ABSTRACT

SUPPLE, MEGAN ANN. Phenotypic Evolution across the *Heliconius* Color Pattern Radiation. (Under the direction of Dr. W. Owen McMillan.)

In nature there exists a tremendous amount of morphological variation, both within species and between species. One key to understanding the origins of this biodiversity is to understand how genetic change translates to phenotypic variation. The extensive diversity within species and the striking mimicry between species of *Heliconius* butterflies makes it an exceptional system to study the evolution of adaptive traits. Here I use natural phenotypic variation across the *Heliconius* speciation continuum to understand the genetic basis and evolutionary history of variation in adaptive wing color patterns.

Using whole genome resequencing data of divergently colored races from hybrid zones of *H. erato*, I identify a 65-kb putative regulatory region modulating expression of the gene *optix*, which controls the spatial distribution of red across the forewings and hindwings in *Heliconius*. I then analyze hybrid zone genomic data from *H. melpomene*, a distantly related co-mimic of *H. erato*, and determine that the two species use the same genomic architecture to generate their mimetic phenotypes. Using additional phenotypic and phylogenetic sampling across the broader *H. erato* radiation, I further investigate the genomic architecture of this 65-kb region. I identify three distinct modules within this region that are associated with distinct red color pattern elements. These modules are likely enhancers of the gene *optix* that modulate distinct expression domains across the wing, resulting in the red color pattern elements. Within this modular architecture, I further narrow down functional regions and identify candidate transcription factors acting upstream of *optix* that modulate the presence of absence of the hindwing rays. These candidates show

differential binding affinities in the rays enhancer region and expression in the hindwings of *Heliconius*, as well as interesting known expression patterns across the *Drosophila* wing.

The identification of the functional regions enables the exploration of the evolutionary history of the adaptive alleles. I show that the rayed allele had a single origin within each of the mimetic species *H. erato* and *H. melpomene*, but it evolved independently between the two species. I demonstrate that the phenotypically recombinant race, *H. e. amalfreda*, evolved through reshuffling of pattern element specific enhancers between the traditional postman and rayed phenotypes. I see the same evidence of enhancer shuffling in generating the wing color pattern of *H. himera*, an incipient species of *H. erato*.

I then examine the processes that drive the genomic patterns of divergence that are used to identify functional variation. I demonstrate that peaks of divergence can be driven by selection within a population, rather than divergent selection between populations. This result suggests caution when interpreting peaks of divergence. Additionally, I show that *H. himera* is an incipient species from within the *H. erato* radiation that evolved due to selection on multiple loci, not just due to color pattern divergence. Combining new genomic techniques with extensive natural variation, I provide insights into how genomic changes can drive convergent and divergent phenotypes across a classic adaptive radiation.

Phenotypic Evolution across the *Heliconius* Color Pattern Radiation

by
Megan Ann Supple

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Biomathematics

Raleigh, North Carolina

2014

APPROVED BY:

_____	_____
W. Owen McMillan Co-chair	James Gilliam Co-chair
_____	_____
Dahlia Nielsen	Alison Motsinger-Reif
_____	_____
Nadia Singh	Brian Counterman

DEDICATION

To Kelly, Jill, Sabrina, Bosley, and of course, Charlie.

BIOGRAPHY

From the time I learned to ride horses when I was a kid, I spent as much time as possible at the barn. I would ride any horse I could get my hands on—including the unbroke and wild ones. My sister and I rode a pony named Sparkle (aka The Little Witch) who was being given away because she had the intelligence and athleticism to dump any rider. Our parents, not being knowledgeable about horses, thought keeping Sparkle was a good idea. Sparkle excelled at jumping, bucking, and keeping her rider humble...the kind of pony every kid should learn to ride on.

After growing up in sunny southern California, I headed to the University of Michigan for my undergraduate education. The snow and cold was not as much fun as I had envisioned. Apparently the summers in Michigan shift to the other thermal extreme, but I wouldn't know. Every summer I took the opportunity to seek adventure elsewhere, including the Sierra Nevada Mountains and Yellowstone National Park. After four years, I graduated from with a degree in Aerospace Engineering. Unlike most rocket scientists, I did not to put my degree to work. Instead I thru-hiked the Appalachian Trail, spending six months walking from Georgia to Maine.

After a stint living in a tent in Alaska, I moved to Washington State to pursue a career as a stable hand. While in Washington, I also worked as an EMT on an ambulance, retrained problem horses, and worked at a fecal DNA lab. I took up foxhunting and I am proud to say that, while in pursuit of the fox (or the hare or the dragged scent), I have fallen off horses on three different continents. Due to such accomplishments, I am one of the founders of

Hillbilly Farms, a world famous website devoted to our battle against perfection in the equestrian world. In addition to contributing fodder for the website and our blog, The Road Apple, I also write the ``Jumping Clinic'' column under the pseudonym George Morris.

