used for PCA, which was performed using Plink v1.9.

1.22 Supplementary Information

1.22.1 Supplementary Methods

1.22.2 Testing the performance of SNP imputation with STITCH

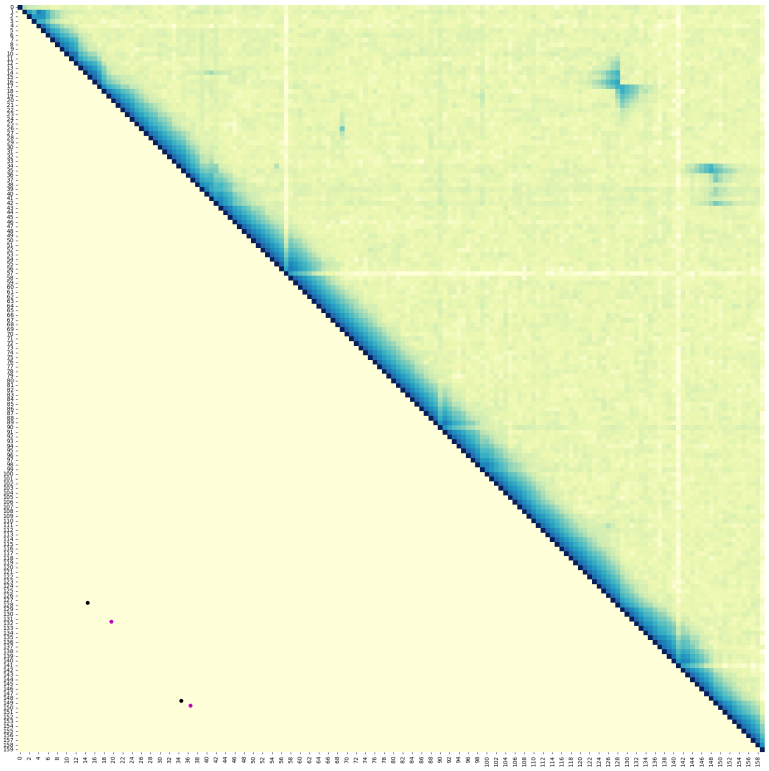
Comparisons of the performance of SNP imputation of STITCH (Davies et al., 2016) with different parameter values were produced following methods from Meier, et al., 2021 (Supplementary Figure 4.2). To evaluate the performance, SNPs were called at the locus associated with forewing pattern and evaluated using their genotype inferred from the phenotype. Comparisons using and not using STITCH were performed. If using STITCH, comparisons with or without BX tag (which integrates linked read data) were performed.

If using STITCH without BX tag:

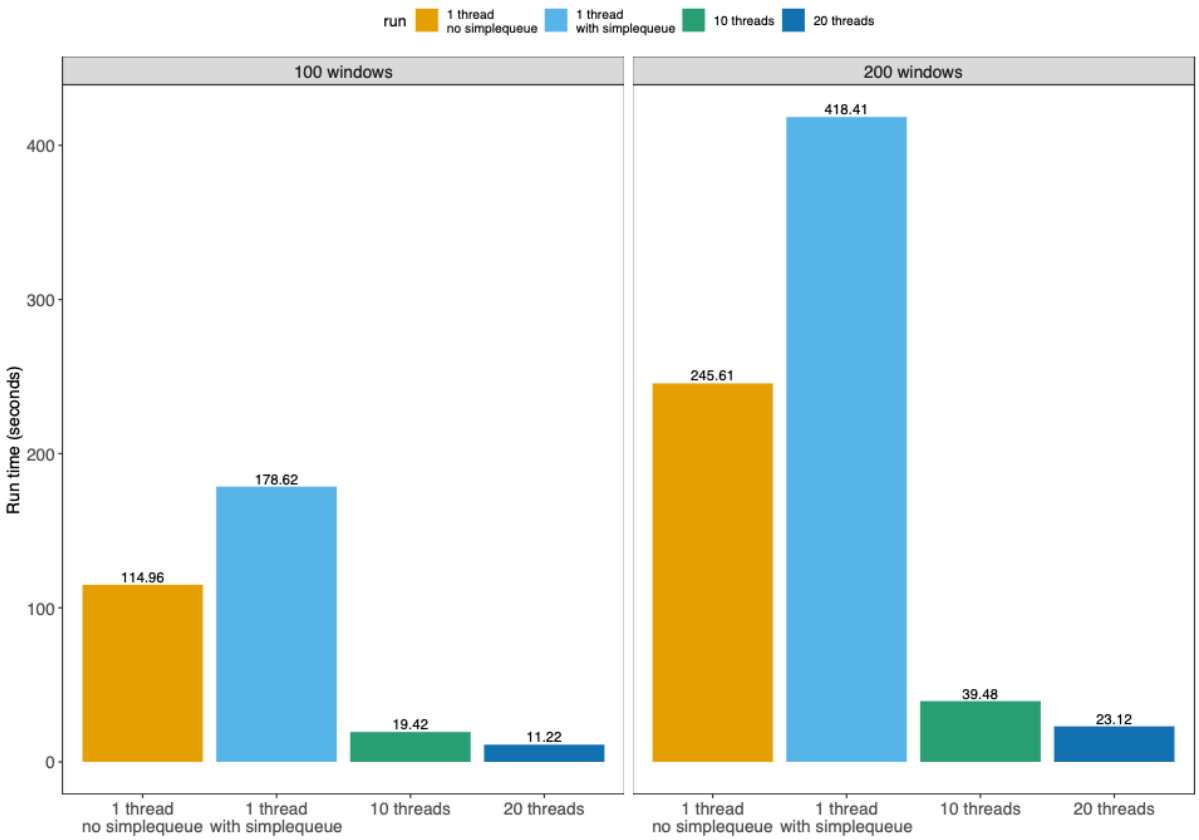
- different values of k : number of founder / mosaic haplotypes to use
- different values of nGen: Estimated number of generations since founding. In uncertain, estimate using 4 * Ne / K, where Ne is the effective population size

Performance was measured by: Sensitivity = TP/(TP+FP). Where TP is True Positive, and FP is False Positive. Using the phenotype of the samples and the M locus in *H. misippus* we can calculate Sensitivity. We know that *inaria* and *immima* individuals are double recessive, while *misippus* individuals can be homozygote dominants or heterozygotes. Samples were coded as: *misippus* phenotype with 0/0 or 0/1 genotypes are TP, *inaria*/*immima* with 1/1 are also TP, *misippus* phenotype with 1/1 genotypes are FP and *inaria*/*immima* with 0/1 or 0/0 are also FP. This will be true for SNPs that are in linkage disequilibrium with the causal mutation. From previous studies we know that over a hundred SNPs could be perfectly associated with phenotype at the associated locus (VanKuren et al., 2019).

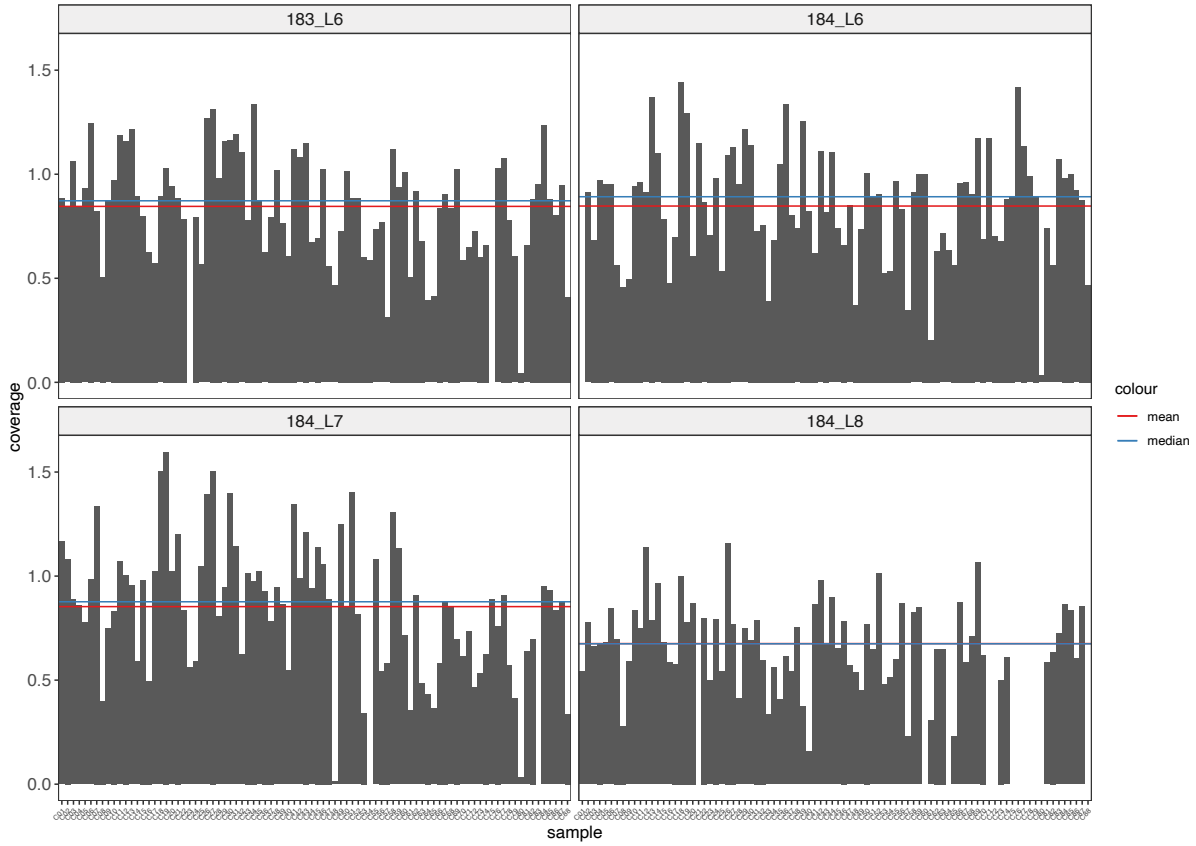
Supplementary Figures



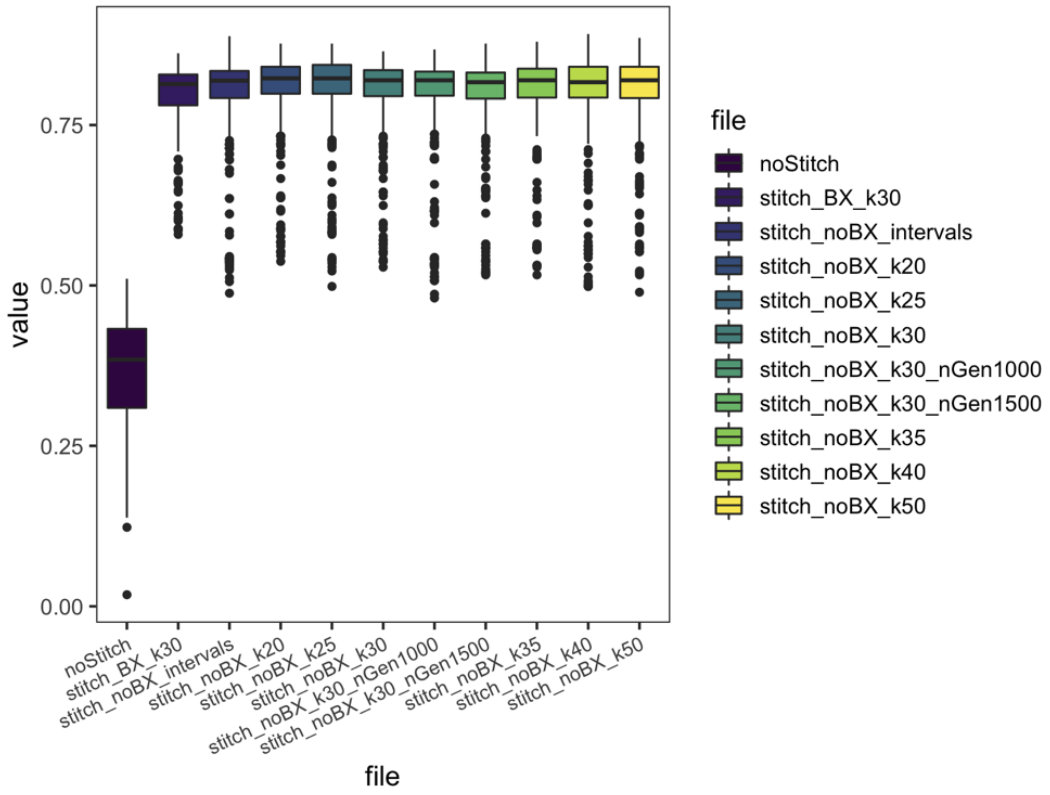
Supplementary Figure 4.1. Wrath output of scaffold Herato0204 in 10 kb windows identifies a known inversion. A bowtie pattern of barcode sharing can be observed off the diagonal. All *Heliconius erato* samples from Meier, et al., 2021, were used for the analysis.



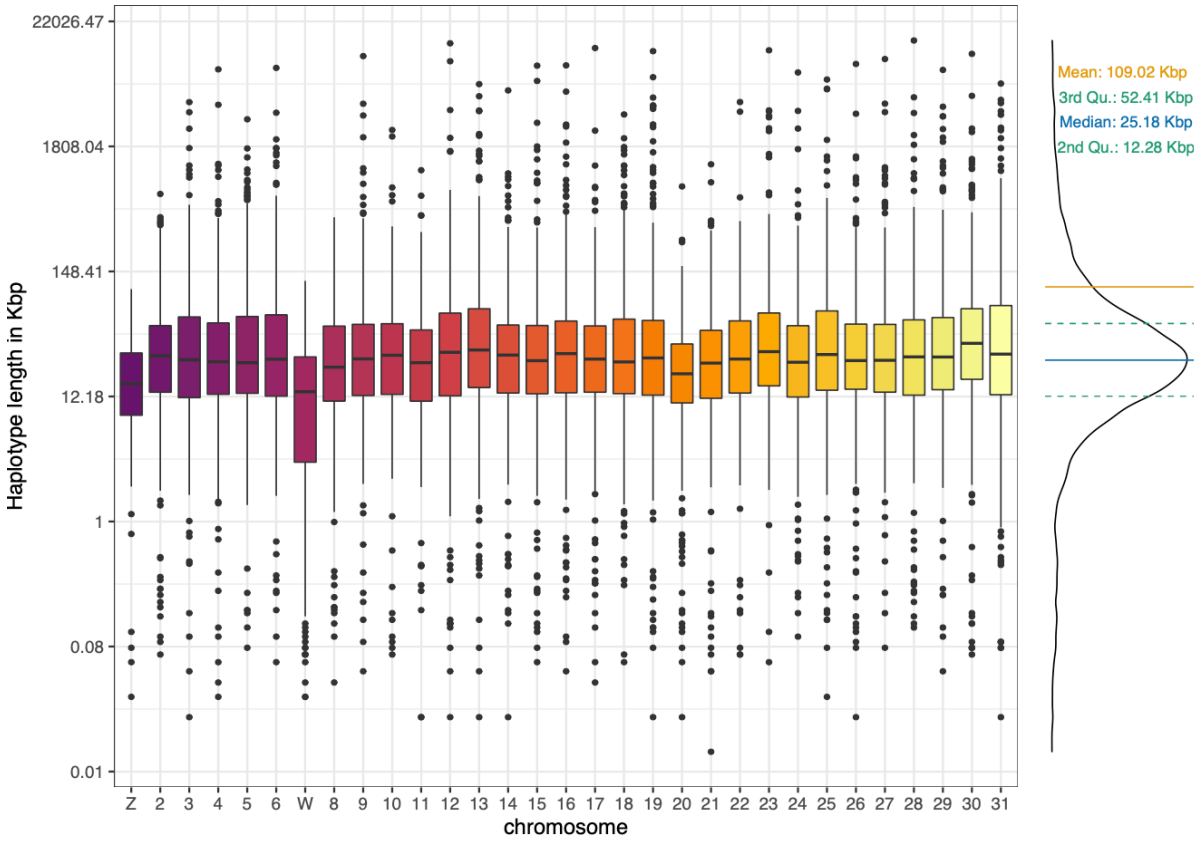
Supplementary Figure 4.2. Performance tests show that Wrath is 10x faster with 20 threads than with a single thread. Performance tests were performed using a subset of genomic windows (100 and 200). 4 scenarios were compared and the run time calculated. To implement multithreading Simplequeue in Python3 was used. A single threaded implementation of Wrath without Simplequeue ran faster than a single thread run with Simplequeue. However, increasing the number of threads decreases run time considerably.



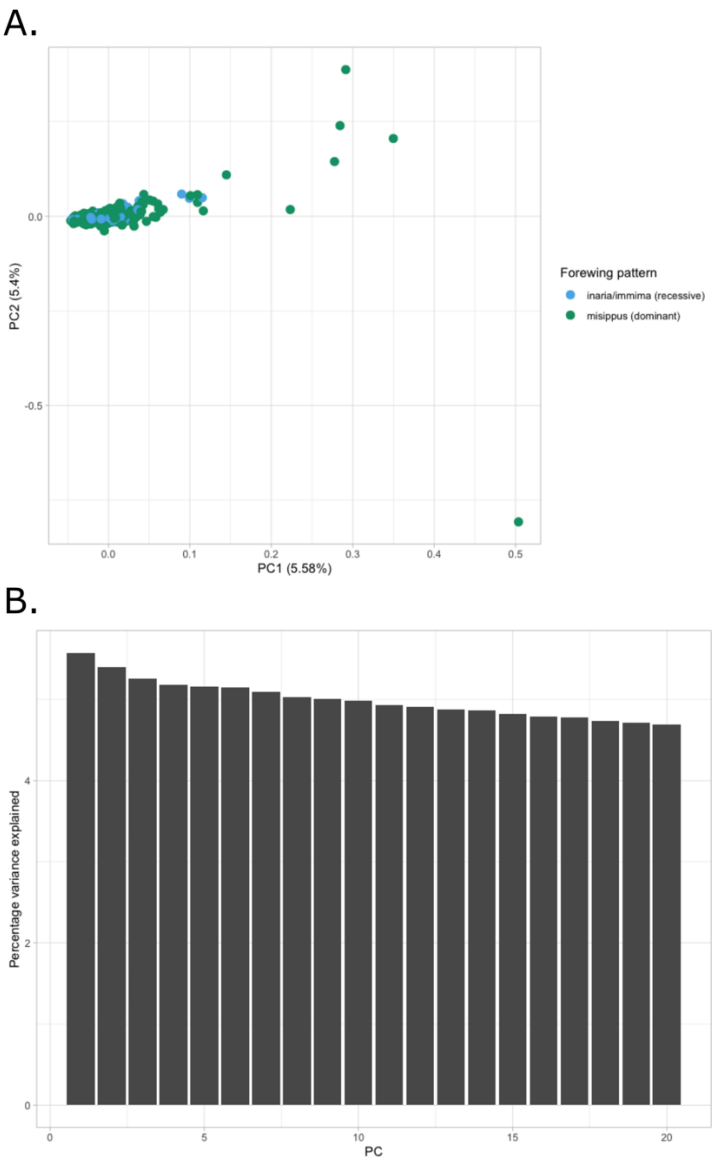
Supplementary Figure 4.3. Read coverage of *Hypolimnas misippus* haplotagging samples per barcode. *H. misippus* individuals were sequenced in four lanes, here shown each in a plot. Mean and median coverage per lane are shown in red and blue respectively.



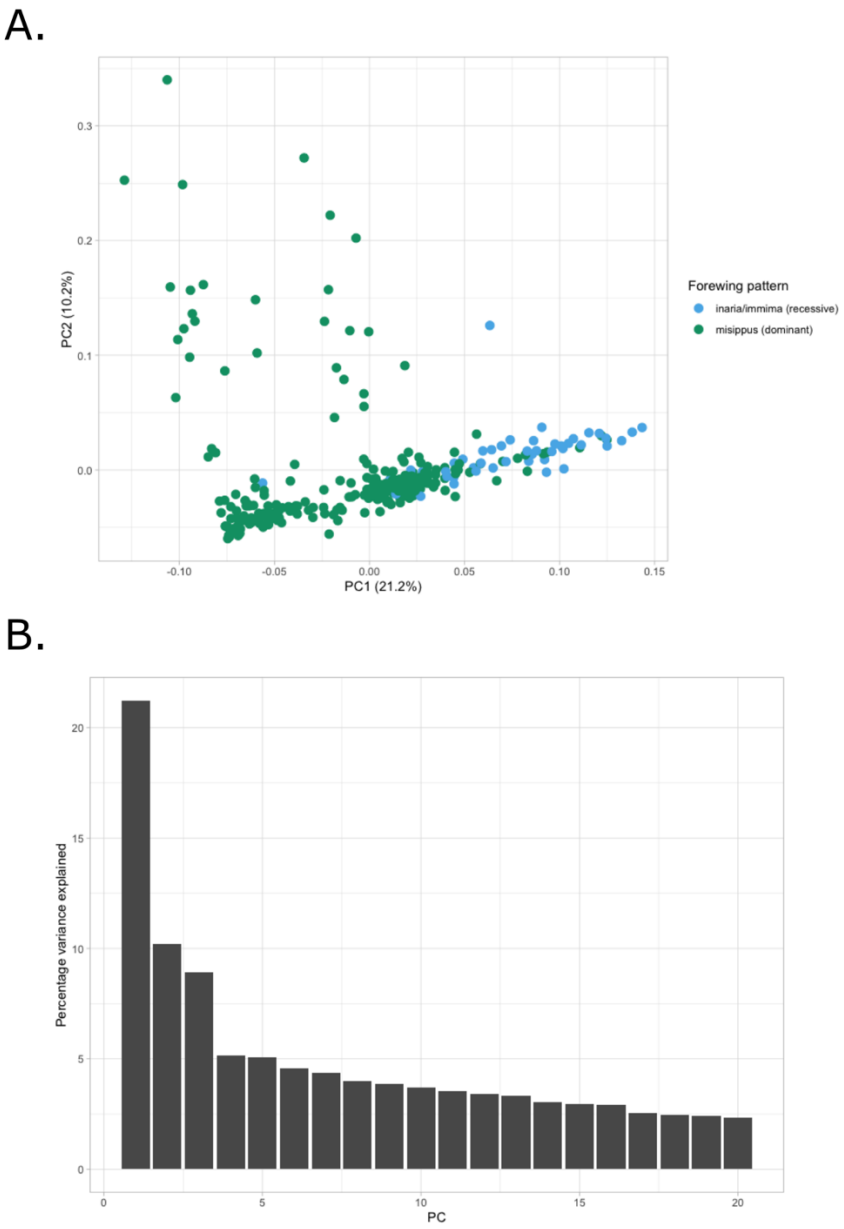
Supplementary Figure 4.4. Testing the performance of SNP imputation of STITCH with different parameters. Y-axis shows the value of Sensitivity as calculated by TP/(TP+FN+FP). Tests using STITCH had higher sensitivity than without using STITCH. And STITCH without BX tag had higher sensitivity than with BX tag (including linked read information). Although difference is minimal and only significant for some comparisons. Other parameters don't make a difference.



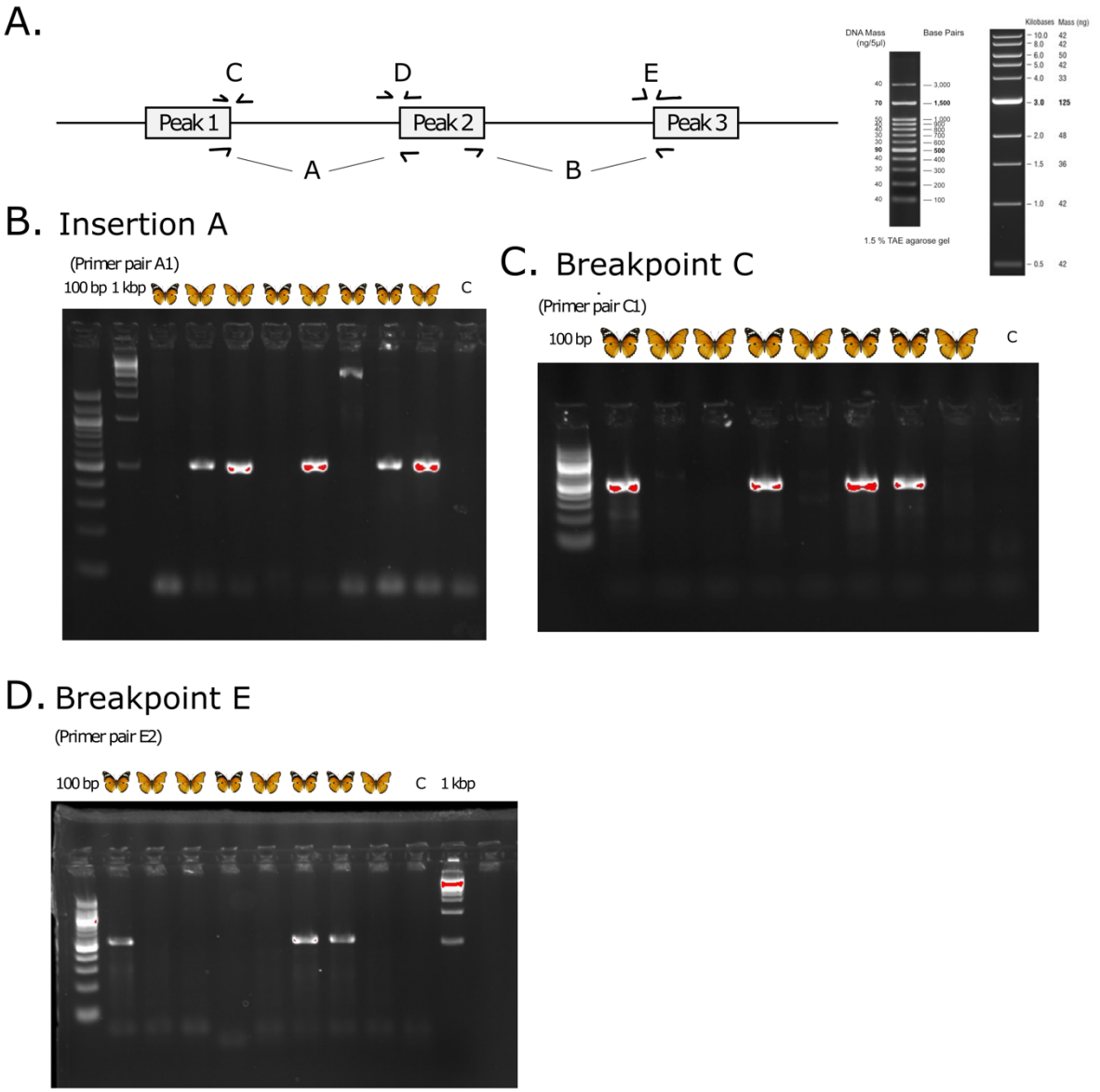
Supplementary Figure 4.5. Haplotype length by chromosome of *H. misippus* samples after phasing with HAPCUT2 (Edge et al., 2017)



Supplementary Figure 4.6. Principal component Analysis (PCA) of *H. misippus* samples using the whole chromosome 29. PCA of the whole chromosome 29 reveals no structure by forewing phenotype (top). The percentage of variance explained by each principal component is low and constant, consistent with the absence of population structure (bottom).



Supplementary Figure 4.7. Principal component Analysis (PCA) of *H. misippus* samples using the region associated with forewing phenotype. PCA of the region associated with forewing phenotype reveals structure by forewing phenotype (top). The percentage of variance explained by each principal component is higher for the first three components, consistent with the presence of population structure (bottom).



Supplementary Figure 4.8. PCR of the insertions at the locus associated with forewing phenotype. A. Diagram of the region associated with forewing phenotype with drawings of the primers used and ladders used in PCRs. **B.** PCR of insertion A shows that *mm* individuals (all orange forewings) have no insertion, as the PCR product is small (500 bp). One *M-* is heterozygote and thus also presents a band. **C.** PCR of breakpoint C shows that *M-* individuals possess insertion A, while *mm* do not. **D.** PCR of breakpoint E shows that *M-* individuals possess insertion B, while *mm* do not. No PCR product was amplificated at breakpoint D and insertion B.

Supplementary Tables

Supplementary Table 1. Read coverage statistics per lane

Lane	Mean coverage	Median	Standard deviation
183_L6	0.846	0.872	0.249
184_L6	0.847	0.892	0.270
184_L7	0.853	0.877	0.320
184_L8	0.676	0.675	0.206

Supplementary Table 2. BLAST results of genes around the region around the associated locus with forewing phenotype against *Drosophila melanogaster*

H.misippus **transcript**	***D.melanogaster*** **protein**	***D. melanogaster*** **prot. description**	**E-value**	**Start pos.**	**End pos.**
g1717.t1	NP_609354.1	uncharacterized protein	0	6001632	6069730
g1717.t1	NP_725571.1	uncharacterized protein%2C isoform B	0	6001632	6069730
g1717.t1	NP_611119.1	uncharacterized protein%2C isoform A	0	6001632	6069730
g1719.t1	NP_649078.1	uncharacterized protein%2C isoform A	5.56E-59	6073069	6074392
g1719.t1	NP_001262027.1	uncharacterized protein%2C isoform B	5.56E-59	6073069	6074392
g1721.t1	NP_649853.1	uncharacterized protein%2C isoform A	4.68E-64	6086308	6100583
g1721.t1	NP_001262394.1	uncharacterized protein%2C isoform B	4.68E-64	6086308	6100583
g1723.t1	NP_609418.1	lipase 4%2C isoform A	1.75E-109	6107821	6110369
g1727.t1	NP_648319.1	uncharacterized protein%2C isoform A	5.28E-44	6119757	6124412
g1727.t1	NP_001137926.1	uncharacterized protein%2C isoform D	5.28E-44	6119757	6124412
g1729.t1	NP_610796.1	Cyp301a1	0	6128617	6145667
g1729.t1	NP_995803.1	Cyp49a1%2C isoform D	2.61E-155	6128617	6145667
g1729.t1	NP_610588.2	Cyp49a1%2C isoform A	2.61E-155	6128617	6145667
g1729.t1	NP_001246256.1	Cyp49a1%2C isoform E	2.61E-155	6128617	6145667
g1730.t1	NP_995803.1	Cyp49a1%2C isoform D	3.7E-166	6158862	6200842
g1730.t1	NP_610588.2	Cyp49a1%2C isoform A	3.7E-166	6158862	6200842
g1730.t1	NP_001246256.1	Cyp49a1%2C isoform E	3.7E-166	6158862	6200842
g1730.t1	NP_610796.1	Cyp301a1	1.53E-154	6158862	6200842
g1731.t1	NP_649220.1	zye	0	6190393	6207540
g1731.t1	NP_649220.1	zye	8.13E-45	6190393	6207540
g1732.t1	NP_001137909.1	down syndrome cell adhesion molecule 4%2C isoform H	5.58E-35	6294473	6308283
g1732.t1	NP_001137909.1	down syndrome cell adhesion molecule 4%2C isoform H	4.7E-31	6294473	6308283
g1732.t1	NP_001137909.1	down syndrome cell adhesion molecule 4%2C isoform H	2.87E-16	6294473	6308283
g1732.t1	NP_001137908.1	down syndrome cell adhesion molecule 4%2C isoform D	7.04E-35	6294473	6308283
g1732.t1	NP_001137908.1	down syndrome cell adhesion molecule 4%2C isoform D	6.61E-31	6294473	6308283
g1732.t1	NP_001137908.1	down syndrome cell adhesion molecule 4%2C isoform D	3.08E-16	6294473	6308283
g1732.t1	NP_001137908.1	down syndrome cell adhesion molecule 4%2C isoform D	7.05E-15	6294473	6308283
g1732.t1	NP_001036596.2	down syndrome cell adhesion molecule 4%2C isoform E	7.04E-35	6294473	6308283
g1732.t1	NP_001036596.2	down syndrome cell adhesion molecule 4%2C isoform E	6.61E-31	6294473	6308283
g1732.t1	NP_001036596.2	down syndrome cell adhesion molecule 4%2C isoform E	3.08E-16	6294473	6308283
g1732.t1	NP_001036596.2	down syndrome cell adhesion molecule 4%2C isoform E	7.05E-15	6294473	6308283
g1732.t1	NP_001036506.1	down syndrome cell adhesion molecule 1%2C isoform BE	1.95E-30	6294473	6308283
g1732.t1	NP_001036506.1	down syndrome cell adhesion molecule 1%2C isoform BE	9.84E-29	6294473	6308283
g1732.t1	NP_001036506.1	down syndrome cell adhesion molecule 1%2C isoform BE	1.69E-19	6294473	6308283
g1732.t1	NP_001036506.1	down syndrome cell adhesion molecule 1%2C isoform BE	1.26E-15	6294473	6308283

g1732.t1	NP_649078.1	uncharacterized protein%2C isoform A	3.9E-26	6294473	6308283
g1732.t1	NP_001262027.1	uncharacterized protein%2C isoform B	3.9E-26	6294473	6308283
g1733.t1	NP_001036506.1	down syndrome cell adhesion molecule 1%2C isoform BE	3.8E-42	6310968	6333469
g1733.t1	NP_001036506.1	down syndrome cell adhesion molecule 1%2C isoform BE	1.75E-22	6310968	6333469
g1733.t1	NP_001036506.1	down syndrome cell adhesion molecule 1%2C isoform BE	6.91E-20	6310968	6333469
g1733.t1	NP_001036506.1	down syndrome cell adhesion molecule 1%2C isoform BE	1.56E-19	6310968	6333469
g1733.t1	NP_001137909.1	down syndrome cell adhesion molecule 4%2C isoform H	1.15E-36	6310968	6333469
g1733.t1	NP_001137909.1	down syndrome cell adhesion molecule 4%2C isoform H	1.75E-23	6310968	6333469
g1733.t1	NP_001137909.1	down syndrome cell adhesion molecule 4%2C isoform H	1.84E-21	6310968	6333469
g1733.t1	NP_001137908.1	down syndrome cell adhesion molecule 4%2C isoform D	1.63E-36	6310968	6333469
g1733.t1	NP_001137908.1	down syndrome cell adhesion molecule 4%2C isoform D	1.84E-23	6310968	6333469
g1733.t1	NP_001137908.1	down syndrome cell adhesion molecule 4%2C isoform D	2.88E-21	6310968	6333469
g1733.t1	NP_001036596.2	down syndrome cell adhesion molecule 4%2C isoform E	1.63E-36	6310968	6333469
g1733.t1	NP_001036596.2	down syndrome cell adhesion molecule 4%2C isoform E	1.84E-23	6310968	6333469
g1733.t1	NP_001036596.2	down syndrome cell adhesion molecule 4%2C isoform E	2.88E-21	6310968	6333469
g1733.t1	NP_649078.1	uncharacterized protein%2C isoform A	2.05E-23	6310968	6333469
g1733.t1	NP_001262027.1	uncharacterized protein%2C isoform B	2.05E-23	6310968	6333469
g1735.t1	NP_001137908.1	down syndrome cell adhesion molecule 4%2C isoform D	5.34E-33	6351841	6361517
g1735.t1	NP_001137908.1	down syndrome cell adhesion molecule 4%2C isoform D	4.59E-21	6351841	6361517
g1735.t1	NP_001137908.1	down syndrome cell adhesion molecule 4%2C isoform D	3.73E-14	6351841	6361517
g1735.t1	NP_001036596.2	down syndrome cell adhesion molecule 4%2C isoform E	5.34E-33	6351841	6361517
g1735.t1	NP_001036596.2	down syndrome cell adhesion molecule 4%2C isoform E	4.59E-21	6351841	6361517
g1735.t1	NP_001036596.2	down syndrome cell adhesion molecule 4%2C isoform E	3.73E-14	6351841	6361517
g1735.t1	NP_001137909.1	down syndrome cell adhesion molecule 4%2C isoform H	5.48E-33	6351841	6361517
g1735.t1	NP_001137909.1	down syndrome cell adhesion molecule 4%2C isoform H	4.64E-21	6351841	6361517
g1735.t1	NP_001137909.1	down syndrome cell adhesion molecule 4%2C isoform H	4.01E-14	6351841	6361517
g1735.t1	NP_001036506.1	down syndrome cell adhesion molecule 1%2C isoform BE	9.74E-27	6351841	6361517
g1735.t1	NP_001036506.1	down syndrome cell adhesion molecule 1%2C isoform BE	4.86E-17	6351841	6361517
g1735.t1	NP_001036506.1	down syndrome cell adhesion molecule 1%2C isoform BE	5.77E-16	6351841	6361517
g1736.t1	NP_649810.1	pink%2C isoform A	3.19E-63	6363126	6372758
g1736.t1	NP_001303456.1	pink%2C isoform C	3.19E-63	6363126	6372758
g1736.t1	NP_001303455.1	pink%2C isoform B	3.19E-63	6363126	6372758

g1737.t1	NP_726612.1	Sox102F%2C isoform A	1.43E-69	6378212	6488360
g1737.t1	NP_001259074.1	Sox102F%2C isoform D	1.43E-69	6378212	6488360
g1737.t1	NP_001259073.1	Sox102F%2C isoform C	1.43E-69	6378212	6488360
g1743.t1	NP_001163715.2	uncharacterized protein%2C isoform C	1.4E-75	6514980	7122953
g1743.t1	NP_651143.2	uncharacterized protein%2C isoform B	3.45E-71	6514980	7122953
g1743.t1	NP_001287484.1	uncharacterized protein%2C isoform D	3.45E-71	6514980	7122953
g1743.t1	NP_001287483.1	uncharacterized protein%2C isoform C	3.45E-71	6514980	7122953
g1741.t1	NP_001259889.1	earmuff%2C isoform B	6.19E-40	6918165	6931190
g1741.t1	NP_001259889.1	earmuff%2C isoform B	9.93E-19	6918165	6931190
g1742.t1	NP_651143.2	uncharacterized protein%2C isoform B	2.06E-82	6995476	7053560
g1742.t1	NP_001287484.1	uncharacterized protein%2C isoform D	2.06E-82	6995476	7053560
g1742.t1	NP_001287483.1	uncharacterized protein%2C isoform C	2.06E-82	6995476	7053560
g1745.t1	NP_524638.3	twin of eyeless%2C isoform A	8.07E-167	7135073	7189742
g1745.t1	NP_524638.3	twin of eyeless%2C isoform A	1.11E-28	7135073	7189742
g1745.t1	NP_524628.2	eyeless%2C isoform A	2.58E-70	7135073	7189742
g1745.t1	NP_524628.2	eyeless%2C isoform A	4.95E-43	7135073	7189742
g1751.t1	NP_732660.1	torso-like%2C isoform B	2.52E-123	7256921	7344844
g1751.t1	NP_524440.2	torso-like%2C isoform A	2.52E-123	7256921	7344844
g1751.t1	NP_001262810.1	torso-like%2C isoform C	2.52E-123	7256921	7344844
g1753.t1	NP_732660.1	torso-like%2C isoform B	4.77E-55	7362526	7369847
g1753.t1	NP_524440.2	torso-like%2C isoform A	4.77E-55	7362526	7369847
g1753.t1	NP_001262810.1	torso-like%2C isoform C	4.77E-55	7362526	7369847
g1758.t1	NP_524225.1	ripped pocket%2C isoform A	5.35E-57	7390792	7398754
g1762.t1	NP_524628.2	eyeless%2C isoform A	3.06E-69	7454928	7460431
g1762.t1	NP_524638.3	twin of eyeless%2C isoform A	3.3E-65	7454928	7460431

Supplementary Table 3. BLAST results of genes around the region around the associated locus with forewing phenotype against *Heliconius melpomene v2.5*

qseqid	sseqid	pident	evalue	start	end
g1717	HMEL032184g1	95.672	0	6001632	6069730
g1719	HMEL012581g1	72.099	0	6073069	6074392
g1720	HMEL003053g1	87.963	5.04E-148	6082880	6085879
g1721	HMEL003054g1	92.26	0	6086308	6100583
g1722	HMEL003055g1	75.676	6.74E-36	6100590	6104698
g1723	HMEL003056g3	78.543	0	6107821	6110369
g1723	HMEL003056g2	73.711	0	6107821	6110369
g1723	HMEL003056g1.t2	54.251	0	6107821	6110369
g1723	HMEL003056g1	65.405	0	6107821	6110369
g1724	HMEL031494g1	30.182	3.24E-34	6114466	6115794
g1725	HMEL036165g1	39.234	1.46E-38	6115855	6116574
g1726	HMEL034362g1	31.751	3.66E-77	6116731	6118584
g1727	HMEL003057g1	56.294	9.59E-106	6119757	6124412
g1728	HMEL003057g2	62.5	2.12E-102	6125223	6127267
g1729	HMEL003058-RA	93.384	0	6128617	6145667
g1730	HMEL007856-RA	80.074	0	6158862	6200842
g1731	HMEL022357g1	71.288	0	6190393	6207540
g1731	HMEL022357g1	67.955	0	6190393	6207540
g1731	HMEL022357g1	71.563	0	6190393	6207540
g1731	HMEL022357g1	68.632	0	6190393	6207540
g1731	HMEL022357g1	60.864	0	6190393	6207540
g1731	HMEL022357g1	63.368	0	6190393	6207540
g1732	HMEL032188g1	47.758	0	6294473	6308283
g1733	HMEL032188g1	38.809	0	6310968	6333469
g1734	HMEL032188g1	37.571	8.9E-68	6343149	6347605
g1735	HMEL032189g1	53.026	0	6351841	6361517
g1736	HMEL016538g2	73.567	0	6363126	6372758
g1737	HMEL020265g1	91.374	0	6378212	6488360
g1743	HMEL032196g1	74.827	0	6514980	7122953
g1740	HMEL016757g1	29.257	1.35E-50	6907246	6912130
g1741	HMEL011763g1	65.421	0	6918165	6931190
g1742	HMEL011806g1.t2	94.333	0	6995476	7053560
g1742	HMEL011806g1	93.333	0	6995476	7053560
g1744	HMEL012216g1	73.496	0	7121438	7133011
g1745	HMEL012214g1	96.928	0	7135073	7189742
g1746	HMEL032198g1	51.485	4.6E-28	7249146	7250042
g1747	HMEL005533g1	61.538	2.63E-19	7251519	7251945
g1748	HMEL005532g1	29.921	3.97E-21	7253053	7256112
g1751	HMEL032199g1	88.75	0	7256921	7344844
g1751	HMEL032199g1.t2	77.808	0	7256921	7344844
g1749	HMEL005007g1	63.087	3.14E-159	7269375	7274356
g1753	HMEL032200g1	81.953	0	7362526	7369847
g1754	HMEL032201g1	78.761	0	7370488	7382484
g1755	HMEL032202g1	56.25	2.16E-70	7382770	7384085
g1756	HMEL032203g1	65.726	1.54E-126	7384808	7387175
g1757	HMEL032204g1	59.112	0	7387805	7390605
g1758	HMEL032205g1	72.213	0	7390792	7398754
g1759	HMEL009698g1	72.848	3.05E-79	7402208	7409244
g1761	HMEL032207g1	69.565	3.7E-13	7448877	7449965
g1762	HMEL021801g1	88.272	7.78E-101	7454928	7460431

Supplementary Table 4. Transposable elements at the locus associated with forewing pattern identified by RepeatMasker

Chr.	Start	End	Family	Matching repeat	Position
29	6729052	6729213	RC/Helitron	rnd-4_family-1108	-
29	6730618	6730842	RC/Helitron	rnd-5_family-121	-
29	6730827	6730869	RC/Helitron	rnd-6_family-235	-
29	6731047	6731220	Unknown	rnd-4_family-146	Peak 1
29	6731227	6731396	Unknown	rnd-4_family-279	Peak 1
29	6732514	6732589	RC/Helitron	rnd-6_family-52	Peak 1
29	6733032	6733181	RC/Helitron	rnd-6_family-681	Peak 1
29	6733276	6733314	Simple_repeat	(TATT)n	Peak 1
29	6733396	6733458	LINE/L2	rnd-6_family-627	Peak 1
29	6733459	6734488	RC/Helitron	rnd-4_family-122	Insertion A
29	6734490	6735220	RC/Helitron	rnd-4_family-122	Insertion A
29	6735155	6735671	Unknown	rnd-4_family-53	Insertion A
29	6735672	6735711	Simple_repeat	(GACA)n	Peak 2
29	6735712	6735757	Unknown	rnd-6_family-258	Peak 2
29	6735753	6735835	LINE/L2	rnd-6_family-139	Peak 2
29	6735762	6735870	LINE/L2	rnd-4_family-271	Peak 2
29	6736485	6736644	LINE/R1	rnd-5_family-302	Peak 2
29	6736763	6736792	Simple_repeat	(AT)n	Peak 2
29	6736903	6737006	RC/Helitron	rnd-6_family-235	Insertion B
29	6736923	6737070	RC/Helitron	rnd-4_family-489	Insertion B
29	6736926	6737107	Unknown	rnd-4_family-219	Insertion B
29	6737135	6737534	Unknown	rnd-5_family-759	Insertion B
29	6737537	6738061	LINE/CR1	rnd-6_family-305	Insertion B
29	6738127	6738890	LINE/CR1	CR1-4_Hmel_A	Insertion B
29	6738890	6739670	LINE/CR1	rnd-6_family-79	Insertion B
29	6739680	6740646	LINE/CR1	rnd-6_family-79	Insertion B
29	6740805	6740954	RC/Helitron	rnd-4_family-166	Insertion B
29	6741294	6741310	Unknown	rnd-5_family-927	Peak 3
29	6741311	6741424	Unknown	rnd-5_family-442	Peak 3
29	6741847	6741879	Simple_repeat	(AATAA)n	Peak 3
29	6742085	6742124	Simple_repeat	(ATA)n	Peak 3
29	6742596	6742667	Simple_repeat	(AT)n	Peak 3
29	6742789	6742968	RC/Helitron	rnd-4_family-489	Peak 3
29	6743031	6743064	Unknown	rnd-6_family-306	Peak 3
29	6743065	6743213	RC/Helitron	rnd-4_family-183	Peak 3
29	6743302	6743327	Simple_repeat	(TAA)n	Peak 3
29	6743615	6743692	RC/Helitron	rnd-3_family-741	-
29	6743763	6743827	RC/Helitron	rnd-5_family-201	-
29	6744341	6744593	RC/Helitron	rnd-4_family-148	-
29	6744779	6745035	RC/Helitron	rnd-4_family-149	-
29	6745354	6745470	RC/Helitron	rnd-5_family-178	-
29	6745905	6746033	RC/Helitron	rnd-5_family-897	-
29	6746111	6746336	RC/Helitron	rnd-3_family-741	-
29	6746825	6747024	DNA/TcMar-Mariner	rnd-2_family-183	-
29	6747111	6747156	Simple_repeat	(AT)n	-
29	6747210	6747247	Simple_repeat	(AATA)n	-
29	6747468	6747524	Unknown	rnd-6_family-278	-
29	6747949	6748103	RC/Helitron	rnd-4_family-183	-
29	6748747	6748803	Unknown	rnd-6_family-105	-
29	6748927	6749058	Unknown	rnd-6_family-515	-
29	6749787	6749927	RC/Helitron	rnd-6_family-691	-

29	6749812	6749930	RC/Helitron	rnd-4_family-489	-
29	6749843	6749941	Unknown	rnd-6_family-352	-

5

Cortex again: the repeated use of a mimicry hotspot gene

1.23 Abstract

The repeatability of evolution has been a widely debated topic among evolutionary biologists. Although rewinding the tape of life perhaps would not lead to the same outcome every time, the repeated use of the same genes for similar functions seems to be a common theme in evolution. Wing patterns of butterflies and moths have provided a wealth of examples of gene reuse, with certain 'hotspot genes' controlling wing patterning in multiple taxa. Here, I show that Batesian wing mimicry in two *Hypolimnas* butterfly species is likely to be controlled by *cortex. Cortex* has been shown to regulate wing pattern formation in many Lepidoptera species by affecting scale type development. In *H. misippus* and *H. bolina* mimicry is female limited and polymorphic and are an ideal case for the study of evolution of adaptive phenotypes. By dissecting the genetic architecture of the loci, I present evidence that suggests that distinct cis-regulatory elements control the development of white pattern elements in the forewing and hindwing of *H. misippus* and *H. bolina* and that no structural variation is found at the locus. Finally, I also show that orange colouration in *H. bolina* is associated with *optix,* a well-known patterning gene. Overall, my study shows that once more variation near the hotspot genes *cortex* and *optix* is associated with wing mimicry and highlights the repeatability of the evolution of wing phenotypes in butterflies.

1.24 Introduction

Convergent evolution, defined as the independent evolution of similar traits in different lineages, often in response to similar environmental pressures, has long fascinated evolutionary biologists (Stern, 2013). Genetic analysis has revealed that convergent phenotypes fairly commonly involve similar genetic changes (Conte et al., 2012; Gompel and Prud'homme, 2009). Many examples of this come from colouration phenotypes, where certain 'hotspot genes' or homologous loci have been repeatedly linked to both similar and divergent phenotypes (reviewed in Chapter 1). These genetic hotspots are hypothesised to represent loci that maximise changes to the trait while minimising pleiotropic effects (Stern, 2013). However, the apparent repeatability of evolution can be the result of different mechanisms that might affect the likelihood of those events, as parallel changes can originate from independent mutations at the same gene or locus (for example (Nadeau et al., 2016)) but also through other processes. Introgression of adapted alleles from other lineages (for example (Jones et al., 2018)) and selection on shared ancestral variation present at the locus after lineage divergence (i.e. standing genetic variation; for example (Jones et al., 2012)) can also lead to repeated evolution without independent mutational events in distinct lineages (Martin and Orgogozo, 2013; Stern, 2013).

Butterfly wing phenotypes are a well-studied system for understanding the evolution of adaptive traits and the genetic basis of convergent phenotypes (Benson, 1972; Jiggins et al., 2017; Mallet and Barton, 1989; Turner, 1987). Much attention has focussed on mimicry in tropical butterflies such as *Heliconius* and *Papilio* species. *Heliconius* is a genus of tropical butterflies with striking Mullerian mimicry, in which multiple sympatric and unpalatable species evolve to resemble one other, thereby sharing the costs of teaching predators (Turner, 1981). Four major effect genes have been associated with wing convergent phenotypes in several *Heliconius* species: *cortex, aristalless1, WntA* and *optix* (Mazo-Vargas et al., 2017; Nadeau et al., 2016; Reed et al., 2011; Westerman et al., 2018). Interestingly, *cortex*, which controls scale morphology and the switch between white/yellow and black/red in *Heliconius* (Livraghi et al., 2021; Nadeau et al., 2016), has also been implicated in colour phenotypes in other divergent Lepidoptera species (Beldade et al., 2009; Ito et al., 2016; van der Burg et al., 2020; van't Hof et al., 2019, 2016; Wang et al., 2022). *WntA* and *optix* are also

involved in wing patterning in several other Lepidoptera highlighting the repeatability of the genetic control of wing phenotypes in butterflies and moths (Mazo-Vargas et al., 2017; Reed et al., 2011; Zhang et al., 2017b). However, the extent of the re-use of similar genes for wing phenotypes is still largely unexplored. Dissecting the genetic architecture of adaptive wing phenotypes in other species will improve our understanding of the evolution of adaptive alleles.

Two species of *Hypolimnas* butterflies, *H. misippus* and *H. bolina*, show female limited polymorphic Batesian mimicry (Clarke and Sheppard, 1975; Gordon et al., 2010; Smith, 1976). In Southeast Asia *H. bolina* is a recognised Batesian mimic of multiple *Euploea* species, while in the south Pacific Islands and Australia it has several non-mimetic forms. The genetic basis of this wing pattern variation was extensively studied by Clarke and Sheppard, who hypothesised that three of the main wing morphs, *nerina*, *naresi* and *eupoeoides*, are determined by two autosomal loci (Clarke and Sheppard, 1975). They hypothesised that the E locus controls the differences between the all-brown mimetic morph *euploeoides* and the *naresi* morph, which presents a hindwing white spot and a subapical white band in the forewing (Figure 5.1). While the N locus controls the presence of forewing orange patch seen in *nerina* wings.

Similarly, polymorphic *H. misippus* females show detailed resemblance to the four morphs of the unpalatable *Danaus chrysippus*. Genetic variation in *D. chrysippus* has been mapped to three loci, offering an opportunity to identify the genetic basis for wing pattern variation on both a model and its Batesian mimic species. In *H. misippus* forewing and hindwing phenotypes are controlled by independent loci and the matching mimetic morphs are achieved by different combinations of alleles in the forewing and hindwing (Gordon et al., 2010; Gordon and Smith, 1989; Smith, 1976; VanKuren et al., 2019). The M locus controls the differences in forewing phenotype, with the dominant allele *M* producing the black-and-white *misippus* wings, while the recessive *m* allele produces all-orange wings and intermediates known as *inaria* and *immima* (Figure 5.1)(VanKuren et al., 2019). Interestingly, the M locus has been narrowed down to an intergenenic region in which no known colour genes are found (see Chapter 3 and (VanKuren et al., 2019)). This shows that wing colour phenotypes can be controlled by novel genes not described in other species. Hindwing colour in *H. misippus* varies continuously from orange to white and is controlled by the A locus,

whose dominant allele produces hindwing white and presents incomplete dominance (Gordon and Smith, 1989). A suppressor locus, the S locus, has been hypothesised to counteract the effect of the A locus by limiting the presence of white in the hindwing. Furthermore, the A locus has been hypothesised to be epistatic with the M locus, controlling the switch between *inaria* and *immima* forewing phenotypes. In summary, *H. misippus* and *H. bolina* both have a simple genetic basis for polymorphic female limited Batesian mimicry. Crucially, white colouration in the hindwing is continuous but controlled by one or two loci of major effect of two alleles each and additional modifier loci (Gordon and Smith, 1989). Separated by 8 million years of evolution (Sahoo et al., 2018), *H. misippus* and *H. bolina* are a good case study to explore the genetics of wing mimicry and the extent of gene reuse in the evolution of these phenotypes.

Here, I investigate the genetic basis of Batesian mimicry in *H. misippus* and *H. bolina*. I use whole genome analysis of linked-read and short read sequencing, synteny and phylogeny to investigate the genetic control and evolution of hindwing and forewing white colouration in the genus and compare it to other Lepidoptera.

1.25 Materials and Methods

1.25.1 Sample collection, processing, and analysis of *H. misippus*

To explore the genetic basis of hindwing colouration in *Hypolimnas misippus*, the samples described in Chapter 4 were used (Supplementary Table 4). Briefly, 335 individuals were collected in different parts of Africa and preserved in 100% ethanol or sun dried. DNA was then extracted from the samples and libraries prepared using custom protocols (described in Chapter 4) and sequenced using haplotagging a linked-read sequencing technique. BX tags including barcode information of the linked reads were included in the read information field (see Chapter 4). Low quality ends and adapters were trimmed using TRIMMOMATIC (Bolger et al., 2014). Then, reads were mapped to the *H. misippus* reference genome (*HypMisi_v2,* see Chapter 3) using BWA-mem (Li, 2013) and PCR duplicated marked using Picard (broadinstitute.github.io/picard). SNPs were called using bcftools v1.11 (Danecek et al., 2021), imputed using STITCH (Davies et al., 2016) and phased using HapCut2 (Edge et al., 2017).

1.25.2 Collection and processing of wild *H. bolina* samples

214 wild *Hypolimnas bolina* were collected from the island of Rurutu in French Polynesia in 2004, 2005, 2007 and 2013, Moorea in 2005 and 2010 and Australia and Samoa 2018 and used to identify the genetic basis of Batesian mimicry in this species (Supplementary Table 3). First, DNA was extracted of the *H. bolina* samples following a custom protocol that uses PureLink buffers and homemade magnetic beads. To do that, a small piece of thorax tissue (1/10) was dissected and placed in a 8-tube PCR strip, to which 45 uL of PureLink Digestion buffer and 10 uL of Proteinase K (20mg/mL) was added and the samples incubated for 2-3 hours at 58˚C with shaking (500rpm), manually inverting them vigorously every 30 minutes. After that, 2uL of RNAse (DNAse free) were added to each sample, mixed by inversion and incubated 10 minutes at room temperature. Tubes were then spined briefly and 45uL of PureLink Lysis Buffer added before mixing and incubating the samples for 30 min at 58˚C with shaking (500 rpm). Afterwards, to pellet any undigested solids, the samples were span at 4000g for 10 at room temperature. Following that, the DNA was extracted from the lysate using a homemade magnetic bead mix. First, 37.5uL of magnetic bead mix was added in each

well of a 96-well plate. Then, 75uL of lysate was transferred to the well-plate and mixed by pipetting. Two rounds of 80% ethanol clean-ups were then performed, placing the well-plate in a magnetic stan. After the second round of clean-up, 50uL of 10mM Tris at pH 8 were added to elute the DNA the mix was incubated for 15 minutes at 45˚C, then the samples were mixed and incubated for 20 minutes at room temperature. Finally, the samples were placed on the magnet stand and transferred the clean DNA to a fresh strip tube.

From the extracted DNA, libraries were prepared following a method based on Nextera DNA Library Prep (Illumina, Inc.) with purified Tn5 transposase (Picelli et al., 2014). PCR extension with the N701–N800 i7-index primer and the N501-N508 and N5017 i5-index primers was performed to barcode the samples. Library purification and size selection was done using the same homemade beads as above. Pooled libraries were sequenced to ~6.5X coverage on average by Novogene Cambridge, UK.

1.25.3 Sample phenotyping

Once samples had been collected, forewings and hindwings of *H. misippus* individual were photographed in a standardised set-up consisting of a CS-920S Copy Stand holding a Cannon EOS 700D camera with a Cannon EFS 60mm macro lens 43 cm above the wings. Two Godox SK400 lights were used, and the wings were placed in green background with a white (Ocean Optics, Inc. WS-1) and a grey balance checker (Grey White Balance Colour Cards). Phenotypes of *H. misippus* were then scored from the photographs following (Gordon and Smith, 1989), in which hindwings are classified according to the number of sections of the wing (interveins) containing white scales. In this scale, 8 is the maximum amount of white possible and 0 the minimum, with the wing being fully orange. For the categorical classification, 0 was assigned to any individual with a 0 in the continuous scale and 1 to all the other scores (1-8). *H. bolina* samples were phenotyped by the presence (1) or absence (0) of white patched in the forewing and hindwing separately. Also, forewings were also scored by the presence (1) or absence (0) of an orange patch. All phenotypes are included in Supplementary Table 1-3.

1.25.4 Analysis of *H. bolina* reared individuals and Quantitative Trait Locus (QTL) mapping

To identify the genetic basis of wing mimicry in *H. bolina,* a QTL mapping analysis was performed using family samples from Chapter 3 (Supplementary Table 2). Briefly, two families were reared two related individuals mated to a wild one and offspring reared. DNA was extracted and libraries prepared using the same custom protocols as for the wild individuals. Samples were then sequenced to ~11X in the offspring and ~20X in the parents. Low quality ends and adapters were trimmed using TrimGalore! (Krueger, 2015), reads mapped using BWA-mem (Li, 2013) and PCR duplicated were marked with Picard tools. SNPs were called using bcftools v1.11 (Danecek et al., 2021) and QTL mapping performed using the R package qtl2 (Broman et al., 2019).

1.25.5 Genome wide analysis of Samoan *H. bolina* (dataset 1)

To confirm and further explore the results of the QTL mapping, wild *H. bolina* from (Hornett et al., 2014) were used for a preliminary study together with the reared individuals (Supplementary Table 2). GWAS correcting for population structure and relatedness using GEMMA. First, a PCA was generated using Plink v 1.9 (Purcell et al., 2007). Then the relatedness matrix was built with GEMMA (Zhou and Stephens, 2012) and used, together with the first 20 PCs, as input for the linear mixed model utility of GEMMA. This approach has been previously used with good results (VanKuren et al., 2019).

1.25.6 Genome wide analyses of *H. bolina* (dataset 2)

The results from QTL mapping and the preliminary GWAS produced a broad associated region, to narrow it down the 214 wild *H. bolina* samples were used. Reads were first trimmed using fastp (Chen et al., 2018), which performs quality control and trims low quality ends and adapters. Then, processed reads were mapped to the two reference genomes produced in Chapter 2 *HypMisi_v2* and *HypBol_v1* using BWA-MEM2 (Vasimuddin et al., 2019) and marked duplicates produced by PCR using the MarkDuplicatesSpark from GATK (Auwera and O'Connor, 2020). SNPs were called on each chromosome separately using bcftools v1.11 (Danecek et al., 2021) mpileup requesting the INFO/AD,AD,DP,DV,DPR,INFO/DPR,DP4,SP tags to output (-a option), setting the minimum mapping quality to 10 (-q) and the minimum base

quality to 20 (-Q), ignoring Read Group tags (--ignore-RG) and removing duplicates (-F 1024). The output was piped of bcftools mpileup directly to bcftools call to obtain the final vcf files of called SNPs using the alternative model for multiallelic and rare-variant calling (--multiallelic-caller), including only variants in the output (--variants-only) and the fields GQ and GP (-f GQ,GP). After that, the data were filtered based on genotype quality (>30) and depth (>2 and <12). Thresholds were set after exploring a subset of the data.

1.25.7 Genome Wide Association Studies (GWAS)

GWAS was used to identify regions associated with wing phenotype. To do that, a principal component analysis (PCA) analysis in *H. bolina* (dataset 2) and *H. misippus* was first carried out using Plink v1.9 (Purcell et al., 2007)(Supplementary Figure 3) and the first 5 principal components used as covariates in the association test, which was performed with Plink using the –assoc option. The -log10 significance levels were calculated using Bonferroni correction to account for multiple testing, which were 8.75 and 9.11 in *H. bolina* mapped to the *H. bolina* reference and to the *H. misippus* reference respectively, and 8.96 in *H. misippus*, calculated from the 27,889,110, 64,636,142 and 46,088,305 SNPs used.

1.25.8 Chromosome naming

The chromosomes for *H. misippus* and *H. bolina* were named based on homology to *M. cinxia* as in Chapter 3. Briefly, BUSCO (Seppey et al., 2019) matches using the odb_insecta10 gene set are used to infer homology. The homology between *H. misippus* and *H. bolina* chromosomes with *M. cinxia* is summarised in Supplementary Table 1. Unless specified, all *H. bolina* and *H. misippus* chromosomes mentioned refer to the names based on homology to *M. cinxia.*

1.25.9 Reference genome alignments

To investigate the origin of the adapted alleles, homology between regions associated with wing phenotype in each species was explored. To identify putative orthologous regions between the reference genomes of *H. bolina* and *H misippus*, the two references, *HypMisi_v2* and *HypBol_v1,* were aligned with Satsuma2 (Grabherr et al., 2010) using default parameters. The resulting alignments were visualised using the asynt R functions (Kim et al., 2022).

1.25.10 Read Depth analysis and identification of indels

Identification of indels putatively associated with wing phenotype was performed by calculating read depth from the bam files using the depth utility from samtools (Danecek et al., 2021) with the –a option to output depth for all sites, including those with no reads mapping to them. The output was visualised in R using the ggplot2 package.

Individual BAM files with marked duplicates were subset for the region of interest using tabix and merged using samtools merge. These merged BAM files were then visualised using IGV. The candidate indels were identified through visual inspection. SNPs associated with the indel were used as proxies to identify the individuals carrying the deletion and insertion.

1.25.11 Phylogenetic trees

To explore the evolution of the alleles, phylogenetic trees were generated in genomic windows. To do that, genotype files were first produced using the parseVCF.py from the genomics_general toolkit using the phased vcf files produced by HAPCUT2 as input. Then phylogenetic trees of the cortex locus were generated in windows of 50 SNPs using the phyml_sliding_windows.py of the genomics_general toolkit setting the options –windType sites –model GTR –optimise n.

1.26 Results

1.26.1 Forewing white and hindwing white are controlled by independent functional elements at the E locus in *H. bolina*

Clarke and Sheppard hypothesised that the differences between the all-brown *euploeoides* morph and the white-spotted *naresi* morph were controlled by a single locus of major effect, the E locus (Clarke and Sheppard, 1975). A single cross described here provides evidence for recombination between forewing and hindwing elements within this major locus. Given that polymorphism is female-limited, genotypic information from crosses comes only from female phenotypes. A first cross was performed between a female presenting a forewing white band and a large hindwing white patch, which would be considered a *naresi* morph. Then, a female offspring of this cross with a forewing white band but only a reduced hindwing white patch was mated to a wild male (Family 1, Supplementary Table 2). All the offspring of this cross had a forewing white band, while they varied in the presence and size of the hindwing patch, with 34 individuals having all-brown wings and 18 individuals with a white patch (varying in size). The segregation of hindwing but not forewing white in this cross suggests that there are two functionally distinct linked elements. Tight physical linkage between the two elements would explain the joint inheritance of the traits but also the existence of recombinants.

1.26.2 GWAS for white colouration points at *cortex* as the main candidate

Next, I identified the region of the genome controlling the presence of white elements in *H. bolina* using two datasets. Using the reared families from Chapter 2, I performed a quantitative trait mapping (QTL) analysis, which showed that the locus associated with hindwing white variation is found on chromosome 8 (Supplementary Figure 1, Supplementary Table 2). In *H. bolina,* males are uninformative for the E locus genotype, as they are monomorphic. This together with the fact that there is substantial continuous variation in hindwing white hinders the correct genotyping of samples from phenotype information. Given the segregation pattern of hindwing phenotype (1:2), I deduced that one of the parents was heterozygote for the E locus and the other homozygous. I performed two QTL analyses assuming heterozygosity of either parent producing the same results, an association peak at chromosome 8. Given that there is no recombination in female butterflies (Turner and

Sheppard, 1975), when the mothers of the crosses are assumed to be heterozygous, the association can only be narrowed down to the chromosomal level (Supplementary Figure 1).

To confirm this result, I used these families together with a dataset of 45 sequenced wild individuals publicly available from (dataset 1)(Hornett et al., 2014) and performed a GWAS correcting for population structure and relatedness using GEMMA. This approach has been previously used with good results (VanKuren et al., 2019). This confirmed the association of hindwing white and forewing white on chromosome 8 but given the sample size and high relatedness, the peak was broad (Supplementary Figure 2, Supplementary Table 2). Photos of these samples are not available, thus, to identify and narrow down the loci associated with morph variation in *H. bolina* I sequenced a dataset of wild individuals.

To narrow down the associated locus, I sequenced whole genomes from 214 *H. bolina* individuals (dataset 2) varying in hindwing colouration and performed a GWAS for each of the traits of interest (Supplementary Table 3). Similarly, I identified the locus associated with hindwing white variation in *H. misippus* using a dataset of genome wised sequences of 335 wild *H. misippus* (Supplementary Table 3). These analyses identified a region of chromosome 8 associated with hindwing white in both species and forewing white in *H. bolina* (Figure 5.1A-B). As chromosomes in both genomes were named based on similarity to the *Melitaea cinxia* genome, these are homologous chromosomes, but coordinates within those chromosomes are not comparable between the two reference genomes. To explore possible candidate genes for the control of the trait and clarify homology between the two regions, I surveyed the genes annotated near each of the associated regions, and *cortex* was a clear candidate in both cases (Figure 5.1C-D). SNPs associated with hindwing polymorphism in both species fall in the non-coding sequence around *cortex*. This is reminiscent of other studies that have linked *cortex* to white/yellow colouration, such as in *Heliconius melpomene* and *Heliconius erato,* in which cis-regulatory variation near *cortex* has been found to cause the switch between colour phenotypes. On the other hand, the region associated with forewing phenotype in *H. bolina* is broad and covers the upstream and downstream regions and *cortex* itself.

A second peak of association is seen in the Z chromosome the *H. misippus* GWAS. This could be a second major effect allele contributing to variation in the trait, which would fit the

previous hypothesis of two loci, the S and A loci, controlling hindwing colour variation in *H. misippus.* The top 6 SNPs fall in the intron of a gene (g10224), which could be linked to this phenotype but is uncharacterised in *D. melanogaster* and other Lepidoptera species (similarity assessed by BLASTp). A gene in the vicinity, g10223, is identified as a Cytochrome b-c1 complex subunit Rieske (*HMEL010374g1* in *H. melpomene*), which could be a better candidate. However, further work is necessary to explore the association peak and its candidate genes.

Variation in forewing phenotype in *H. misippus* is controlled by a major effect Mendelian locus, the M locus (Smith, 1976; VanKuren et al., 2019)(Chapter 4). The dominant allele produces the black-and-white *misippus* wings, while recessive homozygotes can have either the all-orange *inaria* wings and *immima* intermediates. The switch between *inaria* and *immima* wings has been hypothesised to be caused by epistasis with the A locus, one of the major effect loci contributing to hindwing variation. To identify the locus controlling the switch between *inaria* and *immima* morphs, I performed a GWAS but found no genomic regions associated with the two morphs (Supplementary Figure 4). Furthermore, to explore whether the genomic region associated with hindwing variation has an epistatic effect of forewing phenotype as hypothesised for the A locus, I analysed the relationship between haplotypes at that region. Using phylogenetic trees of varying window size, I found no structure among the haplotypes distinguishing the *inaria* and *immima* morphs, which suggest that the region might not be involved in the control of forewing phenotype or that the differences in the alleles are small. Alternatively, it could be that the low coverage (average of 1X) of the samples impedes any meaningful phylogenetic analysis of the alleles, even with imputation and phasing aided by long-read information.

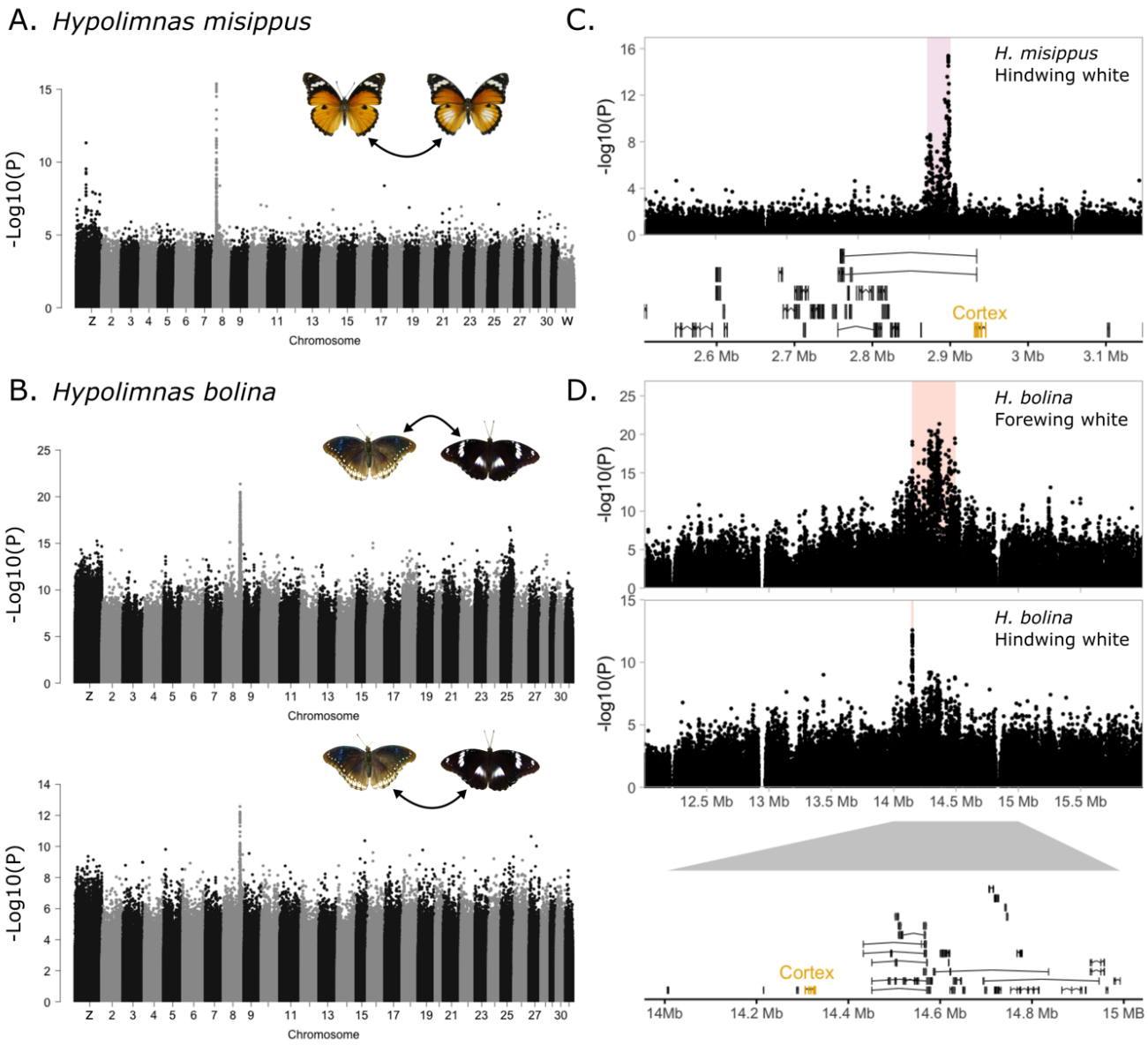


Figure 5.1 | Variation near *Cortex* is associated with white coloration in *Hypolimnas misippus* and *Hypolimnas bolina.* **A.** GWAS for the presence/absence of a hindwing white spot in *H. misippus* reveals an association peak on chromosome 8. **B.** Similarly, GWAS for the presence/absence of a hindwing white spot and a forewing subapical band, both reveal an association peak in the same region of chromosome 8. **C.** Zooming into the associated region in *H. misippues* shows that the gene *cortex* (negative strand) is just downstream of the association peak. **D.** Similarly, zooming into the two association peaks for hindwing and forewing white in *H. bolina* shows that *cortex* (positive strand) is found by the association peak.

1.26.3 No large structural variants are present at the associated locus in *H. misippus*

In some cases in which *cortex* has been associated with wing phenotype morphs, structural variants such as inversions have been shown to be present. For example, multiple inversions around (and including) *cortex* are involved in morph diversity in *Heliconius numata* and in oakleaf butterflies of the *Kallima* genus (Joron et al., 2011; Wang et al., 2022). I examined whether there was any structural variation associated with hindwing white using Wrath (Chapter 4). I analysed barcode sharing between genomic windows and generated heatmaps to explore the presence of SVs in the dataset. First, I used a small window size (100 bp) to explore the region along and around the associated locus and did not detect any signs of structural variants present in the data (Supplementary Figure 5). Using a small window size allows for a fine-grained detection of SVs, but computational requirements limit the application to only small genomic regions. Then, I analysed the whole chromosome 8 with a larger window size, with which I would be able to detect larger SVs in the dataset, and also did not detect any SVs near *cortex* (Supplementary Figure 5). The number of samples used of each phenotype, 104 of individuals with white hindwings and 228 with orange ones, should be enough to detect any SVs present at the associated locus as shown for the M locus in Chapter 4. However, an insertion in the non-reference allele cannot be ruled out, as SVs can only be detected through the mapping signatures present in the reference genome.

1.26.4 The elements controlling white colouration in *Hypolimnas*

Next, I explored the evolutionary history of hindwing colour in the two species. More specifically, I wanted to test whether haplotypes associated with hindwing colouration in *H. bolina* and *H. misippus* were homologous. I first defined regions around the top associated SNPs in each species for hindwing and forewing white (Figure 5.1C-D). In *H. bolina*, I defined a region of 343,476 bp (Chromosome 8:14,149,517-14,492,993) for the forewing white association and a region of 6,149 bp (Chromosome 8:14,145,391-14,151,540) for hindwing white, while for *H. misippus* the region defined was of 33,000 bp (Chromosome 8:2,897,000-2,930,000) in the hindwing white GWAS. Then, to infer sequence homology, I aligned the two reference genomes using Satsuma2 and looked for alignment tracks overlapping the defined regions of association (Figure 5.2A). First, the regions of highest association with hindwing

white did not seem to be homologous based on the alignment. That is, the region containing the top associated SNPs in *H. bolina* maps to a region of chromosome 8 ~125 kbp away from the region of top associated SNPs in *H. misippus.* Contrastingly, the region of highest association with forewing white in *H. bolina*, overlaps with the region of top associated SNPs with hindwing white in *H. misippus* (Figure 5.2A-B). However, homology of the two regions cannot be determined as the region associated with forewing white in *H. bolina* is too broad (~343 kb, Figure 5.2B). Next, I mapped the *H. bolina* sequence reads to the *H. misippus* reference genome and carried out the GWAS again. This showed that the three regions of top associated SNPs are not identical in the *H. misippus* reference genome, but the association peak for *H. bolina* forewing variation is broad and overlaps the region of association in *H. misippus.* Interestingly, the associated loci for hindwing white in both species are situated at opposite sides of cortex: the *H. misippus* associated locus is downstream of *cortex* (3' end), while the *H. bolina* one is upstream (5') of *cortex* (Figure 51C-D).

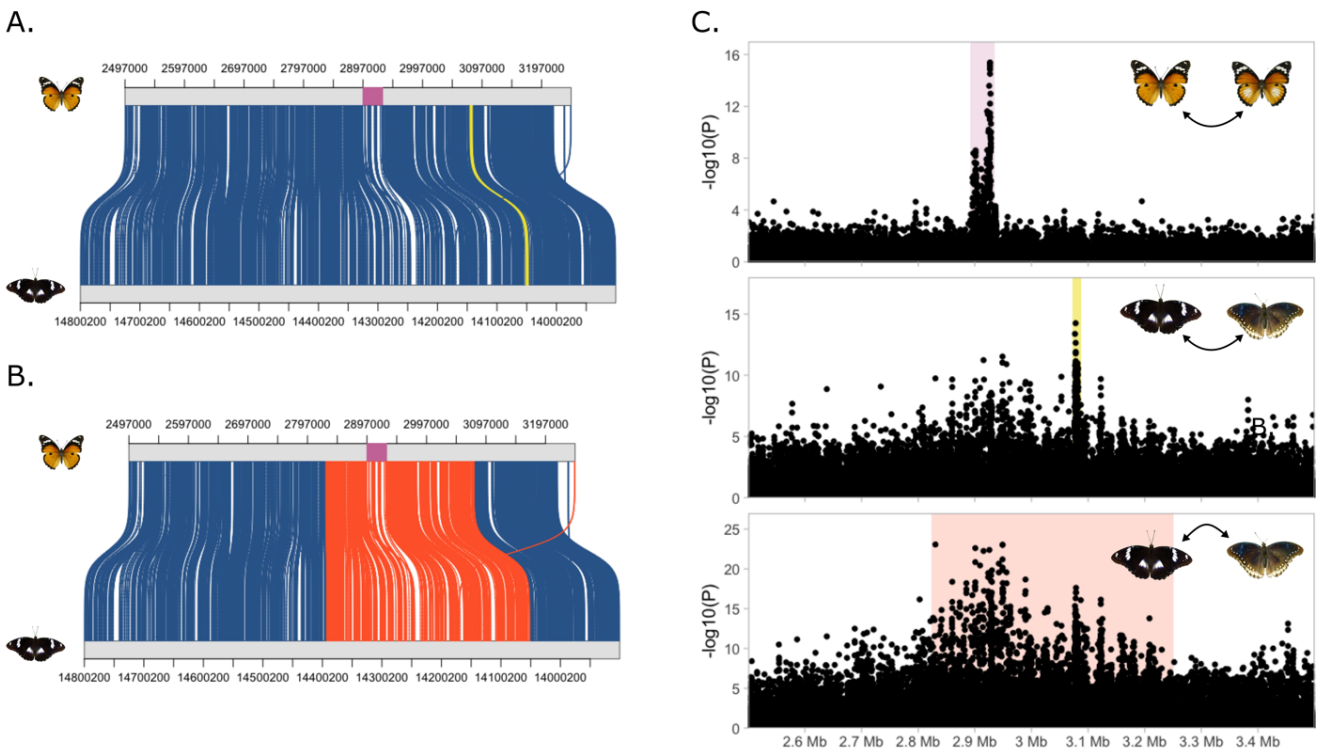


Figure 5.2 | Cis-regulatory elements (CREs) associated hindwing white are not homologous between *H. bolina* and *H. misippus.* **A.** Genome alignments between *H. misippus* and *H. bolina* show that the CRE associated with hindwing white in *H. bolina* (yellow) does not overlap with the locus associated with hindwing white in *H. misippus* (pink). **B.** Contrastingly, the region associated with forewing white in *H. bolina* (orange) overlaps with the locus associated with forewing white in *H. misippus*, suggesting possible homology. **C.** Mapping to the *H. misippus* reference genome (*HypMisi_v2)* and generating a GWAS for hindwing and forewing white in the two species shows that the loci associated with hindwing white in the two species are not overlapping, while the locus associated with forewing white in *H. bolina* is broad and overlaps the locus associated with hindwing white in *H. misippus.*

1.26.5 Two transposable element insertions at the candidate locus in *H. misippus* are associated with continuous variation in hindwing white

Transposable element insertions at the cis-regulatory region of *cortex* have been shown to be involved in wing pattern variation in *Heliconius* butterflies (Livraghi et al., 2021) and in the peppered moth *Biston betularia* (van't Hof et al., 2016). Crucially, I have found that TE insertions are found in the M locus determining forewing phenotype in *H. misippus* (Chapter 4). To explore whether TE insertions are involved in hindwing colour variation in *Hypolimnas*, I calculated read depth at the region of association using BAM files and pooling *H. misippus* individuals by hindwing phenotype. I observed two regions with differential read depth between pooled individuals with all-brown hindwings and those with white spots, I named those insertion A (downstream) and insertion B (upstream; Figure 5.3A). Insertion A contains an insertion of a LINE TE, while Insertion B contains three consecutive TE insertions, two Helitrons and an unknown TE.

Then, I determined the insertion genotype of each of the individuals by using a SNP in linkage with each insertion, I identified two SNPs each associated with one of the insertions. Insertion A was associated with a SNP at 2,923,994 on chromosome 8, which had two alleles A (n=238) and G (n=130). All reads covering the SNP site and across insertion breakpoint (HiC_scaffold_29:2,924,003) had the A allele (n=231), while all reads presenting a G at the SNP site mapped only until the breakpoint, except for one read carrying a G and covering the first nucleotide of the insertion. The SNP at 2,927,362 on chromosome 8 presented two alleles G (n=386) and T (n=118). 99.2% (n=256) of reads covering the SNP site and the breakpoint presented a G, while the 0.8% (n=2) presented a T. 98.3% of reads carrying a T at the SNP site did not cover the insertion breakpoint, while 66.3% of reads carrying a G covered the breakpoint.

Once I had determined the insertion genotype of the individuals, I quantified the association of the insertions with continuous variation in colouration in the hindwing (from orange to having a large white spot). I observed that individuals that are homozygous for either of the insertions have wings that are orange or have a reduced white spot than individuals not carrying them, while heterozygotes have intermediate phenotypes when analysing each insertion separately (ANOVA 2 d.f. X2927362 p-value = 1.234e-07 and X2923994 p value =

2.016e-08; Figure 5.3B). This inheritance fits with Gordon and Smiths' hypothesis for the A/S loci, as both loci were hypothesised to have incomplete dominance and variation in their penetrance (Gordon and Smith, 1989). I then quantified the association of hindwing phenotype with both insertions combined and observed that individuals that are homozygous for at least one of the insertions have darker wings (i.e. have a reduced or absent white spot) than those being homozygous for the deletion in at least one of the sites, while double heterozygotes presented intermediate phenotypes (ANOVA 4 d.f. p-value 6.185e-09; Figure 5.3B). Interestingly, out of the 128 individuals with genotypes at both diagnostic sites, none of them was homozygote for the deletion at one site and homozygote for the insertion at the other.

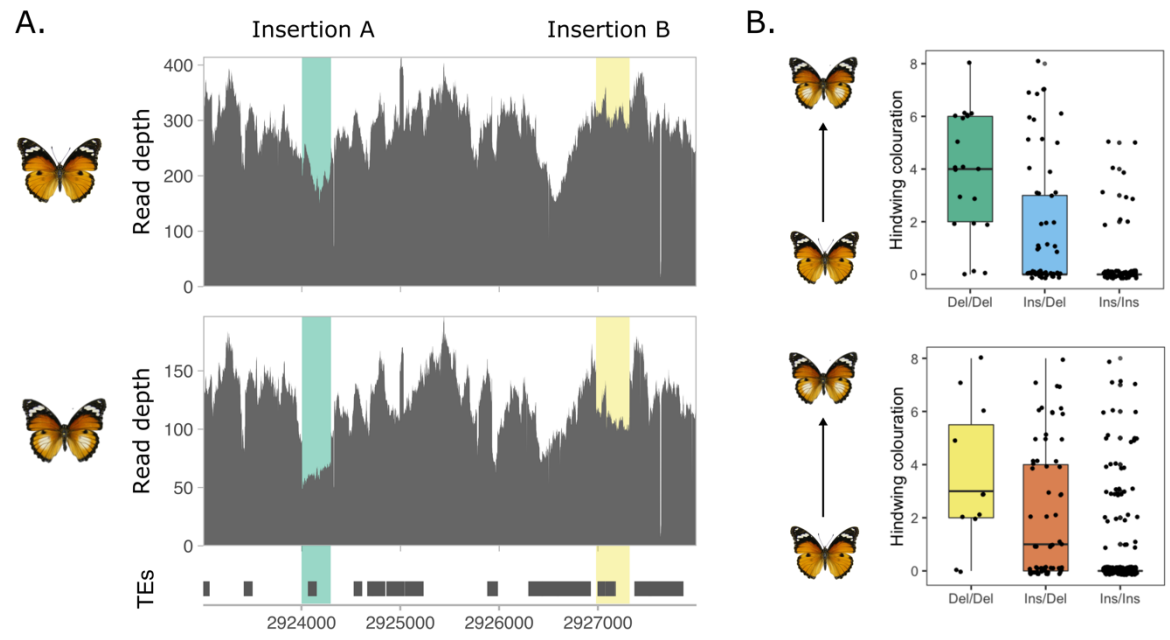


Figure 5.3 | Two insertions are associated with continuous variation in hindwing white in *H. misippus*. **A.** Read coverage at chromosome 8 shows two distinct regions of lower coverage in individuals with a hindwing white spot (middle track) compared to individuals with orange hindwings (top track). Transposable elements are found in both insertions (bottom track). Insertion A is indicated in green and insertion B in yellow. **B.** Continuous variation in hindwing white is associated with the genotype of the insertions, X2923994 (top) and X2927362 (bottom) where homozygous individuals for the insertion (Ins/Ins) have a smaller or absent white spot in the hindwing, homozygous individuals without the insertion (Del/Del) have larger white spots and heterozygotes are intermediates.

1.26.6 GWAS for orange colouration in *H. bolina* points to *optix* as the main candidate

Wing pattern variation in *H. bolina* also includes variation in orange colouration in the forewing (Figure 5.4C). The *nerina* morphs have an orange spot in the forewing, while the other two morphs, *euploeoides* and *naresi,* do not (Figure 5.4C)(Clarke and Sheppard, 1975). This orange element is genetically determined by the N locus, whose dominant allele produces the *nerina* phenotype. Combinations between the two loci determining wing pattern in *H. bolina* generate all possible phenotypes, out of which *E_nn* produce *euploeoides,* the only mimetic form, *eenn* produce *naresi*, *eeN_* generate *nerina* and the rest are intermediates. To identify the N locus, I performed a GWAS on the wild *H. bolina* sequenced samples. This analysis revealed a significant association peak in chromosome 14 (Figure 5.4A). From the genes around the associated region, one clear candidate stood out, the gene *optix* (Figure 4B). *optix* has been linked to orange and red colouration in multiple species of butterflies including some *Heliconius* species and *Vanessa cardui* using association studies and functional testing with CRISPR, and thus is a strong candidate for the control of this phenotype (Reed et al., 2011; Zhang et al., 2017b).

Finally, a fourth morph exists, *pallescens,* whose phenotype might be controlled by a different allele at the N locus or by a third distinct locus (Clarke and Sheppard, 1975). However, no individuals of this phenotype were sampled and thus no evaluation of it was possible.

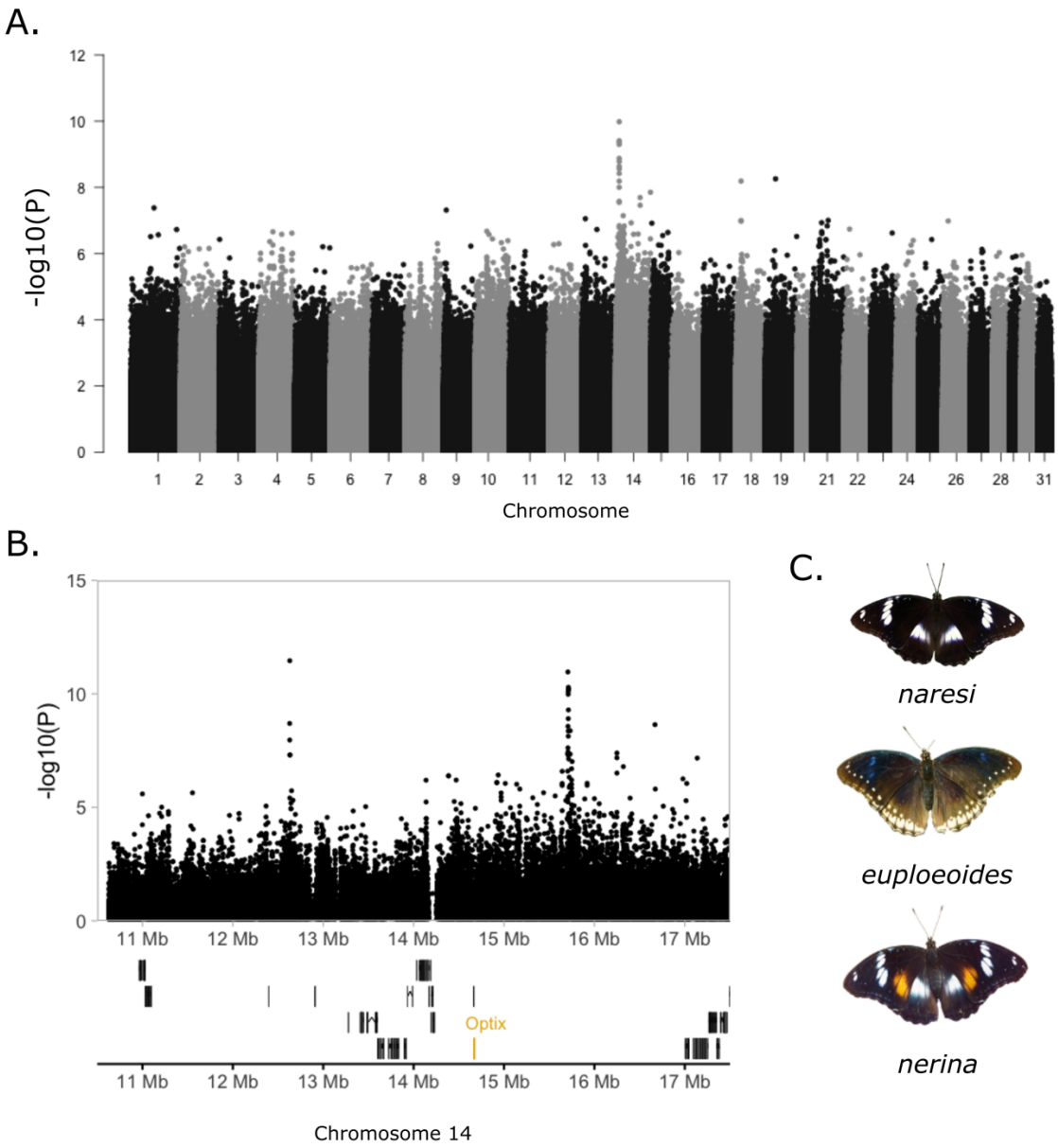


Figure 5.4 | Orange forewing colouration is associated with variation near *optix* in *H. bolina.* **A.** Genome wide association study of orange colouration shows the highest peak of association in chromosome 14. **B.** A zoom in of the associated region reveals two peaks of association in non-coding loci, between which the gene *optix* is found. **C.** The different morphs of *H. bolina: naresi, euploeoides* and *nerina.*

1.27 Discussion

Studies of convergent evolution have revealed the repeated use of the same 'hotspot genes' when dissimilar species evolve similar phenotypes. Here, I have used whole genome analysis of 645 individual butterflies to demonstrate that putative cis-regulatory regions near two well-known wing patterning genes, *cortex* and *optix,* are associated with differences between Batesian mimicry morphs in *Hypolimnas* butterflies. I have found that an intergenic region near *cortex* is associated with variation in white pattern elements in the forewing and hindwing of *H. bolina* and in *H. misippus* hindwings. The *cortex* gene has now been implicated in controlling crypsis, warning colour, Batesian and Mullerian mimicry patterns across the Lepidoptera, making this gene a genuine 'hotspot' for genetic change.

Cortex is in a family of cell cycle regulator genes and has been shown to determine scale identity, resulting in changes in pattern and colouration (Livraghi et al., 2021). It has been repeatedly linked to switches between melanic and white/yellow wing pattern elements in multiple Lepidoptera including *Heliconius* species and other butterflies and moths. For example, a single TE insertion in an intron of *cortex* has been shown to cause the switch between the peppered and melanic morphs of the peppered moth, *Biston betularia* (van't Hof et al., 2016), while in the Batesian mimic *Papilio clytia, cortex* has been associated with the differences between mimetic morphs; one with brown wings and reduced white elements in the apex mimicking *Euploea* models like *H. bolina,* and another with melanic black and pigmented white scales in a pattern resembling a toxic tiger butterflies (VanKuren et al., 2019). *Cortex* has also been linked to changes in colour phenotypes in other moths such as the silk moth *Bombyx mori* and some geometrids, and butterflies such as *Junonia coenia* and *Bicyclus anynana* (Beldade et al., 2009; Ito et al., 2016; van der Burg et al., 2020; van't Hof et al., 2019). Crucially, in some of these cases, cis-regulatory variation around *cortex* has been shown to be the cause of the phenotypic changes, with two cases in which TE insertions at regulatory regions have been implicated (Livraghi et al., 2021; van der Burg et al., 2020; van't Hof et al., 2016). In *Heliconius, cortex* controls the switch between type I scales, which can be either yellow or white, and type II scales, which can be black or red (Livraghi et al., 2021). The added effect of two other genes, *optix* and *aristaless1,* determines the final colouration of

each scale type (Reed et al., 2011; Westerman et al., 2018). Thus, *cortex* is a strong candidate for controlling development of white pattern elements in *Hypolimnas* species.

Alternatively, it is possible that other genes around the associated locus could have a role in wing phenotype in *H. bolina* and *H. misippus.* Crucially, the genes domeless and washout found next to *cortex* have been suggested to be involved in wing phenotype determination in *Heliconius* (Livraghi et al., 2021; Moest et al., 2020). Similarly, evidence from *H. numata,* in which wing polymorphism is controlled by a supergene around *cortex* containing three inversions, suggests that other genes around *cortex* also have a function in wing phenotype (Saenko et al., 2019).

The data in *H. bolina* suggest that there may be distinct cis-regulatory elements (CRE) controlling the presence of the forewing white band and the hindwing white spot. In reared families, I have shown that these elements can segregate independently, and they show distinct albeit partially overlapping association peaks. Strong linkage disequilibrium due to physical linkage and potentially also recent selection, could lead to longer haplotypes and a broad association peak that could contribute to this overlap. Interestingly, it seems likely that the CREs causing differences in white colouration in the hindwing are not homologous between the two species, as they map to slightly different locations. These results highlight the complexity of the region around *cortex* and suggest that modular CREs have spatially restricted effects in *Hypolimnas bolina.* These results strongly parallel those of *Heliconius*, where evidence for recombination within the locus between different elements in the same species is also found (Ferguson et al., 2010), and adjacent but distinct CREs have been implicated in convergence of mimetic species (Livraghi et al., 2021) and modular variation at two narrow regions upstream and downstream of *cortex* have been associated with hindwing yellow pattern variation (Enciso-Romero et al., 2017).

Hindwing phenotype in *H. misippus* varies continuously, and wings can be completely orange or have a white spot with highly variable size. Two major effect loci, the A and S loci, have been hypothesised to control such variation. However, using GWAS, I find evidence of only one of those loci. This could be due to the samples present in the dataset. Furthermore, the A locus has been hypothesised to be a supergene with effects in hindwing colouration, forewing pattern (differences between *inaria* and *immima* morphs) and body size (Gordon

and Smith, 1989). However, I find no genomic region associated with the differences between *inaria* and *immima* morphs. Nonetheless, the GWAS for hindwing white in *H. misippus* reveals a second peak of association in the Z chromosome, which could be the second locus hypothesised (A/S).

Similar to *cortex*, cis-regulatory variation around the transcription factor *optix* has been associated with colour pattern differences in *Heliconius* (Martin et al., 2014; Reed et al., 2011; Wallbank et al., 2016). CRISPR-cas9 knock-outs have shown that *optix* has an effect on structural colouration and on the red and orange pattern differences in *Heliconius* species as well as in *Junonia coenia*, *Agraulis vanilla* and *Vanessa cardui* (Zhang et al., 2017b). This highlights the widespread importance of *cortex* and *optix* in the evolution of wing pattern across diverse Lepidoptera species.

In 1945, Goldschmidt proposed that mimicry could be favoured by shared developmental systems, in which single mutations could activate ancestral developmental pathways to create the same phenotype in model and mimic (Goldschmidt, 1945). In *Heliconius,* it has been shown that convergent mimicry does indeed result from allelic variation at the same few hotspot loci (Livraghi et al., 2021; Mazo-Vargas et al., 2017; Reed et al., 2011). However, as the convergence is between species in the same genus, perhaps it is not surprising that these species show a similar developmental basis. *Hypolimnas* offers an opportunity to study Batesian mimicry between far more distantly related species. Unlike in Müllerian systems, Batesian mimicry has a clear model and mimic, so there is a clear hypothesis for the order in which evolutionary divergence has occurred. The fact that *H. misippus* mimics the four morphs of *D. chrysippus* makes it an ideal system to investigate if convergence in phenotype results from similar molecular changes in more distantly related species (84 MYA, (Kumar et al., 2022)).

Crucially and similarly to Goldschmidt, Bernardi (Bernardi, 1974) and Pierre (Pierre, 1980) suggested that mimetic patterns of female *H. misippus* are ancestral and the male pattern derived and propose that the female phenotypes of *H. misippus* are homologous to those of *D. chrysippus,* dating back to a time of common ancestry. The identification of the loci controlling wing phenotypes in both species is crucial to shed light to these hypotheses. In *Danaus chrysippus,* wing phenotype is controlled by three main loci A, B and C (Smith, 1975;

Smith et al., 1993). Whilst the B and C loci control forewing phenotype and are part of a supergene found in chromosome 15, the A locus controlling hindwing variation is found in chromosome 4 (Martin et al., 2020). Crucially, none of these loci are close to *cortex* (chromosome 8) or *optix* (chromosome 14). This contrasts with my results showing that hindwing variation in white colouration is likely controlled by *cortex* at chromosome 8 and possibly by a locus at the Z chromosome, and that forewing mimicry is associated with a locus at chromosome 29 (Chapter 4). Overall, these results indicate that the convergence in wing phenotype seen in *H. misippus* to mimic *D. chrysippus* does not have an homologous genetic basis and oppose the ideas of Goldschmidt, Bernardi and Pierre (Goldschmidt, 1945)

The putative ancestral phenotype of *Hypolimnas* butterflies has been suggested to have black wings with white elements similar to those seen in the *naresi* morph of *H. bolina* and *H. misippus* and *H. bolina* males, as these are common among *Hypolimnas* species (Swinhoe, 1896). Given that the same locus at chromosome 8 seems to be controlling the presence of white pattern elements in *H. bolina* and *H. misippus* at least in the hindwing, it could be that this control is ancestral in the genus, with *cortex* having a common function. Furthermore, while hindwing white colouration has a clear mimetic benefit in *H. misippus*, that is not the case in *H. bolina*, in which the only morph, *euploeoides*, does not present hindwing white colouration or a subapical forewing band. This could suggest that the mimetic morph is derived in *H. bolina* and that the reduction of hindwing white could have evolved independently in the two species.

Taken together, my results highlight the importance of 'genetic hotspots' in the evolution wing phenotypes and particularly of *cortex* and *optix* in determining wing pattern in diverse Lepidoptera species. More generally, they add to the evidence showing that convergent phenotypes are often the result of repeated evolution at the genetic level (i.e., genetic parallelism or convergence)(Conte et al., 2012; Stern, 2013). Exploring the genetic basis of other adaptive phenotypes in diverse clades is necessary to clarify if this genetic reuse is as generalised as the current evidence suggests.

1.28 Supplementary Information

Supplementary Table 5.1 | Chromosome homology between *H. misippus* and *H. bolina* chromosomes and *M. cinxia.*

M. cinxia	***H. bolina***	***H. misippus***
1	31	1
2	13	18
3	2	32
4	8	27
5	7	30
6	5	17
7	14	19
8	11	29
9	10	24
10	1	23
11	4	12
12	15	25
13	12	6
14	6	28
15	3	10
16	18	11
17	16	16
18	20	15
19	9	13
20	21	3
21	19	31
22	22	26
23	23	9
24	25	5
25	24	4
26	17	14
27	26	22
28	27	8
29	28	2
30	29	21
31	30	20

Supplementary Table 5.2 | Phenotypes and origin of reared *H. bolina* samples and samples collected in Samoa in 2010 and 2001 used for QTL mapping and preliminary GWAS.

Sample	Hindwing white 0 = no white; 1 = white patch	Forewing Orange 0 = no orange; 1 = orange patch	Family
CAM035065	0	1	Family 1
CAM035066	0	1	Family 1
CAM035067	0	1	Family 1
CAM035068	1	1	Family 1
CAM035069	1	1	Family 1
CAM035070	0	1	Family 1
CAM035071	0	1	Family 1
CAM035072	0	1	Family 1
CAM035073	0	1	Family 1
CAM035074	1	1	Family 1
CAM035075	0	1	Family 1
CAM035076	1	1	Family 1
CAM035077	0	1	Family 1
CAM035078	0	1	Family 1
CAM035079	0	1	Family 1
CAM035080	1	1	Family 1
CAM035081	0	1	Family 1
CAM035082	0	1	Family 1
CAM035083	0	1	Family 1
CAM035084	1	1	Family 1
CAM035085	0	1	Family 1
CAM035086	1	1	Family 1
CAM035087	1	1	Family 1
CAM035088	1	1	Family 1
CAM035089	0	1	Family 1
CAM035090	1	1	Family 1
CAM035091	0	1	Family 1
CAM035092	0	1	Family 1
CAM035093	0	1	Family 1
CAM035094	0	1	Family 1
CAM035095	1	1	Family 1
CAM035096	1	1	Family 1
CAM035097	1	0	Family 2
CAM035098	1	0	Family 2
CAM035099	0	0	Family 2
CAM035100	0	0	Family 2
CAM035101	1	0	Family 2
CAM035102	0	0	Family 2
CAM035103	1	0	Family 2
CAM035104	1	0	Family 2
CAM035105	0	0	Family 2
CAM035106	1	0	Family 2
CAM035107	0	0	Family 2
CAM035108	1	0	Family 2
CAM035109	1	0	Family 2
CAM035110	0	0	Family 2
CAM035111	0	0	Family 2
CAM035112	1	0	Family 2
CAM035113	1	0	Family 2

CAM035114	1	0	Family 2
CAM035115	1	0	Family 2
CAM035116	0	0	Family 2
CAM035117	0	0	Family 2
CAM035118	0	0	Family 2
CAM035119	1	0	Family 2
CAM035120	0	0	Family 2
CAM035121	0	0	Family 2
CAM035122	0	0	Family 2
CAM035123	0	0	Family 2
CAM035124	0	0	Family 2
CAM035125	1	0	Family 2
CAM035126	1	0	Family 2
CAM035127	0	0	Family 2
CAM035128	1	0	Family 2
CAM035129	0	0	Family 2
CAM035130	1	0	Family 2
CAM035131	1	0	Family 2
CAM035132	0	0	Family 2
CAM035133	0	0	Family 2
CAM035134	0	0	Family 2
CAM035135	0	0	Family 2
CAM035136	0	0	Family 2
CAM035137	1	0	Family 2
CAM035138	0	0	Family 2
CAM035139	0	0	Family 2
CAM035140	0	0	Family 2
CAM035141	0	0	Family 2
CAM035142	0	0	Family 2
CAM035186	0	0	Mother Family 1
CAM035187	NA	NA	Father Family 2
CAM035188	NA	NA	Father Family 1
CAM035189	0	0	Mother Family 2
Sample_54-SAM1001	0	1	Samoa2010
Sample_55-SAM1002	1	1	Samoa2010
Sample_58-SAM1005	0	1	Samoa2010
Sample_60-SAM1007	0	1	Samoa2010
Sample_61-SAM1008	1	1	Samoa2010
Sample_63-SAM1010	1	1	Samoa2010
Sample_70-SAM1017	0	1	Samoa2010
Sample_72-SAM1019	1	1	Samoa2010
Sample_74-SAM1021	1	1	Samoa2010
Sample_77-SAM1024	0	1	Samoa2010
Sample_79-SAM1026	1	1	Samoa2010
Sample_85-SAM1034	0	1	Samoa2010
Sample_88-SAM1037	1	1	Samoa2010
Sample_96-SAM1045	0	1	Samoa2010
Sample_10-SAM10	1	1	Samoa2001
Sample_11-SAM11	1	1	Samoa2001
Sample_12-SAM13	1	1	Samoa2001
Sample_13-SAM14	1	1	Samoa2001
Sample_14-SAM15	1	0	Samoa2001
Sample_15-SAM16	1	1	Samoa2001
Sample_16-SAM17	1	1	Samoa2001
Sample_17-SAM19	1	0	Samoa2001
Sample_18-SAM20	1	1	Samoa2001
Sample_1-SAM1	1	0	Samoa2001

Sample_22-SAM24	1	1	Samoa2001
Sample_2-SAM2	1	0	Samoa2001
Sample_39-SAM41	1	1	Samoa2001
Sample_3-SAM3	1	1	Samoa2001
Sample_40-SAM43	1	1	Samoa2001
Sample_41-SAM44	1	1	Samoa2001
Sample_42-SAM45	1	0	Samoa2001
Sample_43-SAM46	1	1	Samoa2001
Sample_44-SAM48	1	0	Samoa2001
Sample_45-SAM49	1	1	Samoa2001
Sample_46-SAM50	1	1	Samoa2001
Sample_47-SAM51	1	1	Samoa2001
Sample_48-SAM52	1	1	Samoa2001
Sample_4-SAM4	1	1	Samoa2001
Sample_50-SAM54	1	0	Samoa2001
Sample_51-SAM55	1	1	Samoa2001
Sample_52-SAM56	1	0	Samoa2001
Sample_5-SAM5	1	1	Samoa2001
Sample_6-SAM6	1	0	Samoa2001
Sample_8-SAM8	1	1	Samoa2001
Sample_9-SAM9	1	1	Samoa2001

Supplementary Table 5.3 | Phenotypes and origin of reared and wild *H. bolina* samples used for GWAS of hindwing and forewing white and forewing orange pattern elements.