I continued my progression down the latitudes to Raleigh, North Carolina to pursue higher education and more horses, opening Ponies on Probation, a branch of Hillbilly Farms. I spent my free time volunteering for a local equine rescue organization. I was the proud foster mom to a string of naughty ponies who needed to learn to behave themselves before anyone will adopt them. This involved regularly performing involuntary dismounts, lots of bruises, and at least one broken bone. I continued to partake in other outdoor sports as well, including rock climbing and canoeing, which is much easier on crutches than backpacking.

The southerly pull of gravity became too strong and I left North Carolina and moved to the humid jungles of Panama. I live in a small town full of monkeys and sloths, where the quiet is most often disturbed by the chatter of parrots and the maddening song of the cicadas. I have taken up extreme running—not extreme because of the distance or speed, but rather because of the hazards. In my single most dangerous week, I had close encounters with a crocodile, the poisonous fer-de-lance, and the notorious Gamboa unicorn. I now wander the jungles with a butterfly net in search of the elusive chupacabra.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 2

CHAPTER 3

CHAPTER 4

LIST OF FIGURES

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 1

The Genomics of an Adaptive Radiation: Insights Across the *Heliconius* Speciation Continuum

Published in *Ecological Genomics*:
Supple MA, Papa R, Counterman BA, and McMillan WO. 2014. The genomics of an adaptive radiation: Insights across the Heliconius speciation continuum. In C. Landry and N. Aubin-Horth (editors), *Ecological Genomics: Ecology and the Evolution of Genes and Genomes*. Springer, New York, NY.

The Genomics of an Adaptive Radiation: Insights Across the *Heliconius* Speciation Continuum

13

Megan Supple, Riccardo Papa, Brian Counterman, and W. Owen McMillan

Abstract

Fueled by new technologies that allow rapid and inexpensive assessment of fine scale individual genomic variation, researchers are making transformational discoveries at the interface between genomes and biological complexity. Here we review genomic research in *Heliconius* butterflies – a radiation characterized by extraordinary phenotypic diversity in warningly colored wing patterns and composed of a continuum of taxa across the stages of speciation. These characteristics, coupled with a 50-year legacy of ecological and behavioral research, offer exceptional prospects for genomic studies into the nature of adaptive differences and the formation of new species. Research in *Heliconius* provides clear connections between genotype, phenotype, and fitness of wing color patterns shown to underlie adaptation and speciation. This research is challenging our perceptions about how speciation occurs in the presence of gene flow and the role of hybridization in generating adaptive novelty. With the release of the first *Heliconius* genome assembly, emerging genomic studies are painting a dynamic picture of the evolving species boundary. As the field of speciation genomics moves beyond describing patterns, towards a more integrated understanding of the process of speciation, groups such as *Heliconius*, where there is a clear speciation continuum and the traits underlying adaptation and speciation are known, will provide a roadmap for identifying variation crucial in the origins of biodiversity.

M. Supple
Smithsonian Tropical Research Institute, Panama City, Panama

Biomathematics Program, North Carolina State University, Raleigh, NC 27695, USA

W.O. McMillan (✉)
Smithsonian Tropical Research Institute, Panama City, Panama
e-mail: McMillanO@si.edu

R. Papa
Department of Biology and Center for Applied Tropical Ecology and Conservation, University of Puerto Rico, Rio Piedras, San Juan 00921, Puerto Rico

B. Counterman
Department of Biological Sciences, Mississippi State University, Mississippi State, MS 39762, USA

C.R. Landry and N. Aubin-Horth (eds.), *Ecological Genomics: Ecology and the Evolution of Genes and Genomes*, Advances in Experimental Medicine and Biology 781,
DOI 10.1007/978-94-007-7347-9_13, © Springer Science+Business Media Dordrecht (outside the USA) 2014

Keywords
Adaptation • Association mapping • Genomic divergence • Hybridization • Introgression • Phenotypic evolution • Speciation

13.1 Introduction

Over 150 years ago, Henry Walter Bates published his first observations of butterfly diversity in the New World Tropics (Bates 1862). Exploring deep within the Amazon basin for over a decade, Bates documented extraordinary cases of mimicry in the vivid wing patterns of distantly related butterfly species. His writings provided Darwin with some of the most visually stunning examples of evolution by natural selection and the best evidence of a link between natural selection and speciation. Thanks to Bates and Fritz Müller, who arrived in Brazil a few years after Bates, butterflies, arguably more than any other group, contributed to the early establishment and acceptance of evolutionary theory (Carroll et al. 2009). Research on butterflies continues to be as relevant today as it was 150 years ago. Using modern technologies, there is an active research community encompassing most areas of ecology and evolution, ranging from the molecular details of vision to the analysis of human impact on biodiversity (Boggs et al. 2003; Kotiaho et al. 2005; Briscoe et al. 2010). This includes a vibrant genomics research community working on a number of different species, including passion-vine butterflies (*Heliconius* spp., Heliconius Genome Consortium 2012), monarchs (*Danaus plexippus*, Zhan et al. 2011), swallowtails (*Papilio* spp., O'Neil et al. 2010), the Glanville fritillary (*Melitae acinxia*, Hanski 2011), *Bicyclus anynana* (Brakefield et al. 2009), and *Lycaeides* (Gompert et al. 2012). The past several years have seen remarkable progress in the development of genomic resources in these species, culminating in the publication of the first two butterfly reference genomes (Zhan et al. 2011; Heliconius Genome Consortium 2012), with a number of additional genomes scheduled to be released in the coming year.

In this review, we examine emerging ecological and evolutionary genomic research addressing adaptation and speciation in *Heliconius* butterflies. Genomic research is ongoing in several butterfly species, but genomic studies on *Heliconius* are arguably further developed. Research on *Heliconius* provides a foundation to discuss larger issues relating to the nature of adaptive differences and the formation of new species. We begin with an overview of the *Heliconius* radiation – a radiation that has produced an extraordinary evolutionary continuum composed of divergent races and species at different stages of speciation (Mallet et al. 1998; Mallet 2008; Merrill et al. 2011a). Using this continuum as a backdrop, we review recent progress to (i) identify functional variation in the group and reconstruct the history of adaptive alleles, (ii) understand the importance of hybridization in speciation, and (iii) explore the genomic architecture that allows speciation to proceed in the face of gene flow. We conclude with a discussion of how to move beyond patterns of genomic variation to gain a deeper understanding of the processes that drive divergence and speciation in nature.

13.2 The *Heliconius* Radiation: A Primer

The butterfly subtribe Heliconiina (Lepidoptera: Nymphalidae: Heliconiinae) is restricted to the New World tropics and has a host of life history, ecological, and phenotypic traits that have long fascinated biologists and naturalists. Heliconiines get their common name, passion-vine butterfly, from their strong association with the passion flower family (Passifloraceae). Passion vines are protected by a diverse arsenal of cyanogenic compounds that are likely a by-product of an evolutionary arms race with heliconiines (Spencer 1988). Heliconiines have adopted this defensive tactic–evolving the ability to make and, in some cases sequester, cyanogenic glycosides (Engler et al. 2000; Engler-Chaouat and Gilbert 2007). These compounds render the bearer highly

distasteful and avian predators quickly learn to associate a wing color pattern with unpalatability (Chai 1986).

Within the subtribe, the genus *Heliconius* is characterized by an ecological shift to pollen feeding. Unlike most butterflies, which feed only on fluids (e.g. nectar, decomposing animals and fruits, and dung), *Heliconius* actively collect and process pollen (Gilbert 1972). The origin of pollen feeding is associated with subtle changes in morphology (Gilbert 1972; Krenn and Penz 1998; Eberhard et al. 2009), coupled with more profound changes in a host of life history and behavioral traits. In particular, the transition to pollen feeding is hypothesized to be important in the butterflies' ability to synthesize toxic compounds and to enable a very long adult life – one of the longest recorded for a butterfly (Gilbert 1972). Pollen feeding is also associated with a rapid increase in brain size (Sivinski 1989), the evolution of slower and more maneuverable flight, the development of large eyes with accentuated ultraviolet color vision (Briscoe et al. 2010), and the evolution of a suite of complex behaviors, including trapline feeding, gregarious roosting, and elaborate mating strategies (Brown 1981).

The *Heliconius* genus is best known for the extraordinary mimicry-related wing pattern diversity seen among its 43 species (Fig. 13.1). With over 400 distinct color pattern varieties, the group represents one of the most striking adaptive radiations in the animal kingdom. Repeated convergent and divergent evolution creates a colorful tapestry where distantly related species often look identical and closely related species or races can look strikingly different (Fig. 13.1). The complexity of this tapestry is exemplified by the parallel radiations of *H. erato* and *H. melpomene*, which, although phylogenetically distant and unable to hybridize, have converged on 25 different mimetic color patterns across the Neotropics (Fig. 13.2). Most of the diversity in these two species can be partitioned into two major phenotypic groups: (i) "postman" phenotypes, which have red on their forewing and either possess or lack a yellow hindwing bar; and (ii) "rayed" phenotypes, which have a yellow forewing band, a red patch on the proximal area of the forewing, and red rays on the hindwing. Variations on these themes generate the abundant pattern diversity we see in nature (Fig. 13.2b). In addition to the postman and rayed phenotypes, there are also a number of tiger striped *Heliconius*. For example, *H. numata* shows numerous sympatric color patterns races that are likely a result of strong selection pressure to mimic distantly related *Melinaea* species (Nymphalidae: Ithomiinae) (Brown and Benson 1974; Joron et al. 1999), which can vary dramatically in abundance over small spatial and temporal scales. Geographic variation and convergent evolution are common across *Heliconius* and the wing patterns of most species converge onto a handful of common color patterns, so called mimicry rings, which coexist locally (Mallet and Gilbert 1995). This convergence between species led to the original hypothesis of mimicry (Bates 1862) and *Heliconius* is now a classic example of Müllerian mimicry, where distantly related, but similarly distasteful, species converge on the same warningly colored pattern.