Sample	Origin	Year	Forewing white	Forewing orange	Hindwing white	Reared or wild
CAM035507	Moorea	2005	0	0	1	offspring of FP17MO05-F1-02
CAM035508	Moorea	2005	0	0	1	offspring of FP17MO05-F1-02
CAM035509	Moorea	2005	1	1	1	offspring of FP17MO05-F1-02
CAM035510	Moorea	2005	0	0	1	offspring of FP17MO05-F1-02
CAM035511	Moorea	2005	0	1	1	offspring of FP17MO05-F1-02
CAM035512	Moorea	2005	0	0	1	offspring of FP17MO05-F1-02
CAM035717	Moorea	2005	0	1	1	offspring of FP17MO05-F1-02
CAM035513	Moorea	2005	0	1	1	offspring of FP17MO05-F1-02
CAM035514	Moorea	2005	0	0	1	offspring of FP17MO05-F1-02
CAM035515	Moorea	2005	0	1	1	offspring of FP17MO05-F1-02
CAM035516	Moorea	2005	0	1	1	offspring of FP17MO05-F1-02
CAM035517	Moorea	2005	0	0	1	offspring of FP17MO05-F1-02
CAM035518	Moorea	2005	1	0	1	offspring of FP17MO05-F1-02
CAM035519	Moorea	2010	0	0	1	offspring of FP17MO10-FI-001
CAM035520	Moorea	2010	0	1	1	offspring of FP17MO10-FI-001
CAM035521	Moorea	2010	0	0	1	offspring of FP17MO10-FI-001
CAM035522	Moorea	2010	0	0	1	offspring of FP17MO10-FI-001
CAM035523	Moorea	2010	0	1	1	offspring of FP17MO10-FI-001
CAM035524	Moorea	2010	0	0	1	offspring of FP17MO10-FI-001
CAM035525	Moorea	2010	0	0	1	offspring of FP17MO10-FI-001
CAM035716	Moorea	2010	0	0	1	Wild
CAM035526	Moorea	2010	0	0	1	Wild
CAM035527	Moorea	2010	0	0	1	Wild
CAM035528	Moorea	2005	1	0	1	Offspring of FP17MO05
CAM035529	Moorea	2005	1	1	1	Offspring of FP17MO05
CAM035530	Moorea	2010	1	0	0	offspring of FP17MO10 - parent of family 2017MO1001
CAM035531	Moorea	2010	1	0	0	offspring of FP17MO10
CAM035532	Moorea	2012	1	1	1	offspring of FP17MO12
CAM035533	Australia	2018	1	1	0	female parent of family 2018-fam8
CAM035534	Australia	2018	1	1	0	female parent of family 2018-fam10
CAM035535	Australia	2018	1	1	0	female parent of family 2018-fam12
CAM035536	Australia	2018	0	1	1	Wild
CAM035537	Australia	2018	0	0	1	male parent of family 2018-fam13
CAM035538	Australia	2018	1	1	0	Hb-2018-fam14-05
CAM035539	Australia	2018	1	1	0	parent of 2018-fam19
CAM035540	Australia	2018	1	1	0	Hb-2018-fam14-09
CAM035541	Australia	2018	1	1	0	Hb-2018-fam14-13
CAM035542	Australia	2018	1	1	0	Hb-2018-fam14-14
CAM035543	Australia	2018	1	1	0	Hb-2018-fam14-17
CAM035544	Australia	2018	1	0	0	Hb-2018-fam15-01
CAM035545	Australia	2018	1	1	0	parent of 2018-fam14
CAM035546	Australia	2018	1	1	0	parent of 2018-family 15
CAM035547	Samoa	2018	0	0	0	parent of family 2018-fam1
CAM035548	Samoa	2018	1	0	0	parent of family 2018-fam2
CAM035549	Samoa	2018	1	0	0	parent of family 2018-fam3
CAM035550	Samoa	2018	0	0	0	parent of family 2018-fam5
CAM035551	Samoa	2018	1	0	0	parent of family 2018-fam6
CAM035552	Samoa	2018	0	0	0	parent of family 2018-fam7
CAM035553	Samoa	2018	0	0	1	parent of family 2018-fam9

CAM035554	Samoa	2018	1	0	0	parent of family 2018-fam11
CAM035715	Ruturu	2004	1	0	1	Wild
CAM035555	Ruturu	2004	1	1	0	Wild
CAM035556	Ruturu	2004	1	1	0	Wild
CAM035557	Ruturu	2004	1	1	0	Wild
CAM035558	Ruturu	2004	1	1	0	Wild
CAM035559	Ruturu	2004	1	0	0	Wild
CAM035560	Ruturu	2004	1	1	0	Wild
CAM035561	Ruturu	2004	1	0	0	Wild
CAM035562	Ruturu	2004	1	0	0	Wild
CAM035563	Ruturu	2004	1	0	0	Wild
CAM035564	Ruturu	2004	1	1	0	Wild
CAM035565	Ruturu	2004	0	0	1	Wild
CAM035566	Ruturu	2004	1	0	0	Wild
CAM035567	Ruturu	2004	1	0	1	Wild
CAM035568	Ruturu	2004	1	1	0	Wild
CAM035569	Ruturu	2004	1	1	0	Wild
CAM035570	Ruturu	2004	1	1	0	Wild
CAM035571	Ruturu	2004	1	0	1	Wild
CAM035572	Ruturu	2004	0	1	1	Wild
CAM035573	Ruturu	2004	1	1	1	Wild
CAM035574	Ruturu	2004	1	1	0	Wild
CAM035575	Ruturu	2004	1	1	0	Wild
CAM035576	Ruturu	2004	1	1	0	Wild
CAM035577	Ruturu	2004	1	0	0	Wild
CAM035578	Ruturu	2004	1	1	1	Wild
CAM035579	Ruturu	2004	1	0	0	Wild
CAM035580	Ruturu	2004	1	1	1	Wild
CAM035581	Ruturu	2004	1	1	1	Wild
CAM035582	Ruturu	2004	1	1	0	Wild
CAM035583	Ruturu	2004	1	1	0	Wild
CAM035584	Ruturu	2004	0	0	1	Wild
CAM035585	Ruturu	2005	0	0	0	Wild
CAM035586	Ruturu	2005	1	1	0	Wild
CAM035587	Ruturu	2005	1	1	1	Wild
CAM035588	Ruturu	2005	1	1	1	Wild
CAM035589	Ruturu	2005	1	1	0	Wild
CAM035590	Ruturu	2005	1	1	0	Wild
CAM035591	Ruturu	2005	1	1	0	Wild
CAM035592	Ruturu	2005	1	1	1	Wild
CAM035593	Ruturu	2005	1	1	1	Wild
CAM035712	Ruturu	2005	1	0	1	Wild
CAM035594	Ruturu	2005	0	0	1	Wild
CAM035595	Ruturu	2005	1	1	1	Wild
CAM035596	Ruturu	2005	1	1	0	Wild
CAM035597	Ruturu	2005	1	1	1	Wild
CAM035598	Ruturu	2005	1	1	1	Wild
CAM035599	Ruturu	2005	1	1	1	Wild
CAM035720	Ruturu	2005	1	1	0	Wild
CAM035600	Ruturu	2005	1	0	0	Wild
CAM035601	Ruturu	2005	1	1	1	Wild
CAM035713	Ruturu	2005	1	1	0	Wild
CAM035602	Ruturu	2005	1	0	1	Wild
CAM035603	Ruturu	2005	1	0	1	Wild
CAM035604	Ruturu	2005	1	0	1	Wild
CAM035605	Ruturu	2005	1	0	1	Wild
CAM035606	Ruturu	2005	1	1	0	Wild

CAM035607	Ruturu	2005	1	1	0	Wild
CAM035608	Ruturu	2005	1	1	0	Wild
CAM035609	Ruturu	2005	1	1	0	Wild
CAM035610	Ruturu	2005	1	1	0	Wild
CAM035611	Ruturu	2005	1	1	0	Wild
CAM035612	Ruturu	2005	1	1	0	Wild
CAM035613	Ruturu	2007	0	0	1	Wild
CAM035614	Ruturu	2007	1	0	0	Wild
CAM035615	Ruturu	2007	1	1	1	Wild
CAM035616	Ruturu	2007	1	1	1	Wild
CAM035617	Ruturu	2007	1	1	0	Wild
CAM035618	Ruturu	2007	1	0	0	Wild
CAM035619	Ruturu	2007	1	1	0	Wild
CAM035620	Ruturu	2007	1	1	0	Wild
CAM035621	Ruturu	2007	1	1	0	Wild
CAM035622	Ruturu	2007	1	1	0	Wild
CAM035623	Ruturu	2007	1	1	0	Wild
CAM035624	Ruturu	2007	1	1	0	Wild
CAM035625	Ruturu	2007	1	1	1	Wild
CAM035626	Ruturu	2007	1	0	1	Wild
CAM035627	Ruturu	2007	1	1	0	Wild
CAM035628	Ruturu	2007	1	0	1	Wild
CAM035629	Ruturu	2007	1	1	1	Wild
CAM035630	Ruturu	2007	1	1	0	Wild
CAM035631	Ruturu	2007	0	1	1	Wild
CAM035632	Ruturu	2007	1	0	1	Wild
CAM035633	Ruturu	2007	0	0	1	Wild
CAM035634	Ruturu	2007	1	1	0	Wild
CAM035635	Ruturu	2007	1	1	0	Wild
CAM035636	Ruturu	2007	1	0	1	Wild
CAM035637	Ruturu	2007	1	0	0	Wild
CAM035638	Ruturu	2007	1	0	1	Wild
CAM035639	Ruturu	2007	1	1	0	Wild
CAM035640	Ruturu	2007	1	1	1	Wild
CAM035641	Ruturu	2013	0	0	1	Wild
CAM035642	Ruturu	2013	1	1	0	Wild
CAM035643	Ruturu	2013	1	1	0	Wild
CAM035644	Ruturu	2013	1	1	0	Wild
CAM035645	Ruturu	2013	0	0	1	Wild
CAM035646	Ruturu	2013	1	0	0	Wild
CAM035647	Ruturu	2013	1	1	0	Wild
CAM035648	Ruturu	2013	1	1	0	Wild
CAM035649	Ruturu	2013	0	0	1	Wild
CAM035650	Ruturu	2013	0	0	1	Wild
CAM035651	Ruturu	2013	0	0	1	Wild
CAM035652	Ruturu	2013	1	1	0	Wild
CAM035653	Ruturu	2013	0	0	1	Wild
CAM035654	Ruturu	2013	1	0	0	Wild
CAM035655	Ruturu	2013	1	1	0	Wild
CAM035656	Ruturu	2013	1	1	1	Wild
CAM035657	Ruturu	2013	1	0	0	Wild
CAM035658	Ruturu	2013	1	1	0	Wild
CAM035659	Ruturu	2013	0	0	1	Wild
CAM035660	Ruturu	2013	1	1	0	Wild
CAM035661	Ruturu	2013	1	1	0	Wild
CAM035662	Ruturu	2013	1	1	0	Wild
CAM035663	Ruturu	2013	1	1	0	Wild

CAM035664	Ruturu	2013	1	0	1	Wild
CAM035665	Ruturu	2013	1	1	0	Wild
CAM035666	Ruturu	2013	1	1	0	Wild
CAM035667	Ruturu	2013	0	0	1	Wild
CAM035668	Ruturu	2013	1	1	0	Wild
CAM035669	Ruturu	2013	0	1	1	Wild
CAM035670	Ruturu	2013	1	1	0	Wild
CAM035671	Ruturu	2013	1	1	0	Wild
CAM035672	Ruturu	2013	1	1	0	Wild
CAM035673	Ruturu	2013	0	0	1	Wild
CAM035674	Ruturu	2013	1	1	0	Wild
CAM035675	Ruturu	2013	0	1	1	Wild
CAM035676	Ruturu	2013	1	1	0	Wild
CAM035677	Ruturu	2013	1	0	0	Wild
CAM035678	Ruturu	2013	1	1	0	Wild
CAM035679	Ruturu	2013	1	1	0	Wild
CAM035680	Ruturu	2013	1	1	0	Wild
CAM035714	Ruturu	2013	1	0	0	Wild
CAM035681	Ruturu	2013	1	1	0	Wild
CAM035682	Ruturu	2013	0	0	1	Wild
CAM035683	Ruturu	2013	1	1	0	Wild
CAM035684	Ruturu	2013	1	1	0	Wild
CAM035685	Ruturu	2013	1	0	1	Wild
CAM035686	Ruturu	2013	1	1	1	Wild
CAM035687	Ruturu	2013	1	1	0	Wild
CAM035688	Ruturu	2013	1	0	0	Wild
CAM035689	Ruturu	2013	0	1	1	Wild
CAM035690	Ruturu	2013	1	0	0	Wild
CAM035691	Ruturu	2013	1	1	0	Wild
CAM035692	Ruturu	2013	1	0	0	Wild
CAM035719	Ruturu	2013	1	1	0	Wild
CAM035718	Ruturu	2013	1	1	0	Wild
CAM035693	Ruturu	2013	1	1	0	Wild
CAM035694	Ruturu	2013	1	1	1	Wild
CAM035695	Ruturu	2013	1	1	0	Wild
CAM035696	Ruturu	2013	1	0	0	Wild
CAM035697	Ruturu	2013	1	1	1	Wild
CAM035698	Ruturu	2013	1	0	1	Wild
CAM035699	Ruturu	2013	1	1	1	Wild
CAM035700	Ruturu	2013	0	1	0	Wild
CAM035701	Ruturu	2013	1	1	0	Wild
CAM035702	Ruturu	2013	1	0	0	Wild
CAM035703	Ruturu	2013	0	1	1	Wild
CAM035704	Ruturu	2013	0	0	0	Wild
CAM035705	Ruturu	2013	1	1	0	Wild
CAM035706	Ruturu	2013	1	1	0	Wild
CAM035707	Ruturu	2013	1	1	0	Wild
CAM035708	Ruturu	2013	1	0	0	Wild
CAM035709	Ruturu	2013	0	0	1	Wild
CAM035710	Ruturu	2013	1	1	0	Wild
CAM035711	Ruturu	2013	1	1	0	Wild

Supplementary Table 5.4 | Phenotypes and origin of the wild *H. misippus* samples used for GWAS of hindwing and forewing white pattern elements. Genotype information at the two loci used to genotype indels is also included.

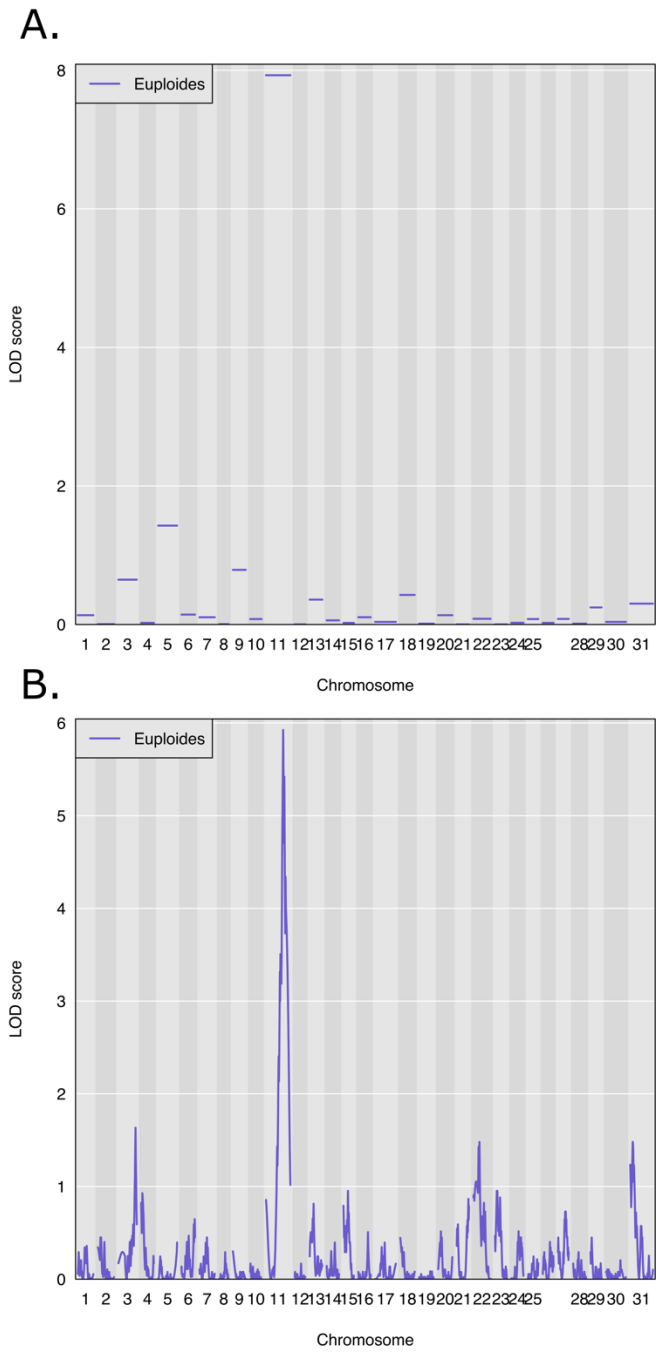
Sample	Genotype at Chr8:2927362	Genotype at Chr8:2923994	Forewing phenotype (i=*immima/inaria,* m=*misippus)*	Hindwing white (o=orange=1=aa, w=white=0=A)	Hindwing white continuous	Immima (1) or inaria (0)
CAM035929	G/G	A/A	i	o	0	1
CAM035418	G/G	A/A	i	o	0	0
CAM035232	G/G	A/A	i	o	0	1
CAM035422	G/G	A/A	i	o	0	1
CAM035443	G/G	A/A	i	o	0	1
CAM035505	G/G	A/A	i	o	0	0
CAM035253	G/G	A/A	i	o	0	1
CAM035275	G/G	A/A	i	o	0	1
CAM035285	G/G	A/A	i	o	0	1
CAM035287	G/G	A/A	i	o	0	1
CAM035328	G/G	A/A	i	o	0	1
CAM035145	G/G	A/A	m	o	0	NA
CAM035156	G/G	A/A	m	o	0	NA
CAM035161	G/G	A/A	m	o	0	NA
CAM035174	G/G	A/A	m	o	0	NA
CAM035205	G/G	A/A	m	o	0	NA
CAM035213	G/G	A/A	m	o	0	NA
CAM035216	G/G	A/A	m	o	0	NA
CAM035225	G/G	A/A	m	o	0	NA
CAM035233	G/G	A/A	m	o	0	NA
CAM035235	G/G	A/A	m	o	0	NA
CAM035241	G/G	A/A	m	o	0	NA
CAM035246	G/G	A/A	m	o	0	NA
CAM035268	G/G	A/A	m	o	0	NA
CAM035276	G/G	A/A	m	o	0	NA
CAM035303	G/G	A/A	m	o	0	NA
CAM035308	G/G	A/A	m	o	0	NA
CAM035330	G/G	A/A	m	o	0	NA
CAM035346	G/G	A/A	m	o	0	NA
CAM035348	G/G	A/A	m	o	0	NA
CAM035349	G/G	A/A	m	o	0	NA
CAM035352	G/G	A/A	m	o	0	NA
CAM035356	G/G	A/A	m	o	0	NA
CAM035365	G/G	A/A	m	o	0	NA
CAM035368	G/G	A/A	m	o	0	NA
CAM035384	G/G	A/A	m	o	0	NA
CAM035405	G/G	A/A	m	o	0	NA
CAM035419	G/G	A/A	m	o	0	NA
CAM035429	G/G	A/A	m	o	0	NA
CAM035431	G/G	A/A	m	o	0	NA
CAM035434	G/G	A/A	m	o	0	NA
CAM035436	G/G	A/A	m	o	0	NA
CAM035440	G/G	A/A	m	o	0	NA
CAM035442	G/G	A/A	m	o	0	NA
CAM035452	G/G	A/A	m	o	0	NA
CAM035454	G/G	A/A	m	o	0	NA
CAM035455	G/G	A/A	m	o	0	NA
CAM035460	G/G	A/A	m	o	0	NA
CAM035487	G/G	A/A	m	o	0	NA
CAM035501	G/G	A/A	m	o	0	NA
CAM035503	G/G	A/A	m	o	0	NA
CAM035247	G/G	A/A	m	w	0	NA
CAM035257	G/G	A/A	m	w	0	NA
CAM035269	G/G	A/A	m	w	4	NA
CAM035271	G/G	A/A	m	w	5	NA
CAM035292	G/G	A/A	m	w	3	NA
CAM035301	G/G	A/A	m	w	5	NA
CAM035307	G/G	A/A	m	w	0	NA
CAM035322	G/G	A/A	m	w	4	NA
CAM035334	G/G	A/A	m	w	3	NA
CAM035383	G/G	A/A	m	w	0	NA
CAM035389	G/G	A/A	m	w	2	NA
CAM035393	G/G	A/A	m	w	2	NA
CAM035445	G/G	A/A	m	w	0	NA
CAM035461	G/G	A/A	m	w	3	NA
CAM035488	G/G	A/A	m	w	2	NA

CAM035470	G/T	A/A	m	o	0	NA
CAM035181	N/N	A/A	i	o	0	0
CAM035339	N/N	A/A	m	o	0	NA
CAM035298	N/N	A/A	m	w	0	NA
CAM035219	T/G	A/A	m	o	0	NA
CAM035446	G/G	A/G	i	o	0	1
CAM035251	G/G	A/G	i	o	0	1
CAM035279	G/G	A/G	i	o	0	1
CAM035404	G/G	A/G	i	o	0	1
CAM035414	G/G	A/G	i	o	0	1
CAM035167	G/G	A/G	m	o	0	NA
CAM035211	G/G	A/G	m	o	0	NA
CAM035317	G/G	A/G	m	o	0	NA
CAM035423	G/G	A/G	m	o	0	NA
CAM035437	G/G	A/G	m	o	0	NA
CAM035481	G/G	A/G	m	o	0	NA
CAM035502	G/G	A/G	m	o	0	NA
CAM035217	G/G	A/G	m	w	3	NA
CAM035316	G/G	A/G	m	w	3	NA
CAM035441	G/G	A/G	m	w	1	NA
CAM035169	G/T	A/G	m	o	0	NA
CAM035286	G/T	A/G	m	o	0	NA
CAM035342	G/T	A/G	m	o	0	NA
CAM035347	G/T	A/G	m	o	0	NA
CAM035382	G/T	A/G	m	o	0	NA
CAM035310	G/T	A/G	i	w	1	1
CAM035492	G/T	A/G	i	w	3	1
CAM035350	G/T	A/G	i	w	2	1
CAM035265	G/T	A/G	m	w	7	NA
CAM035280	G/T	A/G	m	w	4	NA
CAM035332	G/T	A/G	m	w	1	NA
CAM035341	G/T	A/G	m	w	8	NA
CAM035358	N/N	A/G	m	o	0	NA
CAM035367	N/N	A/G	m	o	0	NA
CAM035411	N/N	A/G	m	o	0	NA
CAM035493	N/N	A/G	m	o	0	NA
CAM035379	N/N	A/G	i	w	4	1
CAM035306	N/N	A/G	i	w	0	1
CAM035324	N/N	A/G	i	w	5	1
CAM035240	N/N	A/G	m	w	7	NA
CAM035395	N/N	A/G	m	w	5	NA
CAM035408	N/N	A/G	m	w	1	NA
CAM035165	T/G	A/G	m	w	6	NA
CAM035451	T/G	A/G	m	w	1	NA
CAM035243	T/T	A/G	m	w	7	NA
CAM035148	G/G	G/A	i	o	0	0
CAM035417	G/G	G/A	i	o	0	1
CAM035333	G/G	G/A	m	o	0	NA
CAM035354	G/G	G/A	m	o	0	NA
CAM035302	G/G	G/A	i	w	6	1
CAM035313	G/G	G/A	m	w	0	NA
CAM035315	G/G	G/A	m	w	6	NA
CAM035337	G/G	G/A	m	w	7	NA
CAM035385	G/G	G/A	m	w	2	NA
CAM035327	G/T	G/A	m	o	3	NA
CAM035366	G/T	G/A	m	o	0	NA
CAM035338	G/T	G/A	m	w	5	NA
CAM035335	N/N	G/A	m	o	0	NA
CAM035430	N/N	G/A	m	w	0	NA
CAM035498	T/G	G/A	i	o	0	0
CAM035204	T/G	G/A	m	o	0	NA
CAM035370	T/G	G/A	m	w	1	NA
CAM035428	T/G	G/A	m	w	2	NA
CAM035344	G/T	G/G	m	w	4	NA
CAM035360	G/T	G/G	m	w	6	NA
CAM035387	G/T	G/G	m	w	6	NA
CAM035409	G/T	G/G	m	w	4	NA
CAM035506	G/T	G/G	m	w	6	NA
CAM035176	N/N	G/G	m	o	0	NA
CAM035293	N/N	G/G	i	w	2	1
CAM035238	N/N	G/G	m	w	6	NA
CAM035245	N/N	G/G	m	w	4	NA
CAM035433	N/N	G/G	m	w	4	NA
CAM035462	N/N	G/G	m	w	6	NA
CAM035453	T/G	G/G	i	o	0	0
CAM035151	T/T	G/G	m	o	0	NA

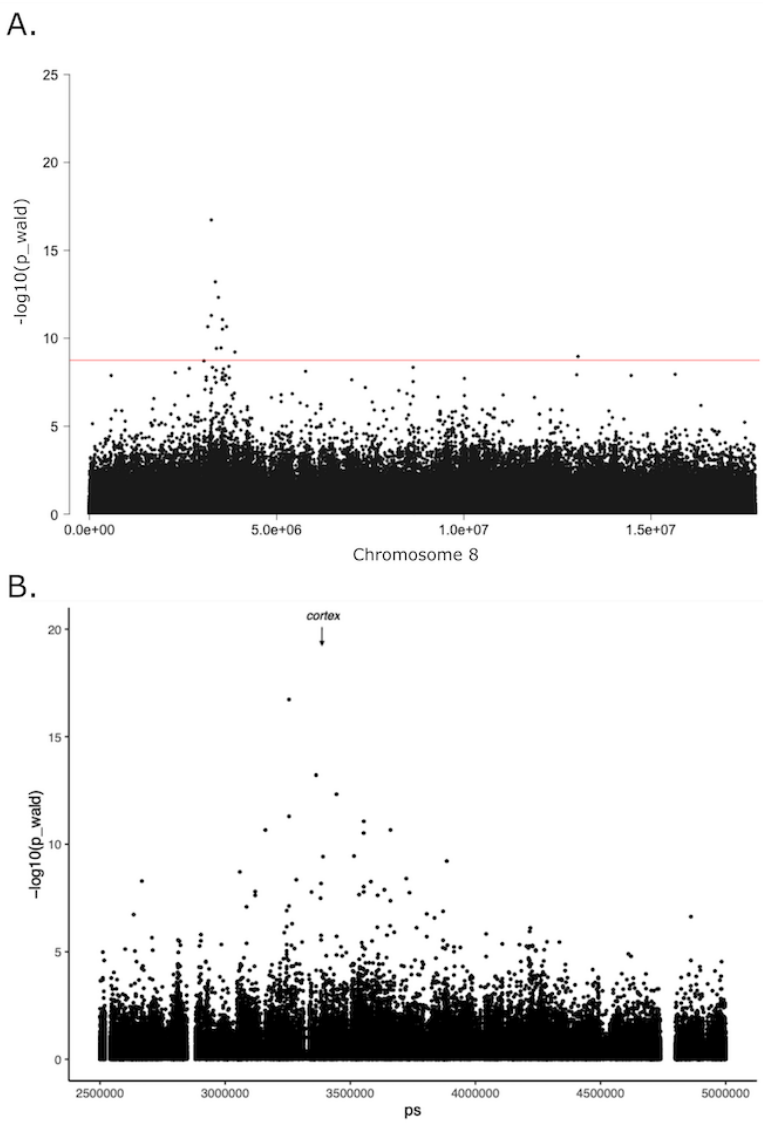
CAM035166	T/T	G/G	i	w	5	1
CAM035202	T/T	G/G	i	w	3	1
CAM035144	T/T	G/G	m	w	3	NA
CAM035378	T/T	G/G	m	w	2	NA
CAM035390	T/T	G/G	m	w	2	NA
CAM035396	T/T	G/G	m	w	6	NA
CAM035424	T/T	G/G	m	w	8	NA
CAM035264	G/G	N/N	i	o	0	0
CAM035439	G/G	N/N	i	o	0	0
CAM035489	G/G	N/N	i	o	0	0
CAM035496	G/G	N/N	i	o	0	0
CAM035157	G/G	N/N	i	o	0	1
CAM035203	G/G	N/N	i	o	0	0
CAM035474	G/G	N/N	i	o	0	1
CAM035155	G/G	N/N	i	o	0	1
CAM035239	G/G	N/N	i	o	0	0
CAM035252	G/G	N/N	i	o	0	1
CAM035254	G/G	N/N	i	o	0	1
CAM035274	G/G	N/N	i	o	0	1
CAM035325	G/G	N/N	i	o	0	1
CAM035386	G/G	N/N	i	o	0	1
CAM035407	G/G	N/N	i	o	0	1
CAM035146	G/G	N/N	m	o	0	NA
CAM035149	G/G	N/N	m	o	0	NA
CAM035154	G/G	N/N	m	o	0	NA
CAM035159	G/G	N/N	m	o	0	NA
CAM035179	G/G	N/N	m	o	0	NA
CAM035180	G/G	N/N	m	o	0	NA
CAM035182	G/G	N/N	m	o	0	NA
CAM035208	G/G	N/N	m	o	0	NA
CAM035209	G/G	N/N	m	o	0	NA
CAM035218	G/G	N/N	m	o	0	NA
CAM035223	G/G	N/N	m	o	0	NA
CAM035226	G/G	N/N	m	o	0	NA
CAM035227	G/G	N/N	m	o	0	NA
CAM035228	G/G	N/N	m	o	0	NA
CAM035230	G/G	N/N	m	o	0	NA
CAM035234	G/G	N/N	m	o	0	NA
CAM035249	G/G	N/N	m	o	0	NA
CAM035255	G/G	N/N	m	o	0	NA
CAM035258	G/G	N/N	m	o	0	NA
CAM035261	G/G	N/N	m	o	0	NA
CAM035270	G/G	N/N	m	o	0	NA
CAM035273	G/G	N/N	m	o	0	NA
CAM035277	G/G	N/N	m	o	0	NA
CAM035278	G/G	N/N	m	o	0	NA
CAM035281	G/G	N/N	m	o	0	NA
CAM035283	G/G	N/N	m	o	0	NA
CAM035284	G/G	N/N	m	o	0	NA
CAM035288	G/G	N/N	m	o	0	NA
CAM035314	G/G	N/N	m	o	0	NA
CAM035323	G/G	N/N	m	o	0	NA
CAM035329	G/G	N/N	m	o	0	NA
CAM035336	G/G	N/N	m	o	0	NA
CAM035345	G/G	N/N	m	o	0	NA
CAM035351	G/G	N/N	m	o	0	NA
CAM035353	G/G	N/N	m	o	0	NA
CAM035359	G/G	N/N	m	o	0	NA
CAM035361	G/G	N/N	m	o	0	NA
CAM035371	G/G	N/N	m	o	0	NA
CAM035380	G/G	N/N	m	o	0	NA
CAM035381	G/G	N/N	m	o	0	NA
CAM035391	G/G	N/N	m	o	0	NA
CAM035392	G/G	N/N	m	o	0	NA
CAM035399	G/G	N/N	m	o	0	NA
CAM035400	G/G	N/N	m	o	0	NA
CAM035401	G/G	N/N	m	o	0	NA
CAM035402	G/G	N/N	m	o	0	NA
CAM035403	G/G	N/N	m	o	0	NA
CAM035406	G/G	N/N	m	o	0	NA
CAM035410	G/G	N/N	m	o	0	NA
CAM035415	G/G	N/N	m	o	0	NA
CAM035416	G/G	N/N	m	o	0	NA
CAM035421	G/G	N/N	m	o	0	NA
CAM035425	G/G	N/N	m	o	0	NA
CAM035426	G/G	N/N	m	o	0	NA

CAM035427	G/G	N/N	m	o	0	NA
CAM035435	G/G	N/N	m	o	0	NA
CAM035438	G/G	N/N	m	o	0	NA
CAM035444	G/G	N/N	m	o	0	NA
CAM035456	G/G	N/N	m	o	0	NA
CAM035459	G/G	N/N	m	o	0	NA
CAM035467	G/G	N/N	m	o	0	NA
CAM035471	G/G	N/N	m	o	0	NA
CAM035473	G/G	N/N	m	o	0	NA
CAM035475	G/G	N/N	m	o	0	NA
CAM035476	G/G	N/N	m	o	0	NA
CAM035477	G/G	N/N	m	o	0	NA
CAM035478	G/G	N/N	m	o	0	NA
CAM035482	G/G	N/N	m	o	0	NA
CAM035495	G/G	N/N	m	o	0	NA
CAM035497	G/G	N/N	m	o	0	NA
CAM035500	G/G	N/N	m	o	0	NA
CAM035504	G/G	N/N	m	o	0	NA
CAM035222	G/G	N/N	i	w	7	1
CAM035244	G/G	N/N	i	w	0	1
CAM035312	G/G	N/N	i	w	1	1
CAM035150	G/G	N/N	m	w	2	NA
CAM035152	G/G	N/N	m	w	0	NA
CAM035153	G/G	N/N	m	w	0	NA
CAM035160	G/G	N/N	m	w	8	NA
CAM035162	G/G	N/N	m	w	0	NA
CAM035164	G/G	N/N	m	w	0	NA
CAM035168	G/G	N/N	m	w	0	NA
CAM035173	G/G	N/N	m	w	0	NA
CAM035236	G/G	N/N	m	w	5	NA
CAM035242	G/G	N/N	m	w	4	NA
CAM035259	G/G	N/N	m	w	3	NA
CAM035263	G/G	N/N	m	w	2	NA
CAM035266	G/G	N/N	m	w	7	NA
CAM035290	G/G	N/N	m	w	1	NA
CAM035291	G/G	N/N	m	w	6	NA
CAM035295	G/G	N/N	m	w	5	NA
CAM035296	G/G	N/N	m	w	4	NA
CAM035299	G/G	N/N	m	w	4	NA
CAM035305	G/G	N/N	m	w	3	NA
CAM035326	G/G	N/N	m	w	5	NA
CAM035369	G/G	N/N	m	w	1	NA
CAM035373	G/G	N/N	m	w	5	NA
CAM035375	G/G	N/N	m	w	5	NA
CAM035448	G/G	N/N	m	w	1	NA
CAM035465	G/G	N/N	m	w	3	NA
CAM035484	G/G	N/N	m	w	3	NA
CAM035490	G/G	N/N	m	w	3	NA
CAM035183	G/G	N/N	NA	NA	NA	NA
CAM035184	G/G	N/N	NA	NA	NA	NA
CAM035212	G/T	N/N	i	o	0	0
CAM035177	G/T	N/N	m	o	0	NA
CAM035178	G/T	N/N	m	o	0	NA
CAM035207	G/T	N/N	m	o	0	NA
CAM035294	G/T	N/N	m	o	0	NA
CAM035319	G/T	N/N	m	o	0	NA
CAM035331	G/T	N/N	m	o	0	NA
CAM035357	G/T	N/N	m	o	0	NA
CAM035374	G/T	N/N	m	o	0	NA
CAM035397	G/T	N/N	m	o	0	NA
CAM035398	G/T	N/N	m	o	0	NA
CAM035432	G/T	N/N	m	o	0	NA
CAM035468	G/T	N/N	m	o	0	NA
CAM035143	G/T	N/N	m	w	0	NA
CAM035147	G/T	N/N	m	w	4	NA
CAM035158	G/T	N/N	m	w	7	NA
CAM035210	G/T	N/N	m	w	1	NA
CAM035260	G/T	N/N	m	w	1	NA
CAM035262	G/T	N/N	m	w	4	NA
CAM035267	G/T	N/N	m	w	6	NA
CAM035282	G/T	N/N	m	w	5	NA
CAM035304	G/T	N/N	m	w	5	NA
CAM035362	G/T	N/N	m	w	2	NA
CAM035363	G/T	N/N	m	w	4	NA
CAM035364	G/T	N/N	m	w	3	NA
CAM035394	G/T	N/N	m	w	4	NA

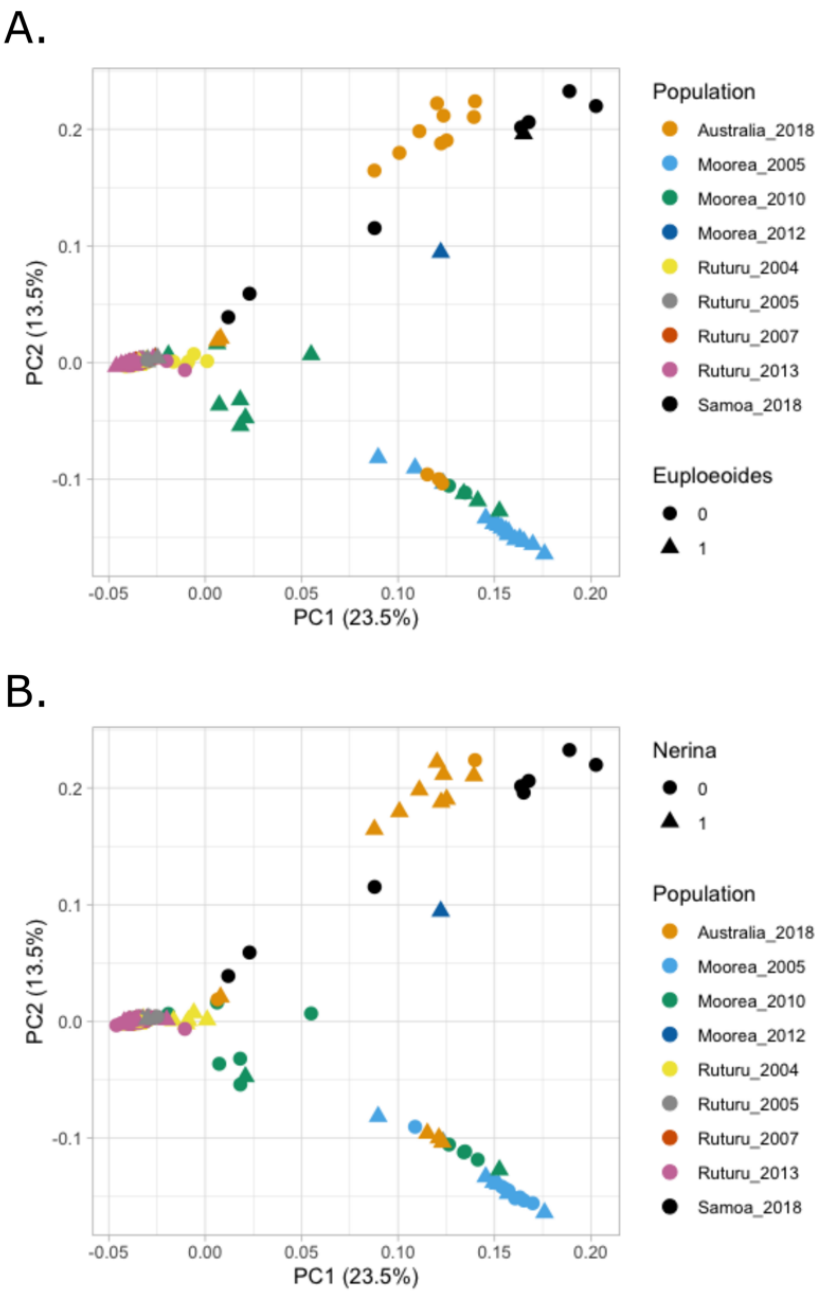
CAM035450	G/T	N/N	m	w	1	NA
CAM035479	G/T	N/N	m	w	7	NA
CAM035206	N/N	N/N	i	o	0	0
CAM035472	N/N	N/N	i	o	0	1
CAM035175	N/N	N/N	m	o	0	NA
CAM035220	N/N	N/N	m	o	0	NA
CAM035229	N/N	N/N	m	o	0	NA
CAM035297	N/N	N/N	m	o	0	NA
CAM035311	N/N	N/N	m	o	0	NA
CAM035343	N/N	N/N	m	o	0	NA
CAM035376	N/N	N/N	m	o	0	NA
CAM035412	N/N	N/N	m	o	0	NA
CAM035420	N/N	N/N	m	o	0	NA
CAM035458	N/N	N/N	m	o	0	NA
CAM035469	N/N	N/N	m	o	0	NA
CAM035483	N/N	N/N	m	o	0	NA
CAM035486	N/N	N/N	m	o	0	NA
CAM035215	N/N	N/N	i	w	1	0
CAM035485	N/N	N/N	i	w	3	0
CAM035256	N/N	N/N	i	w	4	1
CAM035377	N/N	N/N	i	w	0	1
CAM035237	N/N	N/N	m	w	0	NA
CAM035289	N/N	N/N	m	w	8	NA
CAM035309	N/N	N/N	m	w	7	NA
CAM035340	N/N	N/N	m	w	8	NA
CAM035355	N/N	N/N	m	w	6	NA
CAM035388	N/N	N/N	m	w	4	NA
CAM035457	N/N	N/N	m	w	2	NA
CAM035463	N/N	N/N	m	w	2	NA
CAM035491	N/N	N/N	m	w	1	NA
CAM035185	N/N	N/N	NA	NA	NA	NA
CAM035321	N/N	N/N	NA	NA	0	NA
CAM035318	T/G	N/N	m	o	0	NA
CAM035413	T/G	N/N	m	o	0	NA
CAM035447	T/G	N/N	m	o	0	NA
CAM035449	T/G	N/N	m	o	0	NA
CAM035480	T/G	N/N	m	o	0	NA
CAM035272	T/G	N/N	m	w	5	NA
CAM035250	T/G	N/N	NA	NA	6	NA
CAM035214	T/T	N/N	m	o	0	NA
CAM035372	T/T	N/N	m	w	2	NA



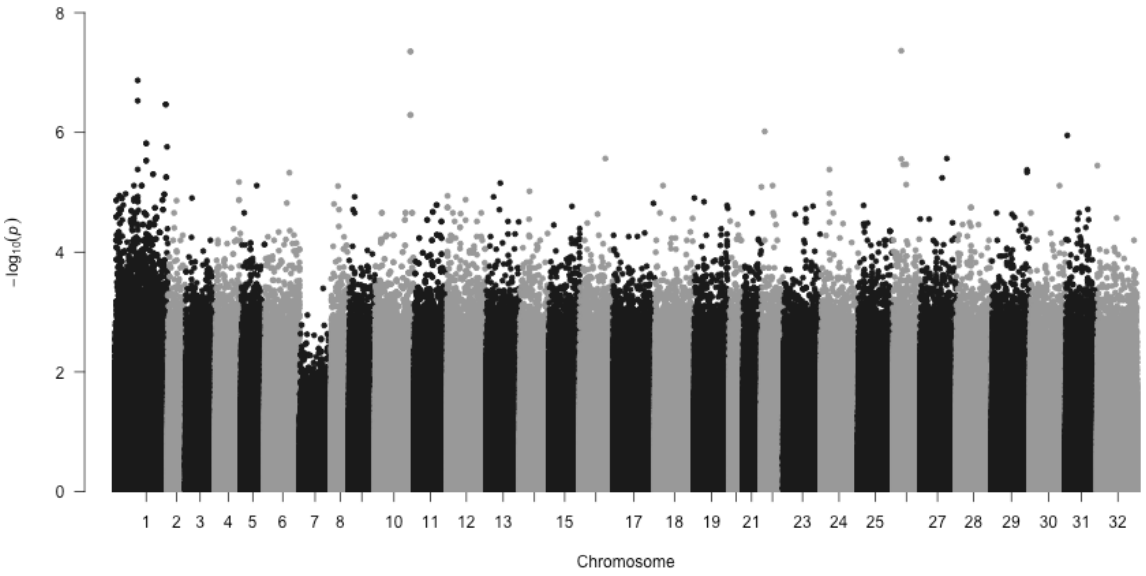
Supplementary Figure 5.1 | QTL mapping of hindwing white using a backcross design with heterozygous mother (**A**) and heterozygous father (**B**). Chromosome have not been named by homology to *M. cinxia*, but by size. Chromosome 11 in the figure is chromosome 8 by homology to *M. cinxia*.



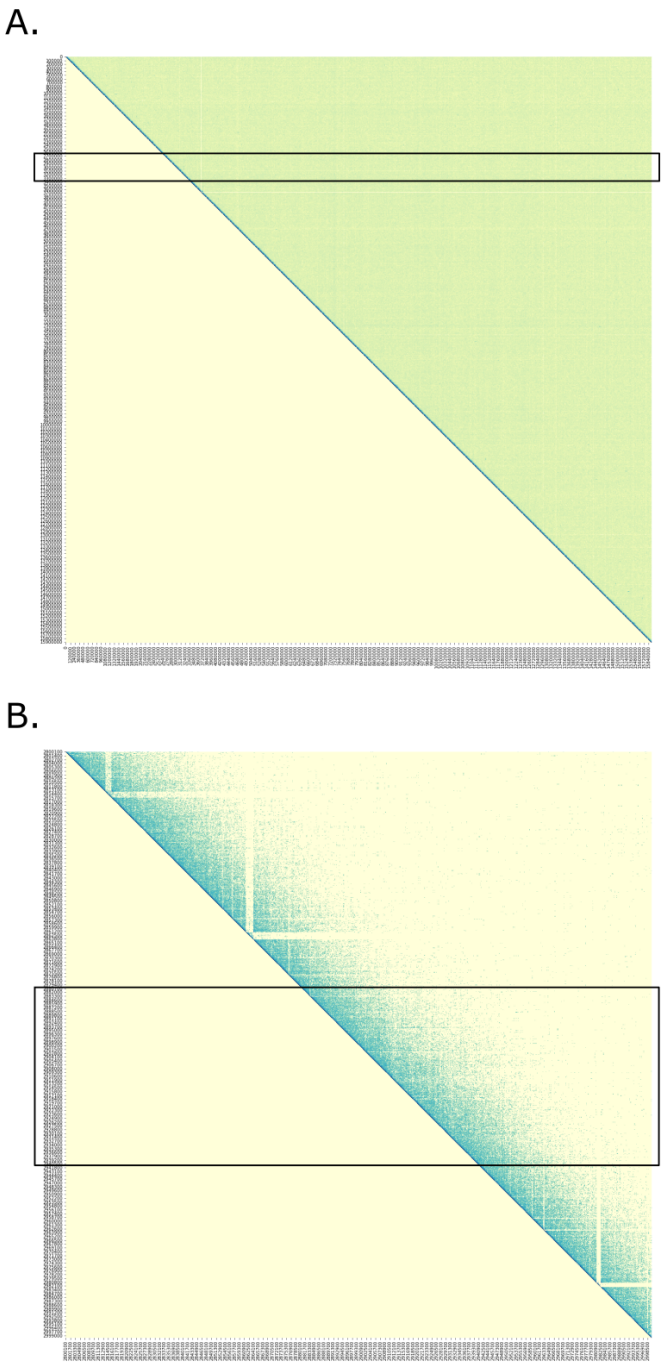
Supplementary Figure 5.2 | GWAS of hindwing white using the two reared families and the published Samoan Samples (Hornett et al., 2014). **A.** Whole chromosome 8. **B.** Zoom in of the association peak.



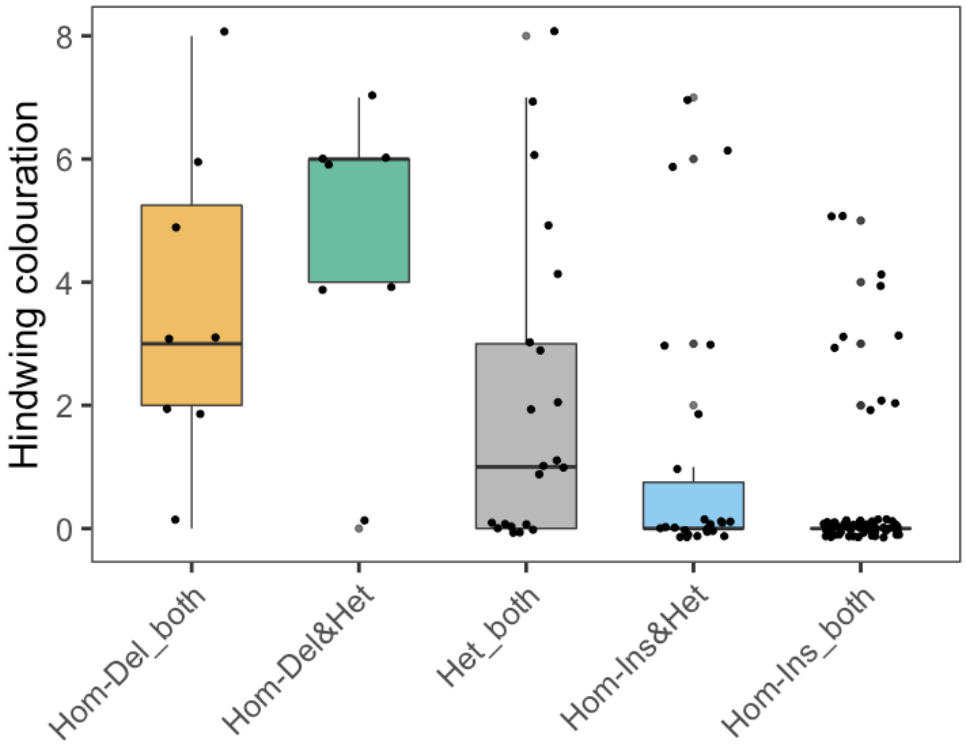
Supplementary Figure 5.3 | Principal Component Analysis of the *H. bolina* samples used for GWAS.



Supplementary Figure 5.4 | GWAS of *H. misippus immima* and *inaria* individuals does not show any clear association peak. Chromosome are numbered based on the original *H. misippus* numbering (unrelated to Merian units; see Supplementary Table 1).



Supplementary Figure 5.5 | No large structural variants are found around *cortex* in *H. misippus.* Wrath analysis of haplotagging data of a large dataset (335 samples), shows no evidence of SVs around *cortex* when analysing the whole chromosome 8 with a window size of 10000 bp (**A**) or a subset of it Chr8:2800100-2930000 with a window size of 50 bp (**B**).



Supplementary Figure 5.6 | Continuous variation of hindwing colouration associated with genotype at the two insertions, when both insertions are homozygous for the deletion (Hom-Del_both), one is homozygous for the deletion and the other heterozygous (Hom-Del&Het), both are heterozygous (Het_both), one is homozygous for the insertion and the other heterozygous (Hom-Ins&Het) and both are homozygous for the insertion (Hom-Ins_both).

2

A *yellow* gene controls male colour polymorphism in the wood tiger moth

2.1 Abstract

Unpalatable species often use colouration to advertise their toxicity. These cases of warning colouration are expected to be under positive frequency dependent selection, with populations converging onto one aposematic phenotype to share the costs of teaching predators. However, there are several puzzling cases of aposematic polymorphisms found in the wild. The wood tiger moth (*Arctia plantaginis)* is a palearctic species that uses bright hindwing colouration to advertise its toxicity, in which males have two distinct morphs that are maintained in the populations. The male morphs have yellow or white hindwings and vary in other phenotypic traits such as behaviour, chemical defences, and physiology. This complex phenotype is maintained at stable population frequencies and is determined by a single Mendelian locus with two alleles. Here, in collaboration with Dr Melanie Brien, I show that a white specific duplication of a *yellow* gene is associated with morph determination. Using gene expression analyses of *A. plantaginis* individuals across development, I show that the duplicated gene, which we name *valkea,* is highly expressed in the dominant morph and absent in the recessive. The *yellow* gene family is involved in melanin pigmentation and behaviour in other species of insects and is thus a strong candidate for this phenotype. This study adds to the evidence suggesting that multiple co-adapted traits can have a simple genetic architecture and will serve as the basis for future work on complex phenotypes.

2.2 Introduction

Understanding how phenotypic diversity is generated and maintained in the wild is a key question in evolutionary biology. Polymorphisms are of special interests to evolutionary biologists, as selection and drift are expected to reduce genetic diversity, pushing the alleles with the highest fitness to fixation. However, there are multiple examples of stable polymorphisms in the wild, from *Heliconius* butterflies to wall lizards (Andrade et al., 2019; Brown and Benson, 1974).

A particularly interesting case are complex polymorphisms, adaptive combinations of multiple traits that are inherited together and that are maintained in the population. The stability of complex polymorphisms is dependent on its genetic architecture, as linkage between the various co-adapted traits is necessary to ensure their joint inheritance. In the literature, two possible architectures have been hypothesised, a single regulatory gene hypothesis and supergenes (reviewed in (Thompson and Jiggins, 2014)). Under the single regulatory gene hypothesis, a single gene interacts with multiple downstream targets affecting their expression or function. This gene is therefore the regulator of all the co-adapted traits in the complex polymorphism. It seems plausible that a gene such as a transcription factor could play such a role, but it is unclear how other genes could have such a broad phenotypic effect (Andrade et al., 2019). In contrast, in the supergene hypothesis, co-adapted alleles controlling each trait of the polymorphism are in linkage disequilibrium, maintaining the co-inheritance of all traits. Often, supergenes contain structural variation such as inversions and deletions that reduce the recombination rate across the region and ensure the joint inheritance of divergent alleles (Küpper et al., 2015; Lamichhaney et al., 2016; Schwander et al., 2014; Wang et al., 2013). Unravelling the genetic basis of complex polymorphisms can therefore shed light on how phenotypic diversity is maintained in the wild.

Some of the most prominent examples of polymorphisms come from the Lepidoptera. The genetic basis of warning colouration in tropical *Heliconius* butterflies has been extensively studied (Hines et al., 2011; Nadeau et al., 2016; Reed et al., 2011; Westerman et al., 2018; Zhang et al., 2017b). Although historically moths have received less attention, some of the most remarkable examples are found in this group, such as the stable polymorphism in the scarlet tiger moth (O'Hara, 2005) and melanism in the peppered moth (van't Hof et al., 2016).

The wood tiger moth (*Arctia plantaginis*) is a particularly attractive system to investigate the evolution of colour polymorphism. Males of this species are polymorphic in warning colouration, both locally and on broader geographical scales. Discrete hindwing morphs are found across Europe, with yellow and white being the most common (Hegna et al., 2015). Such polymorphism is not expected in aposematic species, as selection is stronger against rare forms creating positive frequency dependence. However, in the wood tiger moth, geographic and temporal variation in predation pressure produces changes in morph fitness and maintains the polymorphism (Galarza et al., 2014). Colour morphs are associated with a suit of other traits such as chemical defences (Rojas et al., 2017), flight activity (Rojas et al., 2015), sexual selection (Gordon et al., 2015; Nokelainen et al., 2012) and immune response (Nokelainen et al., 2013), and are determined by a single Mendelian locus with two alleles, the dominant white (W) and the recessive yellow allele (y). Although much is known about the ecology of male wing colour polymorphism in wood tiger moth, its genetic underpinnings remain a mystery.

This chapter represents a collaboration in which we have used QTL, GWAS and RNAseq approaches to identify a single gene from the *yellow* gene family as the gene controlling the development of male wing colour morphs in the wood tiger moth. Previous work had carried out a GWAS and QTL mapping analysis of male polymorphism. I used these results to determine a list of candidate genes, obtain RNA-seq data from several stages of development, describe the gene expression patterns observed in pupal wings and pinpoint the genes presenting differential gene expression between morphs. I also examined the functions of the differentially expressed genes between morphs throughout development and performed an enrichment analysis. Finally, I described the gene expression patterns observed in tiger moth wings throughout development and compared them to the Nymphalid butterfly *Vanessa cardui*.

2.3 Methods

2.3.1 Rearing and dissections

A. plantaginis were reared at the University of Jyväskylä in Finland by Prof Johanna Mappes research group, where lines of homozygous dominant white males (WW) and homozygous recessive yellow males (yy) were established. Pupae and larvae were sent to Cambridge, UK, where they were kept between 22 and 30˚C. Larvae were fed fresh dandelions (*Taraxacum sp*.) leaves until pupation. Pupae were sexed and only males were used.

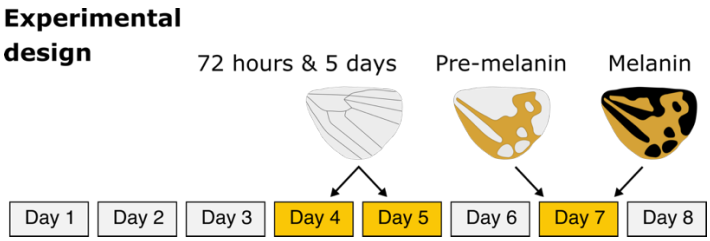


Figure 6.1 | Experimental design used to identify candidate genes associated with colour morphs. Four sampling points were used along pupal development. Two early stages, 72 hours and 5 days, in which no colouration was visible, and two later ones, pre-melanin and melanin, in which yellow and black pigmentation had been deposited.

To explore which genes are involved in colour formation in *A. plantaginis,* I collected hindwing tissue across development. Given that yellow genes and other melanin pathway genes are highly expressed in pupal stages (Ferguson et al., 2011; Zhang et al., 2017a), I focussed on these (see section 6.3.3). I sampled 4 developmental times using morphological differences to define those (Figure 6.1 and Supplementary Figure 6.1); 72-hours and 5 days, where no colouration has appeared; pre-melanin stage, where yellow colouration is visible in yy individuals; and melanin stage, where black melanin s deposited. I sampled 5 individuals per genotype and stage (40 individuals in total). Hindwings and forewings were stored separately in RNA-later (Sigma-Aldrich) at 4˚C for 2 weeks and later transferred to -20 ˚C, while the rest of the body was stored in pure ethanol at -20˚C.

2.3.2 RNA extractions

Once the dissections had been performed, I extracted total RNA from hindwing tissue using a standard hybrid protocol. First, I transferred the wing tissue into Trizol Reagent (Invitrogen) and homogenised it using dounce tissue grinders (Sigma-Aldrich). Then, I performed a chloroform phase extraction, followed by DNase treatment (Ambion) for 30 mins at 37˚C. I measured the concentration of total RNA using Qubit Fluorometric Quantitation (Therofisher) and performed a quality check of the RNA integrity number (RIN) using an Agilent 4200 TapeStation (Agilent). The extracted total RNA was stored at -20˚C before being sent to Novogene UK for library preparation and sequencing.

2.3.3 Differential gene expression of *A. plantaginis*

First, I performed quality control and low quality base and adapter trimming using *TrimGalore!* (Krueger, 2015). Using STAR (Dobin et al., 2013), I then mapped the trimmed reads to the *A. plantaginis* reference genome assembled using the homozygous W individual. I performed a second round of mapping (2pass) including as input the output splice junctions from the first round. The *A. plantaginis* genome annotation iArcPla.TrioW.Curated.20190705 (Yen et al., 2020) was included in each round of mapping. I then used *FeatureCounts* to summarise read counts by gene (Liao et al., 2013).

To identify the gene or genes controlling the development of wing colour in *A. plantaginis,* I performed a genome wide differential expression analysis using limma-voom in R (Ritchie et al., 2015). First, I defined a categorical variable, 'GenStage', with 8 levels containing the genotype and stage information of every individual sample (e.g., yy72h, WW72h, yy5days, etc.). Then, I built the design matrix fitting a model with GenStage as the only fixed effect factor contributing to the variance in gene expression and included family as a random effect factor (gene expression ~ 0 + GenStage + (1|Family)). After that, I filtered lowly expressed genes, which resulted in a reduction of the number of tested genes from 17,930 to 10,920 in the W-mapped one. MDS plots were generated using the plotMDS function from limma. Finally, I extracted the list of genes that are differentially expressed in each stage using the Benjamini-Hochberg procedure to correct for multiple testing.

2.3.4 Orthology Assignment

To infer orthology between *A. plantaginis* and *V. cardui, D. melanogaster* and *H. melpomene*, I used OrthoFinder (Emms and Kelly, 2019), a comparative genomics tools that finds orthologs and infers rooted trees among protein sequences. To do that, I used protein sequences from 6 Lepidoptera species, *Plutela xylostella* (GCA_905116875_2)*, Bombyx mori* (GCF_014905235_1)*, Spodoptera frugiperda* (GCF_011064685_1)*, Parnassius apollo* (GCA_907164705_1)*, Pieris macdunnoughi* (GCA_905332375_1) *amd Pararge aegeria* (GCF_905163445_1) and *Drosophila melanogaster* (GCF_000001215_4). I ran the primary_transcript.py utility from OrthoFinder to extract only one transcript per protein, and then ran OrthoFinder with default settings.

Not all genes were assigned to orthogroups. To get the most likely candidate homologue for remaining genes, BLASTp was used between *A. plantaginis* and *H. melpomene* and *D. melanogaster* (BLAST, 2013). First, I extracted the amino acid sequences for all predicted proteins in the *A. plantaginis* and used BLASTp to find the most similar match in a database of *H. melpomene* predicted proteins. All matches were required to have an e-value < 1e-5 and only the top result was used for each protein, selecting the match with the minimum e-value and in case of two matches having the same e-value, I selected the one with the highest identity percentage.

2.3.5 Gene Ontology analysis

Once all orthologues had been identified, I extracted the GO terms for each gene in the genome using the interprot database (Blum et al., 2021). After that, I performed a Gene Onthology (GO) analysis using the package TopGo in R (Alexa and Rahnenführer, 2022). GO terms enrichment was analysed separately for each developmental stage.

2.3.6 Analysis of V. cardui RNA-seq data and genome-wide gene expression comparison of V. cardui and A. plantaginis

I downloaded *V. cardui* developmental RNA-seq data from SRA, project id PRJNA293289. I analysed the reads following the same methods used for the *A. plantaginis* data (see above), using the *V.cardui* reference genome ASM2240509v1 (Zhang et al., 2021) for mapping. I analysed the count data by fitting a model with 'developmental stage' as the only factor (gene expression ~ 0 + developmental stage). Then, using a contrast matrix, I compared the expression between all pairs of stages. Only genes with p-value < 0.001 in any of the comparisons were retained. To be able to compare datasets, I re-analysed the W-mapped *A. plantaginis* data fitting a model with two factors, genotype and developmental stage (gene expression ~ 0 + genotype + developmental stage) and extracting the appropriate contrasts. For the comparison with *V. cardui*, I retained only the top differentially expressed genes (p-value < 5e-07)

2.4 Results

2.4.1 yellow family genes are found at the GWAS association peak

Previous results had shown that male colour polymorphism in the wood tiger moth is controlled by a narrow region of the genome situated in scaffold 206 of the reference genome iArcPla.TrioY.Curated.20190705 (Figure 6.2A)(Yen et al., 2020). To identify candidate genes for the control of morph development, I searched for orthologous protein sequences in the *Drosophila melanogaster* and *Heliconius melpomene* genome and identified the genes in and around the GWAS association peak. The association peak is 98,836 bp long and includes two genes, *yellow-e* and *yellow-g2* (Figure 6.2B-C). Both genes belong to the *yellow* gene family, which has been previously associated with melanin wing colour development in Lepidoptera and *Drosophila* (Zhang et al., 2017a). These two genes found within the association peak are thus strong candidates for the control of male wing polymorphisms in the *A. plantaginis*. However, another possibility could be that the associated region contains a regulatory element that modifies the expression pattern of a gene found outside the association peak. 15 genes are found in the broader candidate region (9,700,000 to 10,150,000 in Scaffold 206). Of those, 5 are *yellow* family genes, including the 2 genes found in the association peak, and 6 are ATP binding cassette subfamily A members (Supplementary Table 6.1.). To pinpoint which of these candidate genes is controlling male wing polymorphism development in *A. plantaginis* I obtained RNA samples and performed gene expression analysis.

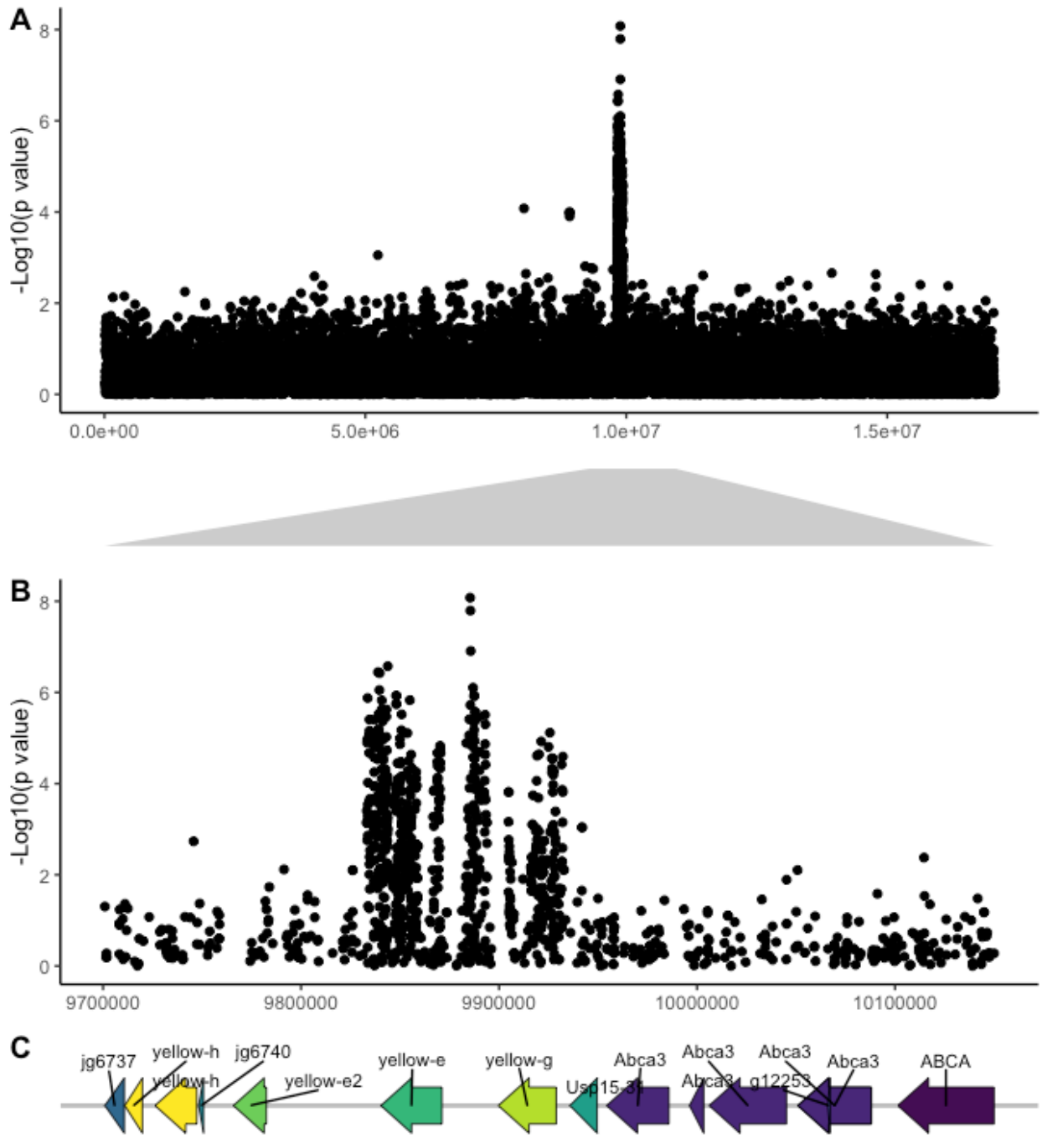


Figure 6.2 | A region in scaffold 206 is associated with variation between male colour morphs. **A.** A GWAS between yellow and white males shows a unique peak in scaffold 206. **B and C.** The associated peak of ~98kb includes two yellow family genes.

2.4.2 The association peak contains a duplication

The trio binning method used by Yen et al. 2020 to produce the *A. plantaginis* reference genome produced two reference sequences, one for a white and one for a yellow allele. In collaboration with Dr Melanie Brien, I extracted the region containing the QTL interval from the yellow reference and aligned it with the white reference. This showed a duplicated region approximately 100kb long in the white reference, containing 5 genes (Supplementary Figure 6.1). Using BLASTp searches and pairwise alignments, we determined that the *yellow-e* gene is duplicated in the white reference. One copy (jg1310) has 7 exons and is similar to the *yellow-e* gene in the yellow reference (99.7% identity in coding sequences). The second copy unique to the white scaffold (jg1308) has only the first 5 exons (81.8% identical to the gene in the yellow reference). We named this gene *valkea,* meaning 'white' in Finish. In the light of these results, I performed differential gene expression analysis on the W reference genome, as it contains the duplicated region that includes *valkea.*

2.4.3 Developmental stage is the main driver of genome wide gene expression

To pinpoint which of these candidate genes controls male wing polymorphism*,* I performed gene expression analysis of *A. plantaginis* individuals across development (Figure 6.1). Based on knowledge of the expression patterns of *yellow* genes (Ferguson et al., 2011) and other colouration genes particularly involved in melanin synthesis in Lepidoptera (Zhang et al., 2017a), I hypothesised that changes in gene regulation that control the development of wing colour morphs in the tiger moth most likely occurs during pupal development. Therefore, I generated RNA-seq libraries for wing tissue from wood tiger moth males across pupal development. I first described pupal development of the wood tiger moth, in order to choose the developmental stages for transcriptomic analysis. Pupal development in the wood tiger moth lasts for 8 days. Yellow colouration appears first at day 7, and a few hours after that, black melanin is deposited. I used these morphological differences between stages to define my sampling developmental times (Figure 6.1 and Supplementary Figure 6.1). I sampled 4 developmental times, two stages early in development, 72-hours and day 5, in which no colouration is present in the wings; and two stages later in development, once yellow coloration has appeared (pre-melanin, 7 days old pupae) and the other after black melanin has also been deposited (melanin, 7-8 days old).

Then I obtained RNAseq of the samples and analysed the whole genome pattern of expression. I filtered out lowly expressed genes and retained 10,920 genes for DE analysis. Then, I examined the genome-wide variation in gene expression between samples using a multidimensional scaling (MDS) and determined that developmental stage was the main factor explaining most of the variation in gene expression (Figure 6.3). Such a pattern would be expected as many genes are involved in development and thus are likely to be differentially expressed between developmental stages. No apparent clustering can be observed among samples of the same colour morph, perhaps consistent with the fact that the phenotype is controlled by a single Mendelian locus such that few genes are DE between morphs. Although that would not be necessarily expected if the expression of many downstream genes is also affected. These results suggest that more genes are involved in expression profiles specific to the developmental stage, rather than the wing phenotype.

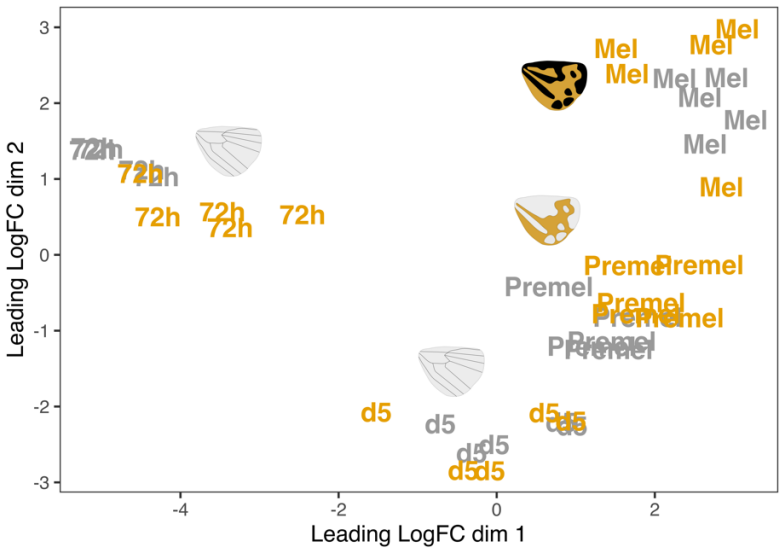


Figure 6.3 | Genome wide expression patterns are shaped by developmental stage. Multidimensional scaling of gene expression shows that samples cluster by developmental stage rather than male colour morph. Developmental stage names have been abbreviated: 72h=72 hours, 5d=5 days, Premel=pre-melanin and Mel=melanin.

2.4.4 *Valkea* is highly expressed in the white morph at the colour forming stage

Following that, I examined differential expression between colour morphs. Overall, 99 genes were differentially expressed (FDR < 0.05) between the two morphs (Figure 6.4). Of these, 49 were upregulated in the yy morph, while the remaining 50 were upregulated in WW individuals. The earliest developmental stage, 72-hours, was the stage with the highest number of DE genes (n = 48), while the 5-day old stage had the fewest (n=7). One gene was over-expressed in yy in the three first stages (see section 6.3.5).

From the 22 genes identified within the GWAS and QTL candidate region, only two, *yellow-e* and *valkea*, were differentially expressed in any of the comparisons. *valkea* was overexpressed in white individuals in the pre-melanin stage with a Log Fold Change of 10.32 and a p-value of 2.18e-06 (Figure 6.4B-C). Since *valkea* is only present in the W genome any expression in the Y genome must result from mismapped reads. *yellow-e* was also overexpressed in white individuals with a Log Fold Change of 3.86 and adjusted p-value of 5.62e-06. In other developmental stages, neither *valkea* nor *yellow-e* showed differences in expression between morphs (Figure 6.4B).

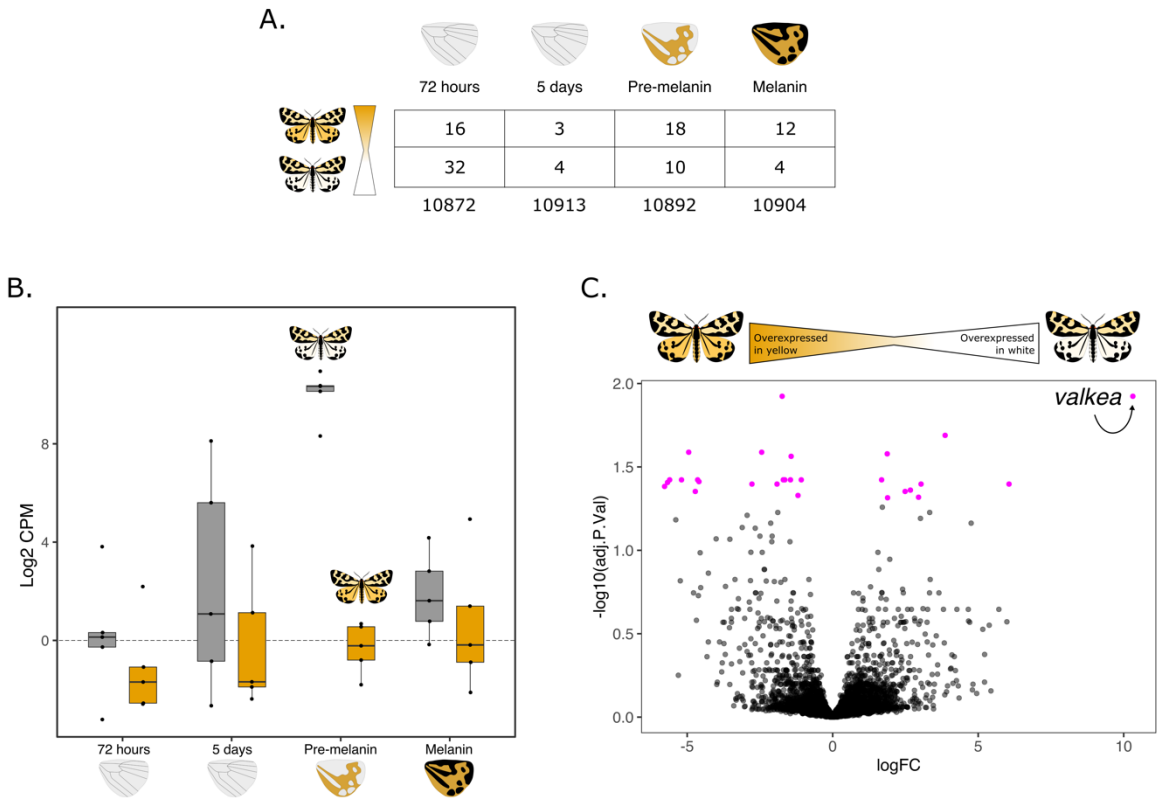


Figure 6.4 | *Valkea* is overexpressed in white males in the pre-melanin stage. **A.** Genome wide differential expression (DE) analysis between yellow and white males shows that the 72 hours stage has the most DE genes, followed by the pre-melanin, melanin, and 5 days. **B.** Expression of *valkea* across developmental timepoints shows that it has higher expression measured in Log2 CPM (Counts Per Million) in white individuals compared to yellow ones. **C.** DE analysis of all genes at the pre-melanin stage shows that *valkea* is the most DE gene (i.e. gene with the highest Log Fold Change).

Finally, it could also be that the GWAS and QTL peak contain a cis-regulatory element (CRE) that regulates the expression of a gene outside the peak but in the same linkage group (chromosome 12). 10 genes are differentially expressed between colour morphs apart from *yellow-e* and *valkea* and are located on chromosome 12 of the WW assembly (chromosome 9 of the yy). These genes are differentially expressed at different stages, 1 gene at 72 hours, 2 at 5-days, 4 at pre-melanin and 3 at the melanin stage. 5 of these genes have orthologues in *D. melanogaster* and of particular interest is *cryptochrome* (jg2035), which is a circadian clock regulator (Supplementary Table 6.2).

2.4.5 The transcription factor, *CG2129*, is overexpressed in white individuals in three consecutive developmental stages

When looking at the expression patterns of genes outside the region of association, I identified one gene, CG2129 (jg15945, identified by orthology with *D. melanogaster*) found in chromosome 25, that is significantly overexpressed in yellow individuals in three different consecutive stages, 72 hours, 5 days and pre-melanin (Supplementary Figure 6.2 and 6.3). *CG2129* encodes a zinc finger transcription factor in *D. melanogaster* expressed during embryo development (Brown et al., 2014), which makes it a good candidate for male colour morph development. One possibility is that a cis-regulatory element at the W-specific insertion controls the expression of a regulatory gene early in development, which regulates the expression of *CG2129*. In turn, *CG2129* regulates the expression of other downstream genes.

2.4.6 Gene Ontology analysis

To explore the functional differences between the transcriptomic profiles of the two morphs, I performed a gene ontology (GO) analysis. 99 genes showed DE between morphs (p value < 0.05), of those, 49 were overexpressed in yy individuals and 50 in WW. To find the GO terms associated with these genes, I first found their orthologues in the *Heliconius melpomene* genome and then retrieved the list of GO terms associated with each gene. Out of the 99 genes annotated in the *A. plantaginis* genome, 90 genes had orthologues in the *H. melpomene* genome, all of which had annotated GO terms. From the GO terms strongly enriched (elimKS < 0.01), none was shared between stages (Supplementary Figure 6.4). Each stage presented 1 to 3 significantly enriched GO terms.

2.4.7 Comparison of gene expression through development of the wood tiger moth and *Vanessa cardui*

Finally, I wanted to explore further the genome wide expression patterns through development in *A. plantaginis* and to compare them to other Lepidoptera species. If development pathways have been conserved across the Lepidoptera, I would expect to find genes with similar patterns of expression through time. To do that, I analysed an RNA-seq dataset of *Vanessa cardui* from (Zhang et al., 2021), which includes dissected hindwings across development, including the same developmental timepoints as my dataset (72 hours,

5 days, pre-melanin and melanin). *V. cardui* is a Nymphalid butterfly distantly related to *A. plantaginis* (99 MYA; (Kumar et al., 2022)). Crucially, colour pattern development in *V. cardui* parallels that of *A. plantaginis.* Development lasts for 8 days in both species, with early developmental stages (up to 6 days) showing no pigment deposition but white scales being visible (Connahs et al., 2016). After that, orange/yellow pigments are deposited at day 6-7, followed by melanin deposition. Despite the parallels, orange colouration is formed by ommochrome pigments in *V. cardui,* while pheomelanins are the most likely pigment that produce yellow colouration in *A. plantaginis* (Appendix B).

The two datasets require the use of different reference genomes. Thus, to be able to compare the results, I identified orthology relationships between the two annotations using OrthoDB. I included 6 Lepidoptera species and *Drosophila melanogaster* to maximise the chances of inferring the correct orthologue pairs (see Methods). First, I analysed the two datasets separately, producing lists of differentially expressed genes in each species, comparing all pairs of developmental stages. Then, I merged the results using the orthology relationships previously produced.

The genome wide patterns of expression recapitulated differences between species (p-value < 0.05, MDS). However, when focusing on the top 91 differentially expressed genes between stages (p-value < 5e-07 for *A. plantaginis* and p-value < 0.001 for *V. cardui*), most of the variation in gene expression was explained by stage rather than species (Figure 6.6.A-B). That means, that there is a set of genes with a pattern of gene expression that is stage specific and conserved between the two species. Clustering analysis using k-means groups samples by the 4 developmental stages, except for the two pre-melanin *V. cardui* samples which are grouped with all the melanin samples. There are no differentially expressed genes between the pre-melanin and melanin *V. cardui* samples, which explains why the samples group together in the k-means clustering and MDS. These 91 genes differentially expressed in both species are broadly linked to various gene ontology terms relating to metabolic processes, such as organic substances, nitrogen and chitin metabolic processes, and transcriptional processes (Supplementary Table 6.3).

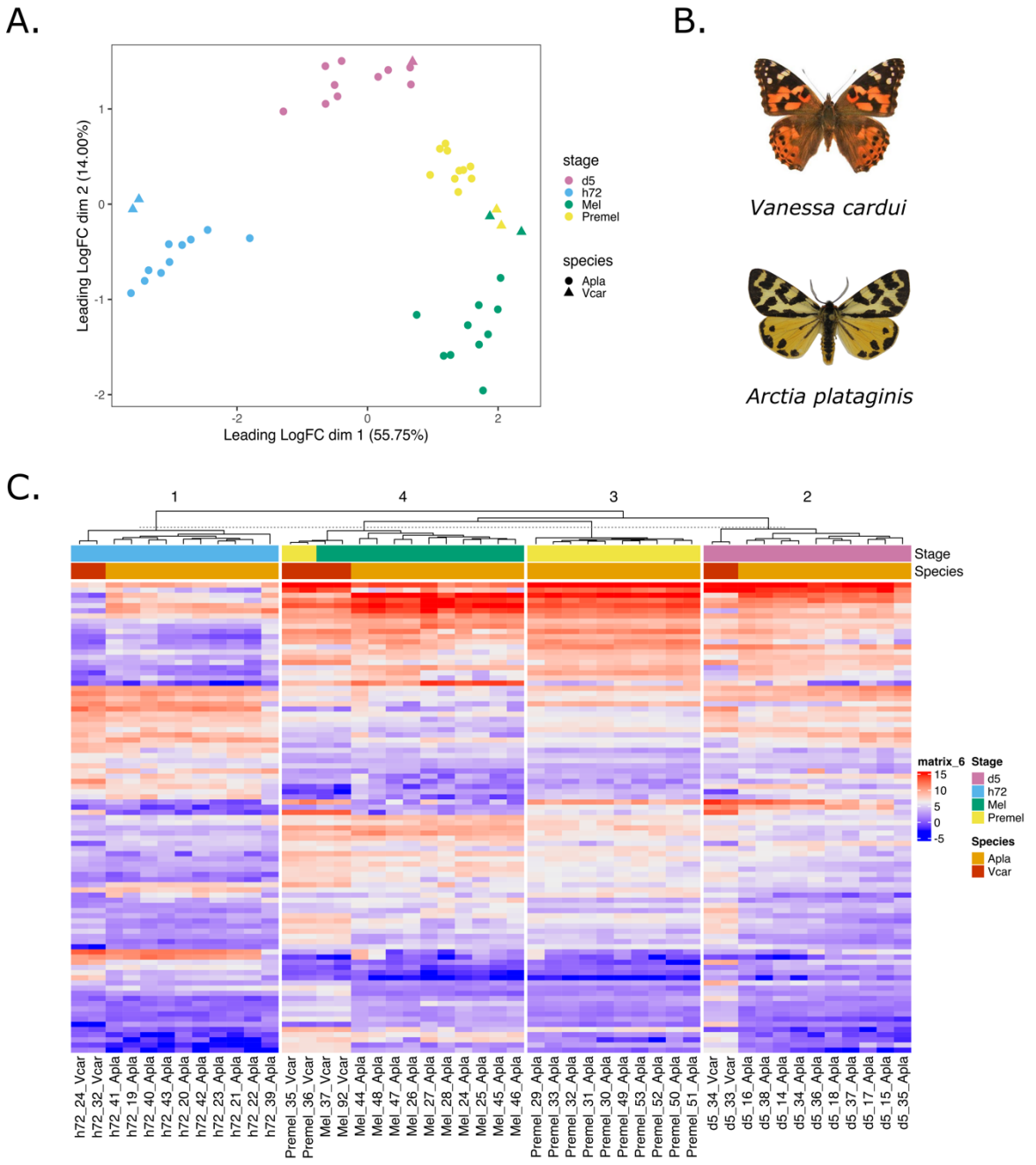


Figure 6.5 | 91 genes show stage-specific expression patterns that are common between *A. plantaginis* and *V. cardui.* **A.** Multidimensional scaling of Log Fold Change of the top 91 DE genes between stages shows clustering by developmental stage, instead of species. **B.** The study systems *V. cardui* (top) and *A. plantaginis* (bottom). **C.** Clustering analysis shows that the 91 candidate genes have stage specific expression that is common in both species. Samples are clustered by similarities in their expression profiles.

2.5 Discussion

Male *Arctia plantaginis* present distinct hindwing colour morphs that also co-vary in other life-history, physiological and behavioural traits, making it a complex polymorphism. Using gene expression data, I have shown that a *yellow* family gene, *valkea,* is the most likely controller of morph development. *valkea* is situated at the centre of the association peak identified though GWAS and QTL mapping and is a recent duplication of *yellow-e*, which has been linked to colour polymorphisms in other Lepidoptera (Ferguson et al., 2011). A duplication of *yellow-e* and its flanking region produced an insertion that is only found in white alleles and contains *valkea*, and such asymmetry between morphs (i.e. presence/absence of the insert) reduces recombination and maintains the joint inheritance of the loci in the insert. Furthermore, *valkea* is highly expressed in white individuals in the pre-melanin stage of development, and not present in yellow individuals. This pattern of expression would be expected from the gene responsible for morph development as colour is inherited as a Mendelian single locus with two alleles, where the white allele is dominant over yellow.

The genetic control of complex polymorphisms has been a long-standing puzzle for evolutionary biologists. Historically they have been explained by invoking 'supergenes', a tightly linked cluster of genes that control multiple phenotypic traits and are inherited together. Recently multiple spectacular examples of supergenes controlling complex polymorphisms have been described (Küpper et al., 2015; Lamichhaney et al., 2016; Yan et al., 2020); reviewed in (Schwander et al., 2014). However, in the absence of a supergene, it is unclear how a single gene, i.e. *valkea*, can control the development of a such broad array of phenotypic traits. A possible explanation is that *valkea* has a similar function to input-output genes; these are genes that sit at the centre of developmental pathways, having many input sources and interacting with multiple downstream gene targets. By influencing the expression of several genes, input-output genes can control the development of complex characters. A potential example of this has been identified in *Colias* butterflies, in which two alternative life history strategies are controlled by a single gene, *BarH-1* (Woronik et al., 2019). Nonetheless, BarH-1 is a transcription factor, and it is thus more feasible that it could interact with multiple downstream targets and with that have a broad phenotypic effect. It is less clear, though,

how a gene that does not code for a transcription factor and that is situated downstream of a colour pathway, such as in the case of *valkea*, can also alter the development of other characteristics apart from colour.

Crucially, hypothesising that *valkea* is the central gene controlling the complex phenotype in *A. plantaginis* assumes that it can alter the expression of multiple downstream genes and influence several phenotypic traits. The *yellow* family of genes, to which *valkea* belongs, have a highly conserved function in melanin colouration throughout insects (Ferguson et al., 2011; Miyazaki et al., 2014; Wittkopp et al., 2002; Zhang et al., 2017a). Interestingly, *yellow* has also been linked to male courtship in *D. melanogaster* and *Bicyclus anynana* (Wilson et al., 1976). Furthermore, *yellow* genes share a common origin with Major Royal Jelly Protein (MRJP) genes, which are crucial for caste regulation in honey bees (Drapeau et al., 2006). Thus, it could be that *valkea* has a similar function to MRJP genes regulating behaviour and other phenotypic traits. However, the differences in courtship behaviour in *D. melanogaster* males have been shown to be a direct effect of the variation in pigmentation in *yellow* mutants, as reduced melanisation results in changes in the structure of sex combs, which are necessary for mating (Massey et al., 2019). Thus, the function of *yellow* genes in colouration in insects is well-defined, but it is less clear if and how they can influence other phenotypic traits.

An alternative explanation of how the duplication is determining morph identity is that a cis-regulatory element (CRE) found in the white specific flanking regions of *valkea* controls the expression of a gene situated outside the region. That gene could regulate the expression of many downstream targets that produce the range of linked traits in the morphs, including the high expression of *valkea* in white individuals. This could explain the over-expression of *CG2129* in yellow individuals in the early pupal stages (72 hours, 5 days and pre-melanin), which would be a target gene. *CG2129* is a transcription factor and could regulate other downstream genes. The reduction in recombination caused by the presence-absence of the insertion in the white and yellow individuals would ensure the joint inheritance of the CRE and *valkea* and guarantee the correct development of the white morph with all its associated traits. In such a case, the insertion should be considered a supergene.

Furthermore, there is the possibility that differences in expression of *valkea* or other genes have been missed due to the experimental design. Differences in gene expression happening

either earlier in development (e.g. larval or very early pupa) or in other tissues would have been missed in the analysis. Nevertheless, a gene with such a broad effect and controlling a developmental cascade is likely to be expressed widely in the body, including the wings. Lastly, it could be that *valkea* has a tissue and/or stage specific effect—that it is differentially expressed in other tissues and/or stages, but that in those it has another function other than colouration.

Finally, I explored gene expression through development in *A. plantaginis* and I have shown that the expression of a core set of 91 genes is conserved through development compared to *V. cardui,* which are 99 MYA. These genes produce a stage specific expression pattern that is common to both species and can be recognised by clustering and dimensionality reduction analyses and are related to broad transcriptional and metabolic processes. These results show that a core group of genes is conserved in Lepidoptera species, which could be used for comparative staging of wing development in other species.

In summary, I have identified a yellow family gene, *valkea,* as the main candidate for the control of a complex polymorphism in *A. plantaginis.* However, further functional tests are necessary to prove causality, explore how *valkea* is regulating the co-adapted traits and evaluate whether a CRE present in the duplicated region is controlling the expression of multiple downstream genes that regulate the co-adapted traits. This study is an intriguing example showing that polymorphisms involving multiple co-adapted traits can have a simple genetic basis and provides the basis for further work on the genetic architecture of complex phenotypes.

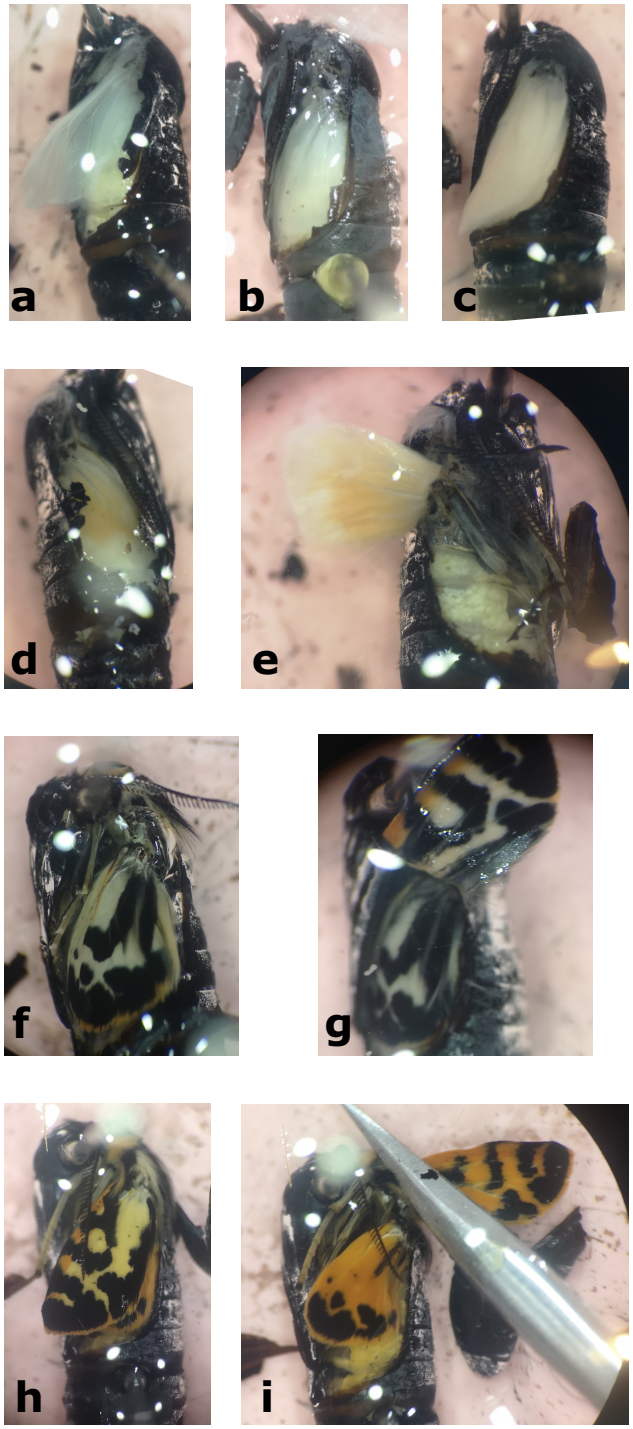
2.6 Supplementary Information

Start Position	End Position	A.pla ID	D.mel ID	Gene Name D.mel	% identity	Length	Start Position
257	1494	jg6745.t1	NP_001259764.1	Abca3	34.97	1264	257
257	1494	jg6745.t1	NP_001259765.1	Abca3	34.97	1264	257
94	1808	jg6747.t1	NP_001259764.1	Abca3	36.72	1792	94
94	1808	jg6747.t1	NP_001259765.1	Abca3	36.72	1792	94
6	541	jg6746.t1	NP_608445.2	ABCA	42.17	543	6
6	541	jg6746.t1	NP_001259765.1	Abca3	42.17	543	6
6	541	jg6746.t1	NP_001259764.1	Abca3	42.17	543	6
23	410	jg6738.t1	NP_651912.3	yellow-h	38.69	398	23
14	393	jg6742.t1	NP_524344.1	yellow-e	47.66	384	14
35	1767	jg6749.t1	NP_608445.2	ABCA	39.36	1771	35
58	366	jg6743.t1	NP_523888.1	yellow-g	29.69	320	58
17	1625	jg6748.t1	NP_001259764.1	Abca3	41.04	1657	17
17	1625	jg6748.t1	NP_608445.2	ABCA	41.04	1657	17
17	1625	jg6748.t1	NP_001259765.1	Abca3	41.04	1657	17
72	241	jg6741.t1	NP_650289.2	yellow-e2	42.78	180	72
65	781	jg6744.t1	NP_001286856.1	Usp15-31	43.76	825	65
65	781	jg6744.t1	NP_001286855.1	Usp15-31	43.76	825	65
113	497	jg6739.t1	NP_651912.3	yellow-h	42.71	391	113

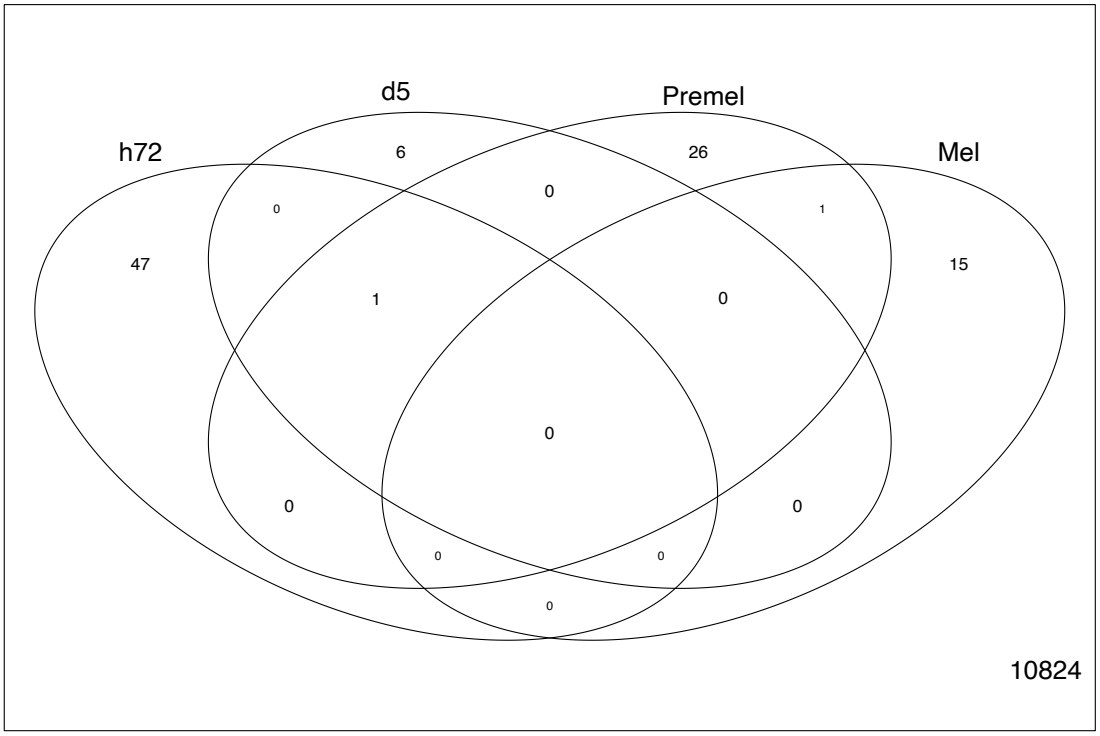
Supplementary Table 6.1 | BLAST results of *A. plantaginis* genes found in and around the QTL and GWAS peak. All BLAST hits were filtered by e-value < 1e-5. If two matches had the same e-value, the one with the highest % identity was chosen.

Supplementary Table 6.2 | List of differentially expressed genes found in the linkage group containing scaffold 419 (WW reference).

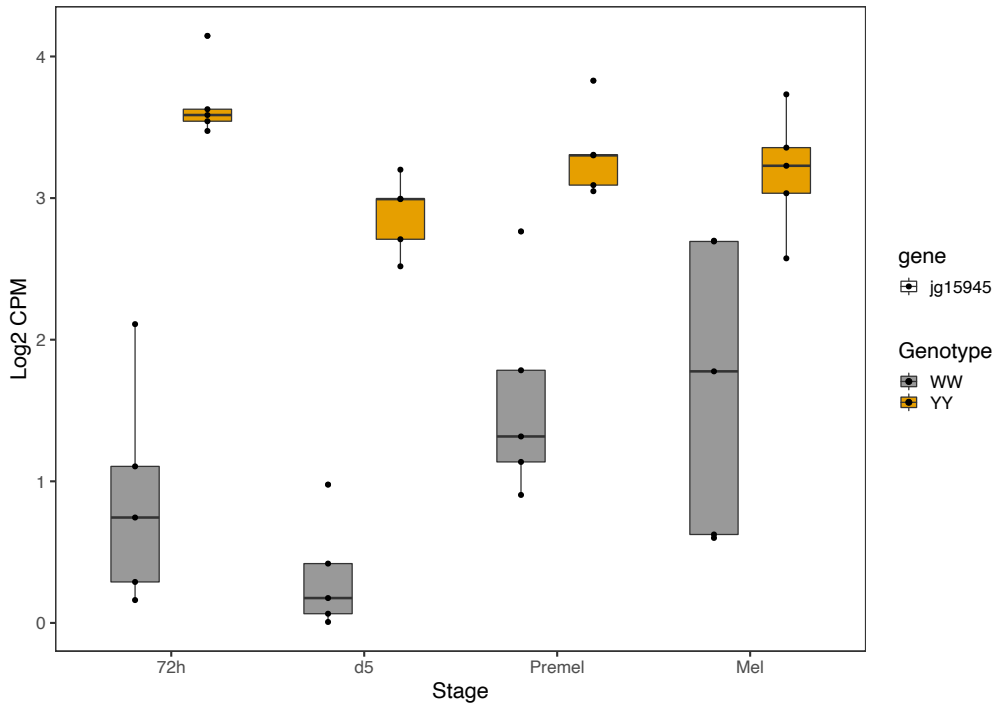
Genes	Log Fold Change	Average Expr.	t	P. Value	Adjusted P.Value	B	Stage
jg1153	-6.95	-3.46	-7.64	5.37E-09	5.86E-05	8.08	mel
jg1308	10.31	1.41	5.64	2.18E-06	1.19E-02	4.58	premel
jg1310	3.86	4.91	5.33	5.62E-06	2.05E-02	3.96	premel
jg14032	-1.70	6.31	-4.70	3.81E-05	3.78E-02	2.14	premel
jg14802	3.04	0.14	4.49	7.21E-05	4.01E-02	1.53	premel
jg15101	-3.88	-1.38	-6.56	1.34E-07	4.88E-04	6.41	mel
jg15103	1.64	2.04	4.73	3.54E-05	1.68E-02	2.26	72h
jg15168	2.21	1.02	5.19	8.67E-06	3.15E-02	3.25	5days
jg2034	2.49	1.94	4.38	1.02E-04	4.44E-02	1.29	premel
jg2035	2.95	1.93	4.32	1.19E-04	4.81E-02	1.15	premel
jg8680	3.42	-0.10	6.69	9.09E-08	9.46E-04	7.72	5days
jg9028	3.24	0.51	4.67	4.22E-05	3.89E-02	2.02	mel



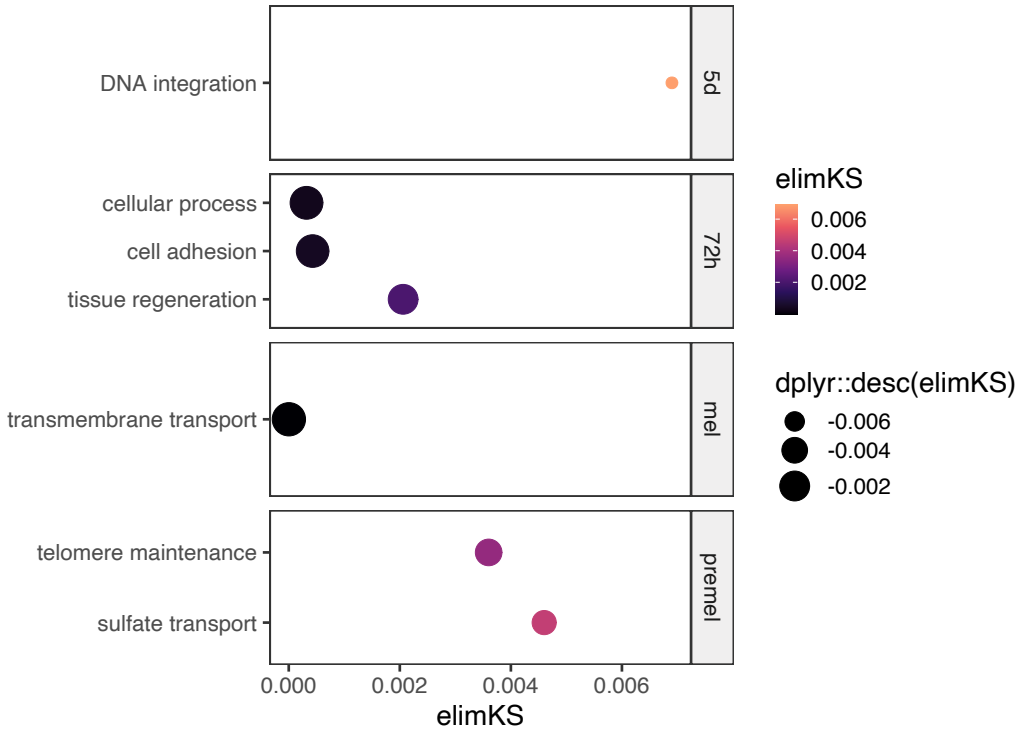
Supplementary Figure 6.3 | Wing development in *A. plantaginis.* At 5 days of pupal development, no colour has deposited in white (WW, **a**) or yellow individuals (yy, **b**)**.** At 7 days, white individuals show no colour, while yellow individuals present yellow colouration (**c**). After 8 days of development, white (**f-g**) and yellow (**h-i**) males show fully formed wings.