Divergence in wing color pattern is also associated with speciation due to the dual role of color pattern in signaling to predators and in mate selection. For example, *H. melpomene* and *H. cydno* are very closely related, broadly sympatric, and occasionally hybridize in nature. In this case, speciation is associated with, and reinforced by, divergence in mimetic color patterns. *Heliconius melpomene* is generally black with red and yellow markings and mimics *H. erato* (Flanagan et al. 2004) (Fig. 13.2); whereas, *H. cydno* is black with white or yellow markings and typically mimics *H. sapho* and *H. eleuchia* (see Fig. 13.1). Mate recognition involves visual attraction of males to females, which leads to strong color-pattern based assortative mating and disruptive sexual selection against hybrids (Jiggins et al. 2001b; Naisbit et al. 2001). Furthermore, there is ecological post-mating isolation that results from increased predation on hybrids due to their non-mimetic wing patterns (Merrill et al. 2012). Species boundaries are often associated with mimetic color pattern shifts, highlighting the pervasive role that color pattern evolution plays in reproductive isolation across the radiation (Mallet et al. 1998).

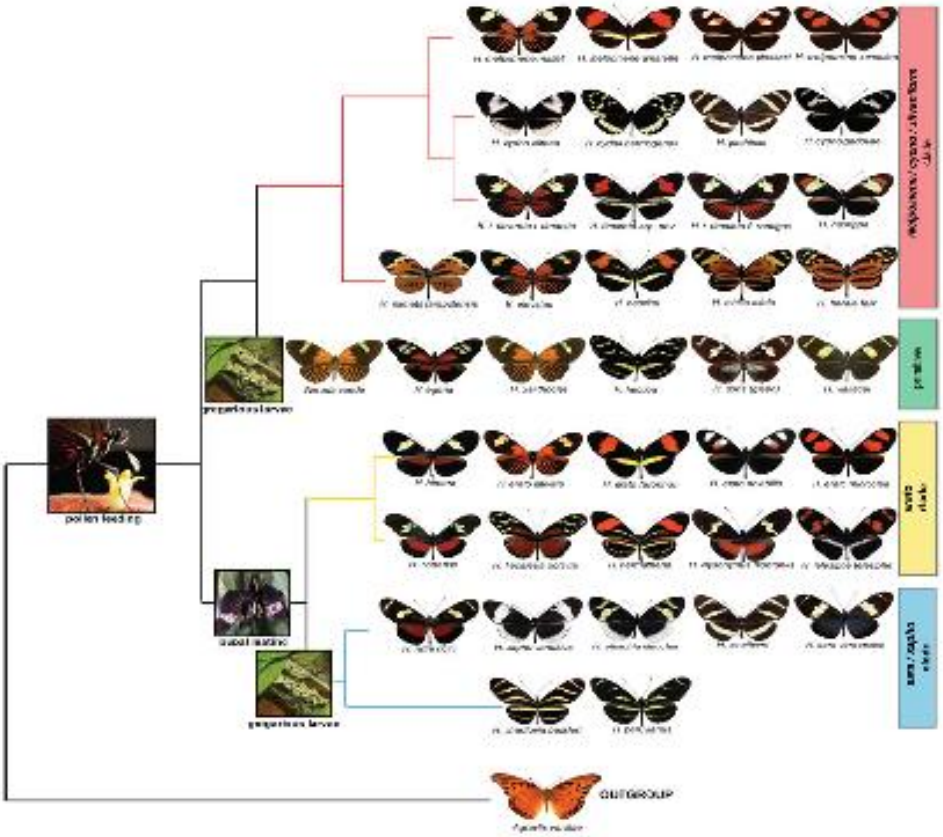


Fig. 13.1 Nature's palette – the *Heliconius* radiation. This representation of the major *Heliconius* clades incorporates over half of the 43 different *Heliconius* species and a large portion of wing pattern variation. Note the reuse of phenotypes across the phylogeny and the extreme phenotypic variation that exists within and between closely related species. The origins of several major ecological and behavioral innovations are indicated, including pollen feeding, pupal mating, and gregarious larvae. Pupal mating is the reproductive behavior where males search and guard pupae and mate with females as they eclose (Deinert et al. 1994). Pollen feeding and pupal mating likely evolved only once in the radiation. In contrast, gregarious aggregations of larvae evolved several times including the lineage containing the primitive *Heliconius*. Branch length does not necessarily reflect divergence time (Beltrán et al. 2007)

13.3 The Genetics of Color Pattern in *Heliconius*

Given the visually dynamic nature and clear importance of coloration in the *Heliconius* radiation, much effort has focused on characterizing the genetic basis of wing pattern variation in the group. The spectrum of pattern diversity in *Heliconius* is largely controlled by variation at a handful of loci of large phenotypic effect (Sheppard et al. 1985; Joron et al. 2006, 2011; Reed et al. 2011; Martin et al. 2012; Papa et al. 2013). These loci are phylogenetically conserved "hotspots" of phenotypic evolution, responsible for color pattern evolution in both convergent and divergent *Heliconius* species. Allelic variation at these loci produce complex phenotypic changes