Supplementary Figure 6.2 | Venn-diagram of differentially expressed genes between male colour morphs in each developmental stage.



Supplementary Figure 6.3 | Gene expression in Log2 Counts Per Million of the gene jg15945 (*CG2129*). Expression between white (WW) and yellow (yy) morphs is significantly different (p.value<0.05) at 72 hours (72h), 5 days (5d) and pre-melanin stage (Premel), but nor at the melanin stage (Mel).



Supplementary Figure 6.4 Gene Ontology (GO) term analysis of gene expression by developmental stage in *A. plataginis* shows that there are not any GO terms shared between developmental stages.

7

Conclusions

Studies of colouration have been crucial to improve our understanding of adaptive evolution (Chapter 1). Wing phenotypes of butterflies and moths have provided compelling examples of adaptations, such as mimicry and warning colouration, and of their genetic basis (reviewed in (Jiggins et al., 2017). Through multiple studies of Lepidoptera species, two general patterns have started to emerge. First, that gene re-use for adaptive wing phenotypes is widespread in the Lepidoptera (Livraghi et al., 2021; Martin et al., 2014; Martin and Reed, 2014; Palmer and Kronforst, 2020; van't Hof et al., 2016). And second, that structural variation is key in the evolution of adaptive wing phenotypes (Jay et al., 2018; Joron et al., 2011; Kim et al., 2022; Livraghi et al., 2021; Martin et al., 2020; van't Hof et al., 2016; Woronik et al., 2019). In this thesis, I have used large genomic and transcriptomic datasets of *Hypolimnas* butterflies and the wood tiger moth, *Arctia plantaginis,* to explore these two generalities.

2.7 Gene re-use for adaptive wing phenotypes is widespread in the Lepidoptera

Repeated phenotypic evolution has been a topic of interest for evolutionary biologists for a long time, particularly with the aim of clarifying if convergent phenotypes are indeed the result of similar genetic changes (reviewed in (Stern, 2013)). Theory predicts that the most likely genetic changes that will spread through a population under directional selection, and thus to underlie adaptations, are those that produce the advantageous phenotype with the minimum number of changes and that have low pleiotropic effects (Gompel and Prud'homme, 2009; Stern, 2013, 2011; Stern and Orgogozo, 2008). Genes with a small number of functions in several species might have limited pleiotropic effects and evolve independently to control convergent phenotypes. This echoes the speculations of Haldane, who suggested that repeated evolution arose by a combination of similar selective pressures and shared genetic biases in the production of genotypic variations (Haldane, 1932).

In line with this, the shape of developmental pathways has been argued to be crucial in repeated evolution at the genetic level with 'input-output' genes sitting at central points of developmental pathways and acting as switches for more complex phenotypes (Chapter 1)(Gompel and Prud'homme, 2009; Martin and Courtier-Orgogozo, 2017; Stern, 2013, 2011; Stern and Orgogozo, 2008). This has been claimed to be of particular importance in phenotypes encompassing multiple traits and thus requiring the coordinated action of multiple genes. In those cases, 'input-output' genes have minimum pleiotropy, as they can alter the expression of the necessary downstream genes concertedly. Contrastingly, changes in patterning genes upstream of input-output genes would have effects that would be too broad, while mutations at genes downstream would have effects that would be too limited. The genes *shavenbaby* and *scute* in *Drosophila* species are considered examples of input-output genes, as they integrate patterning information to direct trichome and bristle development, which requires the coordinated effect of several genes (Stern and Orgogozo, 2008). Another possible example of input-output genes are the Bab transcription factors, which integrate segment and sex specific information to direct melanisation pattern in *Drosophila* species (Dembeck et al., 2015). In chapter 5, I show that convergent wing mimicry in two species of *Hypolimnas* butterflies is likely to be controlled by *cortex,* a gene that has been repeatedly linked to wing pattern phenotypes in Lepidoptera (Beldade et al., 2009; Hanly et al., 2022; Ito et al., 2016; Livraghi et al., 2021; Nadeau et al., 2016; van der Burg et al., 2020; van't Hof et al., 2019). *cortex* controls scale morphology in *Heliconius* butterflies acting as a switch between type II and type II scale types, like an input-output gene (Livraghi et al., 2021). Importantly, temporal and spatial expression differences in *cortex* have been hypothesised to create the differences in wing pattern. Given the number of species in which *cortex* controls wing phenotype, I speculate that *cortex* acts as an 'input-output' gene and this capacity to determine scale identity locally through cis-regulatory changes limits pleiotropic effects and thus favours the co-option of *cortex* for wing phenotypes. However, testing this hypothesis would be challenging. Crucially, I have shown that the loci associated with hindwing white colouration in *H. misippus* and *H. bolina* are not homologous, which suggests that different putative CRE are determining wing pattern. This is in line with results in *Heliconius,* in which distinct modular enhancers drive *cortex* expression and determine wing phenotype (Enciso-Romero et al., 2017; Livraghi et al., 2021). Finally, although *cortex* is a

strong candidate, the evidence I have provided is only correlational and thus it is possible that other genes around *cortex* are functionally important.

Similar to *cortex*, the transcription factor *optix* has also been linked to wing phenotypes in multiple species. In Chapter 5, I show that variation near *optix* is associated with orange colouration in *H. bolina. optix* is a likely candidate for the control of this phenotype as it has been shown to be involved in orange and red colour formation in multiple species of Lepidoptera, including several *Heliconius* species (Reed et al., 2011), *Vanessa cardui, Junonia coenia* and *Agraulis vanillae* (Zhang et al., 2017b). Interestingly, female *H. bolina* morphs showing a forewing orange spot are non-mimetic, which raises the question of what forces have led to the appearance and maintenance of this phenotype (Clarke and Sheppard, 1975). In *H. misippus* the presence of orange colouration is sexually dimorphic in clearly linked to mimicry and I hypothesise that it might have evolved by regulatory changes of *optix.* However, I could not test this hypothesis with the datasets produced in this thesis, as all *H. misippus* females are orange and thus no association studies can be performed.

The evolution of convergent phenotypes can also be the result of different genetic changes, particularly when similar phenotypes are generated by distinct molecular processes or pathways. For example, red colouration in animals can be produced by ommochromes as well as phaeomelanins (Chapter 1) and given that different pathways are involved in the production of these pigments, different genes are expected to be the target of adaptations relating to variation on these phenotypes. This would lead to convergence in phenotype but not in the genetic mechanisms. In chapter 5, I show that a *yellow* family gene, *valkea,* is associated with alternative reproductive morphs in the wood tiger moth, in which morphs differ in several traits including wing colouration, physiology, and behaviour (Gordon et al., 2015; Nokelainen et al., 2013, 2012; Rojas et al., 2017, 2015). *yellow* is known to be involved in melanin production in *Drosophila melanogaster* (Wittkopp et al., 2002) and has been associated with melanin pigments butterflies including in *V. cardui* and *Papilio xutus* (Zhang et al., 2017a), and moths including *Bicyclus anynana* (Matsuoka and Monteiro, 2018; Zhang et al., 2017a) and the silk moth, *Bombyx mori* (Ito et al., 2010), which would explain the differences in colouration in the wood tiger moth as those are pheomelanin based. This contrasts with the control of yellow pattern elements in *Heliconius,* which are controlled by

the joint effect of *cortex* and *aristaless1* (Livraghi et al., 2021; Westerman et al., 2018). However, yellow colouration in *Heliconius* is produced by 3-OH-K pigments, which requires the action of a different molecular pathway than melanins.

Crucially, it is clear how a *valkea,* a *yellow* family gene can be linked to colouration differences seen in males, but it is harder to explain how it can be involved in the determination of traits unrelated to colour such as the behavioural and physiological differences between the male morphs. One option could be that *valkea* acts as an 'input-output' gene controlling multiple traits through changes in expression of downstream genes. However, it is hard to understand how a colouration gene that is not a transcription factor could regulate the expression of target genes. Another option could be that the morphs are involved in an energetical trade off stemming from the production of pigment that has knock-on effects on other traits. An example of this is seen in *Colias* butterflies, where behavioural and physiological differences between the white *alba* and orange morphs might be a result of an energetical trade-off (Woronik et al., 2019). Finally, another option is that a cis-regulatory element near *valkea* controls the expression of a regulatory gene nearby, which in turn regulates *valkea.* Interestingly, a similar case is found in wall-lizards, in which differences between sympatric morphs that vary in colouration and in their ecology and behaviour are associated with two narrow regions near a pterin and carotenoid metabolism genes (Andrade et al., 2019).

Finally, in Chapter 1, I speculate that changes in complex phenotypes such as pattern are expected to happen more often in cis-regulatory regions, while changes on simpler phenotypes such as colour might happen in coding regions (Chapter 1). This is reminiscent of the cis-regulatory hypothesis that proposes that most mutations causing morphological variation are expected to reside in cis-regulatory, rather than coding regions of developmental genes (Stern and Orgogozo, 2008). Disruption of coding sequences are more likely to control colour rather than pattern, as the presence or absence of a pigment can be easily produced by a loss-of-function mutation. Examples of this are the multiple cases of coding mutations in the melanin gene *MC1R,* a transmembrane transporter gene that activates the pathway leading to the production of black and brown eumelanin (Gross et al., 2009; Mundy Nicholas I, 2005). In island flycatchers, coding mutations in the *MC1R* gene or *Agouti,* which encodes MC1R's antagonist, result in fully melanised plumage(Uy J. Albert C. et

al., 2016). Similarly, in budgerigars, variation in feather colouration is associated with a coding mutation in the gene *MuPKS,* which has been shown to produce yellow pigments when heterologously expressed in yeast (Cooke et al., 2017).

In contrast, complex changes in pattern, might require the fine tuning of the expression of regulatory genes, rather than an on and off switch. For example, variation in stripe patterning in two closely related species of cichlid fish, *Pundamilia nyererei* and *Haplochromis sauvagei,* are determined by a 1.1 kbp cis-regulatory element controlling the expression of *Agrp2,* a gene of the *Agouti* family that regulates melanosome aggregation (Kratochwil et al., 2018). Similarly, flower colour pattern differences in two species of monkeyflowers, *Mimulus lewisii* and *Mimulus cardinalis,* is controlled by a cis-regulatory element at the gene *LAR1*, which is involved in the switch between the synthesis of pink anthocyanins and colourless flavonols (Yuan et al., 2016). In line with this, in chapter 4 and 5, I show that regulatory changes at the M and A/S loci are associated with Batesian mimicry in *H. misippus.* Similarly, regulatory changes at the E and N loci are associated with white hindwing colouration and orange forewing colouration in *H. bolina.*

2.8 Batesian mimicry as a system for the study of convergent evolution

Mimicry systems offer an ideal opportunity to analyse the repeatability of evolution. In 1945, Goldschmidt speculated that mimicry could be produced by shared developmental systems between model and mimic, in which single changes could activate ancestral developmental pathways to generate the same phenotype in both species (Goldschmidt, 1945). Studies in *Heliconius* butterflies have shown that the multiple events of convergent evolution in wing phenotype in the genus are controlled by four main genes *cortex* (Nadeau et al., 2016)*, optix* (Reed et al., 2011)*, WntA* (Mazo-Vargas et al., 2017) and *Aristales1* (Westerman et al., 2018)*.* These results agree with Goldschmidt's hypothesis on the evolution of mimicry, but the phenotypic closeness of the co-mimics makes the repeated evolution of the control of mimicry unsurprising. Also, on many occasions, the adaptive alleles have been shared between species through introgression, rather than evolved independently multiple times (Jay et al., 2018; Kronforst et al., 2006; Pardo-Diaz et al., 2012; Zhang et al., 2016). Contrastingly,

Exploring cases of Batesian mimicry, where mimic and model are phylogenetically further apart, is crucial to understand if similar phenotypes are always made by similar genetic mechanisms as it has been suggested. In chapter 4 and 5, I identified the loci determining wing phenotype in the pantropical butterfly in *Hypolimnas misippus,* which is a Batesian mimic of the defended *Danaus chrysippus.* First, I have shown that the M locus that controls the switch between black and white apex and orange forewings is found in chromosome 29. The dominant allele *M* that generates black-and-white wing tips presents two large insertions of transposable elements that might be disrupting the function of a cis-regulatory element (CRE) of a nearby gene. Contrastingly, forewing phenotype in *D. chrysippus* is controlled by the C locus found in chromosome 15 as part of the BC supergene (Martin et al., 2020). An inversion of the BC supergene has reduced recombination between the B locus determining brown scales in the wings and the C locus. Second, I have identified the locus controlling hindwing white colouration to a region in chromosome 8 near the gene *cortex.* The same phenotype in *D. chrysippus* is controlled by the A locus found in chromosome 7 (Martin et al., 2020), which suggests that convergent wing phenotypes in *D. chrysippus* and *H. misippus* are not produced by homologous loci. This contrasts with the results in *Heliconius* butterflies where different co-mimic species have reached convergent phenotypes by changes in the same hotspot loci. Interestingly, brown background colour in *D. chrysippus* seems to be controlled by a gene of the yellow family, the same family as the gene determining the switch between morphs in the wood tiger moth. These suggests that brown colouration in *D. chrysippus* could be produced by melanin pigments, as *yellow* genes have been repeatedly shown to be involved in their production, and highlights the importance of understanding the pigment basis of colouration.

2.9 Structural variation is key in the evolution of adaptive phenotypes

Another generality that has started to emerge from studies of wing phenotype in butterflies and moths is the involvement of structural variation (SV) in these phenotypes (Joron et al., 2011; Kim et al., 2022; Livraghi et al., 2021; van't Hof et al., 2016; Woronik et al., 2019). More broadly, structural variants such as inversions have been shown to be crucial in the evolution of adaptive phenotypes. For example, alternative reproductive strategies in the ruff, which

show differences in plumage colouration, behaviour and body size, have been shown to be determined by a supergene containing around a 100 genes. An inversion in the alleles of two of the morphs ensures the joint inheritance of the co-adapted loci and leads to the maintenance of the polymorphism. Similarly, the two social forms of the fire ant showing differences in X and Y differ in a large inversion containing X genes (Wang et al., 2013). Smaller SVs, like transposable element insertions, have also been associated with adaptive evolution. For example, plumage variation between two subspecies of European crows, *Corvus corvus corvis* and *C. corvus corone, is associated in a 2.25 kb retrotransposon insertion that affects plumage (Kim et al., 2019). Similarly, TE activity has been shown to be crucial in the evolution of stress response in yeast (Weissensteiner et al., 2020) and have been suggested to play a role in adaptation in the invasive ant, Cardiocondyla obscurior* (Schrader et al., 2014). In Chapter 4 and 6 I show that SVs are found at the loci associated with wing polymorphism in both *H. misippus* and *A. plantaginis.* It is possible that reduced recombination caused by the duplication enhances the maintenance of the polymorphism in *A. plantaginis*. A similar case has been seen in *Primula* plants, in which flower morphology is determined by the S locus, a supergene containing five genes that evolved via stepwise duplications (Huu et al., 2020; Li et al., 2016). The thrum flower form carries the S locus, while it is absent in pin flowers, which has led to the reduction in recombination and the maintenance of the polymorphism (Li et al., 2016). Similarly, the TE insertions in the M locus of *H. misippus* identified in Chapter 4 might also be reducing recombination, which could lead to the accumulation of deleterious mutations through reduced purifying selection. This accumulation of deleterious mutations could be leading to associative overdominance, in which heterozygotes are favoured, which is seen in the supergene controlling wing phenotype in *H. numata* (Jay et al., 2021). However, it is unclear if such small SVs as the ones in the M locus could reduce recombination to the extent of limiting the effect of purifying selection in the region.

2.10 Future directions

My thesis has laid the foundation for future work in the evolutionary genetics of wing polymorphism in *Hypolimnas* butterflies and in the wood tiger moth, *A. plantaginis.* Firstly, the first future directions in *Hypolimnas* would be to functionally test the involvement of *cortex* in hindwing phenotype. Crucially, the putative CRE could be identified and targeted

following studies in *Heliconius*, in which a chromatin accessibility analysis was performed (Livraghi et al., 2021). That could help confirm the modularity of the CREs in *H. bolina* in controlling forewing and hindwing white separately. A crucial step would be to clarify the evolutionary history of white colouration in *H. misippus* and *H. bolina.* I have provided evidence that the loci controlling hindwing colouration are not homologous, but it is unclear if that is also the case for forewing white patterns elements. Furthermore, to fully understand the role of *cortex* in the evolution of wing phenotypes, scale structure and pigment analyses should be performed in *Hypolimnas.* This would clarify whether *cortex* has a similar function as in *Heliconius* in determining scale identity and would shed light on the forces behind repeated evolution of phenotypes.

A crucial aspect of the evolution of *H. misippus* is the maintenance of mimetic polymorphism in the population in the apparent absence of selection for mimicry. Now that the loci determining wing mimicry have been identified, long-term population studies could be carried out to estimate selection at these loci. Furthermore, I have hypothesised that forewing polymorphism in *H. misippus* might be maintained through associative overdominance. This hypothesis could be tested by exploring fitness differences in the distinct M locus genotypes. Future work could also explore how TE insertions can cause such differences in wing phenotype and clarify if they are disrupting a genetic network and reversing the phenotype to a more ancestral one. Finally, in Chapter 4, I hypothesised that the first mimetic morph to evolve in *H. misippus* is the all-orange *inaria* morph. By comparison to *H. bolina* in which orange colouration are associated with an inter-genic region near *optix,* I speculate that changes at *optix* are controlling orange pigmentation in *H. misippus* too. Identifying the genes determining orange colouration in *H. misippus* would help clarify how mimicry evolved in this species.

My results shown in Chapter 6 on the genetic basis of wing phenotype in the wood tiger moth are only the first steps towards understanding the molecular mechanisms controlling such a complex polymorphism. Performing functional testing and evaluating the effects on the co-adapted traits would enhance our understanding of the joint inheritance of the phenotypes and on the role of *valkea* in morph determination. Future work could bring together our

understanding of the genetic basis of the polymorphism with the ecological observations to explore its evolutionary history.

2.11 Concluding remarks

In summary, in this thesis, I have explored the extent of repeated evolution in the Lepidoptera and shown that often the same genes underlie adaptive phenotypes, such as in white pattern elements in *Hypolimnas* and male wing colouration in the wood tiger moth. These results suggest that biases in the repeatability of evolution might exist, with genes with certain developmental roles being repeatedly targeted, but this hypothesis remains to be tested. Particularly, more evidence and concrete examples are necessary on the role of 'input-output' genes in convergent evolution. Furthermore, I have shown that structural variants are central to wing pattern evolution, with TEs being repeatedly associated with wing phenotypes. These results add to the growing evidence from multiple other systems such as ants (Schrader et al., 2014), yeasts (Esnault et al., 2019), and birds (Weissensteiner et al., 2020) that TE insertions play a crucial role in adaptive evolution. Future population genetic studies testing the link between small SVs and TE insertions and adaptive loci are necessary to evaluate the statistical significance of this hypothesis. Finally, my results support the intriguing possibility of small SVs driving the maintenance of a polymorphism through associative overdominance as a result of reduced recombination. Assessments of the impact of small SVs on the local recombination rate in butterflies and other systems will clarify if small SVs can have a large enough effect to lead to the significant accumulation of deleterious mutations at a locus. All in all, this thesis has contributed to our understanding of the evolution of polymorphisms, exposing that adaptive evolution often relies in the same genes and that small structural variants can have large adaptive effects.

Bibliography

Albert, N.W., Davies, K.M., Lewis, D.H., Zhang, H., Montefiori, M., Brendolise, C., Boase, M.R., Ngo, H., Jameson, P.E., Schwinn, K.E., 2014. A Conserved Network of Transcriptional Activators and Repressors Regulates Anthocyanin Pigmentation in Eudicots. Plant Cell 26, 962–980. https://doi.org/10.1105/tpc.113.122069

Alexa, A., Rahnenführer, J., 2022. Gene set enrichment analysis with topGO. R package version 2.48.0. 26.

Anderson, T.M., vonHoldt, B.M., Candille, S.I., Musiani, M., Greco, C., Stahler, D.R., Smith, D.W., Padhukasahasram, B., Randi, E., Leonard, J.A., Bustamante, C.D., Ostrander, E.A., Tang, H., Wayne, R.K., Barsh, G.S., 2009. Molecular and Evolutionary History of Melanism in North American Gray Wolves. Science 323, 1339–1343. https://doi.org/10.1126/science.1165448

Ando, T., Matsuda, T., Goto, K., Hara, K., Ito, A., Hirata, J., Yatomi, J., Kajitani, R., Okuno, M., Yamaguchi, K., Kobayashi, M., Takano, T., Minakuchi, Y., Seki, M., Suzuki, Y., Yano, K., Itoh, T., Shigenobu, S., Toyoda, A., Niimi, T., 2018. Repeated inversions within a pannier intron drive diversification of intraspecific colour patterns of ladybird beetles. Nat Commun 9, 3843. https://doi.org/10.1038/s41467-018-06116-1

Andrade, P., Pinho, C., Lanuza, G.P. i de, Afonso, S., Brejcha, J., Rubin, C.-J., Wallerman, O., Pereira, P., Sabatino, S.J., Bellati, A., Pellitteri-Rosa, D., Bosakova, Z., Bunikis, I., Carretero, M.A., Feiner, N., Marsik, P., Paupério, F., Salvi, D., Soler, L., While, G.M., Uller, T., Font, E., Andersson, L., Carneiro, M., 2019. Regulatory changes in pterin and carotenoid genes underlie balanced color polymorphisms in the wall lizard. PNAS 116, 5633–5642. https://doi.org/10.1073/pnas.1820320116

Auton, A., Abecasis, G.R., Altshuler, D.M., Durbin, R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., Flicek, P., Gabriel, S.B., Gibbs, R.A., Green, E.D., Hurles, M.E., Knoppers, B.M., Korbel, J.O., Lander, E.S., Lee, C., Lehrach, H., Mardis, E.R., Marth, G.T., McVean, G.A., Nickerson, D.A., Schmidt, J.P., Sherry, S.T., Wang, J., Wilson, R.K., Gibbs, R.A., Boerwinkle, E., Doddapaneni, H., Han, Y., Korchina, V., Kovar, C., Lee, S., Muzny, D., Reid, J.G., Zhu, Y., Wang, J., Chang, Y., Feng, Q., Fang, X., Guo, X., Jian, M., Jiang, H., Jin, X., Lan, T., Li, G., Li, J., Li, Y., Liu, S., Liu, X., Lu, Y., Ma, X., Tang, M., Wang, B., Wang, G., Wu, H., Wu, R., Xu, X., Yin, Y., Zhang, D., Zhang, W., Zhao, J., Zhao, M., Zheng, X., Lander, E.S., Altshuler, D.M., Gabriel, S.B., Gupta, N., Gharani, N., Toji, L.H., Gerry, N.P., Resch, A.M., Flicek, P., Barker, J., Clarke, L., Gil, L., Hunt, S.E., Kelman, G., Kulesha, E., Leinonen, R., McLaren, W.M., Radhakrishnan, R., Roa, A., Smirnov, D., Smith, R.E., Streeter, I., Thormann, A., Toneva, I., Vaughan, B., Zheng-Bradley, X., Bentley, D.R., Grocock, R., Humphray, S.,

James, T., Kingsbury, Z., Lehrach, H., Sudbrak, R., Albrecht, M.W., Amstislavskiy, V.S., Borodina, T.A., Lienhard, M., Mertes, F., Sultan, M., Timmermann, B., Yaspo, M.-L., Mardis, E.R., Wilson, R.K., Fulton, L., Fulton, R., Sherry, S.T., Ananiev, V., Belaia, Z., Beloslyudtsev, D., Bouk, N., Chen, C., Church, D., Cohen, R., Cook, C., Garner, J., Hefferon, T., Kimelman, M., Liu, C., Lopez, J., Meric, P., O'Sullivan, C., Ostapchuk, Y., Phan, L., Ponomarov, S., Schneider, V., Shekhtman, E., Sirotkin, K., Slotta, D., Zhang, H., McVean, G.A., Durbin, R.M., Balasubramaniam, S., Burton, J., Danecek, P., Keane, T.M., Kolb-Kokocinski, A., McCarthy, S., Stalker, J., Quail, M., Schmidt, J.P., Davies, C.J., Gollub, J., Webster, T., Wong, B., Zhan, Y., Auton, A., Campbell, C.L., Kong, Y., Marcketta, A., Gibbs, R.A., Yu, F., Antunes, L., Bainbridge, M., Muzny, D., Sabo, A., Huang, Z., Wang, J., Coin, L.J.M., Fang, L., Guo, X., Jin, X., Li, G., Li, Q., Li, Y., Li, Z., Lin, H., Liu, B., Luo, R., Shao, H., Xie, Y., Ye, C., Yu, C., Zhang, F., Zheng, H., Zhu, H., Alkan, C., Dal, E., Kahveci, F., Marth, G.T., Garrison, E.P., Kural, D., Lee, W.-P., Fung Leong, W., Stromberg, M., Ward, A.N., Wu, J., Zhang, M., Daly, M.J., DePristo, M.A., Handsaker, R.E., Altshuler, D.M., Banks, E., Bhatia, G., del Angel, G., Gabriel, S.B., Genovese, G., Gupta, N., Li, H., Kashin, S., Lander, E.S., McCarroll, S.A., Nemesh, J.C., Poplin, R.E., Yoon, S.C., Lihm, J., Makarov, V., Clark, A.G., Gottipati, S., Keinan, A., Rodriguez-Flores, J.L., Korbel, J.O., Rausch, T., Fritz, M.H., Stütz, A.M., Flicek, P., Beal, K., Clarke, L., Datta, A., Herrero, J., McLaren, W.M., Ritchie, G.R.S., Smith, R.E., Zerbino, D., Zheng-Bradley, X., Sabeti, P.C., Shlyakhter, I., Schaffner, S.F., Vitti, J., Cooper, D.N., Ball, E.V., Stenson, P.D., Bentley, D.R., Barnes, B., Bauer, M., Keira Cheetham, R., Cox, A., Eberle, M., Humphray, S., Kahn, S., Murray, L., Peden, J., Shaw, R., Kenny, E.E., Batzer, M.A., Konkel, M.K., Walker, J.A., MacArthur, D.G., Lek, M., Sudbrak, R., Amstislavskiy, V.S., Herwig, R., Mardis, E.R., Ding, L., Koboldt, D.C., Larson, D., Ye, K., Gravel, S., The 1000 Genomes Project Consortium, Corresponding authors, Steering committee, Production group, Baylor College of Medicine, BGI-Shenzhen, Broad Institute of MIT and Harvard, Coriell Institute for Medical Research, European Molecular Biology Laboratory, E.B.I., Illumina, Max Planck Institute for Molecular Genetics, McDonnell Genome Institute at Washington University, US National Institutes of Health, University of Oxford, Wellcome Trust Sanger Institute, Analysis group, Affymetrix, Albert Einstein College of Medicine, Bilkent University, Boston College, Cold Spring Harbor Laboratory, Cornell University, European Molecular Biology Laboratory, Harvard University, Human Gene Mutation Database, Icahn School of Medicine at Mount Sinai, Louisiana State University, Massachusetts General Hospital, McGill University, National Eye Institute, N., 2015. A global reference for human genetic variation. Nature 526, 68–74. https://doi.org/10.1038/nature15393

Auwera, G.A.V. der, O'Connor, B.D., 2020. Genomics in the Cloud: Using Docker, GATK, and WDL in Terra. O'Reilly Media, Inc.

Bachtrog, D., Kirkpatrick, M., Mank, J.E., McDaniel, S.F., Pires, J.C., Rice, W., Valenzuela, N., 2011. Are all sex chromosomes created equal? Trends in Genetics 27, 350–357. https://doi.org/10.1016/j.tig.2011.05.005

Bachtrog, D., Mank, J.E., Peichel, C.L., Kirkpatrick, M., Otto, S.P., Ashman, T.-L., Hahn, M.W., Kitano, J., Mayrose, I., Ming, R., Perrin, N., Ross, L., Valenzuela, N., Vamosi, J.C., Consortium, T.T. of S., 2014. Sex Determination: Why So Many Ways of Doing It? PLOS Biology 12, e1001899. https://doi.org/10.1371/journal.pbio.1001899

Barrett, R.D.H., Laurent, S., Mallarino, R., Pfeifer, S.P., Xu, C.C.Y., Foll, M., Wakamatsu, K., Duke-Cohan, J.S., Jensen, J.D., Hoekstra, H.E., 2019. Linking a mutation to survival in wild mice. Science 363, 499–504. https://doi.org/10.1126/science.aav3824

Beldade, P., Brakefield, P.M., 2018. The genetics and evo–devo of butterfly wing patterns. Nature Reviews Genetics 3, 442–452. https://doi.org/10.1038/nrg818

Beldade, P., Saenko, S.V., Pul, N., Long, A.D., 2009. A Gene-Based Linkage Map for Bicyclus anynana Butterflies Allows for a Comprehensive Analysis of Synteny with the Lepidopteran Reference Genome. PLOS Genetics 5, e1000366. https://doi.org/10.1371/journal.pgen.1000366

Benson, W.W., 1972. Natural Selection for Miillerian Mimicry in Heliconius erato in Costa Rica. Science 176, 936–939. https://doi.org/10.1126/science.176.4037.936

Bernardi, G., 1974. Polymorphisme et mimétisme chez les Lépidoptères Rhopalocères. Mémoires de la Société Zoologique de France, Paris 37, 129–168.

Beukeboom, L.W., Perrin, N., 2014. The Evolution of Sex Determination. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199657148.001.0001

Bilandžija, H., Ma, L., Parkhurst, A., Jeffery, W.R., 2013. A Potential Benefit of Albinism in Astyanax Cavefish: Downregulation of the oca2 Gene Increases Tyrosine and Catecholamine Levels as an Alternative to Melanin Synthesis. PLOS ONE 8, e80823. https://doi.org/10.1371/journal.pone.0080823

BLAST, 2013. BLAST Basic Local Alignment Search Tool. https://doi.org/10.1006/jmbi.1990.9999

Blum, M., Chang, H.-Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., Nuka, G., Paysan-Lafosse, T., Qureshi, M., Raj, S., Richardson, L., Salazar, G.A., Williams, L., Bork, P., Bridge, A., Gough, J., Haft, D.H., Letunic, I., Marchler-Bauer, A., Mi, H., Natale, D.A., Necci, M., Orengo, C.A., Pandurangan, A.P., Rivoire, C., Sigrist, C.J.A., Sillitoe, I., Thanki, N., Thomas, P.D., Tosatto, S.C.E., Wu, C.H., Bateman, A., Finn, R.D., 2021. The InterPro protein families and domains database: 20 years on. Nucleic Acids Research 49, D344–D354. https://doi.org/10.1093/nar/gkaa977

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Brien Melanie N., Enciso-Romero Juan, Parnell Andrew J., Salazar Patricio A., Morochz Carlos, Chalá Darwin, Bainbridge Hannah E., Zinn Thomas, Curran Emma V., Nadeau Nicola J., 2019. Phenotypic variation in Heliconius erato crosses shows that iridescent structural colour is sex-linked and controlled by multiple genes. Interface Focus 9, 20180047. https://doi.org/10.1098/rsfs.2018.0047

Brien, M.N., Orteu, A., Yen, E.C., Galarza, J.A., Kirvesoja, J., Pakkanen, H., Wakamatsu, K., Jiggins, C.D., Mappes, J., 2022. Colour polymorphism associated with a gene duplication in male wood tiger moths. https://doi.org/10.1101/2022.04.29.490025

Broman, K.W., Gatti, D.M., Simecek, P., Furlotte, N.A., Prins, P., Sen, Ś., Yandell, B.S., Churchill, G.A., 2019. R/qtl2: Software for Mapping Quantitative Trait Loci with High-Dimensional Data and Multiparent Populations. Genetics 211, 495–502. https://doi.org/10.1534/genetics.118.301595

Brown, J.B., Boley, N., Eisman, R., May, G.E., Stoiber, M.H., Duff, M.O., Booth, B.W., Wen, J., Park, S., Suzuki, A.M., Wan, K.H., Yu, C., Zhang, D., Carlson, J.W., Cherbas, L., Eads, B.D., Miller, D., Mockaitis, K., Roberts, J., Davis, C.A., Frise, E., Hammonds, A.S., Olson, S., Shenker, S., Sturgill, D., Samsonova, A.A., Weiszmann, R., Robinson, G., Hernandez, J., Andrews, J., Bickel, P.J., Carninci, P., Cherbas, P., Gingeras, T.R., Hoskins, R.A., Kaufman, T.C., Lai, E.C., Oliver, B., Perrimon, N., Graveley, B.R., Celniker, S.E., 2014. Diversity and dynamics of the Drosophila transcriptome. Nature 512, 393–399. https://doi.org/10.1038/nature12962

Brown, K.S., Benson, W.W., 1974. Adaptive Polymorphism Associated with Multiple Müllerian Mimicry in Heliconius numata (Lepid. Nymph.). Biotropica 6, 205–228. https://doi.org/10.2307/2989666

Brůna, T., Hoff, K.J., 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. https://doi.org/10.1093/nargab/lqaa108

Cabanettes, F., Klopp, C., 2018. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. PeerJ 6, e4958. https://doi.org/10.7717/peerj.4958

Caro, T., Allen, W.L., 2017. Interspecific visual signalling in animals and plants: a functional classification. Philosophical Transactions of the Royal Society B: Biological Sciences 372, 20160344. https://doi.org/10.1098/rstb.2016.0344

Carroll, S.B., Grenier, J.K., Weatherbee, S.D., 2005. From DNA to diversity: molecular genetics and the evolution of animal design / Sean B. Carroll, Jennifer K. Grenier, Scott D. Weatherbee., 2nd ed. ed. Blackwell, Malden, Mass. ; Oxford.

Carvalho, C.M.B., Lupski, J.R., 2016. Mechanisms underlying structural variant formation in genomic disorders. Nat Rev Genet 17, 224–238. https://doi.org/10.1038/nrg.2015.25

Casacuberta, E., González, J., 2013. The impact of transposable elements in environmental adaptation. Molecular Ecology 22, 1503–1517. https://doi.org/10.1111/mec.12170

Castro, S.D., Peronnet, F., Gilles, J.-F., Mouchel-Vielh, E., Gibert, J.-M., 2018. bric à brac (bab), a central player in the gene regulatory network that mediates thermal plasticity of pigmentation in Drosophila melanogaster. PLOS Genetics 14, e1007573. https://doi.org/10.1371/journal.pgen.1007573

Challis, R., Richards, E., Rajan, J., Cochrane, G., Blaxter, M., 2020. BlobToolKit – Interactive Quality Assessment of Genome Assemblies. G3 Genes|Genomes|Genetics 10, 1361–1374. https://doi.org/10.1534/g3.119.400908

Charlat, S., Duplouy, A., Hornett, E.A., Dyson, E.A., Davies, N., Roderick, G.K., Wedell, N., Hurst, G.D., 2009. The joint evolutionary histories of Wolbachia and mitochondria in Hypolimnas bolina. BMC Evolutionary Biology 9, 64. https://doi.org/10.1186/1471-2148-9-64

Chen, S., Zhou, Y., Chen, Y., Gu, J., 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34, i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Cheng, H., Concepcion, G.T., Feng, X., Zhang, H., Li, H., 2020. Haplotype-resolved de novo assembly with phased assembly graphs. arXiv:2008.01237 [q-bio].

Chouteau, M., Arias, M., Joron, M., 2016. Warning signals are under positive frequency-dependent selection in nature. Proceedings of the National Academy of Sciences 113, 2164–2169. https://doi.org/10.1073/pnas.1519216113

Chouteau, M., Llaurens, V., Piron-Prunier, F., Joron, M., 2017. Polymorphism at a mimicry supergene maintained by opposing frequency-dependent selection pressures. Proceedings of the National Academy of Sciences 114, 8325–8329. https://doi.org/10.1073/pnas.1702482114

Clarke, C., Sheppard, P.M., 1975. The genetics of the mimetic butterfly Hypolimnas bolina (L.). Philosophical transactions of the Royal Society of London. Series B, Biological sciences. https://doi.org/10.1098/rstb.1975.0084

Connahs, H., Rhen, T., Simmons, R.B., 2016. Transcriptome analysis of the painted lady butterfly, Vanessa cardui during wing color pattern development. BMC Genomics 17, 270. https://doi.org/10.1186/s12864-016-2586-5

Conte, G.L., Arnegard, M.E., Peichel, C.L., Schluter, D., 2012. The probability of genetic parallelism and convergence in natural populations. Proceedings of the Royal Society B: Biological Sciences 279, 5039–5047. https://doi.org/10.1098/rspb.2012.2146

Cooke, T.F., Fischer, C.R., Wu, P., Jiang, T.X., Xie, K.T., Kuo, J., Doctorov, E., Zehnder, A., Khosla, C., Chuong, C.M., Bustamante, C.D., 2017. Genetic Mapping and Biochemical

Basis of Yellow Feather Pigmentation in Budgerigars. Cell 171, 427-439.e21. https://doi.org/10.1016/j.cell.2017.08.016

Corl, A., Bi, K., Luke, C., Challa, A.S., Stern, A.J., Sinervo, B., Nielsen, R., 2018. The Genetic Basis of Adaptation following Plastic Changes in Coloration in a Novel Environment. Current Biology 28, 2970-2977.e7. https://doi.org/10.1016/j.cub.2018.06.075

Crawford, N.G., Kelly, D.E., Hansen, M.E.B., Beltrame, M.H., Fan, S., Bowman, S.L., Jewett, E., Ranciaro, A., Thompson, S., Lo, Y., Pfeifer, S.P., Jensen, J.D., Campbell, M.C., Beggs, W., Hormozdiari, F., Mpoloka, S.W., Mokone, G.G., Nyambo, T., Meskel, D.W., Belay, G., Haut, J., Program, N.C.S., Rothschild, H., Zon, L., Zhou, Y., Kovacs, M.A., Xu, M., Zhang, T., Bishop, K., Sinclair, J., Rivas, C., Elliot, E., Choi, J., Li, S.A., Hicks, B., Burgess, S., Abnet, C., Watkins-Chow, D.E., Oceana, E., Song, Y.S., Eskin, E., Brown, K.M., Marks, M.S., Loftus, S.K., Pavan, W.J., Yeager, M., Chanock, S., Tishkoff, S.A., 2017. Loci associated with skin pigmentation identified in African populations. Science 358. https://doi.org/10.1126/science.aan8433

Cuthill, I.C., Allen, W.L., Arbuckle, K., Caspers, B., Chaplin, G., Hauber, M.E., Hill, G.E., Jablonski, N.G., Jiggins, C.D., Kelber, A., Mappes, J., Marshall, J., Merrill, R., Osorio, D., Prum, R., Roberts, N.W., Roulin, A., Rowland, H.M., Sherratt, T.N., Skelhorn, J., Speed, M.P., Stevens, M., Stoddard, M.C., Stuart-Fox, D., Talas, L., Tibbetts, E., Caro, T., 2017. The biology of color. Science 357. https://doi.org/10.1126/science.aan0221

Dalíková, M., Zrzavá, M., Hladová, I., Nguyen, P., Šonský, I., Flegrová, M., Kubíčková, S., Voleníková, A., Kawahara, A.Y., Peters, R.S., Marec, F., 2017. New Insights into the Evolution of the W Chromosome in Lepidoptera. Journal of Heredity 108, 709–719. https://doi.org/10.1093/jhered/esx063

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., Li, H., 2021. Twelve years of SAMtools and BCFtools. Gigascience 10, giab008. https://doi.org/10.1093/gigascience/giab008

Davey, J.W., Barker, S.L., Rastas, P.M., Pinharanda, A., Martin, S.H., Durbin, R., McMillan, W.O., Merrill, R.M., Jiggins, C.D., 2017. No evidence for maintenance of a sympatric Heliconius species barrier by chromosomal inversions. Evolution Letters 1, 138–154. https://doi.org/10.1002/evl3.12

Davey, J.W., Chouteau, M., Barker, S.L., Maroja, L., Baxter, S.W., Simpson, F., Merrill, R.M., Joron, M., Mallet, J., Dasmahapatra, K.K., Jiggins, C.D., 2016. Major Improvements to the Heliconius melpomene Genome Assembly Used to Confirm 10 Chromosome Fusion Events in 6 Million Years of Butterfly Evolution. G3 (Bethesda) 6, 695–708. https://doi.org/10.1534/g3.115.023655

Davies, R.W., Flint, J., Myers, S., Mott, R., 2016. Rapid genotype imputation from sequence without reference panels. Nat Genet 48, 965–969. https://doi.org/10.1038/ng.3594

Dembeck, L.M., Huang, W., Magwire, M.M., Lawrence, F., Lyman, R.F., Mackay, T.F.C., 2015. Genetic Architecture of Abdominal Pigmentation in Drosophila melanogaster. PLOS Genetics 11, e1005163. https://doi.org/10.1371/journal.pgen.1005163

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R., 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21. https://doi.org/10.1093/bioinformatics/bts635

Drapeau, M.D., Albert, S., Kucharski, R., Prusko, C., Maleszka, R., 2006. Evolution of the Yellow/Major Royal Jelly Protein family and the emergence of social behavior in honey bees. Genome Res. 16, 1385–1394. https://doi.org/10.1101/gr.5012006

Dudchenko, O., Batra, S.S., Omer, A.D., Nyquist, S.K., Hoeger, M., Durand, N.C., Shamim, M.S., Machol, I., Lander, E.S., Aiden, A.P., Aiden, E.L., 2017. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. Science 356, 92–95. https://doi.org/10.1126/science.aal3327

Dudchenko, O., Shamim, M.S., Batra, S.S., Durand, N.C., Musial, N.T., Mostofa, R., Pham, M., Hilaire, B.G.S., Yao, W., Stamenova, E., Hoeger, M., Nyquist, S.K., Korchina, V., Pletch, K., Flanagan, J.P., Tomaszewicz, A., McAloose, D., Estrada, C.P., Novak, B.J., Omer, A.D., Aiden, E.L., 2018. The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. bioRxiv 254797. https://doi.org/10.1101/254797

Durand, N.C., Robinson, J.T., Shamim, M.S., Machol, I., Mesirov, J.P., Lander, E.S., Aiden, E.L., 2016a. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. Cell Systems 3, 99–101. https://doi.org/10.1016/j.cels.2015.07.012

Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S.P., Huntley, M.H., Lander, E.S., Aiden, E.L., 2016b. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. cels 3, 95–98. https://doi.org/10.1016/j.cels.2016.07.002

Dyson, E.A., Kamath, M.K., Hurst, G.D.D., 2002. Wolbachia infection associated with all-female broods in Hypolimnas bolina (Lepidoptera: Nymphalidae): evidence for horizontal transmission of a butterfly male killer. Heredity 88, 166–171. https://doi.org/10.1038/sj.hdy.6800021

Edge, P., Bafna, V., Bansal, V., 2017. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. Genome Res 27, 801–812. https://doi.org/10.1101/gr.213462.116

Elyanow, R., Wu, H.-T., Raphael, B.J., 2018. Identifying structural variants using linked-read sequencing data. Bioinformatics 34, 353–360. https://doi.org/10.1093/bioinformatics/btx712

Emms, D.M., Kelly, S., 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biology 20, 238. https://doi.org/10.1186/s13059-019-1832-y

Enciso-Romero, J., Pardo-Díaz, C., Martin, S.H., Arias, C.F., Linares, M., McMillan, W.O., Jiggins, C.D., Salazar, C., 2017. Evolution of novel mimicry rings facilitated by adaptive introgression in tropical butterflies. Molecular Ecology 26, 5160–5172. https://doi.org/10.1111/mec.14277

Esnault, C., Lee, M., Ham, C., Levin, H.L., 2019. Transposable element insertions in fission yeast drive adaptation to environmental stress. Genome Res. 29, 85–95. https://doi.org/10.1101/gr.239699.118

Espeland, M., Breinholt, J., Willmott, K.R., Warren, A.D., Vila, R., Toussaint, E.F.A., Maunsell, S.C., Aduse-Poku, K., Talavera, G., Eastwood, R., Jarzyna, M.A., Guralnick, R., Lohman, D.J., Pierce, N.E., Kawahara, A.Y., 2018. A Comprehensive and Dated Phylogenomic Analysis of Butterflies. Current Biology 28, 770-778.e5. https://doi.org/10.1016/j.cub.2018.01.061

Faria, R., Johannesson, K., Butlin, R.K., Westram, A.M., 2019. Evolving Inversions. Trends in Ecology & Evolution 34, 239–248. https://doi.org/10.1016/j.tree.2018.12.005

Ferguson, L., Lee, S.F., Chamberlain, N., Nadeau, N., Joron, M., Baxter, S., Wilkinson, P., Papanicolaou, A., Kumar, S., Kee, T.J., Clark, R., Davidson, C., Glithero, R., Beasley, H., Vogel, H., Ffrench-Constant, R., Jiggins, C., 2010. Characterization of a hotspot for mimicry: Assembly of a butterfly wing transcriptome to genomic sequence at the HmYb/Sb locus. Molecular Ecology 19, 240–254. https://doi.org/10.1111/j.1365-294X.2009.04475.x

Ferguson, L.C., Green, J., Surridge, A., Jiggins, C.D., 2011. Evolution of the insect yellow gene family. Molecular Biology and Evolution 28, 257–272. https://doi.org/10.1093/molbev/msq192

Fisher, R.A., 1930. The genetical theory of natural selection. The Clarendon Press, Oxford.

Ford, E.B., 1965. Genetic polymorphism. Genetic polymorphism.

Fraïsse, C., Picard, M.A.L., Vicoso, B., 2017. The deep conservation of the Lepidoptera Z chromosome suggests a non-canonical origin of the W. Nat Commun 8, 1486. https://doi.org/10.1038/s41467-017-01663-5

Fukova, I., Collins, S., Timmermans, M.J.T.N., Jiggins, C.D., Vogel, H., Papanicolaou, A., Clark, R., Thompson, M.J., Vogler, A.P., Joron, M., Baxter, S.W., Heckel, D.G., Ffrench-Constant, R.H., 2014. Comparative genomics of the mimicry switch in Papilio dardanus. Proceedings of the Royal Society B: Biological Sciences 281, 20140465.

Galarza, J.A., Nokelainen, O., Ashrafi, R., Hegna, R.H., Mappes, J., 2014. Temporal relationship between genetic and warning signal variation in the aposematic wood tiger moth (Parasemia plantaginis). Molecular Ecology 23, 4939–4957. https://doi.org/10.1111/mec.12913

Gautier, M., Yamaguchi, J., Foucaud, J., Loiseau, A., Ausset, A., Facon, B., Gschloessl, B., Lagnel, J., Loire, E., Parrinello, H., Severac, D., Lopez-Roques, C., Donnadieu, C., Manno, M., Berges, H., Gharbi, K., Lawson-Handley, L., Zang, L.-S., Vogel, H., Estoup, A., Prud'homme, B., 2018. The Genomic Basis of Color Pattern Polymorphism in the Harlequin Ladybird. Current Biology 28, 3296-3302.e7. https://doi.org/10.1016/j.cub.2018.08.023

Gibert, J.-M., Mouchel-Vielh, E., De Castro, S., Peronnet, F., 2016. Phenotypic Plasticity through Transcriptional Regulation of the Evolutionary Hotspot Gene tan in Drosophila melanogaster. PLoS Genet. 12, e1006218. https://doi.org/10.1371/journal.pgen.1006218

Giska, I., Farelo, L., Pimenta, J., Seixas, F.A., Ferreira, M.S., Marques, J.P., Miranda, I., Letty, J., Jenny, H., Hackländer, K., Magnussen, E., Melo-Ferreira, J., 2019. Introgression drives repeated evolution of winter coat color polymorphism in hares. PNAS 116, 24150–24156. https://doi.org/10.1073/pnas.1910471116

Goldschmidt, R.B., 1945. Mimetic Polymorphism, a Controversial Chapter of Darwinism (Concluded). The Quarterly Review of Biology 20, 205–230. https://doi.org/10.1086/394883

Gompel, N., Prud'homme, B., 2009. The causes of repeated genetic evolution. Developmental Biology, Special Section: Evolution of Developmental Regulatory Systems 332, 36–47. https://doi.org/10.1016/j.ydbio.2009.04.040

Gordon, I.J., 1987. Natural-Selection For Rare and Mimetic Color Pattern Combinations in Wild Populations of the Diadem Butterfly, Hypolimnas-Misippus L. Biological Journal of the Linnean Society 31, 1–23.

Gordon, I.J., Edmunds, M., EDgar, J.A., Lawrence, J., Smith, D.A.S., 2010. Linkage disequilibrium and natural selection for mimicry in the Batesian mimic Hypolimnas misippus (L.) (Lepidoptera: Nymphalidae) in the Afrotropics. Biological Journal of the Linnean Society 100, 180–194. https://doi.org/10.1111/j.1095-8312.2010.01412.x

Gordon, I.J., Smith, D.A.S., 1998. Body size and colour-pattern genetics in the polymorphic mimetic butterfly Hypolimnas misippus (L.). Heredity 80, 62–69. https://doi.org/10.1038/sj.hdy.6882590

Gordon, J.S., Smith, D. a S., 1989. Genetics of the mimetic African butterfly Hypolimnas misippus: hindwing polymorphism. Heredity 63, 409–425.

Gordon, S.P., Kokko, H., Rojas, B., Nokelainen, O., Mappes, J., 2015. Colour polymorphism torn apart by opposing positive frequency-dependent selection, yet maintained in space. Journal of Animal Ecology 84, 1555–1564. https://doi.org/10.1111/1365-2656.12416

Grabherr, M.G., Russell, P., Meyer, M., Mauceli, E., Alföldi, J., Di Palma, F., Lindblad-Toh, K., 2010. Genome-wide synteny through highly sensitive sequence alignment: Satsuma. Bioinformatics 26, 1145–1151. https://doi.org/10.1093/bioinformatics/btq102

Graham, S.M., Watt, W.B., Gall, L.F., 1980. Metabolic resource allocation vs. mating attractiveness: Adaptive pressures on the “alba” polymorphism of Colias butterflies. PNAS 77, 3615–3619. https://doi.org/10.1073/pnas.77.6.3615

Gross, J.B., Borowsky, R., Tabin, C.J., 2009. A Novel Role for Mc1r in the Parallel Evolution of Depigmentation in Independent Populations of the Cavefish Astyanax mexicanus. PLoS Genet 5, e1000326. https://doi.org/10.1371/journal.pgen.1000326

Gu, Z., Gu, L., Eils, R., Schlesner, M., Brors, B., 2014. circlize implements and enhances circular visualization in R. Bioinformatics 30, 2811–2812. https://doi.org/10.1093/bioinformatics/btu393

Guan, D., 2022. Purge_Dups.

Haldane, J.B.S., 1932. The causes of evolution. London : Longmans, 1932., London.

Hanly, J.J., Livraghi, L., Heryanto, C., McMillan, W.O., Jiggins, C.D., Gilbert, L.E., Martin, A., 2022. A large deletion at the cortex locus eliminates butterfly wing patterning. G3 Genes|Genomes|Genetics 12, jkac021. https://doi.org/10.1093/g3journal/jkac021

Hébant, C., Lee, D.W., 1984. Ultrastructural Basis and Developmental Control of Blue Iridescence in Selaginella Leaves. American Journal of Botany 71, 216–219. https://doi.org/10.1002/j.1537-2197.1984.tb12506.x

Hegna, R.H., Galarza, J.A., Mappes, J., 2015. Global phylogeography and geographical variation in warning coloration of the wood tiger moth (Parasemia plantaginis). Journal of Biogeography 42, 1469–1481. https://doi.org/10.1111/jbi.12513

Hickey, G., Paten, B., Earl, D., Zerbino, D., Haussler, D., 2013. HAL: a hierarchical format for storing and analyzing multiple genome alignments. Bioinformatics 29, 1341–1342. https://doi.org/10.1093/bioinformatics/btt128

Hill, Geoffrey E., Hill, Geoffrey Edward, McGraw, K.J., 2006. Bird Coloration: Function and evolution. Harvard University Press.