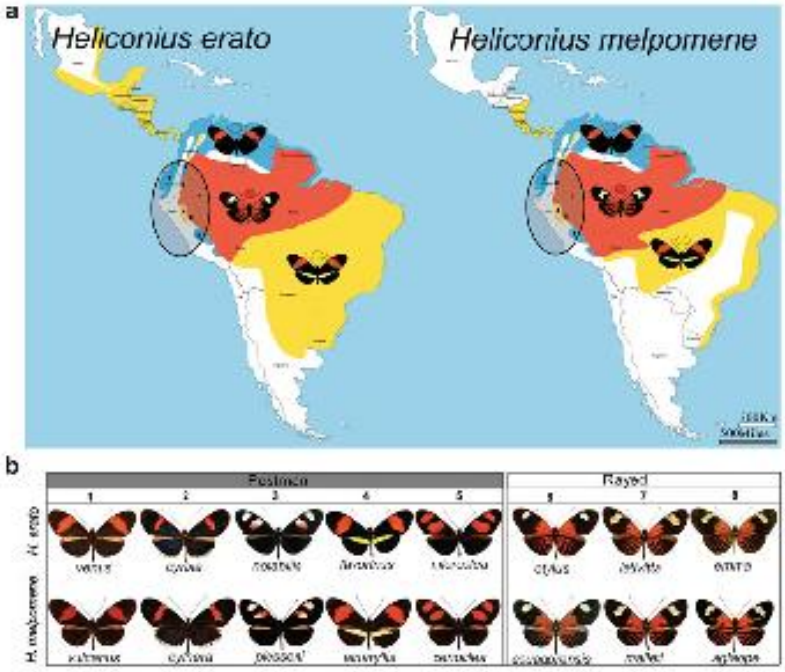


Fig. 13.2 Variations on a theme – parallel color pattern radiations in *Heliconius erato* and *Heliconius melpomene*. (**a**) Geographic distributions of the major phenotypes of *H. erato* and *H. melpomene*, showing the "rayed" phenotype throughout the Amazon basin and disjoint populations of the two variations of the "postman" phenotype. (**b**) Each row shows some of the color pattern divergence within the two concordant radiations, classified by the two major phenotypes ("postman" and "rayed"). Comparing color pattern phenotypes between the two species reflects the convergent color pattern evolution

that show spatial variation across wing surfaces (Fig. 13.3). The identification and characterization of these color pattern alleles has been a major step forward in understanding the evolution and development of lepidopteran wing patterns and provides a unique glimpse into the developmental genetic architecture of pattern evolution in nature.

The nomenclature surrounding the genetics of color pattern variation in *Heliconius* that has developed over the past 50 years is complicated; however, most described color pattern variation is due to the combined effects of four major loci. All four loci contain functional variation that affects distinct color pattern elements. For simplicity, we will refer to them as the "red switch", the "yellow switch", the "forewing melanin shutter", and the "global melanin shutter" (Fig. 13.3). The red switch controls the presence of several discrete red elements, including the forewing band, the proximal forewing "dennis" patch, and the hindwing rays (Fig. 13.3). The yellow switch causes a loss of yellow ommochrome pigments across both wing surfaces, resulting in a switch from yellow to white pattern elements (Linares 1997; Naisbit et al. 2003) (Fig. 13.3). This locus has largely been described from crosses of *H. cydno*, but probably underlies pattern differences in races of *H. sapho* and *H. eleuchia* as well. The two "shutters" (sensu Gilbert 2003) modulate the complex distribution of black/melanin across both wing surfaces. The forewing melanin shutter acts predominately in the central portion of the forewing, generating variation in the shape, position, and size of the forewing band (Fig. 13.3).