Hines, H.M., Counterman, B.A., Papa, R., Albuquerque de Moura, P., Cardoso, M.Z., Linares, M., Mallet, J., Reed, R.D., Jiggins, C.D., Kronforst, M.R., McMillan, W.O., 2011. Wing patterning gene redefines the mimetic history of Heliconius butterflies. Proceedings of the National Academy of Sciences 108, 19666–19671. https://doi.org/10.1073/pnas.1110096108

Ho, S.S., Urban, A.E., Mills, R.E., 2020. Structural variation in the sequencing era. Nat Rev Genet 21, 171–189. https://doi.org/10.1038/s41576-019-0180-9

Hoekstra, H.E., Hirschmann, R.J., Bundey, R.A., Insel, P.A., Crossland, J.P., 2006. A Single Amino Acid Mutation Contributes to Adaptive Beach Mouse Color Pattern. Science 313, 101–104. https://doi.org/10.1126/science.1126121

Hoffmann, A.A., Rieseberg, L.H., 2008. Revisiting the Impact of Inversions in Evolution: From Population Genetic Markers to Drivers of Adaptive Shifts and Speciation? Annual Review of Ecology, Evolution, and Systematics 39, 21–42. https://doi.org/10.1146/annurev.ecolsys.39.110707.173532

Hopkins, R., Rausher, M.D., 2012. Pollinator-Mediated Selection on Flower Color Allele Drives Reinforcement. Science 335, 1090–1092. https://doi.org/10.1126/science.1215198

Hopkins, R., Rausher, M.D., 2011. Identification of two genes causing reinforcement in the Texas wildflower Phlox drummondii. Nature 469, 411–414. https://doi.org/10.1038/nature09641

Hornett, E.A., Charlat, S., Duplouy, A.M.R., Davies, N., Roderick, G.K., Wedell, N., Hurst, G.D.D., 2006. Evolution of male-killer suppression in a natural population. PLoS Biology 4, 1643–1648. https://doi.org/10.1371/journal.pbio.0040283

Hornett, E.A., Moran, B., Reynolds, L.A., Charlat, S., Tazzyman, S., Wedell, N., Jiggins, C.D., Hurst, G.D.D., 2014. The Evolution of Sex Ratio Distorter Suppression Affects a 25 cM Genomic Region in the Butterfly Hypolimnas bolina. PLoS Genetics. https://doi.org/10.1371/journal.pgen.1004822

Huu, C.N., Keller, B., Conti, E., Kappel, C., Lenhard, M., 2020. Supergene evolution via stepwise duplications and neofunctionalization of a floral-organ identity gene. Proceedings of the National Academy of Sciences 117, 23148–23157. https://doi.org/10.1073/pnas.2006296117

Hysi, P.G., Valdes, A.M., Liu, F., Furlotte, N.A., Evans, D.M., Bataille, V., Visconti, A., Hemani, G., McMahon, G., Ring, S.M., Smith, G.D., Duffy, D.L., Zhu, G., Gordon, S.D., Medland, S.E., Lin, B.D., Willemsen, G., Hottenga, J.J., Vuckovic, D., Girotto, G., Gandin, I., Sala, C., Concas, M.P., Brumat, M., Gasparini, P., Toniolo, D., Cocca, M., Robino, A., Yazar, S., Hewitt, A.W., Chen, Y., Zeng, C., Uitterlinden, A.G., Ikram, M.A., Hamer, M.A., Duijn, C.M. van, Nijsten, T., Mackey, D.A., Falchi, M., Boomsma, D.I., Martin, N.G., Hinds, D.A., Kayser, M., Spector, T.D., 2018. Genome-wide association meta-analysis of individuals of European ancestry identifies new loci explaining a substantial fraction of hair color variation and heritability. Nat Genet 50, 652–656. https://doi.org/10.1038/s41588-018-0100-5

Iakovishina, D., Janoueix-Lerosey, I., Barillot, E., Regnier, M., Boeva, V., 2016. SV-Bay: structural variant detection in cancer genomes using a Bayesian approach with correction for GC-content and read mappability. Bioinformatics 32, 984–992. https://doi.org/10.1093/bioinformatics/btv751

Iijima, T., Kajitani, R., Komata, S., Lin, C.-P., Sota, T., Itoh, T., Fujiwara, H., 2018. Parallel evolution of Batesian mimicry supergene in two Papilio butterflies, P. polytes and P. memnon. Science Advances 4, eaao5416. https://doi.org/10.1126/sciadv.aao5416

Ito, K., Katsuma, S., Kuwazaki, S., Jouraku, A., Fujimoto, T., Sahara, K., Yasukochi, Y., Yamamoto, K., Tabunoki, H., Yokoyama, T., Kadono-Okuda, K., Shimada, T., 2016. Mapping and recombination analysis of two moth colour mutations, Black moth and Wild wing spot, in the silkworm Bombyx mori. Heredity 116, 52–59. https://doi.org/10.1038/hdy.2015.69

Ito, K., Katsuma, S., Yamamoto, K., Kadono-Okuda, K., Mita, K., Shimada, T., 2010. Yellow-e Determines the Color Pattern of Larval Head and Tail Spots of the Silkworm Bombyx mori. J Biol Chem 285, 5624–5629. https://doi.org/10.1074/jbc.M109.035741

Järvi, V.V., Burg, K.R.L. van der, Reed, R.D., 2019. Seasonal Plasticity in Junonia coenia (Nymphalidae): Linking Wing Color, Temperature Dynamics, and Behavior. lepi 73, 34–42. https://doi.org/10.18473/lepi.73i1.a5

Jay, P., Chouteau, M., Whibley, A., Bastide, H., Llaurens, V., Parrinello, H., Joron, M., 2019. Mutation accumulation in chromosomal inversions maintains wing pattern polymorphism in a butterfly. bioRxiv 736504. https://doi.org/10.1101/736504

Jay, P., Chouteau, M., Whibley, A., Bastide, H., Parrinello, H., Llaurens, V., Joron, M., 2021. Mutation load at a mimicry supergene sheds new light on the evolution of inversion polymorphisms. Nat Genet. https://doi.org/10.1038/s41588-020-00771-1

Jay, P., Whibley, A., Frézal, L., Rodríguez de Cara, M.Á., Nowell, R.W., Mallet, J., Dasmahapatra, K.K., Joron, M., 2018. Supergene Evolution Triggered by the Introgression of a Chromosomal Inversion. Current Biology 28, 1839-1845.e3. https://doi.org/10.1016/j.cub.2018.04.072

Jiggins, C.D., 2017. The Ecology and Evolution of Heliconius Butterflies. Oxford University Press.

Jiggins, C.D., 2016. A flamboyant behavioral polymorphism is controlled by a lethal supergene. Nat Genet 48, 7–8. https://doi.org/10.1038/ng.3472

Jiggins, C.D., Wallbank, R.W.R., Hanly, J.J., 2017. Waiting in the wings: What can we learn about gene co-option from the diversification of butterfly wing patterns? Philosophical Transactions of the Royal Society B: Biological Sciences 372. https://doi.org/10.1098/rstb.2015.0485

Johanson, H.C., Chen, W., Wicking, C., Sturm, R.A., 2010. Inheritance of a novel mutated allele of the *OCA2* gene associated with high incidence of oculocutaneous albinism in a Polynesian community. Journal of Human Genetics 55, 103–111. https://doi.org/10.1038/jhg.2009.130

Jones, F.C., Grabherr, M.G., Chan, Y.F., Russell, P., Mauceli, E., Johnson, J., Swofford, R., Pirun, M., Zody, M.C., White, S., Birney, E., Searle, S., Schmutz, J., Grimwood, J., Dickson, M.C., Myers, R.M., Miller, C.T., Summers, B.R., Knecht, A.K., Brady, S.D., Zhang, H., Pollen, A.A., Howes, T., Amemiya, C., Lander, E.S., Di Palma, F., Lindblad-Toh, K., Kingsley, D.M., 2012. The genomic basis of adaptive evolution in threespine sticklebacks. Nature 484, 55–61. https://doi.org/10.1038/nature10944

Jones, M.R., Mills, L.S., Alves, P.C., Callahan, C.M., Alves, J.M., Lafferty, D.J.R., Jiggins, F.M., Jensen, J.D., Melo-Ferreira, J., Good, J.M., 2018. Adaptive introgression underlies polymorphic seasonal camouflage in snowshoe hares. Science 360, 1355–1358. https://doi.org/10.1126/science.aar5273

Joron, M., Frezal, L., Jones, R.T., Chamberlain, N.L., Lee, S.F., Haag, C.R., Whibley, A., Becuwe, M., Baxter, S.W., Ferguson, L., Wilkinson, P.A., Salazar, C., Davidson, C., Clark, R., Quail, M.A., Beasley, H., Glithero, R., Lloyd, C., Sims, S., Jones, M.C., Rogers, J., Jiggins, C.D., Ffrench-Constant, R.H., 2011. Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. Nature 477, 203–206. https://doi.org/10.1038/nature10341

Joron, M., Wynne, I.R., Lamas, G., Mallet, J., 1999. Variable Selection and the Coexistence of Multiple mimetic forms of the Butterfly Heliconius numata. Evolutionary Ecology 13, 721–754. https://doi.org/10.1023/A:1010875213123

Jory, A., Estella, C., Giorgianni, M.W., Slattery, M., Laverty, T.R., Rubin, G.M., Mann, R.S., 2012. A Survey of 6,300 Genomic Fragments for cis-Regulatory Activity in the Imaginal Discs of Drosophila melanogaster. Cell Reports 2, 1014–1024. https://doi.org/10.1016/j.celrep.2012.09.010

Kapan, D.D., 2001. Three-butterfly system provides a field test of müllerian mimicry. Nature 409, 338–340. https://doi.org/10.1038/35053066

Kawahara, A.Y., Plotkin, D., Espeland, M., Meusemann, K., Toussaint, E.F.A., Donath, A., Gimnich, F., Frandsen, P.B., Zwick, A., Dos Reis, M., Barber, J.R., Peters, R.S., Liu, S., Zhou, X., Mayer, C., Podsiadlowski, L., Storer, C., Yack, J.E., Misof, B., Breinholt, J.W., 2019. Phylogenomics reveals the evolutionary timing and pattern of butterflies and moths. Proc Natl Acad Sci U S A 116, 22657–22663. https://doi.org/10.1073/pnas.1907847116

Keilwagen, J., Hartung, F., Grau, J., 2019. GeMoMa: Homology-Based Gene Prediction Utilizing Intron Position Conservation and RNA-seq Data, in: Kollmar, M. (Ed.), Gene Prediction: Methods and Protocols, Methods in Molecular Biology. Springer, New York, NY, pp. 161–177. https://doi.org/10.1007/978-1-4939-9173-0_9

Kellenberger, R.T., Byers, K.J.R.P., De Brito Francisco, R.M., Staedler, Y.M., LaFountain, A.M., Schönenberger, J., Schiestl, F.P., Schlüter, P.M., 2019. Emergence of a floral colour

polymorphism by pollinator-mediated overdominance. Nat Commun 10, 63. https://doi.org/10.1038/s41467-018-07936-x

Kent, W.J., 2002. BLAT - The BLAST-like alignment tool. Genome Research. https://doi.org/10.1101/gr.229202. Article published online before March 2002

Kim, K.-W., De-Kayne, R., Gordon, I.J., Omufwoko, K.S., Martins, D.J., ffrench-Constant, R., Martin, S.H., 2022. Stepwise evolution of a butterfly supergene via duplication and inversion. Philosophical Transactions of the Royal Society B: Biological Sciences 377, 20210207. https://doi.org/10.1098/rstb.2021.0207

Kim, K.-W., Jackson, B.C., Zhang, H., Toews, D.P.L., Taylor, S.A., Greig, E.I., Lovette, I.J., Liu, M.M., Davison, A., Griffith, S.C., Zeng, K., Burke, T., 2019. Genetics and evidence for balancing selection of a sex-linked colour polymorphism in a songbird. Nat Commun 10, 1852. https://doi.org/10.1038/s41467-019-09806-6

Kolmogorov, M., Armstrong, J., Raney, B.J., Streeter, I., Dunn, M., Yang, F., Odom, D., Flicek, P., Keane, T.M., Thybert, D., Paten, B., Pham, S., 2018. Chromosome assembly of large and complex genomes using multiple references. Genome Res. https://doi.org/10.1101/gr.236273.118

Kolmogorov, M., Raney, B., Paten, B., Pham, S., 2014. Ragout—a reference-assisted assembly tool for bacterial genomes. Bioinformatics 30, i302–i309. https://doi.org/10.1093/bioinformatics/btu280

Kondrashov, F.A., 2012. Gene duplication as a mechanism of genomic adaptation to a changing environment. Proceedings of the Royal Society B: Biological Sciences 279, 5048–5057. https://doi.org/10.1098/rspb.2012.1108

Kratochwil, C.F., Liang, Y., Gerwin, J., Woltering, J.M., Urban, S., Henning, F., Machado-Schiaffino, G., Hulsey, C.D., Meyer, A., 2018a. Agouti-related peptide 2 facilitates convergent evolution of stripe patterns across cichlid fish radiations. Science 362, 457–460. https://doi.org/10.1126/science.aao6809

Kratochwil, C.F., Liang, Y., Gerwin, J., Woltering, J.M., Urban, S., Henning, F., Machado-Schiaffino, G., Hulsey, C.D., Meyer, A., 2018b. Agouti-related peptide 2 facilitates convergent evolution of stripe patterns across cichlid fish radiations. Science 362, 457–460. https://doi.org/10.1126/science.aao6809

Kronforst, M.R., Young, L.G., Blume, L.M., Gilbert, L.E., 2006. Multilocus Analyses of Admixture and Introgression Among Hybridizing Heliconius Butterflies. Evolution 60, 1254–1268. https://doi.org/10.1111/j.0014-3820.2006.tb01203.x

Krueger, F., 2015. Trim galore. A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files.

Kumar, S., Suleski, M., Craig, J.M., Kasprowicz, A.E., Sanderford, M., Li, M., Stecher, G., Hedges, S.B., 2022. TimeTree 5: An Expanded Resource for Species Divergence

Times. Molecular Biology and Evolution 39, msac174. https://doi.org/10.1093/molbev/msac174

Kunte, K., 2009. The diversity and evolution of batesian mimicry in papilio swallowtail butterflies. Evolution. https://doi.org/10.1111/j.1558-5646.2009.00752.x

Kunte, K., Zhang, W., Tenger-Trolander, A., Palmer, D.H., Martin, A., Reed, R.D., Mullen, S.P., Kronforst, M.R., 2014. Doublesex is a mimicry supergene. Nature 507, 229–232. https://doi.org/10.1038/nature13112

Küpper, C., Stocks, M., Risse, J.E., Dos Remedios, N., Farrell, L.L., McRae, S.B., Morgan, T.C., Karlionova, N., Pinchuk, P., Verkuil, Y.I., Kitaysky, A.S., Wingfield, J.C., Piersma, T., Zeng, K., Slate, J., Blaxter, M., Lank, D.B., Burke, T., 2015a. A supergene determines highly divergent male reproductive morphs in the ruff. Nature Genetics 48, 79–83. https://doi.org/10.1038/ng.3443

Küpper, C., Stocks, M., Risse, J.E., Dos Remedios, N., Farrell, L.L., McRae, S.B., Morgan, T.C., Karlionova, N., Pinchuk, P., Verkuil, Y.I., Kitaysky, A.S., Wingfield, J.C., Piersma, T., Zeng, K., Slate, J., Blaxter, M., Lank, D.B., Burke, T., 2015b. A supergene determines highly divergent male reproductive morphs in the ruff. Nature Genetics 48, 79–83. https://doi.org/10.1038/ng.3443

Lamichhaney, S., Fan, G., Widemo, F., Gunnarsson, U., Thalmann, D.S., Hoeppner, M.P., Kerje, S., Gustafson, U., Shi, C., Zhang, H., Chen, W., Liang, X., Huang, L., Wang, J., Liang, E., Wu, Q., Lee, S.M.-Y., Xu, X., Höglund, J., Liu, X., Andersson, L., 2016. Structural genomic changes underlie alternative reproductive strategies in the ruff (Philomachus pugnax). Nat Genet 48, 84–88. https://doi.org/10.1038/ng.3430

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., Carey, V.J., 2013. Software for Computing and Annotating Genomic Ranges. PLOS Computational Biology 9, e1003118. https://doi.org/10.1371/journal.pcbi.1003118

Layer, R.M., Chiang, C., Quinlan, A.R., Hall, I.M., 2014. LUMPY: a probabilistic framework for structural variant discovery. Genome Biology 15, R84. https://doi.org/10.1186/gb-2014-15-6-r84

Lehtonen, P.K., Laaksonen, T., Artemyev, A.V., Belskii, E., Berg, P.R., Both, C., Buggiotti, L., Bureš, S., Burgess, M.D., Bushuev, A.V., Krams, I., Moreno, J., Mägi, M., Nord, A., Potti, J., Ravussin, P.-A., Sirkiä, P.M., Sætre, G.-P., Winkel, W., Primmer, C.R., 2012. Candidate genes for colour and vision exhibit signals of selection across the pied flycatcher (Ficedula hypoleuca) breeding range. Heredity 108, 431–440. https://doi.org/10.1038/hdy.2011.93

Lewis, J.J., Cicconardi, F., Martin, S.H., Reed, R.D., Danko, C.G., Montgomery, S.H., 2021. The Dryas iulia Genome Supports Multiple Gains of a W Chromosome from a B Chromosome in Butterflies. Genome Biology and Evolution 13. https://doi.org/10.1093/gbe/evab128

Lewis, J.J., Geltman, R.C., Pollak, P.C., Rondem, K.E., Belleghem, S.M.V., Hubisz, M.J., Munn, P.R., Zhang, L., Benson, C., Mazo-Vargas, A., Danko, C.G., Counterman, B.A., Papa, R., Reed, R.D., 2019. Parallel evolution of ancient, pleiotropic enhancers underlies butterfly wing pattern mimicry. PNAS 116, 24174–24183. https://doi.org/10.1073/pnas.1907068116

Li, H., 2022. lh3/yak.

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup, 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, J., Cocker, J.M., Wright, J., Webster, M.A., McMullan, M., Dyer, S., Swarbreck, D., Caccamo, M., Oosterhout, C. van, Gilmartin, P.M., 2016. Genetic architecture and evolution of the S locus supergene in Primula vulgaris. Nature Plants 2, 16188. https://doi.org/10.1038/nplants.2016.188

Liao, Y., Smyth, G.K., Shi, W., 2013. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30, 923–930.

Linnen, C.R., Kingsley, E.P., Jensen, J.D., Hoekstra, H.E., 2009. On the Origin and Spread of an Adaptive Allele in Deer Mice. Science 325, 1095–1098. https://doi.org/10.1126/science.1175826

Linnen, C.R., Poh, Y.-P., Peterson, B.K., Barrett, R.D.H., Larson, J.G., Jensen, J.D., Hoekstra, H.E., 2013. Adaptive evolution of multiple traits through multiple mutations at a single gene. Science 339, 1312–1316. https://doi.org/10.1126/science.1233213

Liu, B.H., 1997. Statistical Genomics: Linkage, Mapping, and QTL Analysis. CRC Press.

Livraghi, L., Hanly, J.J., Van Bellghem, S.M., Montejo-Kovacevich, G., van der Heijden, E.S., Loh, L.S., Ren, A., Warren, I.A., Lewis, J.J., Concha, C., Hebberecht, L., Wright, C.J., Walker, J.M., Foley, J., Goldberg, Z.H., Arenas-Castro, H., Salazar, C., Perry, M.W., Papa, R., Martin, A., McMillan, W.O., Jiggins, C.D., 2021. Cortex cis-regulatory switches establish scale colour identity and pattern diversity in Heliconius. eLife 10, e68549. https://doi.org/10.7554/eLife.68549

Lopes, R.J., Johnson, J.D., Toomey, M.B., Ferreira, M.S., Araujo, P.M., Melo-Ferreira, J., Andersson, L., Hill, G.E., Corbo, J.C., Carneiro, M., 2016. Genetic Basis for Red Coloration in Birds. Current Biology 26, 1427–1434. https://doi.org/10.1016/j.cub.2016.03.076

Lou, R.N., Jacobs, A., Wilder, A.P., Therkildsen, N.O., 2021. A beginner's guide to low-coverage whole genome sequencing for population genomics. Molecular Ecology 30, 5966–5993. https://doi.org/10.1111/mec.16077

Lukhtanov, V.A., 2000. Sex chromatin and sex chromosome systems in nonditrysian Lepidoptera (Insecta). Journal of Zoological Systematics and Evolutionary Research 38, 73–79. https://doi.org/10.1046/j.1439-0469.2000.382130.x

Mahmoud, M., Gobet, N., Cruz-Dávalos, D.I., Mounier, N., Dessimoz, C., Sedlazeck, F.J., 2019. Structural variant calling: the long and the short of it. Genome Biology 20, 246. https://doi.org/10.1186/s13059-019-1828-7

Mallet, J., Barton, N.H., 1989. Strong Natural Selection in a Warning-Color Hybrid Zone. Evolution 43, 421–431. https://doi.org/10.1111/j.1558-5646.1989.tb04237.x

Marks, P., Garcia, S., Barrio, A.M., Belhocine, K., Bernate, J., Bharadwaj, R., Bjornson, K., Catalanotti, C., Delaney, J., Fehr, A., Fiddes, I.T., Galvin, B., Heaton, H., Herschleb, J., Hindson, C., Holt, E., Jabara, C.B., Jett, S., Keivanfar, N., Kyriazopoulou-Panagiotopoulou, S., Lek, M., Lin, B., Lowe, A., Mahamdallie, S., Maheshwari, S., Makarewicz, T., Marshall, J., Meschi, F., O'Keefe, C.J., Ordonez, H., Patel, P., Price, A., Royall, A., Ruark, E., Seal, S., Schnall-Levin, M., Shah, P., Stafford, D., Williams, S., Wu, I., Xu, A.W., Rahman, N., MacArthur, D., Church, D.M., 2019. Resolving the full spectrum of human genome variation using Linked-Reads. Genome Res. 29, 635–645. https://doi.org/10.1101/gr.234443.118

Marques, D.A., Meier, J.I., Seehausen, O., 2019. A Combinatorial View on Speciation and Adaptive Radiation. Trends in Ecology & Evolution 34, 531–544. https://doi.org/10.1016/j.tree.2019.02.008

Marsh, N.A., Clarke, C.A., Rothschild, M., Kellett, D.N., 1977. Hypolimnas bolina (L.), a mimic of danaid butterflies, and its model Euploea core (Cram.) store cardioactive substances. Nature 268, 726–728. https://doi.org/10.1038/268726a0

Marshall, G.A.K., Poulton, E.B., 1902. Five Years' Observations and Experiments (1896-1901) on the Bionomics of South African Insects: Chiefly Directed to the Investigations of Mimicry and Warning Colours. Royal Entomological Society of London.

Martin, A., Courtier-Orgogozo, V., 2017. Morphological Evolution Repeatedly Caused by Mutations in Signaling Ligand Genes, in: Sekimura, T., Nijhout, H.F. (Eds.), Diversity and Evolution of Butterfly Wing Patterns: An Integrative Approach. Springer Singapore, Singapore, pp. 59–87. https://doi.org/10.1007/978-981-10-4956-9_4

Martin, A., McCulloch, K.J., Patel, N.H., Briscoe, A.D., Gilbert, L.E., Reed, R.D., 2014. Multiple recent co-options of Optix associated with novel traits in adaptive butterfly wing radiations. EvoDevo 5, 7. https://doi.org/10.1186/2041-9139-5-7

Martin, A., Orgogozo, V., 2013. The Loci of Repeated Evolution: A Catalog of Genetic Hotspots of Phenotypic Variation. Evolution 67, 1235–1250. https://doi.org/10.1111/evo.12081

Martin, A., Papa, R., Nadeau, N.J., Hill, R.I., Counterman, B.A., Halder, G., Jiggins, C.D., Kronforst, M.R., Long, A.D., McMillan, W.O., Reed, R.D., 2012. Diversification of complex butterfly wing patterns by repeated regulatory evolution of a Wnt ligand. Proceedings of the National Academy of Sciences 109, 12632–12637. https://doi.org/10.1073/pnas.1204800109

Martin, A., Reed, R.D., 2014. Wnt signaling underlies evolution and development of the butterfly wing pattern symmetry systems. Developmental Biology 395, 367–378. https://doi.org/10.1016/j.ydbio.2014.08.031

Martin, S.H., Singh, K.S., Gordon, I.J., Omufwoko, K.S., Collins, S., Warren, I.A., Munby, H., Brattström, O., Traut, W., Martins, D.J., Smith, D.A.S., Jiggins, C.D., Bass, C., ffrench-Constant, R.H., 2020. Whole-chromosome hitchhiking driven by a male-killing endosymbiont. PLOS Biology 18, e3000610. https://doi.org/10.1371/journal.pbio.3000610

Massey, J.H., Chung, D., Siwanowicz, I., Stern, D.L., Wittkopp, P.J., 2019. The yellow gene influences Drosophila male mating success through sex comb melanization. eLife 8, e49388. https://doi.org/10.7554/eLife.49388

Massey, J.H., Wittkopp, P.J., 2016. The Genetic Basis of Pigmentation Differences Within and Between Drosophila Species, in: Current Topics in Developmental Biology. Elsevier, pp. 27–61. https://doi.org/10.1016/bs.ctdb.2016.03.004

Matsuoka, Y., Monteiro, A., 2018. Melanin Pathway Genes Regulate Color and Morphology of Butterfly Wing Scales. Cell Reports 24, 56–65. https://doi.org/10.1016/j.celrep.2018.05.092

Mazo-Vargas, A., Concha, C., Livraghi, L., Massardo, D., Wallbank, R.W.R., Zhang, L., Papador, J.D., Martinez-Najera, D., Jiggins, C.D., Kronforst, M.R., Breuker, C.J., Reed, R.D., Patel, N.H., McMillan, W.O., Martin, A., 2017. Macroevolutionary shifts of WntA function potentiate butterfly wing-pattern diversity. Proceedings of the National Academy of Sciences 114, 10701–10706. https://doi.org/10.1073/pnas.1708149114

McMillan, W.O., Livraghi, L., Concha, C., Hanly, J.J., 2020. From Patterning Genes to Process: Unraveling the Gene Regulatory Networks That Pattern Heliconius Wings. Frontiers in Ecology and Evolution 8.

Meier, J.I., Salazar, P.A., Kučka, M., Davies, R.W., Dréau, A., Aldás, I., Box Power, O., Nadeau, N.J., Bridle, J.R., Rolian, C., Barton, N.H., McMillan, W.O., Jiggins, C.D., Chan, Y.F., 2021. Haplotype tagging reveals parallel formation of hybrid races in two butterfly

species. Proceedings of the National Academy of Sciences 118, e2015005118. https://doi.org/10.1073/pnas.2015005118

Miyazaki, S., Okada, Y., Miyakawa, H., Tokuda, G., Cornette, R., Koshikawa, S., Maekawa, K., Miura, T., 2014. Sexually Dimorphic Body Color Is Regulated by Sex-Specific Expression of Yellow Gene in Ponerine Ant, Diacamma Sp. PLOS ONE 9, e92875. https://doi.org/10.1371/journal.pone.0092875

Moest, M., Belleghem, S.M.V., James, J.E., Salazar, C., Martin, S.H., Barker, S.L., Moreira, G.R.P., Mérot, C., Joron, M., Nadeau, N.J., Steiner, F.M., Jiggins, C.D., 2020. Selective sweeps on novel and introgressed variation shape mimicry loci in a butterfly adaptive radiation. PLOS Biology 18, e3000597. https://doi.org/10.1371/journal.pbio.3000597

Moran, N.A., Jarvik, T., 2010. Lateral Transfer of Genes from Fungi Underlies Carotenoid Production in Aphids. Science 328, 624–627. https://doi.org/10.1126/science.1187113

Morgan, M.D., Pairo-Castineira, E., Rawlik, K., Canela-Xandri, O., Rees, J., Sims, D., Tenesa, A., Jackson, I.J., 2018. Genome-wide study of hair colour in UK Biobank explains most of the SNP heritability. Nat Commun 9, 1–10. https://doi.org/10.1038/s41467-018-07691-z

Morisse, P., Legeai, F., Lemaitre, C., 2021. LEVIATHAN: efficient discovery of large structural variants by leveraging long-range information from Linked-Reads data. https://doi.org/10.1101/2021.03.25.437002

Muller, H.J., 1918. Genetic Variability, Twin Hybrids and Constant Hybrids, in a Case of Balanced Lethal Factors. Genetics 3, 422–499.

Mundy Nicholas I, 2005. A window on the genetics of evolution: MC1R and plumage colouration in birds. Proceedings of the Royal Society B: Biological Sciences 272, 1633–1640. https://doi.org/10.1098/rspb.2005.3107

Mundy, N.I.I., Stapley, J., Bennison, C., Tucker, R., Twyman, H., Kim, K.W., Burke, T., Birkhead, T.R.R., Andersson, S., Slate, J., 2016. Red Carotenoid Coloration in the Zebra Finch Is Controlled by a Cytochrome P450 Gene Cluster. Current Biology 26, 1435–1440. https://doi.org/10.1016/j.cub.2016.04.047

Nadeau, N.J., Pardo-Diaz, C., Whibley, A., Supple, M.A., Saenko, S.V., Wallbank, R.W.R., Wu, G.C., Maroja, L., Ferguson, L., Hanly, J.J., Hines, H., Salazar, C., Merrill, R.M., Dowling, A.J., ffrench-Constant, R.H., Llaurens, V., Joron, M., McMillan, W.O., Jiggins, C.D., 2016. The gene cortex controls mimicry and crypsis in butterflies and moths. Nature 534, 106–110. https://doi.org/10.1038/nature17961

Nadeau, N.J., Ruiz, M., Salazar, P., Counterman, B., Medina, J.A., Ortiz-Zuazaga, H., Morrison, A., McMillan, W.O., Jiggins, C.D., Papa, R., 2014. Population genomics of

parallel hybrid zones in the mimetic butterflies, H. melpomene and H. erato. Genome Res 24, 1316–1333. https://doi.org/10.1101/gr.169292.113

Nisar, N., Li, L., Lu, S., Khin, N.C., Pogson, B.J., 2015. Carotenoid Metabolism in Plants. Molecular Plant, Plant Metabolism and Synthetic Biology 8, 68–82. https://doi.org/10.1016/j.molp.2014.12.007

Nishikawa, H., Iijima, T., Kajitani, R., Yamaguchi, J., Ando, T., Suzuki, Y., Sugano, S., Fujiyama, A., Kosugi, S., Hirakawa, H., Tabata, S., Ozaki, K., Morimoto, H., Ihara, K., Obara, M., Hori, H., Itoh, T., Fujiwara, H., 2015. A genetic mechanism for female-limited Batesian mimicry in Papilio butterfly. Nature Genetics. https://doi.org/10.1038/ng.3241

Nokelainen, O., Galarza, J.A., Kirvesoja, J., Suisto, K., Mappes, J., 2022. Genetic colour variation visible for predators and conspecifics is concealed from humans in a polymorphic moth. Journal of Evolutionary Biology 35, 467–478. https://doi.org/10.1111/jeb.13994

Nokelainen, O., Hegna, R.H., Reudler, J.H., Lindstedt, C., Mappes, J., 2012. Trade-off between warning signal efficacy and mating success in the wood tiger moth. Proceedings of the Royal Society B: Biological Sciences 279, 257–265. https://doi.org/10.1098/rspb.2011.0880

Nokelainen, O., Lindstedt, C., Mappes, J., 2013. Environment-mediated morph-linked immune and life-history responses in the aposematic wood tiger moth. Journal of Animal Ecology 82, 653–662.

Nosil, P., Villoutreix, R., Carvalho, C.F. de, Farkas, T.E., Soria-Carrasco, V., Feder, J.L., Crespi, B.J., Gompert, Z., 2018. Natural selection and the predictability of evolution in Timema stick insects. Science 359, 765–770. https://doi.org/10.1126/science.aap9125

O'Hara, R.B., 2005. Comparing the effects of genetic drift and fluctuating selection on genotype frequency changes in the scarlet tiger moth. Proceedings of the Royal Society B: Biological Sciences 272, 211–217. https://doi.org/10.1098/rspb.2004.2929

Ohta, T., Kimura, M., 1970. Development of associative overdominance through linkage disequilibrium in finite populations*. Genetics Research 16, 165–177. https://doi.org/10.1017/S0016672300002391

Orr, H.A., 2005. The genetic theory of adaptation: a brief history. Nat Rev Genet 6, 119–127. https://doi.org/10.1038/nrg1523

Orr, H.A., 1998. The Population Genetics of Adaptation: The Distribution of Factors Fixed During Adaptive Evolution. Evolution 52, 935–949. https://doi.org/10.1111/j.1558-5646.1998.tb01823.x

Orteu, A., Jiggins, C.D., 2020. The genomics of coloration provides insights into adaptive evolution. Nature Reviews Genetics 21, 461–475. https://doi.org/10.1038/s41576-020-0234-z

Paaby, A.B., Rockman, M.V., 2013. The many faces of pleiotropy. Trends in Genetics 29, 66–73. https://doi.org/10.1016/j.tig.2012.10.010

Palmer, D.H., Kronforst, M.R., 2020. A shared genetic basis of mimicry across swallowtail butterflies points to ancestral co-option of doublesex. Nat Commun 11, 6. https://doi.org/10.1038/s41467-019-13859-y

Pamilo, P., Pálsson, S., 1998. Associative overdominance, heterozygosity and fitness. Heredity 81, 381–389. https://doi.org/10.1046/j.1365-2540.1998.00395.x

Papa, R., Kapan, D.D., Counterman, B.A., Maldonado, K., Lindstrom, D.P., Reed, R.D., Nijhout, H.F., Hrbek, T., McMillan, W.O., 2013. Multi-Allelic Major Effect Genes Interact with Minor Effect QTLs to Control Adaptive Color Pattern Variation in Heliconius erato. PLoS ONE 8, e57033. https://doi.org/10.1371/journal.pone.0057033

Pardo-Diaz, C., Salazar, C., Baxter, S.W., Merot, C., Figueiredo-Ready, W., Joron, M., McMillan, W.O., Jiggins, C.D., 2012. Adaptive Introgression across Species Boundaries in Heliconius Butterflies. PLOS Genetics 8, e1002752. https://doi.org/10.1371/journal.pgen.1002752

Pardo-Diaz, C., Salazar, C., Jiggins, C.D., 2015. Towards the identification of the loci of adaptive evolution. Methods in Ecology and Evolution 6, 445–464. https://doi.org/10.1111/2041-210X.12324

Paten, B., Earl, D., Nguyen, N., Diekhans, M., Zerbino, D., Haussler, D., 2011. Cactus: Algorithms for genome multiple sequence alignment. Genome Res. 21, 1512–1528. https://doi.org/10.1101/gr.123356.111

Payer, L.M., Steranka, J.P., Yang, W.R., Kryatova, M., Medabalimi, S., Ardeljan, D., Liu, C., Boeke, J.D., Avramopoulos, D., Burns, K.H., 2017. Structural variants caused by Alu insertions are associated with risks for many human diseases. Proc Natl Acad Sci U S A 114, E3984–E3992. https://doi.org/10.1073/pnas.1704117114

Picelli, S., Björklund, Å.K., Reinius, B., Sagasser, S., Winberg, G., Sandberg, R., 2014. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. Genome Res. 24, 2033–2040. https://doi.org/10.1101/gr.177881.114

Pickar-Oliver, A., Gersbach, C.A., 2019. The next generation of CRISPR–Cas technologies and applications. Nat Rev Mol Cell Biol 20, 490–507. https://doi.org/10.1038/s41580-019-0131-5

Pierre, J., 1980. Variation géographique du polymorphisme et du mimétisme de Danaus chrysippus et d'Hypolimnas misippus (Lépidoptères Rhopaloceres) en Afrique et en Asie.

Protas, M.E., Patel, N.H., 2008. Evolution of Coloration Patterns. Annual Review of Cell and Developmental Biology 24, 425–446. https://doi.org/10.1146/annurev.cellbio.24.110707.175302

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., Sham, P.C., 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics. https://doi.org/10.1086/519795

Quicke, D.L.J., 2017. Mimicry, Crypsis, Masquerade and other Adaptive Resemblances. John Wiley & Sons.

Rastas, P., 2020. Lep-Anchor: automated construction of linkage map anchored haploid genomes. Bioinformatics 36, 2359–2364. https://doi.org/10.1093/bioinformatics/btz978

Rastas, P., 2017. Lep-MAP3: robust linkage mapping even for low-coverage whole genome sequencing data. Bioinformatics 33, 3726–3732. https://doi.org/10.1093/bioinformatics/btx494

Rausch, T., Zichner, T., Schlattl, A., Stütz, A.M., Benes, V., Korbel, J.O., 2012. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 28, i333–i339. https://doi.org/10.1093/bioinformatics/bts378

Rebeiz, M., Pool, J.E., Kassner, V.A., Aquadro, C.F., Carroll, S.B., 2009. Stepwise Modification of a Modular Enhancer Underlies Adaptation in a Drosophila Population. Science 326, 1663–1667. https://doi.org/10.1126/science.1178357

Reed, R.D., Papa, R., Martin, A., Hines, H.M., Kronforst, M.R., Chen, R., Halder, G., Nijhout, H.F., Mcmillan, W.O., 2011. optix Drives the Repeated. Science 333, 1137–1141. https://doi.org/10.1126/science.1208227

Rice, E.S., Green, R.E., 2019. New Approaches for Genome Assembly and Scaffolding. Annual Review of Animal Biosciences 7, 17–40. https://doi.org/10.1146/annurev-animal-020518-115344

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., Smyth, G.K., 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic acids research 43, e47–e47.

Rogers, W.A., Salomone, J.R., Tacy, D.J., Camino, E.M., Davis, K.A., Rebeiz, M., Williams, T.M., 2013. Recurrent Modification of a Conserved Cis-Regulatory Element Underlies Fruit Fly Pigmentation Diversity. PLOS Genetics 9, e1003740. https://doi.org/10.1371/journal.pgen.1003740

Rojas, B., Burdfield-Steel, E., Pakkanen, H., Suisto, K., Maczka, M., Schulz, S., Mappes, J., 2017. How to fight multiple enemies: target-specific chemical defences in an aposematic moth. Proceedings of the Royal Society B: Biological Sciences 284, 20171424. https://doi.org/10.1098/rspb.2017.1424

Rojas, B., Luis-MartÍnez, A., Mappes, J., 2015. Frequency-dependent flight activity in the colour polymorphic wood tiger moth. Curr Zool 61, 765–772. https://doi.org/10.1093/czoolo/61.4.765

Rojas, B., Mappes, J., Burdfield-Steel, E., 2019. Multiple modalities in insect warning displays have additive effects against wild avian predators. Behav Ecol Sociobiol 73, 37. https://doi.org/10.1007/s00265-019-2643-6

Römpler, H., Rohland, N., Lalueza-Fox, C., Willerslev, E., Kuznetsova, T., Rabeder, G., Bertranpetit, J., Schöneberg, T., Hofreiter, M., 2006. Nuclear Gene Indicates Coat-Color Polymorphism in Mammoths. Science 313, 62–62. https://doi.org/10.1126/science.1128994

Rönkä, K., Valkonen, J.K., Nokelainen, O., Rojas, B., Gordon, S., Burdfield-Steel, E., Mappes, J., 2020. Geographic mosaic of selection by avian predators on hindwing warning colour in a polymorphic aposematic moth. Ecology Letters 23, 1654–1663. https://doi.org/10.1111/ele.13597

Rosenblum, E.B., Rompler, H., Schoneberg, T., Hoekstra, H.E., 2009. Molecular and functional basis of phenotypic convergence in white lizards at White Sands. Proceedings of the National Academy of Sciences 107, 2113–2117. https://doi.org/10.1073/pnas.0911042107

Ruan, J., Li, H., 2020. Fast and accurate long-read assembly with wtdbg2. Nat Methods 17, 155–158. https://doi.org/10.1038/s41592-019-0669-3

Saenko, S.V., Chouteau, M., Piron-Prunier, F., Blugeon, C., Joron, M., Llaurens, V., 2019. Unravelling the genes forming the wing pattern supergene in the polymorphic butterfly Heliconius numata. EvoDevo 10, 16. https://doi.org/10.1186/s13227-019-0129-2

Saenko, S.V., Lamichhaney, S., Barrio, A.M., Rafati, N., Andersson, L., Milinkovitch, M.C., 2015. Amelanism in the corn snake is associated with the insertion of an LTR-retrotransposon in the *OCA2* gene. Scientific Reports 5, 17118. https://doi.org/10.1038/srep17118

Sahara, K., Yoshido, A., Traut, W., 2012. Sex chromosome evolution in moths and butterflies. Chromosome Res 20, 83–94. https://doi.org/10.1007/s10577-011-9262-z

Sahoo, R.K., Lohman, D.J., Wahlberg, N., Müller, C.J., Brattström, O., Collins, S.C., Peggie, D., Aduse-Poku, K., Kodandaramaiah, U., 2018. Evolution of Hypolimnas butterflies (Nymphalidae): Out-of-Africa origin and Wolbachia-mediated introgression.

Molecular Phylogenetics and Evolution 123, 50–58. https://doi.org/10.1016/j.ympev.2018.02.001

Schrader, L., Kim, J.W., Ence, D., Zimin, A., Klein, A., Wyschetzki, K., Weichselgartner, T., Kemena, C., Stökl, J., Schultner, E., Wurm, Y., Smith, C.D., Yandell, M., Heinze, J., Gadau, J., Oettler, J., 2014. Transposable element islands facilitate adaptation to novel environments in an invasive species. Nat Commun 5, 5495. https://doi.org/10.1038/ncomms6495

Schwander, T., Libbrecht, R., Keller, L., 2014. Supergenes and Complex Phenotypes. Current Biology 24, R288–R294. https://doi.org/10.1016/j.cub.2014.01.056

Sedlazeck, F.J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A., Schatz, M.C., 2018. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods 15, 461–468. https://doi.org/10.1038/s41592-018-0001-7

Seimiya, M., Gehring, W.J., 2000. The Drosophila homeobox gene optix is capable of inducing ectopic eyes by an eyeless-independent mechanism. Development 127, 1879–1886.

Seppey, M., Manni, M., Zdobnov, E.M., 2019. BUSCO: Assessing Genome Assembly and Annotation Completeness, in: Kollmar, M. (Ed.), Gene Prediction: Methods and Protocols, Methods in Molecular Biology. Springer, New York, NY, pp. 227–245. https://doi.org/10.1007/978-1-4939-9173-0_14

Sharp, A.J., Locke, D.P., McGrath, S.D., Cheng, Z., Bailey, J.A., Vallente, R.U., Pertz, L.M., Clark, R.A., Schwartz, S., Segraves, R., Oseroff, V.V., Albertson, D.G., Pinkel, D., Eichler, E.E., 2005. Segmental Duplications and Copy-Number Variation in the Human Genome. Am J Hum Genet 77, 78–88.

Shawkey, M.D., Hill, G.E., 2005. Carotenoids need structural colours to shine. Biology Letters 1, 121–124. https://doi.org/10.1098/rsbl.2004.0289

Shawkey, M.D., Morehouse, N.I., Vukusic, P., 2009. A protean palette: colour materials and mixing in birds and butterflies. Journal of The Royal Society Interface 6, S221–S231. https://doi.org/10.1098/rsif.2008.0459.focus

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.M., 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31, 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Sindi, S.S., Önal, S., Peng, L.C., Wu, H.-T., Raphael, B.J., 2012. An integrative probabilistic model for identification of structural variation in sequencing data. Genome Biology 13, R22. https://doi.org/10.1186/gb-2012-13-3-r22

Singh, K.S., De-Kayne, R., Omufwoko, K.S., Martins, D.J., Bass, C., ffrench-Constant, R., Martin, S.H., 2022. Genome assembly of Danaus chrysippus and comparison with the Monarch Danaus plexippus. G3 Genes|Genomes|Genetics 12, jkab449. https://doi.org/10.1093/g3journal/jkab449

Smit, A., Hubley, R., Green, P., 2015. RepeatMasker Open-4.0. 2013–2015.

Smith, D. a. S., 1976. Phenotypic diversity, mimicry and natural selection in the African butterfly Hypolimnas misippus L. (Lepidoptera: Nymphalidae). Biological Journal of the Linnean Society 8, 183–204. https://doi.org/10.1111/j.1095-8312.1976.tb00245.x

Smith, D. a. S., 1975. Genetics of Some Polymorphic Forms of the African Butterfly Danaus chrysippus L. (Lepidoptera: Danaidae). Insect Systematics & Evolution 6, 134–144. https://doi.org/10.1163/187631275X00235

Smith, D. a. S., 1973. Batesian Mimicry between Danaus chrysippus and Hypolimnas misippus (Lepidoptera) in Tanzania. Nature 242, 129–131. https://doi.org/10.1038/242129a0

Smith, D.A.S., 1976. Phenotypic diversity, mimicry and natural selection in the African butterfly Hypolimnas misippus L. (Lepidoptera: Nymphalidae). Biological Journal of the Linnean Society 8, 183–204. https://doi.org/10.1111/j.1095-8312.1976.tb00245.x

Smith, D.A.S., Gordon, J.S., 1987. The genetics of the butterfly Hypolimnas misippus (L.): The classification of phenotypes and the inheritance of forms misippus and inaria. Heredity 59, 467–475.

Smith, D.A.S., Owen, D.F., Gordon, I.J., Owiny, A.M., 1993. Polymorphism and evolution in the butterfly danaus chrysippus (L.) (lepidoptera: Danainae). Heredity 71, 242–251. https://doi.org/10.1038/hdy.1993.132

Spies, N., Weng, Z., Bishara, A., McDaniel, J., Catoe, D., Zook, J.M., Salit, M., West, R.B., Batzoglou, S., Sidow, A., 2017. Genome-wide reconstruction of complex structural variants using read clouds. Nat Methods 14, 915–920. https://doi.org/10.1038/nmeth.4366

Stern, D.L., 2013. The genetic causes of convergent evolution. Nat Rev Genet 14, 751–764. https://doi.org/10.1038/nrg3483

Stern, D.L., 2011. Evolution, development, & the predictable genome. Roberts and Co. Publishers.

Stern, D.L., Orgogozo, V., 2009. Is Genetic Evolution Predictable? Science 323, 746–751. https://doi.org/10.1126/science.1158997

Stern, D.L., Orgogozo, V., 2008. The Loci of Evolution: How Predictable Is Genetic Evolution? Evolution 62, 2155–2177. https://doi.org/10.1111/j.1558-5646.2008.00450.x

Storer, J., Hubley, R., Rosen, J., Wheeler, T.J., Smit, A.F., 2021. The Dfam community resource of transposable element families, sequence models, and genome annotations. Mobile DNA 12, 2. https://doi.org/10.1186/s13100-020-00230-y

Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M., Konkel, M.K., Malhotra, A., Stütz, A.M., Shi, X., Paolo Casale, F., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., Malig, M., Chaisson, M.J.P., Walter, K., Meiers, S., Kashin, S., Garrison, E., Auton, A., Lam, H.Y.K., Jasmine Mu, X., Alkan, C., Antaki, D., Bae, T., Cerveira, E., Chines, P., Chong, Z., Clarke, L., Dal, E., Ding, L., Emery, S., Fan, X., Gujral, M., Kahveci, F., Kidd, J.M., Kong, Y., Lameijer, E.-W., McCarthy, S., Flicek, P., Gibbs, R.A., Marth, G., Mason, C.E., Menelaou, A., Muzny, D.M., Nelson, B.J., Noor, A., Parrish, N.F., Pendleton, M., Quitadamo, A., Raeder, B., Schadt, E.E., Romanovitch, M., Schlattl, A., Sebra, R., Shabalin, A.A., Untergasser, A., Walker, J.A., Wang, M., Yu, F., Zhang, C., Zhang, J., Zheng-Bradley, X., Zhou, W., Zichner, T., Sebat, J., Batzer, M.A., McCarroll, S.A., Mills, R.E., Gerstein, M.B., Bashir, A., Stegle, O., Devine, S.E., Lee, C., Eichler, E.E., Korbel, J.O., 2015. An integrated map of structural variation in 2,504 human genomes. Nature 526, 75–81. https://doi.org/10.1038/nature15394

Suomalainen, E., 1938. Die Erblichkeitsverhältnisse Des Männlichen Dimorphismus Bei Parasemia Plantaginis. Hereditas 24, 386–390. https://doi.org/10.1111/j.1601-5223.1938.tb02676.x

Swinhoe, C., 1896. On Mimicry in Butterflies of the Genus Hypolimnas. Journal of the Linnean Society of London, Zoology 25, 339–348.

Tanaka, S., Harano, K., Nishide, Y., Sugahara, R., 2016. The mechanism controlling phenotypic plasticity of body color in the desert locust: some recent progress. Current Opinion in Insect Science, Global change biology * Molecular physiology 17, 10–15. https://doi.org/10.1016/j.cois.2016.05.011

Thompson, M.J., Jiggins, C.D., 2014. Supergenes and their role in evolution. Heredity 113, 1–8. https://doi.org/10.1038/hdy.2014.20

Tian, L., Rahman, S.R., Ezray, B.D., Franzini, L., Strange, J.P., Lhomme, P., Hines, H.M., 2019. A homeotic shift late in development drives mimetic color variation in a bumble bee. Proc Natl Acad Sci USA 201900365. https://doi.org/10.1073/pnas.1900365116

Toews, D.P.L., Taylor, S.A., Vallender, R., Brelsford, A., Butcher, B.G., Messer, P.W., Lovette, I.J., 2016. Plumage Genes and Little Else Distinguish the Genomes of Hybridizing Warblers. Current Biology 26, 2313–2318. https://doi.org/10.1016/j.cub.2016.06.034

Toomey, M.B., Lopes, R.J., Araújo, P.M., Johnson, J.D., Gazda, M.A., Afonso, S., Mota, P.G., Koch, R.E., Hill, G.E., Corbo, J.C., Carneiro, M., 2017. High-density lipoprotein

receptor SCARB1 is required for carotenoid coloration in birds. PNAS 114, 5219–5224. https://doi.org/10.1073/pnas.1700751114

Toomey, M.B., Marques, C.I., Andrade, P., Araújo, P.M., Sabatino, S., Gazda, M.A., Afonso, S., Lopes, R.J., Corbo, J.C., Carneiro, M., 2018. A non-coding region near Follistatin controls head colour polymorphism in the Gouldian finch. Proceedings of the Royal Society B: Biological Sciences 285, 20181788. https://doi.org/10.1098/rspb.2018.1788

Trimen, R., Bowker, J.H., 1889. South-African butterflies: a monograph of the extra-tropical species. Trübner.

Turner, J.R.G., 1987. The evolutionary dynamics of batesian and muellerian mimicry: similarities and differences. Ecological Entomology 12, 81–95. https://doi.org/10.1111/j.1365-2311.1987.tb00987.x

Turner, J.R.G., 1981. Adaptation and Evolution in Heliconius: A Defense of NeoDarwinism. Annu. Rev. Ecol. Syst. 12, 99–121. https://doi.org/10.1146/annurev.es.12.110181.000531

Turner, J.R.G., 1967. On Supergenes. I. The Evolution of Supergenes. The American Naturalist 101, 195–221.