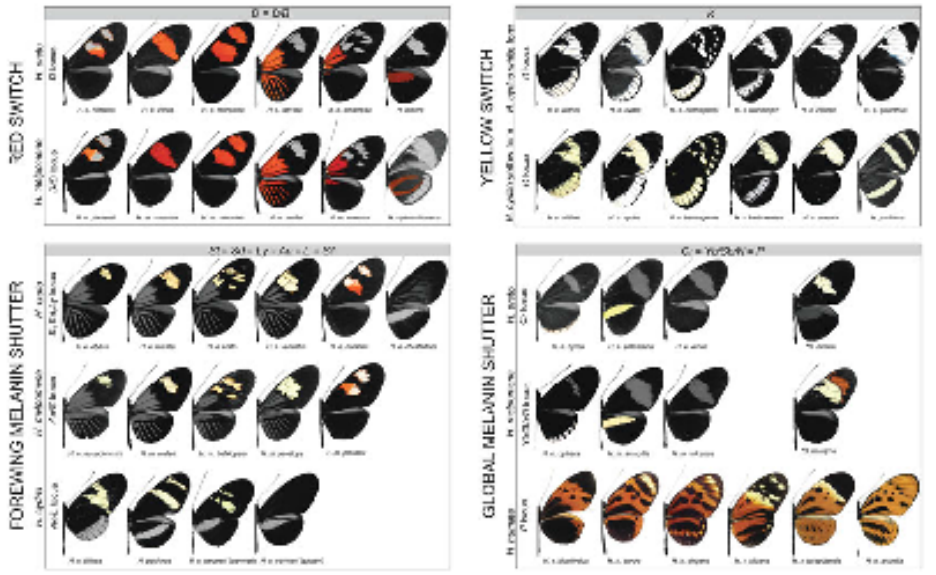


Fig. 13.3 Hotspots of phenotypic evolution. Although more than two dozen loci have been described, color pattern variation across *Heliconius* is largely regulated by four major effect loci – two color "switch" loci and two melanin "shutter" loci. These loci interact to create the natural diversity of *Heliconius* color patterns. The "red switch" controls various red elements across the wing surfaces and was originally described as multiple loci (Sheppard et al. 1985). The "yellow switch" changes forewing and hindwing pattern elements between white and yellow. The "forewing melanin shutter" controls the distribution of black/melanic wing scale cells on the forewing to either expose or cover white or yellow pattern elements and generates variation in forewing band shape, size and position. This locus also originally described as at least five distinct loci in different *Heliconius* species. Finally, the "global melanin shutter" is similar to the forewing melanin shutter, but it acts more broadly across the wing. It was similarly described as a number of distinct loci in different *Heliconius* species. Allelic variation at this locus can have multiple phenotypic effects across the wing including: (i) creating the white fringes on the fore wing and hindwing, (ii) causing the presence or absence of the yellow hindwing bar, and (iii) changing the shape and color of the forewing band in some species, including *H. himera* and *H. heurippa*. In addition, allelic variation at this locus, in the form of a series of localized inversions, controls all color and pattern variation in *H. numata* (see Joron et al. 2011)

The global melanin shutter affects the distribution of melanin across both wing surfaces to create variation in a number of red, yellow, and white pattern elements (Fig. 13.3). Due to the broad phenotypic effects of these color pattern loci, most were originally described as supergenes (Mallet 1989), or clusters of linked genes that facilitate co-segregation of adaptive variation (sensu Mather 1950).

The above synthesis is an oversimplification in several ways. First, there are other loci that have moderate size effects that contribute to pattern variation in important ways (Papa et al. 2013). This includes completely unexplored loci that alter the nanostructures of wing scale surfaces, causing light to scatter, which results in iridescent wing surfaces. Second, all of the major loci have pleiotropic effects that extend beyond the neat categorizations described above. Furthermore, most interact epistatically with each other to generate additional pattern diversity. For example, in *H. melpomene* the global melanin shutter interacts with the red switch and the forewing melanin shutter to finely control the color, size, and position of the forewing band. Finally, the magnitude of color pattern variation generated by allelic differences within these loci can be extreme. This is best exemplified in *H. numata*,

where the entire spectrum of wing pattern variation is controlled by variation at a single locus – the global melanin shutter (Fig. 13.3) (Joron et al. 2006).

13.4 Genomic Divergence Across the Speciation Continuum

A major strength of *Heliconius* as a genomic model system is that the radiation has produced a continuum of divergent races and species, which provide an excellent opportunity to study taxa at different stages of speciation (Fig. 13.4). At one end of the continuum, many divergent color pattern races freely hybridize, showing only weak reproductive isolation from each other. These racial boundaries are maintained primarily by selection on wing color patterns, with free gene flow across the rest of the genome. This heterogeneity in gene flow across the genome is ideal for identifying functional variation, as the genomes of divergent, hybridizing races should only differ at regions responsible for phenotypic differences. For some pairs of *Heliconius* taxa, speciation has progressed further. For example, the hybrid zone between *H. erato* and *H. himera* in Ecuador is characterized by the evolution of strong pre-mating isolation through assortative mating, but there is no evidence for hybrid inviability or sterility (McMillan et al. 1997). Both prezygotic and postzygotic isolation have been shown to contribute to isolation in