Turner, J.R.G., Sheppard, P.M., 1975. Absence of crossing-over in female butterflies (Heliconius). Heredity 34, 265–269. https://doi.org/10.1038/hdy.1975.29

Tuttle, E.M., Bergland, A.O., Korody, M.L., Brewer, M.S., Newhouse, D.J., Minx, P., Stager, M., Betuel, A., Cheviron, Z.A., Warren, W.C., Gonser, R.A., Balakrishnan, C.N., 2016. Divergence and functional degradation of a sex chromosome-like supergene. Current Biology 26, 344–350. https://doi.org/10.1016/j.cub.2015.11.069

Twomey, E., Vestergaard, J.S., Venegas, P.J., Summers, K., 2015. Mimetic Divergence and the Speciation Continuum in the Mimic Poison Frog Ranitomeya imitator. The American Naturalist 187, 205–224. https://doi.org/10.1086/684439

Uy J. Albert C., Cooper Elizabeth A., Cutie Stephen, Concannon Moira R., Poelstra Jelmer W., Moyle Robert G., Filardi Christopher E., 2016. Mutations in different pigmentation genes are associated with parallel melanism in island flycatchers. Proceedings of the Royal Society B: Biological Sciences 283, 20160731. https://doi.org/10.1098/rspb.2016.0731

Uy, J.A.C., Cooper, E.A., Cutie, S., Concannon, M.R., Poelstra, J.W., Moyle, R.G., Filardi, C.E., 2016. Mutations in different pigmentation genes are associated with parallel melanism in island flycatchers. Proc. Biol. Sci. 283. https://doi.org/10.1098/rspb.2016.0731

van der Burg, K.R.L., Lewis, J.J., Brack, B.J., Fandino, R.A., Mazo-Vargas, A., Reed, R.D., 2020. Genomic architecture of a genetically assimilated seasonal color pattern. Science 370, 721–725. https://doi.org/10.1126/science.aaz3017

Vane-Wright, R.I., 1976. A unified classification of mimetic resemblances. Biological Journal of the Linnean Society 8, 25–56. https://doi.org/10.1111/j.1095-8312.1976.tb00240.x

Vane-Wright, R.I., Ackery, P.R., Smiles, R.L., 1977. The polymorphism, mimicry, and host plant relationships of Hypolimnas butterflies. Biological Journal of the Linnean Society 9, 285–297. https://doi.org/10.1111/j.1095-8312.1977.tb00271.x

VanKuren, N.W., Massardo, D., Nallu, S., Kronforst, M.R., 2019. Butterfly Mimicry Polymorphisms Highlight Phylogenetic Limits of Gene Reuse in the Evolution of Diverse Adaptations. Mol Biol Evol 36, 2842–2853. https://doi.org/10.1093/molbev/msz194

van't Hof, A.E., Campagne, P., Rigden, D.J., Yung, C.J., Lingley, J., Quail, M.A., Hall, N., Darby, A.C., Saccheri, I.J., 2016. The industrial melanism mutation in British peppered moths is a transposable element. Nature 534, 102–105. https://doi.org/10.1038/nature17951

van't Hof, A.E., Reynolds, L.A., Yung, C.J., Cook, L.M., Saccheri, I.J., 2019. Genetic convergence of industrial melanism in three geometrid moths. Biology Letters 15, 20190582. https://doi.org/10.1098/rsbl.2019.0582

Vasimuddin, Md., Misra, S., Li, H., Aluru, S., 2019. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems, in: 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS). Presented at the 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS), pp. 314–324. https://doi.org/10.1109/IPDPS.2019.00041

Vítková, M., Fuková, I., Kubíčková, S., Marec, F., 2007. Molecular divergence of the W chromosomes in pyralid moths (Lepidoptera). Chromosome Res 15, 917–930. https://doi.org/10.1007/s10577-007-1173-7

Vukusic, P., Sambles, J.R., Lawrence, C.R., Wootton, R.J., 1999. Quantified Interference and Diffraction in Single Morpho Butterfly Scales. Proceedings: Biological Sciences 266, 1403–1411.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., Earl, A.M., 2014. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLOS ONE 9, e112963. https://doi.org/10.1371/journal.pone.0112963

Wallbank, R.W.R., Baxter, S.W., Pardo-Diaz, C., Hanly, J.J., Martin, S.H., Mallet, J., Dasmahapatra, K.K., Salazar, C., Joron, M., Nadeau, N., McMillan, W.O., Jiggins, C.D.,

2016. Evolutionary Novelty in a Butterfly Wing Pattern through Enhancer Shuffling. PLoS Biology. https://doi.org/10.1371/journal.pbio.1002353

Wang, J., Wurm, Y., Nipitwattanaphon, M., Riba-Grognuz, O., Huang, Y.-C., Shoemaker, D., Keller, L., 2013. A Y-like social chromosome causes alternative colony organization in fire ants. Nature 493, 664–668. https://doi.org/10.1038/nature11832

Wang, S., Teng, D., Li, X., Yang, P., Da, W., Zhang, Yiming, Zhang, Yubo, Liu, G., Zhang, X., Wan, W., Dong, Z., Wang, D., Huang, S., Jiang, Z., Wang, Q., Lohman, D.J., Wu, Y., Zhang, Linlin, Jia, F., Westerman, E., Zhang, Li, Wang, W., Zhang, W., 2022. The evolution and diversification of oakleaf butterflies. Cell. https://doi.org/10.1016/j.cell.2022.06.042

Weischenfeldt, J., Symmons, O., Spitz, F., Korbel, J.O., 2013. Phenotypic impact of genomic structural variation: insights from and for human disease. Nat Rev Genet 14, 125–138. https://doi.org/10.1038/nrg3373

Weissensteiner, M.H., Bunikis, I., Catalán, A., Francoijs, K.-J., Knief, U., Heim, W., Peona, V., Pophaly, S.D., Sedlazeck, F.J., Suh, A., Warmuth, V.M., Wolf, J.B.W., 2020. Discovery and population genomics of structural variation in a songbird genus. Nat Commun 11, 3403. https://doi.org/10.1038/s41467-020-17195-4

Wellenreuther, M., Mérot, C., Berdan, E., Bernatchez, L., 2019. Going beyond SNPs: The role of structural genomic variants in adaptive evolution and species diversification. Molecular Ecology 28, 1203–1209. https://doi.org/10.1111/mec.15066

Westerman, E.L., VanKuren, N.W., Massardo, D., Tenger-Trolander, A., Zhang, W., Hill, R.I., Perry, M., Bayala, E., Barr, K., Chamberlain, N., Douglas, T.E., Buerkle, N., Palmer, S.E., Kronforst, M.R., 2018a. Aristaless Controls Butterfly Wing Color Variation Used in Mimicry and Mate Choice. Current Biology 28, 3469-3474.e4. https://doi.org/10.1016/j.cub.2018.08.051

Westerman, E.L., VanKuren, N.W., Massardo, D., Tenger-Trolander, A., Zhang, W., Hill, R.I., Perry, M., Bayala, E., Barr, K., Chamberlain, N., Douglas, T.E., Buerkle, N., Palmer, S.E., Kronforst, M.R., 2018b. Aristaless Controls Butterfly Wing Color Variation Used in Mimicry and Mate Choice. Current Biology 28, 3469-3474.e4. https://doi.org/10.1016/j.cub.2018.08.051

Wick, R., 2022. Porechop.

Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis, 2nd ed. 2016. ed, Use R! Springer International Publishing : Imprint: Springer, Cham. https://doi.org/10.1007/978-3-319-24277-4

Widemo, F., 1998. Alternative reproductive strategies in the ruff,Philomachus pugnax: a mixed ESS? Animal Behaviour 56, 329–336. https://doi.org/10.1006/anbe.1998.0792

Wilson, R., Burnet, B., Eastwood, L., Connolly, K., 1976. Behavioural pleiotropy of the yellow gene in Drosophila melanogaster. Genetics Research 28, 75–88. https://doi.org/10.1017/S0016672300016748

Wilson, R.C., Doudna, J.A., 2013. Molecular Mechanisms of RNA Interference. Annual Review of Biophysics 42, 217–239. https://doi.org/10.1146/annurev-biophys-083012-130404

Wittkopp, P.J., Vaccaro, K., Carroll, S.B., 2002. Evolution of yellow Gene Regulation and Pigmentation in Drosophila. Current Biology 12, 1547–1556. https://doi.org/10.1016/S0960-9822(02)01113-2

Woronik, A., Tunström, K., Perry, M.W., Neethiraj, R., Stefanescu, C., Celorio-Mancera, M. de la P., Brattström, O., Hill, J., Lehmann, P., Käkelä, R., Wheat, C.W., 2019. A transposable element insertion is associated with an alternative life history strategy. Nat Commun 10, 1–11. https://doi.org/10.1038/s41467-019-13596-2

Wright, A.E., Dean, R., Zimmer, F., Mank, J.E., 2016. How to make a sex chromosome. Nat Commun 7, 12087. https://doi.org/10.1038/ncomms12087

Wu, Y.-W., Simmons, B.A., Singer, S.W., 2016. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics 32, 605–607. https://doi.org/10.1093/bioinformatics/btv638

Wulf, A., 2016. The Woman Who Made Science Beautiful [WWW Document]. The Atlantic. URL https://www.theatlantic.com/science/archive/2016/01/the-woman-who-made-science-beautiful/424620/ (accessed 7.7.22).

Yan, Z., Martin, S.H., Gotzek, D., Arsenault, S.V., Duchen, P., Helleu, Q., Riba-Grognuz, O., Hunt, B.G., Salamin, N., Shoemaker, D., Ross, K.G., Keller, L., 2020. Evolution of a supergene that regulates a trans-species social polymorphism. Nat Ecol Evol 4, 240–249. https://doi.org/10.1038/s41559-019-1081-1

Yang, J., Wan, W., Xie, M., Mao, J., Dong, Z., Lu, S., He, J., Xie, F., Liu, G., Dai, X., Chang, Z., Zhao, R., Zhang, R., Wang, S., Zhang, Y., Zhang, W., Wang, W., Li, X., 2020. Chromosome-level reference genome assembly and gene editing of the dead-leaf butterfly Kallima inachus. Molecular Ecology Resources 20, 1080–1092. https://doi.org/10.1111/1755-0998.13185

Yen, E.C., McCarthy, S.A., Galarza, J.A., Generalovic, T.N., Pelan, S., Nguyen, P., Meier, J.I., Warren, I.A., Mappes, J., Durbin, R., Jiggins, C.D., 2020. A haplotype-resolved, de novo genome assembly for the wood tiger moth (Arctia plantaginis) through trio binning. Gigascience 9. https://doi.org/10.1093/gigascience/giaa088

Yoshida, K., Terai, Y., Mizoiri, S., Aibara, M., Nishihara, H., Watanabe, M., Kuroiwa, A., Hirai, H., Hirai, Y., Matsuda, Y., Okada, N., 2011. B Chromosomes Have a Functional Effect

on Female Sex Determination in Lake Victoria Cichlid Fishes. PLOS Genetics 7, e1002203. https://doi.org/10.1371/journal.pgen.1002203

Yoshido, A., Šíchová, J., Kubíčková, S., Marec, F., Sahara, K., 2013. Rapid turnover of the W chromosome in geographical populations of wild silkmoths, Samia cynthia ssp. Chromosome Res 21, 149–164. https://doi.org/10.1007/s10577-013-9344-1

Yuan, Y.-W., Rebocho, A.B., Sagawa, J.M., Stanley, L.E., Bradshaw, H.D., 2016. Competition between anthocyanin and flavonol biosynthesis produces spatial pattern variation of floral pigments between Mimulus species. PNAS 113, 2448–2453. https://doi.org/10.1073/pnas.1515294113

Zeevi, D., Korem, T., Godneva, A., Bar, N., Kurilshikov, A., Lotan-Pompan, M., Weinberger, A., Fu, J., Wijmenga, C., Zhernakova, A., Segal, E., 2019. Structural variation in the gut microbiome associates with host health. Nature 568, 43–48. https://doi.org/10.1038/s41586-019-1065-y

Zhang, C., Song, Y., Thompson, D.A., Madonna, M.A., Millhauser, G.L., Toro, S., Varga, Z., Westerfield, M., Gamse, J., Chen, W., Cone, R.D., 2010. Pineal-specific agouti protein regulates teleost background adaptation. PNAS 107, 20164–20171. https://doi.org/10.1073/pnas.1014941107

Zhang, L., Martin, A., Perry, M.W., Burg, K.R.L. van der, Matsuoka, Y., Monteiro, A., Reed, R.D., 2017a. Genetic Basis of Melanin Pigmentation in Butterfly Wings. Genetics 205, 1537–1550. https://doi.org/10.1534/genetics.116.196451

Zhang, L., Mazo-Vargas, A., Reed, R.D., 2017b. Single master regulatory gene coordinates the evolution and development of butterfly color and iridescence. Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.1709058114

Zhang, L., Steward, R.A., Wheat, C.W., Reed, R.D., 2021. High-Quality Genome Assembly and Comprehensive Transcriptome of the Painted Lady Butterfly Vanessa cardui. Genome Biology and Evolution 13. https://doi.org/10.1093/gbe/evab145

Zhang, W., Dasmahapatra, K.K., Mallet, J., Moreira, G.R.P., Kronforst, M.R., 2016. Genome-wide introgression among distantly related Heliconius butterfly species. Genome Biology 17, 25. https://doi.org/10.1186/s13059-016-0889-0

Zhao, L., Charlesworth, B., 2016. Resolving the Conflict Between Associative Overdominance and Background Selection. Genetics 203, 1315–1334. https://doi.org/10.1534/genetics.116.188912

Zhou, X., Stephens, M., 2012. Genome-wide efficient mixed-model analysis for association studies. Nat Genet 44, 821–824. https://doi.org/10.1038/ng.2310

Appendix A

Paper “The genomics of coloration provides insights into adaptive evolution”

Chapter 1 was published:

Orteu, Anna, and Chris D. Jiggins. "The genomics of coloration provides insights into adaptive evolution." *Nature Reviews Genetics* 21.8 (2020): 461-475.

The genomics of coloration provides insights into adaptive evolution

Anna Orteu ✉ *and Chris D. Jiggins* ✉

Abstract | Coloration is an easily quantifiable visual trait that has proven to be a highly tractable system for genetic analysis and for studying adaptive evolution. The application of genomic approaches to evolutionary studies of coloration is providing new insight into the genetic architectures underlying colour traits, including the importance of large-effect mutations and supergenes, the role of development in shaping genetic variation and the origins of adaptive variation, which often involves adaptive introgression. Improved knowledge of the genetic basis of traits can facilitate field studies of natural selection and sexual selection, making it possible for strong selection and its influence on the genome to be demonstrated in wild populations.

Association studies
Studies that correlate phenotypic variation with genetic variation. The most common methods are genome-wide association studies and quantitative trait locus mapping studies.

Polymorphism
The occurrence of two or more distinct phenotypes or morphs of a species within a population.

Genetic architecture
The genetic basis underlying variation in a phenotypic trait. The main characteristics are the number of loci, their interactions and effect sizes, and their relative positions in the genome.

Department of Zoology, University of Cambridge, Cambridge, UK.

✉*e-mail: af658@cam.ac.uk; c.jiggins@zoo.cam.ac.uk*

https://doi.org/10.1038/s41576-020-0234-z

The study of coloration in animals and plants has long fascinated biologists[1]. Coloration has a key role in modulating animal and plant fitness through its effects on many aspects of phenotype, including courtship and mate preference, predator avoidance through camouflage, photoprotection, structural support, microbial resistance and thermoregulation (reviewed in REF.[2]). Indeed, any given visual cue is commonly under multiple selective pressures, which are often related to both interspecific and intraspecific communication[1,3]. For example, coloration can be adapted to reduce predation through camouflage, warning coloration or mimicry, while also functioning in pollinator attraction or signalling to conspecifics for mate choice or intrasexual competition[4,5].

Identifying and understanding the genetic basis of adaptations has become a key goal of evolutionary genetics. Determining the molecular mechanisms underlying fitness-related traits can reveal novel insights into the evolutionary forces that shape biodiversity and also provide the explicit link between genetic changes and natural selection. Colour traits were well studied in ecological and traditional genetic contexts by early evolutionary biologists, so a great deal is known about both their adaptive value and the nature of their genetic control. However, until recently the precise genes and molecular mechanisms underlying colour variation remained largely unknown.

Ongoing advances in genomic methods are rapidly increasing our understanding of the genetic basis of colour-related traits (BOX 1). In particular, assembly of reference genomes has become much more accessible, both technically and in terms of cost, which has enabled genome-scale analysis of genetic variation for association studies and transcriptomic studies of gene expression. As an example, a recent study of wall lizards generated a reference genome, conducted a genome-scale analysis of a colour polymorphism and identified causal genes, a process that until recently might have taken a decade of work[6]. Genomic methods are therefore permitting a new understanding of the molecular and developmental mechanisms underlying adaptive coloration in a wide variety of systems.

In this Review, we summarize the diverse ways in which organisms produce colours and their adaptive function, and then evaluate how our understanding of evolution has benefited from studies of adaptive coloration, focusing on three major topics. First, we describe the genetic architecture of adaptive coloration and how even complex adaptations can be controlled by genes of large effect. Second, we discuss the importance of developmental processes in constraining and directing the genetic changes that occur during evolution and perhaps even the overall direction of evolutionary change. Finally, we illustrate how genetic studies of coloration help us understand adaptive evolution more broadly.

Coloration and its adaptive value

Broadly, colours produced by organisms can be divided into pigmented and structural colours (FIG. 1). Pigmented colours result from the accumulation of molecules (pigments) capable of selectively absorbing light, and are typically synthesized by the organism in a series of enzymatic steps (FIG. 1a). By contrast, structural colours depend on physical structures that reflect light. In natural systems, colours often have both structural and pigment components[7,8]. Regardless of the type, coloration can be classified as having one of three main functions: signalling, in which visual cues affect interspecific and intraspecific communication; vision, in which pigments are an essential component of the animal's visual system; and physiological, in which pigments are required for

Box 1 | **Genomic methods to study adaptive evolution**

Ongoing development of genomic technologies and the fall in sequencing prices has revolutionized the ways in which we study adaptive evolution, and species that in the past were challenging to study because of a lack of resources are now becoming accessible.

Commonly the first step in any study is to assemble a reference genome, often using a combination of long-read and short-read sequence data[103]. A reference genome permits a genome-wide search for loci underlying phenotype variation. For example, short-read sequence data from individuals collected in wild populations that differ in colour phenotypes can be used to conduct a genome-wide association study (GWAS)[104]. Alternatively, genotype–phenotype associations can be tested using laboratory populations in a quantitative trait locus (QTL) mapping experiment[105]. Typically, two populations or species are crossed and F2 hybrids are analysed for both phenotype and genotype to investigate associations across the genome. Often, GWAS and QTL mapping identify regions that contain multiple genes, and transcriptomic analyses can be used to identify genes that are differentially expressed between variants, thereby narrowing down an initial list of candidates. Transcriptomic analyses are also facilitated by good reference genomes, which ensure more accurate calling of transcript counts.

Once candidate loci have been identified, functional tests can be used to confirm they contribute to the phenotype. Methods that allow in vivo genetic manipulations are now a particularly powerful approach for studying coloration phenotypes. Genome editing technologies, such as those based on the CRISPR–Cas9 system, are widely used[106], most commonly to introduce non-functional mutations into coding sequences (gene knockouts). However, in the future it will be exciting to perform more targeted and specific genetic manipulations that recapitulate evolutionary changes in natural populations, such as introducing or removing specific alleles from the genome, or even making individual mutations at putatively functional sites. RNA interference is another molecular tool that can be used to lower the expression of a gene (knockdown), offering insights into gene function[107]. Genomic data can also be combined with tests for ecological function, such as translocation experiments, to confirm the presence of signatures of selection at the candidate loci (FIG. 6).

a particular task within the organism, such as energy production through photosynthesis using chlorophyll.

Biological pigments. Pigments are highly diverse molecules that often have physiological functions: in addition to chlorophyll, they are found in haemoglobin in vertebrate blood cells and flavins in vitamin B_2, among other compounds. Some pigments are taxonomically restricted, such as the psittacofulvins responsible for the bright red, yellow and orange colours of parrot plumage[9]. Pterins, which can be white, yellow, orange or red, are commonly found in insects, including in the red eyes of the fruit fly and the orange in some *Pieris* butterfly wings, but are also responsible for the bright colours in some vertebrates, such as the wall lizard[6]. Ommochromes are red and yellow pigments that are restricted to invertebrates. They are important for vision in arthropods but have signalling functions in butterfly wings and squid chromatophores. Anthocyanins (red, purple and black) are restricted to plants and are responsible for the coloration of many types of flowers. By contrast, the melanins, eumelanin (brown or black) and phaeomelanin (yellow to red), and carotenoids (yellow, red and orange) are widespread across bacteria, fungi, plants and animals (FIG. 1b). As well as playing a part in signalling, melanins are also important for thermoregulation and in protection against UV damage to tissues. Plants can synthesize carotenoids, which can be crucial in light harvesting during photosynthesis[10]. By contrast, animals were long thought to be unable to synthesize carotenoids, but recent work has shown that aphids can do so using genetic machinery acquired through horizontal transfer from fungi[11], and that birds can transform yellow carotenoids obtained from the diet to red carotenoids through ketolation. CYP2J19, a cytochrome P450 family member, was recently identified as the enzyme responsible for this step[12,13]. In birds, carotenoids have an important role in generating bright signals important in sexual selection and may play an important part in honest signalling of male quality.

Structural coloration. Structural colours can cover a huge variety of different forms (FIG. 1c) and, because they are formed by reflected light, they are commonly iridescent; that is, the colour varies depending on the viewing angle. This variation often makes structural colours harder to quantify and study than pigments and, as a result, their developmental basis and genetic basis remain comparatively poorly understood. Some of the best-known classes of structural colours are multilayer reflectors, surface diffraction gratings and photonic crystals. The shimmering blue iridescence of *Selaginella* plants is the result of a multilayer reflector (FIG. 1c), and is caused by the presence of multiple cuticular layers in the leaf[14]. The distinct layers have different optical properties, which causes part of the light to be reflected in each layer transition. The reflected beams interfere with each other, reducing the spectrum of light reflected to shades of blue. An example of diffraction gratings is given by the wing scales of the *Morpho* butterfly, in which repetitive ridged structures split the incident ray of light into multiple wavelengths with different reflection angles (FIG. 1c). The diffracted light from distinct ridges will interfere, limiting the spectrum of light reflected to blue. The combined effect of the vertical multilayer interference within each ridge and the diffraction among the different ridges produces a remarkably efficient wide-angled reflectance of blue light[15]. Photonic crystals are periodic nanostructures that affect the motion of photons to produce coloured reflectance that can be found, for example, in the yellow and green colours of the Kaiser-i-Hind swallowtail butterfly, *Teinopalpus imperialis*, and in the multicoloured feathers of peacocks (FIG. 1c).

Colour and adaptation. In this Review, we focus on colours used in signalling, where the link between diversity in coloration and its adaptive value is well understood. Signalling through visual cues is essential for interspecific and intraspecific communication in animals, and consequently colour attributes are under multiple selective pressures[1]. For example, coloration commonly plays a role in sexual selection, with bright colours used as indicators of high-quality males (FIG. 2a). Coloration can be involved in many other social functions; for example, female primates can use colour signals to indicate sexual receptivity, such as in the Celebes crested macaque (*Macaca nigra*) (FIG. 2b).

In addition, many predators rely on their vision to detect prey, and thus camouflage can be a successful strategy for predator avoidance. Common forms of camouflage include background matching through cryptic coloration (such as that seen in snowshoe hares

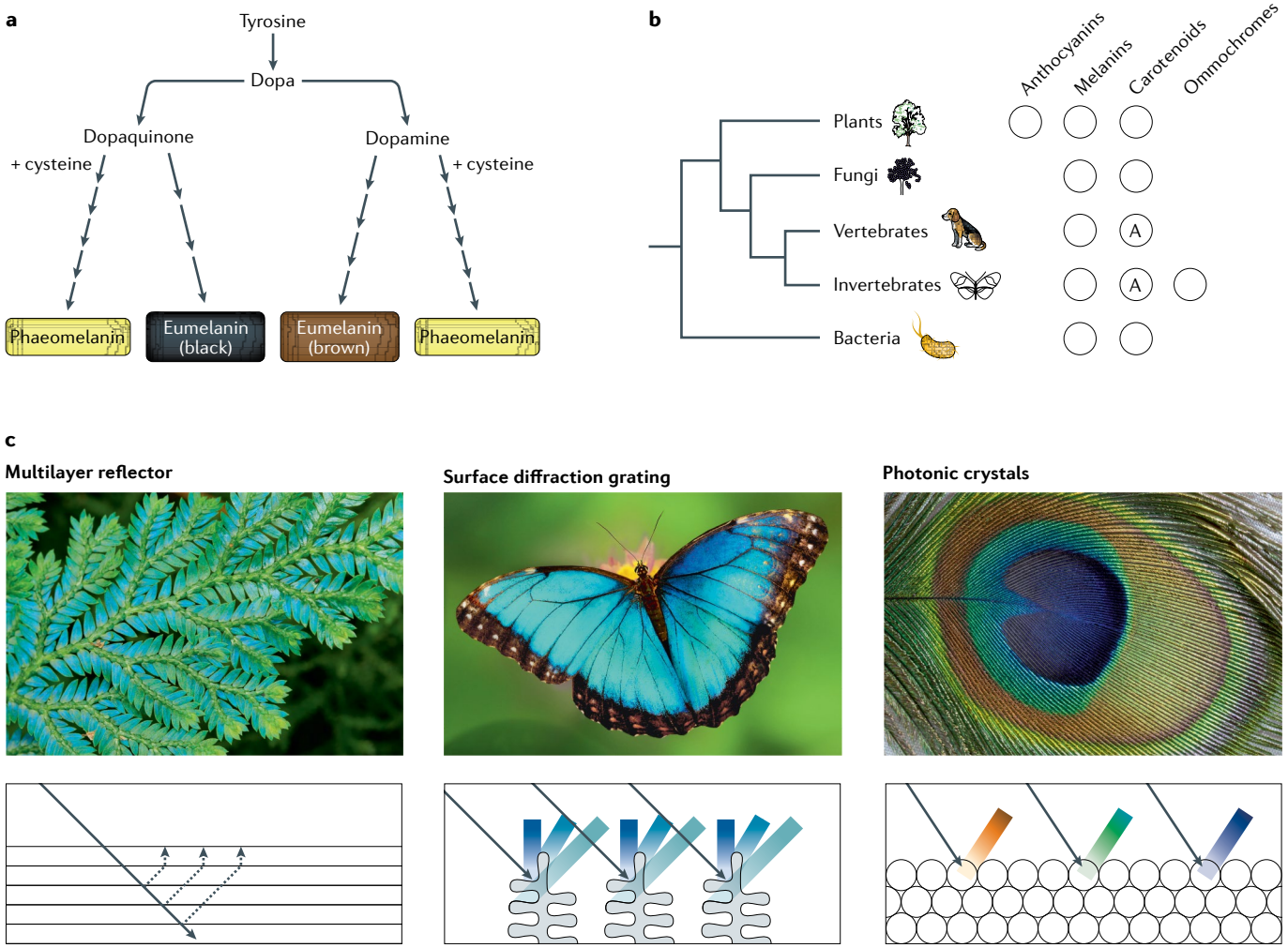


Fig. 1 | **Different ways of making colour. a** | Pigments are compounds that selectively absorb light and therefore produce colour. In most cases, pigments are synthesized biochemically by the organism. A simplified biosynthetic pathway for melanin pigments is shown, in which each arrow represents a biochemical step often catalysed by a different enzyme, providing multiple targets on which selection can act. **b** | The phylogenetic distribution of four common pigment types highlights that some are found in all groups (melanins and carotenoids), whereas others are more restricted (anthocyanins to plants, ommochromes to invertebrates). Colours in the pie charts represent the different colours of each type of pigment. 'A' indicates that the pigment is acquired through the diet instead of being synthesized by the organism. **c** | Three types of physical structures underlying structural coloration. Multilayer reflectors are found in the leaves of *Selaginella* plants[14]. Multiple stacked cuticular layers with different optical properties reflect the incident ray of light (solid black arrow) at each transition zone (dotted black lines). Interference of the reflected beams limits the reflected light to blue-green shades (coloured arrows). Surface diffraction gratings are found in *Morpho* butterflies[15]. The periodic ridged structures found in the scales diffract light into multiple wavelengths. Interference of the diffracted rays from different ridges reduces the light reflected to shades of blue (coloured arrows). Photonic crystals are found in the multicoloured feathers of peacocks. These periodic structures affect the motion of photons such that certain wavelengths are reflected (coloured arrows) rather than transmitted. Part **a** is adapted from REF.[101], CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/). Part **c** credit for *Selaginella vogelii*: Hervé Lenain/Alamy; credit for blue morpho: blickwinkel/H. Schmidbauer/Alamy; credit for peacock feather: David Chapman/Alamy.

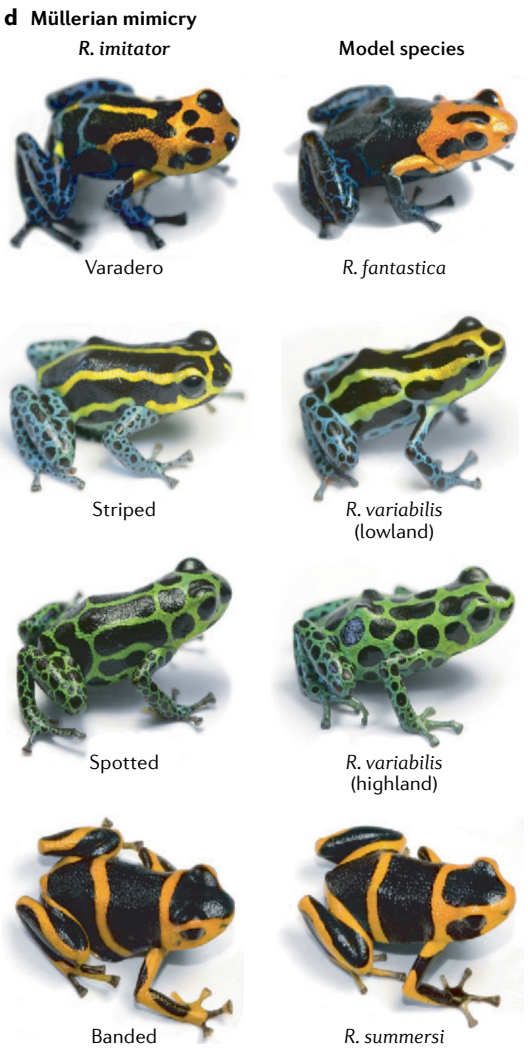
and pygmy seahorses) (FIG. 2c), disruptive coloration to break the body outline (such as zebra stripes and leopard spots) and masquerading as a common inedible object (as is seen in stick insects and leaf-mimicking butterflies)[3]. Escaping predation can also be achieved by predator deterrence through warning coloration or aposematism, in which unpalatable prey announce their toxicity to predators by displaying conspicuous colour patterns. Striking colours increase visibility and help predators remember the toxicity of the prey in future encounters. In some cases, toxic species share aposematic coloration and, by resembling one another, they reduce the cost of teaching predators, a phenomenon known as Müllerian mimicry[16]. An example is the *Ranitomeya* genus of poison dart frogs, in which the four morphs of the toxic *R. imitator* resemble multiple toxic models from the same genus[17] (FIG. 2d). In other cases, palatable species resemble toxic ones to gain protection against predators, which is known as Batesian mimicry. For instance, the palatable butterfly

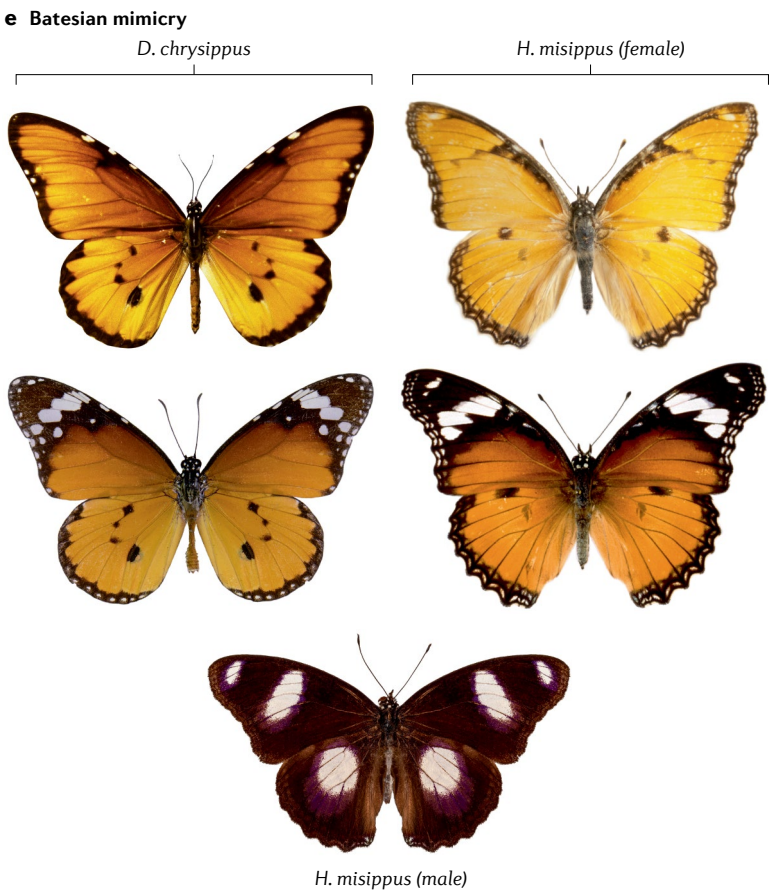
Hypolimnas misippus has very similar colour patterns to morphs of the unpalatable African queen, *Danaus chrysippus*[18] (FIG. 2e).

Genetic architecture of colour traits

The genetic architecture of a trait describes the number and effect size of genes that influence variability within and between populations and its patterns of linkage and interaction with other traits. The genetic architecture can influence the potential for a trait to undergo evolutionary change and the degree to which different traits are genetically associated.

Mutational effect size. One question of general interest to evolutionary geneticists is the effect size of loci involved in adaptation. Mutations of large effect should make evolutionary change easier and faster, but conversely they are arguably more likely to be deleterious[19]. Furthermore, a body of theory predicts that a single bout of adaptation will involve loci with a distribution of effect sizes,





Quantitative trait loci
Genomic regions at which there is a correlation between genetic variation and phenotypic variation in a trait of interest among individuals from a laboratory-generated cross.

Heterozygote advantage
A scenario in which the heterozygous genotype is fitter than either homozygous genotype.

Transposable element
A genetic element that can move from one position in the genome to another.

cis-regulatory element
(CRE). A genetic region that regulates expression of a coding sequence on the same DNA strand.

Selection coefficient
A measure of the difference in fitness between two genotypes, which are a necessary condition for the action of natural selection.

with a few large-effect and many small-effect mutations expected[20,21]. Both association studies and analyses of quantitative trait loci can provide estimates of the effect sizes of individual loci (in terms of the amount of phenotypic variation explained by a particular locus), while genetic manipulation can investigate the importance of individual mutations towards a particular phenotypic change (BOX 1). Such tools have confirmed that many recent adaptive changes in coloration do indeed involve a few loci of large effect, a few of which have been characterized at the level of individual mutations. In particular, genes encoding proteins involved in pigment synthesis pathways provide a large mutational target, and simple genetic changes in these genes can have major effects on phenotype. Indeed, there are now many examples of large-effect coding sequence mutations in pigmentation genes, in both animals and plants, that influence adaptive coloration[22,23]. For example, the genetic basis of feather colour variation in budgerigars has been narrowed down to a single-nucleotide polymorphism in the coding region of a polyketide synthase gene, *MuPKS* (also known as *LOC101880715*), which has been shown to produce yellow pigment when heterologously expressed in yeast[9] (FIG. 3a). Similarly, island populations of flycatchers have independently evolved dark coloration through coding mutations in either the *MC1R* gene (which encodes melanocortin 1 receptor (MC1R)) or the *Agouti* gene (which encodes the agouti signalling protein (Agouti))[24], and in vitro tests have confirmed that the observed coding changes in *MC1R* disrupt protein function[25,26]. MC1R is a transmembrane receptor that, when activated, triggers a signalling cascade that results in the production of brown or black eumelanin. MC1R can be antagonized by Agouti, which reverses the cascade to the default production of yellow phaeomelanin. Similarly, a flower colour polymorphism in the alpine orchid *Gymnadenia rhellicani* is regulated by a mutation that introduces a premature stop codon in a gene encoding an R2R3-MYB transcription factor. Transcription factors of this family have been shown to regulate expression of anthocyanin biosynthesis in multiple plant species[27,28]. In this alpine orchid, the loss of function of the transcription factor results in lower expression of an anthocyanin synthase, which controls the production of dark purple anthocyanidin pigments, and thus results in reduced accumulation of pigment and whiter coloration. This mutation is maintained in the population by heterozygote advantage, with both bee and fly pollinators attracted to the heterozygote red morph, leading to higher fitness of that genotype[29].

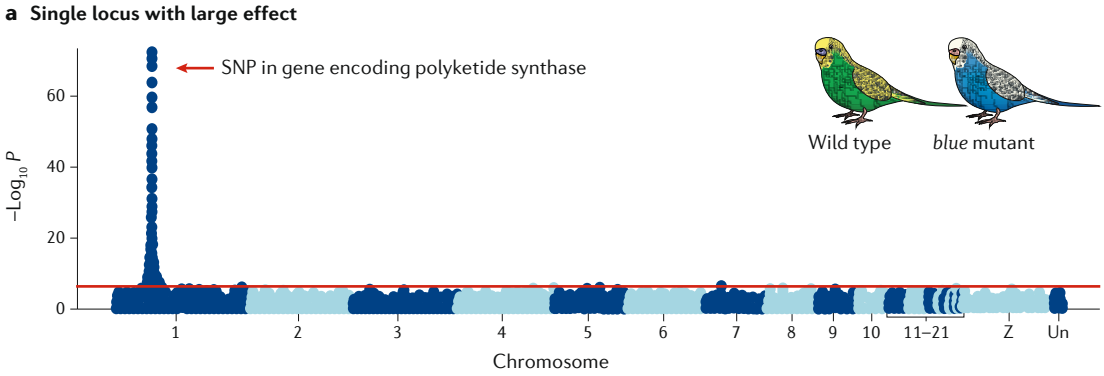
Large-effect mutations are not only found in coding sequences. In common canaries (*Serinus canaria*), a single regulatory mutation in the splice donor site of the *SCARB1* gene causes a change of plumage coloration from wild-type yellow to white[30]; *SCARB1* encodes a membrane receptor for high-density lipoprotein and is a mediator of carotenoid uptake. The SNP in the splice donor site causes abnormal splicing, leading to exon skipping and the loss of gene function. Similarly, compelling population genomic evidence suggests that a single transposable element insertion in an intron of the *cortex* gene gave rise to the British melanic form of the peppered moth (*Biston betularia*), which, driven by natural selection, spread rapidly through industrial Britain; however, further studies are needed to prove the causal role of this mutation[31]. Variation at *cortex* (and two other loci, *WntA* and *optix*) has also been associated with the diverse colour patterns of *Heliconius* butterfly species and races[32–35].

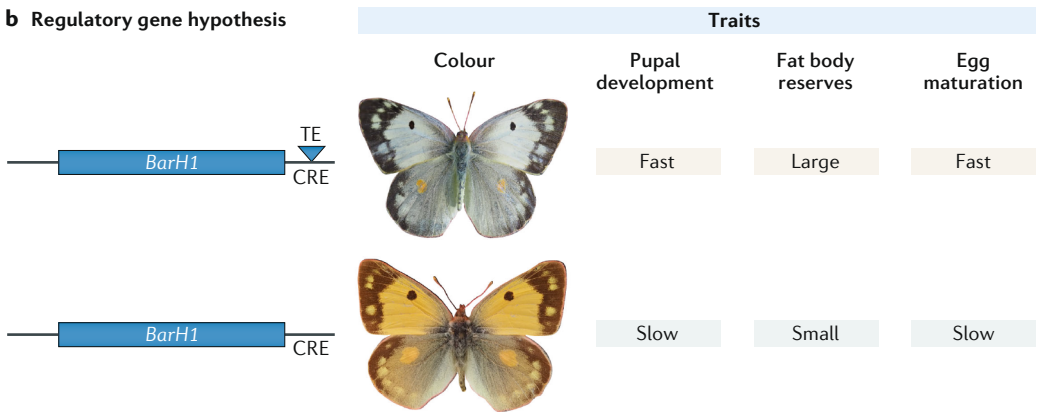
In other cases, alleles with major effects on phenotype can involve *cis*-regulatory alleles that encode complex spatial expression patterns. For example, differences in stripe patterning between two closely related species of the cichlid fish (*Pundamilia nyererei* and *Haplochromis sauvagei)* are controlled by a 1.1-kb *cis*-regulatory element (CRE) at the *agrp2* gene, which encodes agouti-related peptide 2, a protein that belongs to the agouti family and is involved in melanosome aggregation[36,37]. Similarly, *cis*-regulatory changes at the *LAR1* gene underlie spatial pattern variation in floral colours between two species of monkeyflowers, *Mimulus lewisii* and *Mimulus cardinalis*. The R2R3-MYB transcription factor encoded by *LAR1* positively regulates the expression of flavonol synthase, which redirects metabolite flux from the production of pink anthocyanins to the biosynthesis of colourless flavonols, resulting in white petal in areas where *LAR1* is expressed[27].

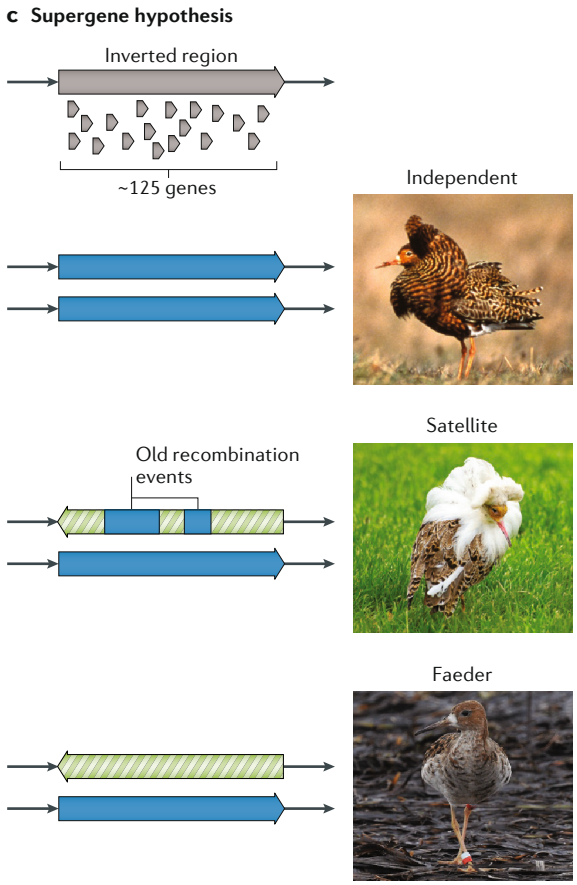
In summary, and somewhat contrary to the gradualist expectations of many evolutionary biologists, it is surprisingly common that evolutionary change is controlled by alleles, and even single mutations, with large effects.

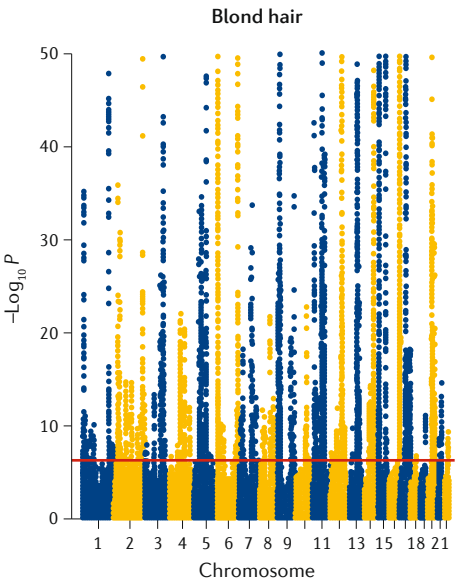
Complex adaptive polymorphisms. Some of the most compelling examples of adaptive coloration controlled by major-effect loci are polymorphisms, particularly those involving complex phenotypes comprising multiple traits that are locally stable. Polymorphisms can be maintained if their selection coefficient varies through

◀ Fig. 2 | **Signalling through colour. a**,**b** | An important function of colour is signalling to conspecifics, especially to possible future mating partners. During the mating season, male superb fairywrens (*Malurus cyaneus*) adopt conspicuous breeding plumage of contrasting shades of iridescent blue and black to attract females (part **a**). The prominent red swelling of the female Celebes crested macaque (*Macaca nigra*) advertises its sexual receptivity (part **b**). **c** | Concealment through background matching allows animals to escape predation. The pygmy seahorse *Hippocampus bargibanti* has detailed resemblances to the gorgonian coral it inhabits. **d**,**e** | Warning coloration or aposematism is another effective way to avoid predation: toxic prey display conspicuous colour patterns that warn predators of their toxicity. Some aposematic species converge to the same colour pattern and therefore share the costs of teaching predators, which is known as Müllerian mimicry[16]. The *Ranitomeya* genus of poison dart frogs provides one example. Multiple toxic *Ranitomeya* species act as models for the four morphs of the similarly toxic *R. imitator*[17] (part **d**). By contrast, in Batesian mimicry, a palatable species mimics a toxic one to gain protection against predators. For example, the butterfly *Hypolimnas misippus* resembles the unpalatable morphs of *Danaus chrysippus* (part **e**). Part **a** credit for superb fairywren: Piia Kopsa/Alamy. Part **b** credit for crested macaque: Ernie Janes/Alamy. Part **c** credit for pygmy seahorse (*Hippocampus bargibanti*): Shahar Shabtai/Alamy. Part **d** used with permission of The University of Chicago Press, from Mimetic divergence and the speciation continuum in the mimic poison frog Ranitomeya imitator. Twomey, E., Vestergaard, J. S., Venegas, P. J. & Summers, K. **187**, 2 (2015)[17]; permission conveyed through Copyright Clearance Center, Inc. Part **e** credit for African queen (*Danaus chrysippus*): Vonkara1/iStock/Getty; credit for Danaid eggfly (*Hypolimnas misippus*): Paolo_Toffanin/iStock/Getty; credit for African queen (*Danaus chrysippus*): komkrit tonusin/Alamy; credit for *Hypolimnas misippus*, Danaid eggfly: The Natural History Museum/Alamy; credit for *Hypolimnas misippus* male Danaid eggfly: Domiciano Pablo Romero Franco/Alamy.


a Single locus with large effect
SNP in gene encoding polyketide synthase
Wild type
blue mutant
−Log10 P
60
40
20
0
1
2
3
4
5
6
7
8
9
10
11–21
Z
Un
Chromosome


b Regulatory gene hypothesis
Traits
Colour
Pupal development
Fat body reserves
Egg maturation
TE
BarH1
CRE
Fast
Large
Fast
BarH1
CRE
Slow
Small
Slow


c Supergene hypothesis
Inverted region
~125 genes
Independent
Old recombination events
Satellite
Faeder


d Polygenic inheritance
Blond hair
−Log10 P
50
40
30
20
10
0
1
2
3
4
5
6
7
8
9
11
13
15
18
21
Chromosome

◀ Fig. 3 | **Genetic architecture of colour loci. a** | In many cases, the genetic basis of colour polymorphism has been narrowed down to single mutations of large effect. Shown are the results of a genome-wide association study to identify single-nucleotide polymorphisms (SNPs) associated with colour morphs in budgerigars. Only one significant peak is detected, located on chromosome 1. The red line indicates the Bonferroni-corrected critical value ($\alpha = 0.05$). Fine mapping and functional analysis confirmed that the causative variant associated with the SNP is a non-synonymous mutation in *MuPKS*, a gene encoding a polyketide synthase, that abolishes yellow pigmentation and produces the blue morph[9]. **b** | The regulatory gene hypothesis for the inheritance of complex adaptive phenotypes states that a single gene regulates all of the multiple co-inherited traits. An example that potentially supports this hypothesis is *Colias* butterflies, which present two morphs that differ in wing colour as well as numerous life history-related traits[46]. These forms are controlled by a single gene, *BarH1*, which is overexpressed in the white morph during development because of a transposable element (TE) insertion in its *cis*-regulatory element (CRE). *BarH1* supresses orange coloration in the wings and, either directly or as energetic side effect, triggers the switch in multiple life-history traits. **c** | The supergene hypothesis states that a complex polymorphism is controlled by a single locus containing numerous genes that are inherited together owing to reduced levels of genetic recombination (often caused by an inversion). In the ruff, three morphs differing in body size, ornamentation and mating behaviour are controlled by a supergene (large grey arrow) that contains approximately 125 genes (small grey arrows). The supergene has three alleles (large blue and green arrows): an ancestral allele (blue) that in homozygotes produces the 'independent' morph; a derived allele (green) that in heterozygotes produces the 'faeder' morph (it is lethal in homozygotes); and a recombinant allele that is also found in heterozygotes and produces 'satellites'. Hatched regions indicate reduced recombination due to the inversion. **d** | In other cases, colour variation is highly polygenic. The plot shows the results of a genome-wide association study for loci associated with blond hair colour in humans, which is influenced by many loci scattered across the genome[64]. The red line indicates the threshold of significance ($\alpha = 5 \times 10^{-8}$). Part **a** is adapted with permission from REF.[9], Elsevier. Part **b** is adapted from REF.[46], CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/). Part **c** is adapted from REF.[102], Springer Nature Limited. Part **d** is reprinted from REF.[64], CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

time, across space or with local allele frequency or population density. For example, alternative mating strategies adopted by different phenotypic morphs can be subject to negative frequency-dependent selection such that no single allele becomes fixed and multiple distinct morphs are maintained, as illustrated by the ruff, a wading bird in which alternative male forms differ in physiology, body size, behaviour and plumage[38–40]. However, the stability of such complex polymorphisms depends on a genetic architecture in which linkage between the multiple co-adapted traits ensures their joint inheritance. Early theoretical discussion focused on two possible genetic architectures that could stabilize such polymorphisms: the 'regulatory gene' hypothesis and the 'supergene' hypothesis (reviewed in REF.[41]).

The regulatory gene hypothesis states that changes in a single 'master' gene, acting at the top of regulatory gene networks and with multiple downstream targets, could lead to morph variation between individuals[41]. Several examples of complex polymorphisms associated with colour traits are now known to be controlled by a single master gene. In *Papilio* butterflies, the *doublesex* locus encodes a transcription factor that regulates both sex determination and a female-limited colour polymorphism[42–44]. *Colias* butterflies exhibit two alternative life-history strategies differing in wing colour (and a suite of other traits) that are maintained by an energetical trade-off[45] (FIG. 3b). White butterflies invest their resources in reproductive and somatic development instead of producing expensive colour pigments, which results in faster pupal development, larger fat body reserves and faster egg maturation compared with orange females. This polymorphism has been mapped to a transposable element insertion in the white morph (also known as *alba*), downstream of *BarH1*, a gene that encodes a homeobox transcription factor that also affects eye colour in *Drosophila melanogaster*[46]. *BarH1* is expressed in white but not orange forewing scales and CRISPR–Cas9 knockout experiments have shown that it supresses orange coloration in the wing. It is not known whether *BarH1* interacts with other genes to control the developmental and physiological traits characteristic of the white form or whether the differences between morphs in life-history traits are a side effect of the suppression of coloration, with the energy saved by reducing the production of pigment granules on the wings reallocated to reproductive and somatic development. In the Gouldian finch, two alternative colour morphs, black and red, differ in multiple traits, including social dominance, stress hormone levels, sperm-length plasticity and sex allocation[47,48]. These morphs have been associated with a small intergenic region (~72 kb) of the Z chromosome that likely regulates follistatin, a glycoprotein that antagonizes transforming growth factor-β (TGFβ) superfamily function and has been indirectly linked to variation in plumage colour in flycatchers and warblers[49,50]. Colour polymorphism in the wall-lizard described earlier also includes behavioural and ecological differences between morphs, and is controlled by two small regulatory regions near genes associated with pterin and carotenoid metabolism[6]. By contrast, in some plants, a single transcription factor can regulate multiple pigmentation enzymes that, for example, control major differences in flower hue[51,52]; however, such transcription factors typically do not control balanced polymorphisms.

The supergene hypothesis proposes that a single locus (that is, the supergene) contains multiple functional elements — potentially multiple genes — that control a complex, adaptive phenotype[41,53]. A key element of the supergene hypothesis is the maintenance of linkage between co-adapted loci. Although it was initially suggested that translocation of genes might have a role in establishing closer linkage[54], empirical evidence suggests that instead the occurrence of successive co-adapted mutations in close proximity is more important, a process known as Turner's sieve[55]. Linkage is therefore assured through physical proximity in the genome. There may also be ongoing selection for further reduction in recombination, either through localization to genomic regions with naturally low recombination rates, such as centromeres[53], or though accumulation of structural variation, such as transposable element insertions, tandem duplication and inversions[42,56]. However, reduced recombination itself could lead to an increase in structural mutations due to the reduced efficacy of purifying selection, so the direction of causality may be unclear. Inversions are the most common mechanism for reducing recombination and are associated with many supergenes underlying colour polymorphisms, such as in birds (including the white-throated sparrow and the ruff) and in butterflies (such as *Heliconius numata*)[38,39,57,58]. Inversions can link multiple traits, such

Negative frequency-dependent selection
An evolutionary process in which the fitness of a genotype or phenotype depends on its frequency in the population relative to other genotypes or phenotypes such that its fitness decreases as its frequency increases.

Purifying selection
The selective removal of deleterious alleles from the population.

as the coloration, developmental, physiological and behavioural traits that differ between ruff alternative mating morphs[38,39] (FIG. 3c).

Understanding supergene architecture can shed light on their long-term maintenance. Polymorphism has generally been explained through ecological mechanisms, such as frequency-dependent fitness and environmental heterogeneity, but an intriguing possibility is that intrinsic genetic mechanisms contribute to the maintenance of variation in ecologically important traits. Balanced lethal systems are genomic regions in which the accumulation of deleterious mutations on alternative haplotypes ensures heterozygote advantage and the long-term maintenance of variation[59]. Supergene alleles associated with colour polymorphisms can similarly accumulate deleterious variation, owing to a small effective population size and low recombination. For example, there is evidence that the supergene alleles in *H. numata* have higher transposable element activity and severe deleterious fitness effects in homozygous genotypes compared with heterozygous genotypes[60]. Similarly, dominant supergene alleles in the ruff are lethal in homozygotes[38,39]. The reduced fitness of homozygotes effectively leads to heterozygote advantage, irrespective of the relative fitness of colour phenotypes, and could play an important part in maintaining dramatic colour polymorphisms. These examples demonstrate how a better understanding of the underlying genetics can inform our interpretation of the evolutionary processes maintaining variation in natural populations.

Polygenic adaptation. The ability to conduct genomic studies on larger populations through developments in high-throughput sequencing has led to increasing interest in the genetic basis for highly polygenic inheritance, but it has generally not been well explored in relation to coloration. In traits where major-effect loci have been identified as underlying adaptive change, there is often also a proportion of phenotypic variation controlled by loci of small effect, such as the distribution of effect sizes among colour forms of *Heliconius* butterflies[61]. In other cases, variation in populations is largely polygenic. For example, a genome-wide association study (GWAS) identified more than 150 genes influencing variation in abdominal pigmentation among lines derived from a single *D. melanogaster* population, including many previously uncharacterized genes that were confirmed to influence pigmentation by use of RNA interference-based targeted gene knockdown[62] (BOX 1). Similarly, structural colours in butterflies have a polygenic basis[63], and in humans, analysis of traits such as hair colour has demonstrated the power of large-scale GWAS to identify hundreds of small-effect loci influencing highly polygenic traits[64,65] (FIG. 3d). Coloration traits can therefore offer a tractable system for the analysis of polygenic architecture, and this is clearly an area with potential for future work. One specific future goal is to understand the underlying cause of different genetic architectures. It is possible that adaptive differences between populations and species in traits under strong directional selection will be more likely to be under the control of major-effect loci compared with traits under stabilizing selection, but this hypothesis will need to be tested by comparative analysis across many species.

In summary, coloration traits demonstrate a wide range of genetic architectures, ranging from highly polygenic through to major-effect loci and adaptive supergenes. While considerable progress has been made in describing these architectures in specific cases, major outstanding questions remain regarding the reasons for differences between species: How common are supergenes and why do they arise in some species but not others? When are major loci expected to regulate adaptive change? And, are there circumstances in which a polygenic architecture is more likely? One way to address these questions is to understand how genetic changes regulate development, and in turn how developmental processes might shape genetic architecture.

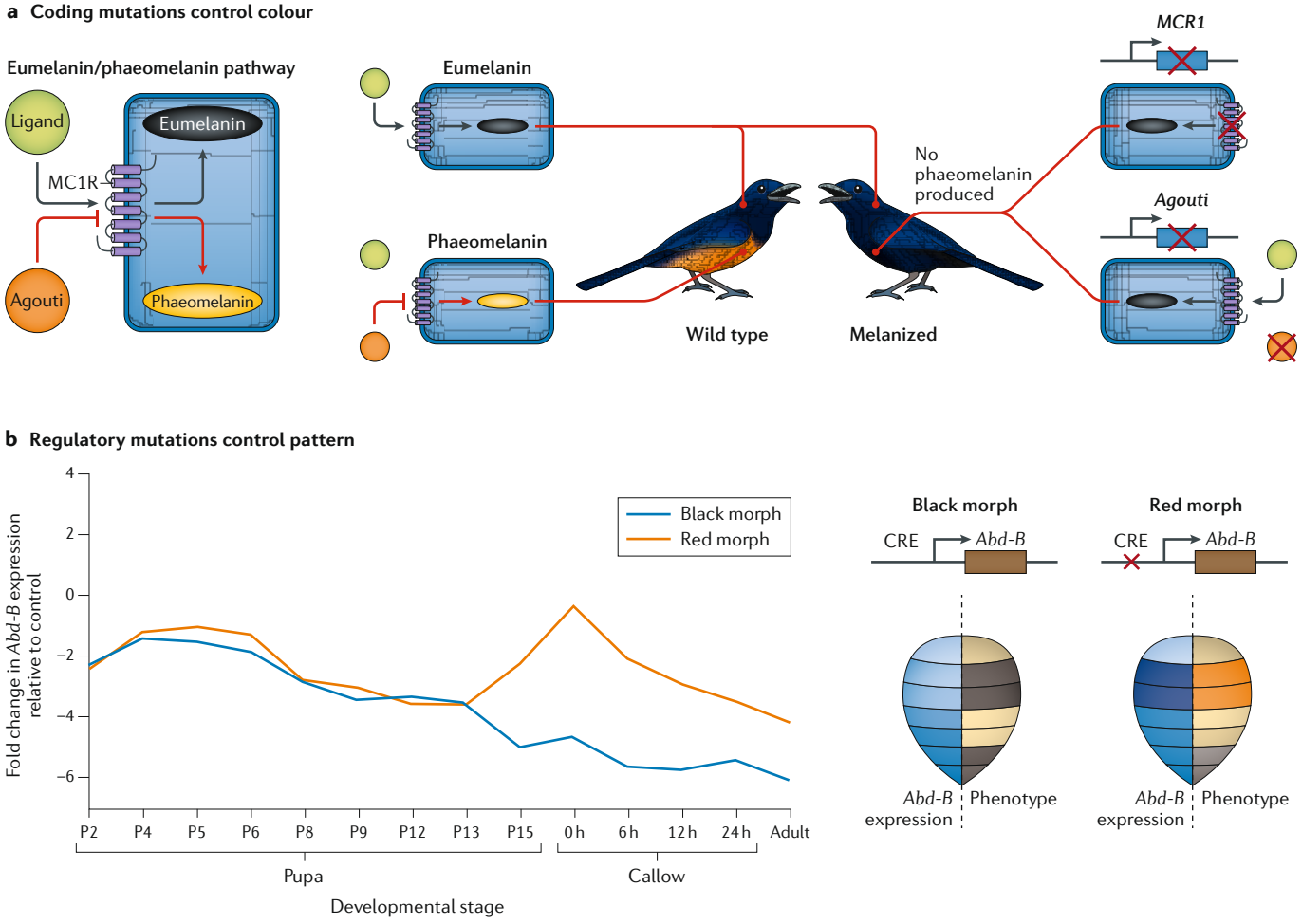
Fig. 4 | **Development can shape evolution. a** | Colour is often controlled by coding mutations, particularly in genes found in the biochemical pathways that lead to the production of pigments. In the melanin synthesis pathway, ligand-bound melanocortin 1 receptor (MC1R) activates the production of black and/or brown eumelanins, whereas if it is bound to its antagonist agouti signalling protein (Agouti), the default production of phaeomelanin is restored. The island flycatcher presents two morphs, one with chestnut belly plumage resulting from the production of phaeomelanin, and one with only black plumage as a consequence of only eumelanin being produced. Two different mutations can produce the black morph: a mutation in *MC1R*, which activates the receptor and leads to the production of eumelanin; and a mutation in *Agouti*, which abrogates the function of Agouti as an MC1R antagonist and allows the ligand to bind to MC1R, activating the production of eumelanin[24]. **b** | Colour pattern is often under the control of regulatory mutations. In bumblebees, a *cis*-regulatory element (CRE) mutation in the *Abd-B* gene changes its pattern of gene expression late in development (at the callow stage), which leads to a change in abdominal colour pattern from black to red[69]. **c** | Input–output genes are found in the middle of developmental networks. They integrate complex spatiotemporal information from patterning genes and, as a result, trigger different cell differentiation programmes by regulating downstream genes, which leads to distinct phenotypes[83]. Although patterning genes can trigger the development of phenotypes, they are likely to have pleiotropic effects. By contrast, some genes (such as input–output genes) may be able to initiate differentiation programmes while showing low pleiotropy, which may make them more likely to be involved in evolutionary change. A potential example of such genes is *bric à brac 1* and *bric à brac 2* (Bab) in *Drosophila melanogaster*, which integrate information about the abdominal segment identity from *Abd-B* and sex information from the female (F)-specific and male (M)-specific isoforms of *doublesex* (*Dsx*), and in response regulate the pigment-related genes *tan* and *yellow* to control melanization[73]. Segments A5 and A6 are melanized in males (Bab expression is repressed, allowing *yellow* and *tan* to be expressed (blue and black arrows)), whereas in females they are not (Bab is expressed and represses expression of *yellow* and *tan* (red arrows)). Solid lines indicate direct interactions, whereas dashed lines indicate indirect interactions. Part **b** is adapted with permission from REF.[69], PNAS.

Effective population size
The number of individuals in an idealized population that would show the same degree of genetic drift as seen in the real population.

Polygenic inheritance
Also known as polygenicity. The genetic control of a phenotype by multiple genes of small effect.

Genome-wide association study
(GWAS). A study that correlates genetic variation between individuals across the genome with phenotypic variation among those same individuals, typically in a wild population. Associated regions are inferred to contain causal variants controlling phenotypic variation. More generally known as 'association studies'.

Stabilizing selection
A selective force that maintains the population phenotypic mean and eliminates extreme values.



How does development shape evolution?

A greater understanding of evolution comes from considering the role of development in shaping and potentially channelling the evolutionary process. Coloration provides multiple examples in which mutations seem to be distributed non-randomly across the genome, with convergence in phenotype often influenced by the same genes or the same types of mutations in different lineages. Studies of these examples provide insight into the relative importance of coding versus regulatory

Convergence
The independent evolution of similar features in different lineages or species.

Pleiotropy
The effect of a single mutation on multiple aspects of the phenotype.

change, the extent to which the same genes are reused in evolution and the influence of the environment in the development of phenotypes.

Coding changes control colour but regulatory changes control pattern. Morphological variation can be produced by changes in coding or *cis*-regulatory regions of genes. It has been proposed that coding sequence changes in genes that have multiple functions, especially those involved in signalling pathways that are deployed in multiple developmental contexts, will commonly be deleterious: even if a mutation is beneficial to one of its functions, it may be detrimental to another[66]. Nonetheless, coding mutations do commonly play a role in coloration, but they tend to have large phenotypic effects because any tissue or developmental stage expressing the gene will be affected. For example, flycatchers homozygous for either of the previously mentioned coding mutations in the *MC1R* and *Agouti* genes lack the ability to produce phaeomelanin and therefore have only black plumage[24] (FIG. 4a).

By contrast, subtle differences in local patterning are likely to be caused by regulatory changes, which can be more modular than coding changes[22]; beneficial mutations that arise are therefore less likely to have negative pleiotropic effects on other functions[66]. Modularity in gene expression can be temporal and/or spatial in nature. For example, the evolution of dark *D. melanogaster* in montane habitats is the result of a tissue-specific regulatory change that reduces expression of *ebony*, a gene that supresses the production of dark melanin[67]. Studies using GFP reporter constructs showed that this dark regulatory allele reduces expression of *ebony* in the abdomen, but not in other tissues, such as the head and halteres, when compared with the light regulatory alleles. Indeed, regulatory mutations are responsible for every evolutionary pigmentation change in *Drosophila* spp. that has been mapped to date[68], supporting the hypothesis that phenotypic variation is, on many occasions, controlled by *cis*-regulatory mutations. In another example, a GWAS with wild samples showed that a red-to-black colour switch in the bumblebee *Bombus melanopygus* is controlled by *cis*-regulatory changes in an ~4-kb intergenic block downstream of *Abd-B. Abd-B* is a Hox gene that determines abdominal segment identity during development, which has been shown to positively regulate the pigmentation gene *yellow* and the transcription factor genes *bric à brac 1* and *bric à brac 2* (together referred to as Bab). Quantitative PCR analysis demonstrated that this regulatory change specifically expands the highly conserved expression domain of the *Abd-B* gene late in development; earlier functions of the transcription factor AbdB in determining abdominal fate are unaffected[69] (FIG. 4b). Similarly, the *optix* gene encodes a transcription factor that is involved in brain and eye development in *D. melanogaster* (and presumably in all insects)[70], but in nymphalid butterflies it is also expressed during wing development and acts to paint red colours onto the wings, with *cis*-regulatory alleles defining the precise patterns to be painted in any particular species. It seems likely that co-option of *optix* into wing patterning has occurred through gene regulatory changes that do not disrupt its existing functions in structures such as the eye and brain.

However, the simple model in which CREs are highly modular and tissue specific is not always confirmed by experimental studies; indeed, many individual CREs control expression in multiple different tissues[71]. For example, *optix* regulatory elements are associated with specific wing colour pattern elements (such as red rays or bands) in *Heliconius* populations, which is consistent with modularity. Yet, knocking out one of these CREs by CRISPR–Cas9 gene editing results in a broad phenotype that involves multiple pattern elements, which suggests pleiotropy rather than modularity[72]. This result is consistent with earlier analysis of pigmentation in *D. melanogaster*, which used reporter transgenes to show that individual mutations within a CRE can have highly tissue-specific effects and avoid deleterious pleiotropy, even when the CRE is multifunctional[73]. This observation highlights the fact that it is the level of pleiotropy of a mutation that affects the likelihood of fixation of a novel variant, not the degree to which a particular gene or CRE has multiple functions. In other words, CREs may be developmentally multifunctional, affecting expression in multiple tissues, but evolutionarily modular in that specific mutations are tissue specific. It is therefore misleading to label a gene or a CRE as being 'highly pleiotropic' — the term should be applied to mutations rather than functional elements themselves; this more restrictive definition of pleiotropy has recently been termed 'selectional pleiotropy'[74].

Developmental networks influence which genes are evolutionary hotspots. A pattern emerging from analysis of the genetic basis for adaptive change in coloration and other phenotypes is the repeated co-option of the same genes into convergent evolutionary change, a phenomenon that implies constraint on which genes are likely to influence a particular phenotype. For example, linkage and association mapping has identified mutations in *OCA2* that underlie amelanic phenotypes in at least two populations of cave fish and in a domesticated corn snake population, and human albinism and skin pigmentation phenotypes in African populations[23,75–77]. *OCA2* encodes a membrane transporter protein thought to be involved in uptake of tyrosine, a melanin precursor molecule and, given it has a specific melanization function, it is perhaps unsurprising that it is repeatedly co-opted to regulate adaptive coloration[78]. Similarly, MC1R and its antagonist Agouti control a signalling cascade leading to melanin production and have been associated with melanin-based coloration in many species, including deer mice[79,80], white-sand lizards[81] and island flycatchers[24] (FIG. 4a). It is more surprising that genes with multiple roles in development, such as transcription factors and signalling molecules, can be repeatedly targeted during evolutionary change – yet there are now many such examples, including WntA, a Wnt-signalling pathway ligand that has been shown by both association studies and functional analysis to be repeatedly involved in establishing colour patterns among butterflies[34,82]. One hypothesis is that the shape of developmental networks might constrain which genes

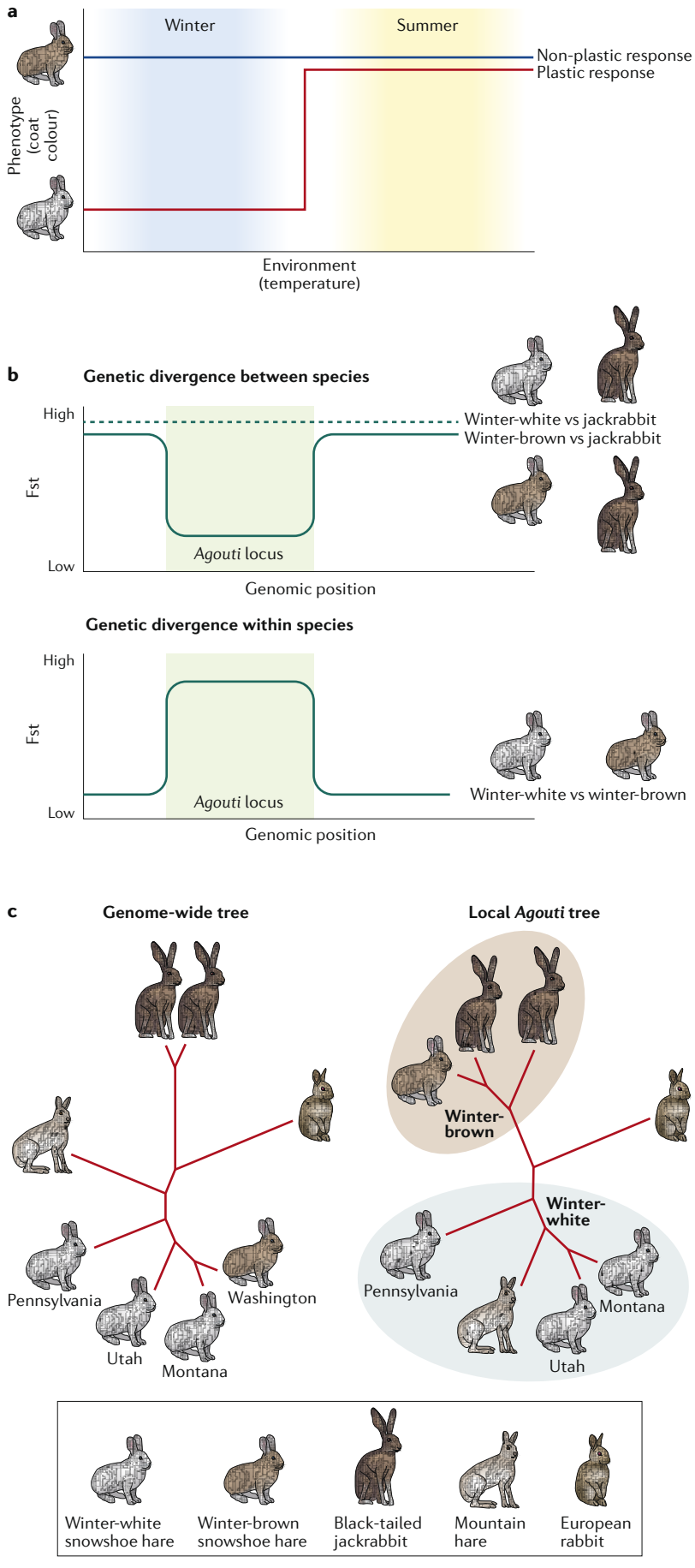


Fig. 5 | **An introgressed allele is responsible for the loss of phenotypic plasticity in snowshoe hares.** **a** | Snowshoe hares can show phenotypic plasticity in coat colour and maintain seasonal camouflage by moulting into a white coat in winter. However, some snowshoe hares have lost this phenotypic plasticity and retain the brown coat all year round. **b** | This winter-brown coat phenotype is controlled by an allele of the *Agouti* locus, and measures of genetic divergence indicate this allele has introgressed from black-tailed jackrabbits. Genetic divergence (Fst) measured between snowshoe hares and jackrabbits (top) shows that winter-brown snowshoe hares are less divergent than winter-white snowshoe hares from jackrabbits at the *Agouti* locus. Genetic divergence within snowshoe hares shows that winter-white and winter-brown snowshoe hares are highly divergent at the *Agouti* locus, but probably not at other loci. **c** | A tree based on genome-wide pairwise sequence similarity between snowshoe hares and related species recapitulates the species phylogeny. By contrast, a tree based on local sequence similarity at the *Agouti* locus indicates that the snowshoe hare winter-brown allele is more similar to the jackrabbit *Agouti* allele than to snowshoe hare winter-white alleles, providing evidence for introgression from jackrabbits into snowshoe hares. Drawings in parts **a**–**c** are adapted with permission from REF.[86], AAAS.

are most free to evolve, with 'input–output' genes lying at developmental switch points being most likely to regulate novelty (FIG. 4c), but the generality of this hypothesis needs to be tested[83]. A likely example of input–output genes are the Bab transcription factor genes, which integrate segment and sex information to trigger sex-specific and segment-specific coloration in various *Drosophila* species[62,68] (FIG. 4c).

Phenotypic plasticity. Colour variation can also occur as a result of phenotypic plasticity, in which environmental cues lead to the same genotype resulting in different phenotypes, often by inducing gene expression changes that alter developmental trajectories (FIG. 5a). Background-dependent and temperature-dependent colour morphs in locusts, temperature-dependent pigmentation in *Drosophila* spp., seasonal polyphenism in butterflies and seasonal camouflage in snowshoe hares are all examples of phenotypic plasticity[84–86]. Epigenetic and gene expression studies have shown that the darker abdominal pigmentation that occurs in *D. melanogaster* in response to lower temperatures involves changes in the transcriptional regulation of the pigmentation gene *tan*[87]; this differential gene expression is part of the phenotype of a plastic response to the environment. In this case dark coloration is likely an adaptive response to optimize heat absorption in colder environments.

A further challenge will be to understand how plasticity itself can evolve, by identifying genes that control differential responses of populations to environmental cues, as depicted by reaction norms. Currently, there are two hypotheses: either plasticity is largely regulated by alterations in the mechanisms that sense environmental cues, which would implicate genes unrelated to coloration; or plasticity could be switched on by changes in downstream genes that respond to the environmental cues, and the genetic loci involved would be similar to

'Input–output' genes
Genes that integrate complex spatiotemporal information and trigger alternative developmental outputs.

Phenotypic plasticity
The ability of a single genotype to produce a range of phenotypes depending on the environmental conditions.

Polyphenism
A type of phenotypic plasticity in which a single genotype can produce two or more discrete alternative phenotypes depending on the environmental conditions.

Reaction norms
Patterns of phenotypic expression of a single genotype across differing environmental conditions.

Balancing selection
Selective processes that maintain multiple alleles in a population, such as negative frequency-dependent selection and heterozygote advantage.

Introgression
The transfer of genetic material from one species to another through hybridization.

those associated with the evolution of colour. Emerging results seem to support the second hypothesis. For example, genetic variation occurs at a CRE of the Bab locus in *D. melanogaster* strains with different reaction norms, and Bab protein is responsible for altering *tan* expression in response to temperature cues[87,88] (FIG. 4c); regulatory alleles at the *Agouti* locus control plasticity in the seasonal coat colour of snowshoe hares[86] (FIG. 5a); and the loss of plasticity in side-blotched lizards is associated with changes at two genes in the melanin pathway (*PREP* and *PRKAR1A*)[89]. Although these early studies suggest that colour plasticity is often regulated by genes in pigmentation pathways, rather than by changes to the genetic mechanisms that underlie detection of environmental cues, more examples will be needed to test whether this pattern holds up as plasticity becomes better understood.

Understanding adaptation through colour

The origins of variation. Mutation and recombination have commonly been considered to be the primary sources of genetic variation required for adaptation to occur. However, an emerging pattern from genomic studies of coloration and other traits is that allelic variants underlying adaptive change are often more ancient than the populations in which they are found. For example, colour morphs of the invasive ladybird, *Harmonia axyridis*, are controlled by regulatory alleles at the locus encoding the transcription factor pannier[90,91]. Genomic comparisons of the two morphs have shown that genetic divergence is far greater at this locus than at surrounding genomic regions, which suggests an ancient origin for these colour alleles. Such patterns of high sequence divergence within species can result from either long-term balancing selection at colour variants or more recent introgression of divergent alleles between species. Phylogenetic reconstruction of *Agouti* alleles show that the loss of plasticity in the winter coat colour of snowshoe hares is the result of introgression of an allele from the black-tailed jackrabbit that gives rise to brown winter coats[86] (FIG. 5b,c), and in mountain hares, which present a winter-white/winter-grey polymorphism, the winter-grey variant introgressed through hybridization with the Iberian hare[92]. Similarly, a dark form of the wolf that is common in forest habitats has alleles very similar to those in domestic dogs, suggesting recent introgression[93]. These examples demonstrate how useful genetic variation can be acquired through introgression from close relatives. At a much finer scale, phylogenetic analysis of two adjacent CREs of *optix* that control red elements in multiple *Heliconius* species has shown a complex history involving multiple introgression events that generated novel combinations of phenotypes[94].

Introgression can also help to explain the evolution of supergene architectures that would otherwise require multiple genetic steps. In butterflies, introgression from *Heliconius pardalinus* has given rise to a novel allele at the *H. numata* supergene locus[60]. The levels of interspecies divergence associated with the introgressed allele were substantially lower than across the genome as a whole, which is consistent with introgression rather than long-term balanced polymorphism. Similarly, in the white-throated sparrow the allele associated with the tan morph is more closely related to variants in a sister species, the Harris sparrow, than to white morph alleles in the same species[58]. Overall, it seems likely that introgression between species is a common and important evolutionary process[95].

Observing and testing evolution in the wild. An understanding of the genetic basis for adaptive traits can facilitate studies of natural selection in the wild. Direct tests for selection in natural environments can now take advantage of the ability to genotype the alleles under selection (FIG. 6). For example, enclosure experiments with deer mice (*Peromyscus maniculatus*) on different soil substrates demonstrated rapid allele frequency changes at the *Agouti* locus that support the hypothesis of selection for crypsis against visual predators[96]. Similarly, experimental field studies that tracked genetic variation in natural populations of *Timema* stick insects over 25 years have provided support for negative frequency-dependent selection on pattern variants (striped and unstriped)[97]. Short-term changes in allele frequencies occurred substantially more rapidly at the patterning locus than at other similarly sized segments of the genome, indicating selection at these loci. In both cases, genomic data add value to selection experiments by facilitating estimates of selection on genotypes rather than solely on phenotype, and by allowing changes in allele frequency at selected loci to be compared with the genomic average.

Natural selection also leaves a signature in the genome of wild populations. Therefore, an alternative approach to studying selection is to search for this genomic signature, so as to estimate the past action of natural selection. Colour variants are especially good for such studies, as there is often a well-understood link between colour variation and genotype. For example, population genetic analysis of deer mice in the Nebraska Sand Hills demonstrated that coat colour in this species is composed of multiple traits and that each of these colour traits is associated with a distinct mutation within the *Agouti* locus[79]. These mutations have been independently selected, with larger-effect variants showing evidence for stronger selection. A recent study explored the extensive radiation of *Heliconius* butterflies and provided evidence for strong natural selection at multiple colour loci across almost all sampled populations[98]. These results support field evidence for consistently strong selection on colour forms in these butterflies[99,100], but additionally show that colour loci are some of the most strongly selected regions in the genome.

Conclusions and future perspectives

In many ways, studies of coloration have been at the forefront of evolutionary genetics and have served as a model for investigating complex traits. Coloration represents some of the most striking and accessible adaptive traits in the natural world and was well understood before the genomics revolution, and colour-related traits are some of the best understood examples of genotype–phenotype links. Nonetheless, the extent to which colour traits are representative of other

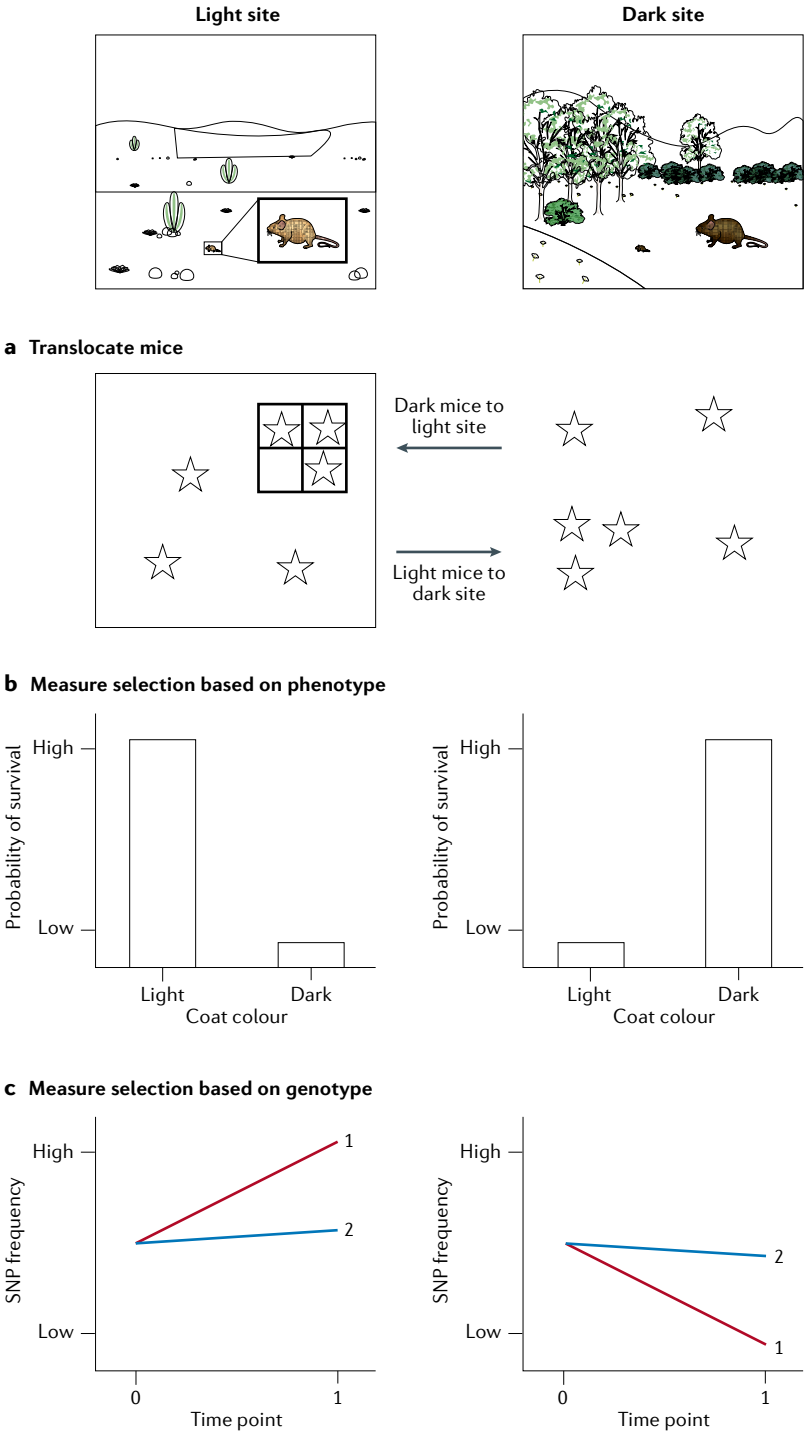


Fig. 6 | **Estimating natural selection.** A mouse species presents two cryptic coat colour phenotypes (light and dark) that inhabit two types of soil colour (light and dark, respectively). **a** | A translocation experiment can be used to test whether the phenotypes are adaptive[96]. Individuals from the light-coloured population on light-coloured soil (sampling sites indicated by blue stars) are translocated to enclosures in a dark-coloured soil area (blue stars in a grid) and dark-coloured individuals (sampling sites indicated by orange stars) are transferred from the dark-coloured site to enclosures in the light-coloured site (orange stars in a grid). **b** | Adult survival is measured and the probability of survival is plotted against coat colour. The plots illustrate the hypothetical case in which coat colour is discrete and adaptive: light-coloured mice have a higher survival rate than dark-coloured mice in light-coloured soil sites, and vice versa. If coat colour were not adaptive, the probability of survival would be similar for either soil type regardless of coat colour. **c** | Selection at single-nucleotide polymorphisms (SNPs) known to be associated with coat colour can be estimated by calculating the change in their frequency between different time points: before selection (0) and after selection (1). The plotted lines indicate frequency changes of individual SNPs. SNP 1 shows positive selection in the light-coloured site and negative selection in the dark-coloured site, whereas SNP 2 is neutral in both sites. Thus, translocation experiments that incorporate genomic data provide the opportunity to track allele frequencies in the wild and measure the direction and strength of selection.

phenotypes remains unclear – for example, loci of large effect might be more common among colour traits if pigmentation has a relatively simple genetic basis compared with other morphological structures. Challenges remain, notably in understanding polygenic inheritance and the genetic control of phenotypic plasticity, but colour phenotypes offer excellent opportunities to study both of these phenomena. The long-term goal will be to synthesize what we learn from individual examples to infer general patterns regarding the evolution of genetic architectures. Finally, while genomic studies have informed our understanding of the action of natural selection in wild populations, there have been fewer attempts to study sexual selection in the wild using genomic approaches. Perhaps in the future, genetic and genomic analyses will provide similar insights into the action of sexual selection on colour phenotypes in natural populations.

Published online 7 May 2020

1. Cuthill, I. C. et al. The biology of color. *Science* **357**, eaan0221 (2017).
2. Protas, M. E. & Patel, N. H. Evolution of coloration patterns. *Annu. Rev. Cell Dev. Biol.* **24**, 425–446 (2008).
3. Quicke, D. L. J. *Mimicry, Crypsis, Masquerade and Other Adaptive Resemblances* (Wiley, 2017).
4. Caro, T. & Allen, W. L. Interspecific visual signalling in animals and plants: a functional classification. *Phil. Trans. R. Soc. B Biol. Sci.* **372**, 20160344 (2017).
5. Hill, G. E., Hill, G. E. & McGraw, K. J. *Bird Coloration: Function and Evolution* (Harvard Univ. Press, 2006).
6. Andrade, P. et al. Regulatory changes in pterin and carotenoid genes underlie balanced color polymorphisms in the wall lizard. *Proc. Natl Acad. Sci. USA* **116**, 5633–5642 (2019).
This study identifies two genes that underlie red, orange and yellow pigmentation in the wall lizard. The colour morphs also differ in their physiology and behaviour and the identified loci are candidates for involvement in those traits.
7. Shawkey, M. D. & Hill, G. E. Carotenoids need structural colours to shine. *Biol. Lett.* **1**, 121–124 (2005).
8. Shawkey, M. D., Morehouse, N. I. & Vukusic, P. A protean palette: colour materials and mixing in birds and butterflies. *J. R. Soc. Interface* **6**, S221–S231 (2009).
9. Cooke, T. F. et al. Genetic mapping and biochemical basis of yellow feather pigmentation in budgerigars. *Cell* **171**, 427–439.e21 (2017).
10. Nisar, N., Li, L., Lu, S., Khin, N. C. & Pogson, B. J. Carotenoid metabolism in plants. *Mol. Plant.* **8**, 68–82 (2015).
11. Moran, N. A. & Jarvik, T. Lateral transfer of genes from fungi underlies carotenoid production in aphids. *Science* **328**, 624–627 (2010).
12. Lopes, R. J. et al. Genetic basis for red coloration in birds. *Curr. Biol.* **26**, 1427–1434 (2016).
13. Mundy, N. I. I. et al. Red carotenoid coloration in the zebra finch is controlled by a cytochrome P450 gene cluster. *Curr. Biol.* **26**, 1435–1440 (2016).
14. Hébant, C. & Lee, D. W. Ultrastructural basis and developmental control of blue iridescence in Selaginella leaves. *Am. J. Botany* **71**, 216–219 (1984).

15. Vukusic, P., Sambles, J. R., Lawrence, C. R. & Wootton, R. J. Quantified interference and diffraction in single *Morpho* butterfly scales. *Proc. Biol. Sci.* **266**, 1403–1411 (1999).
16. Vane-Wright, R. I. A unified classification of mimetic resemblances. *Biol. J. Linn. Soc.* **8**, 25–56 (1976).
17. Twomey, E., Vestergaard, J. S., Venegas, P. J. & Summers, K. Mimetic divergence and the speciation continuum in the mimic poison frog Ranitomeya imitator. *Am. Nat.* **187**, 205–224 (2015).
18. Smith, Da. S. Phenotypic diversity, mimicry and natural selection in the African butterfly Hypolimnas misippus L. (Lepidoptera: Nymphalidae). *Biol. J. Linn. Soc.* **8**, 183–204 (1976).
19. Fisher, R. A. *The Genetical Theory of Natural Selection* (Clarendon, 1930).
20. Orr, H. A. The population genetics of adaptation: the distribution of factors fixed during adaptive evolution. *Evolution* **52**, 935–949 (1998).
21. Orr, H. A. The genetic theory of adaptation: a brief history. *Nat. Rev. Genet.* **6**, 119–127 (2005).
22. Martin, A. & Courtier-Orgogozo, V. in *Diversity and Evolution of Butterfly Wing Patterns: An Integrative Approach* (eds. Sekimura, T. & Nijhout, H. F.) 59–87 (Springer, 2017).
23. Martin, A. & Orgogozo, V. The loci of repeated evolution: a catalog of genetic hotspots of phenotypic variation. *Evolution* **67**, 1235–1250 (2013).
24. Uy, J. A. C. et al. Mutations in different pigmentation genes are associated with parallel melanism in island flycatchers. *Proc. Biol. Sci.* **283**, 20160731 (2016).
25. Hoekstra, H. E., Hirschmann, R. J., Bundey, R. A., Insel, P. A. & Crossland, J. P. A single amino acid mutation contributes to adaptive beach mouse color pattern. *Science* **313**, 101–104 (2006).
26. Römpler, H. et al. Nuclear gene indicates coat-color polymorphism in mammoths. *Science* **313**, 62 (2006).
27. Yuan, Y.-W., Rebocho, A. B., Sagawa, J. M., Stanley, L. E. & Bradshaw, H. D. Competition between anthocyanin and flavonol biosynthesis produces spatial pattern variation of floral pigments between Mimulus species. *Proc. Natl Acad. Sci. USA* **113**, 2448–2453 (2016).
This study identifies a gene encoding an R2R3-MYB transcription factor as the causal gene underlying flower spatial pattern variation in two species of monkeyflowers with different pollinators.
28. Albert, N. W. et al. A conserved network of transcriptional activators and repressors regulates anthocyanin pigmentation in eudicots. *Plant. Cell* **26**, 962–980 (2014).
29. Kellenberger, R. T. et al. Emergence of a floral colour polymorphism by pollinator-mediated overdominance. *Nat. Commun.* **10**, 63 (2019).
30. Toomey, M. B. et al. High-density lipoprotein receptor SCARB1 is required for carotenoid coloration in birds. *Proc. Natl Acad. Sci. USA* **114**, 5219–5224 (2017).
31. van't Hof, A. E. et al. The industrial melanism mutation in British peppered moths is a transposable element. *Nature* **534**, 102–105 (2016).
32. Nadeau, N. J. et al. The gene cortex controls mimicry and crypsis in butterflies and moths. *Nature* **534**, 106–110 (2016).
33. Reed, R. D. et al. Optix drives the repeated convergent evolution of butterfly wing pattern mimicry. *Science* **333**, 1137–1141 (2011).
34. Mazo-Vargas, A. et al. Macroevolutionary shifts of WntA function potentiate butterfly wing-pattern diversity. *Proc. Natl Acad. Sci. USA* **114**, 10701–10706 (2017).
35. Westerman, E. L. et al. Aristaless controls butterfly wing color variation used in mimicry and mate choice. *Curr. Biol.* **28**, 3469–3474.e4 (2018).
36. Kratochwil, C. F. et al. Agouti-related peptide 2 facilitates convergent evolution of stripe patterns across cichlid fish radiations. *Science* **362**, 457–460 (2018).
This article describes an example of phenotypic convergence through independent mutations at the same locus. Different *cis*-regulatory mutations near the gene encoding agouti-related peptide 2 lower its expression and produce convergent striped phenotypes in cichlid fish.
37. Zhang, C. et al. Pineal-specific agouti protein regulates teleost background adaptation. *Proc. Natl Acad. Sci. USA* **107**, 20164–20171 (2010).
38. Küpper, C. et al. A supergene determines highly divergent male reproductive morphs in the ruff. *Nat. Genet.* **48**, 79–83 (2015).
This article, together with reference 39, shows that a complex phenotype encompassing differences in body size, ornamentation and aggressive and mating behaviour in the ruff is controlled by a supergene containing an inversion.
39. Lamichhaney, S. et al. Structural genomic changes underlie alternative reproductive strategies in the ruff (*Philomachus pugnax*). *Nat. Genet.* **48**, 84–88 (2016).
40. Widemo, F. Alternative reproductive strategies in the ruff, Philomachus pugnax: a mixed ESS? *Anim. Behav.* **56**, 329–336 (1998).
41. Thompson, M. J. & Jiggins, C. D. Supergenes and their role in evolution. *Heredity* **113**, 1–8 (2014).
42. Iijima, T. et al. Parallel evolution of Batesian mimicry supergene in two Papilio butterflies, P. polytes and P. memnon. *Sci. Adv.* **4**, eaao5416 (2018).
43. Kunte, K. et al. Doublesex is a mimicry supergene. *Nature* **507**, 229–232 (2014).
44. Nishikawa, H. et al. A genetic mechanism for female-limited Batesian mimicry in *Papilio* butterfly. *Nat. Genet.* **47**, 405–409 (2015).
45. Graham, S. M., Watt, W. B. & Gall, L. F. Metabolic resource allocation vs. mating attractiveness: adaptive pressures on the "alba" polymorphism of *Colias* butterflies. *Proc. Natl Acad. Sci. USA* **77**, 3615–3619 (1980).
46. Woronik, A. et al. A transposable element insertion is associated with an alternative life history strategy. *Nat. Commun.* **10**, 1–11 (2019).
47. Toomey, M. B. et al. A non-coding region near *Follistatin* controls head colour polymorphism in the Gouldian finch. *Proc. Biol. Sci.* **285**, 20181788 (2018).
48. Kim, K.-W. et al. Genetics and evidence for balancing selection of a sex-linked colour polymorphism in a songbird. *Nat. Commun.* **10**, 1852 (2019).
49. Lehtonen, P. K. et al. Candidate genes for colour and vision exhibit signals of selection across the pied flycatcher (Ficedula hypoleuca) breeding range. *Heredity* **108**, 431–440 (2012).
50. Toews, D. P. L. et al. Plumage genes and little else distinguish the genomes of hybridizing warblers. *Curr. Biol.* **26**, 2313–2318 (2016).
51. Hopkins, R. & Rausher, M. D. Identification of two genes causing reinforcement in the Texas wildflower Phlox drummondii. *Nature* **469**, 411–414 (2011).
52. Hopkins, R. & Rausher, M. D. Pollinator-mediated selection on flower color allele drives reinforcement. *Science* **335**, 1090–1092 (2012).
53. Schwander, T., Libbrecht, R. & Keller, L. Supergenes and complex phenotypes. *Curr. Biol.* **24**, R288–R294 (2014).
54. Ford, E. B. *Genetic Polymorphism* (Faber & Faber, 1965).
55. Turner, J. R. G. On supergenes. I. The evolution of supergenes. *Am. Nat.* **101**, 195–221 (1967).
56. Martin, S. H. et al. Whole-chromosome hitchhiking driven by a male-killing endosymbiont. *PLoS Biol.* **18**, e3000610 (2020).
57. Joron, M. et al. Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature* **477**, 203–206 (2011).
58. Tuttle, E. M. et al. Divergence and functional degradation of a sex chromosome-like supergene. *Curr. Biol.* **26**, 344–350 (2016).
59. Muller, H. J. Genetic variability, twin hybrids and constant hybrids, in a case of balanced lethal factors. *Genetics* **3**, 422–499 (1918).
60. Jay, P. et al. Mutation accumulation in chromosomal inversions maintains wing pattern polymorphism in a butterfly. Preprint at *bioRxiv* https://doi.org/10.1101/736504 (2019).
61. Papa, R. et al. Multi-allelic major effect genes interact with minor effect QTLs to control adaptive color pattern variation in heliconius erato. *PLoS One* **8**, e57033 (2013).
62. Dembeck, L. M. et al. Genetic architecture of abdominal pigmentation in drosophila melanogaster. *PLoS Genet.* **11**, e1005163 (2015).
63. Brien Melanie, N. et al. Phenotypic variation in *Heliconius erato* crosses shows that iridescent structural colour is sex-linked and controlled by multiple genes. *Interface Focus* **9**, 20180047 (2019).
64. Morgan, M. D. et al. Genome-wide study of hair colour in UK Biobank explains most of the SNP heritability. *Nat. Commun.* **9**, 1–10 (2018).
65. Hysi, P. G. et al. Genome-wide association meta-analysis of individuals of European ancestry identifies new loci explaining a substantial fraction of hair color variation and heritability. *Nat. Genet.* **50**, 652–656 (2018).
66. Carroll, S. B., Grenier, J. K. & Weatherbee, S. D. *From DNA to Diversity: Molecular Genetics and the Evolution of Animal Design* 2nd edn (Blackwell, 2005).
67. Rebeiz, M., Pool, J. E., Kassner, V. A., Aquadro, C. F. & Carroll, S. B. Stepwise modification of a modular enhancer underlies adaptation in a drosophila population. *Science* **326**, 1663–1667 (2009).
68. Massey, J. H. & Wittkopp, P. J. in *Current Topics in Developmental Biology* Vol. 119 (Orgogozo, V.) 27–61 (Elsevier, 2016).
69. Tian, L. et al. A homeotic shift late in development drives mimetic color variation in a bumble bee. *Proc. Natl Acad. Sci. USA* **116**, 11857–11865 (2019).
By analysing convergent colour phenotypes in bumblebees, this study demonstrates how shifts in spatial and temporal expression can reduce pleiotropy such that a crucial developmental Hox gene can play a role in phenotypic diversification.
70. Seimiya, M. & Gehring, W. J. The Drosophila homeobox gene optix is capable of inducing ectopic eyes by an eyeless-independent mechanism. *Development* **127**, 1879–1886 (2000).
71. Jory, A. et al. A survey of 6,300 genomic fragments for cis-regulatory activity in the imaginal discs of Drosophila melanogaster. *Cell Rep.* **2**, 1014–1024 (2012).
72. Lewis, J. J. et al. Parallel evolution of ancient, pleiotropic enhancers underlies butterfly wing pattern mimicry. *Proc. Natl Acad. Sci. USA* **116**, 24174–24183 (2019).
73. Rogers, W. A. et al. Recurrent modification of a conserved cis-regulatory element underlies fruit fly pigmentation diversity. *PLoS Genet.* **9**, e1003740 (2013).
74. Paaby, A. B. & Rockman, M. V. The many faces of pleiotropy. *Trends Genet.* **29**, 66–73 (2013).
75. Saenko, S. V. et al. Amelanism in the corn snake is associated with the insertion of an LTR-retrotransposon in the *OCA2* gene. *Sci. Rep.* **5**, 17118 (2015).
76. Johanson, H. C., Chen, W., Wicking, C. & Sturm, R. A. Inheritance of a novel mutated allele of the *OCA2* gene associated with high incidence of oculocutaneous albinism in a Polynesian community. *J. Hum. Genet.* **55**, 103–111 (2010).
77. Crawford, N. G. et al. Loci associated with skin pigmentation identified in African populations. *Science* **358**, eaan8433 (2017).
78. Bilandžija, H., Ma, L., Parkhurst, A. & Jeffery, W. R. A potential benefit of albinism in astyanax cavefish: downregulation of the oca2 gene increases tyrosine and catecholamine levels as an alternative to melanin synthesis. *PLoS One* **8**, e80823 (2013).
79. Linnen, C. R. et al. Adaptive evolution of multiple traits through multiple mutations at a single gene. *Science* **339**, 1312–1316 (2013).
80. Linnen, C. R., Kingsley, E. P., Jensen, J. D. & Hoekstra, H. E. On the origin and spread of an adaptive allele in deer mice. *Science* **325**, 1095–1098 (2009).
81. Rosenblum, E. B., Rompler, H., Schoneberg, T. & Hoekstra, H. E. Molecular and functional basis of phenotypic convergence in white lizards at White Sands. *Proc. Natl Acad. Sci. USA* **107**, 2113–2117 (2009).
82. Martin, A. et al. Diversification of complex butterfly wing patterns by repeated regulatory evolution of a Wnt ligand. *Proc. Natl Acad. Sci. USA* **109**, 12632–12637 (2012).
83. Stern, D. L. & Orgogozo, V. Is genetic evolution predictable? *Science* **323**, 746–751 (2009).
84. Tanaka, S., Harano, K., Nishide, Y. & Sugahara, R. The mechanism controlling phenotypic plasticity of body color in the desert locust: some recent progress. *Curr. Opin. Insect Sci.* **17**, 10–15 (2016).
85. Järvi, V. V., Burg van der, K. R. L. & Reed, R. D. Seasonal plasticity in Junonia coenia (Nymphalidae): linking wing color, temperature dynamics, behavior. *J. Lepid. Soc.* **3**, 34–42 (2019).
86. Jones, M. R. et al. Adaptive introgression underlies polymorphic seasonal camouflage in snowshoe hares. *Science* **360**, 1355–1358 (2018).
This article describes the genetic basis of seasonal phenotypic plasticity of coat colour in snowshoe hares.
87. Gibert, J.-M., Mouchel-Vielh, E., De Castro, S. & Peronnet, F. Phenotypic plasticity through transcriptional regulation of the evolutionary hotspot gene tan in Drosophila melanogaster. *PLoS Genet.* **12**, e1006218 (2016).
88. Castro, S. D., Peronnet, F., Gilles, J.-F., Mouchel-Vielh, E. & Gibert, J.-M. bric à brac (bab), a central player in the gene regulatory network that mediates thermal

plasticity of pigmentation in Drosophila melanogaster. *PLoS Genet.* **14**, e1007573 (2018).
89. Corl, A. et al. the genetic basis of adaptation following plastic changes in coloration in a novel environment. *Curr. Biol.* **28**, 2970–2977.e7 (2018).
90. Ando, T. et al. Repeated inversions within a pannier intron drive diversification of intraspecific colour patterns of ladybird beetles. *Nat. Commun.* **9**, 3843 (2018).
91. Gautier, M. et al. The genomic basis of color pattern polymorphism in the harlequin ladybird. *Curr. Biol.* **28**, 3296–3302.e7 (2018).
92. Giska, I. et al. Introgression drives repeated evolution of winter coat color polymorphism in hares. *Proc. Natl Acad. Sci. USA* **116**, 24150–24156 (2019).
93. Anderson, T. M. et al. Molecular and evolutionary history of melanism in north american gray wolves. *Science* **323**, 1339–1343 (2009).
94. Wallbank, R. W. R. et al. Evolutionary novelty in a butterfly wing pattern through enhancer shuffling. *PLoS Biol.* **14**, e1002353 (2016).
This study shows that shuffling of two narrow enhancer regions between *Heliconius* species has led to phenotype diversification, exemplifying the importance of introgression in morphological innovation.
95. Marques, D. A., Meier, J. I. & Seehausen, O. A combinatorial view on speciation and adaptive radiation. *Trends Ecol. Evol.* **34**, 531–544 (2019).
96. Barrett, R. D. H. et al. Linking a mutation to survival in wild mice. *Science* **363**, 499–504 (2019).
97. Nosil, P. et al. Natural selection and the predictability of evolution in Timema stick insects. *Science* **359**, 765–770 (2018).
98. Moest, M. et al. Selective sweeps on novel and introgressed variation shape mimicry loci in a butterfly adaptive radiation. *PLoS Biol.* **18**, e3000597 (2020).
99. Mallet, J. & Barton, N. H. Strong natural selection in a warning-color hybrid zone. *Evolution* **43**, 421–431 (1989).
100. Kapan, D. D. Three-butterfly system provides a field test of müllerian mimicry. *Nature* **409**, 338–340 (2001).
101. Matsuoka, Y. & Monteiro, A. Melanin pathway genes regulate color and morphology of butterfly wing scales. *Cell Rep.* **24**, 56–65 (2018).
102. Jiggins, C. D. A flamboyant behavioral polymorphism is controlled by a lethal supergene. *Nat. Genet.* **48**, 7–8 (2016).
103. Rice, E. S. & Green, R. E. New approaches for genome assembly and scaffolding. *Annu. Rev. Anim. Biosci.* **7**, 17–40 (2019).
104. Pardo-Diaz, C., Salazar, C. & Jiggins, C. D. Towards the identification of the loci of adaptive evolution. *Methods Ecol. Evol.* **6**, 445–464 (2015).
105. Liu, B. H. *Statistical Genomics: Linkage, Mapping, and QTL Analysis* (CRC, 1997).
106. Pickar-Oliver, A. & Gersbach, C. A. The next generation of CRISPR–Cas technologies and applications. *Nat. Rev. Mol. Cell Biol.* **20**, 490–507 (2019).
107. Wilson, R. C. & Doudna, J. A. Molecular mechanisms of RNA interference. *Annu. Rev. Biophys.* **42**, 217–239 (2013).

Acknowledgements
The authors thank everyone at the Butterfly Genetics Group at the University of Cambridge for useful discussion and comments on the manuscript. A.O. was supported by a UK Natural Environment Research Council doctoral training partnership (grant NE/L002507/1) and C.D.J. was supported by a European Research Council (grant 339873 Speciation Genetics) and UK Biotechnology and Biological Sciences Research Council grant BB/R007500/1.

Author contributions
The authors contributed equally to all aspects of the article.

Competing interests
The authors declare no competing interests.

Peer review information
Nature Reviews Genetics thanks M. Carneiro, I. Saccheri and C. Wheat for their contribution to the peer review of this work.

Publisher's note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.