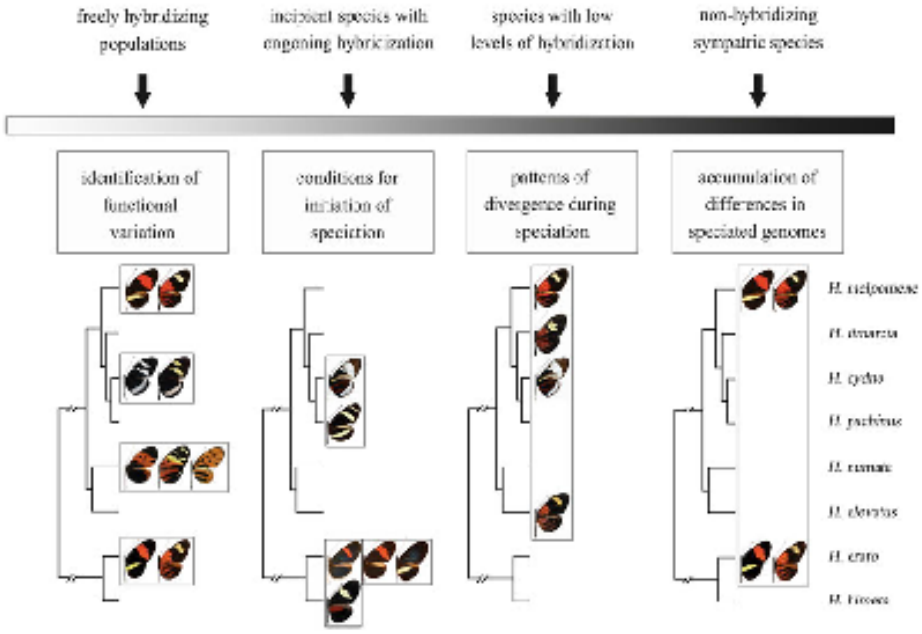


Fig. 13.4 Across the speciation continuum. The *Heliconius* radiation provides an exceptional opportunity to study taxa at different stages of speciation. The continuum, from freely hybridizing populations to non-hybridizing species, provides insight into a variety of key areas of research related to speciation. The genomes of freely hybridizing races should diverge only at genomic regions driving phenotypic divergence, allowing identification of important functional variation. Species pairs at the earliest stages of speciation give insights into how reproductive isolation is established. As reproductive isolation increases, but low levels of hybridization remain, patterns of divergence across the genome reflect the complex interplay between evolutionary forces and genome structure. Studying mimetic species pairs that do not hybridize gives insights into how genomes diverge with complete reproductive isolation and how distantly related species can converge on nearly identical phenotypes

other *Heliconius* hybrid zones, including the Colombian hybrid zone between *H. e. venus* and *H. e. chestertonii* (Arias et al. 2008). Species pairs such as these permit analysis of the earliest stages of speciation, where speciation has begun, but reproductive isolation is not complete, with hybrids making up a small proportion of the population in narrow areas of overlap (McMillan et al. 1997; Arias et al. 2008). Further along the continuum, there are many closely related, sympatric species where hybridization occurs, but it is rare (Mallet et al. 2007; Mallet 2009). For example, *H. melpomene* and *H. cydno* are broadly sympatric in Central America and northern South America, coexisting as a result of several ecological differences, including their mimetic association, host plant association, and habitat preference (Naisbit 2001). There are also more distantly related species in sympatry, such as *H. melpomene* and *H. elevatus*, where the occurrence of very rare hybridization has facilitated adaptive introgression of color pattern alleles and the spread of mimetic patterns across the genus (Heliconius Genome Consortium 2012). Finally, there are non-hybridizing species, such as *H. erato* and *H. melpomene*, which are ecologically and behaviorally distinct, yet share identical mimetic wing patterns. These species provide a comparative framework for exploring the repeatability of evolution and for gaining a more general understanding of how genomic changes influence developmental pathways, phenotype, and ultimately fitness.

13.4.1 Identifying Functional Variation

Decades of research in *Heliconius* has shown that wing color patterns are under strong natural selection (Benson 1972; Mallet and Barton 1989; Mallet et al. 1990; Kapan 2001). In one of the best studied *Heliconius* hybrid zones, the transition between divergent postman and rayed color pattern races of *H. erato* in Northeastern Peru is sharp and occurs across a narrow 10 km transect (Fig. 13.5a). Strong natural selection on color pattern was demonstrated experimentally on either side of the hybrid zone by releasing individuals with the postman pattern within the rayed population, and vice versa, and estimating survivorship (Mallet and Barton 1989). On both sides of the hybrid zone, recapture rates were significantly lower for butterflies with the foreign color pattern, yielding an estimated overall selection coefficient of 0.52. This estimate was comparable to indirect measures of selection based on fitting linkage disequilibria and cline theory models to extensive hybrid zone data (Mallet et al. 1990).

Recent research has focused on identifying and understanding the architecture of color pattern loci in order to connect genotype to phenotype and fitness. Using a combination of traditional genetic and emerging genomic approaches, the genomic regions containing the four major color pattern loci have been localized to small intervals and, in two cases, very narrow genomic regions, with specific genes being strongly implicated (Joron et al. 2006; Counterman et al. 2010; Baxter et al. 2010; Ferguson et al. 2010; Reed et al. 2011; Joron et al. 2011; Martin et al. 2012; Heliconius Genome Consortium 2012; Supple et al. 2013). The most progress has been made in understanding red color pattern variation, with research on the red switch locus serving as an exemplar of how to identify functionally important variation. Allelic variation at the red switch controls multiple distinct red color pattern elements that vary between the two divergent color pattern phenotypes (Fig. 13.3). The genomic interval containing this switch was localized to a 400 kb region on chromosome 18 that contained more than a dozen predicted genes (Counterman et al. 2010). Within this region, microarray expression studies, using probes tiled across this region, identified *optix*, a homeobox transcription factor, as the only gene that was consistently differently expressed between divergent color pattern races – showing high upregulation in regions of the pupal wing fated to become red in the adult wing (Fig. 13.5b). Furthermore, beginning at approximately 60 hours after pupation, *optix* expression perfectly prefigures red pattern elements (Fig. 13.5c) – even

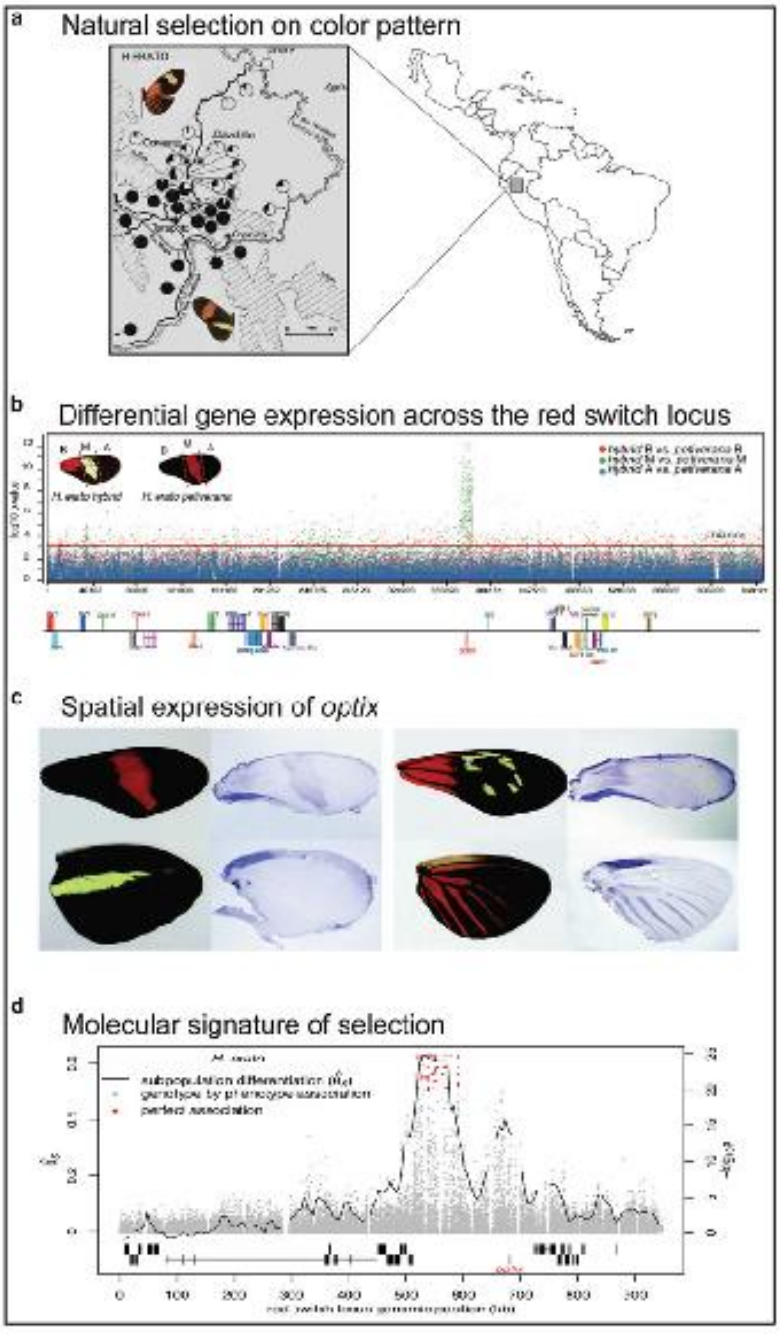
reflecting subtle pattern differences between co-mimics (Reed et al. 2011). The *optix* amino acid sequence is highly conserved within *Heliconius*, which suggests that the control of red pattern variation is due to allelic variation in *cis*-regulatory elements (Hines et al. 2011; Reed et al. 2011).

The prediction that cis-regulatory variation modulates red color pattern variation is supported by population genomic analyses of individuals collected across narrow hybrid zones between divergent color pattern phenotypes. These hybrid zones contain individuals representing many generations of recombination and phenotypically distinct individuals collected within them differ only at genomic regions responsible for the those differences (Counterman et al. 2010; Baxter et al. 2010; Nadeau et al. 2012). Analysis of three replicated hybrid zones between postman and rayed *H. erato* races shows a sharp peak of genomic divergence in a region approximately 100 kb 3′ of *optix*, in a "gene desert" containing no predicted genes nor any transcriptional activity. The peak of divergence is approximately 65 kb wide and contains numerous SNPs perfectly associated with color pattern across the broader *H. erato* pattern radiation (Fig. 13.5d) (Supple et al. 2013).

Relative to other regions of the genome, there is extensive linkage disequilibrium (LD), reduced heterozygosity, reduced nucleotide diversity, and high levels of population differentiation across the 65 kb peak (Supple et al. 2013)—hallmarks of a history of strong selection. A compelling hypothesis is that this region contains a series of modular enhancers that regulate the spatial expression of *optix*, similar to the architecture recently described for genes responsible for morphological variation in melanin and trichome patterns in *Drosophila* (Bickel et al. 2011; Frankel et al. 2011). This model is consistent with phenotypic recombinants occasionally collected in postman/rayed hybrid zones that disassociate the proximal red patch on the forewing from the red hindwing rays, which are red color pattern elements that typically occur together (Mallet 1989). In fact, in both co-mimics, *H. erato* and *H. melpomene*, there is a single race found in a narrow geographic region in the Guianas that have the red forewing patch, but lack the hindwing rays. Additionally, one of the polymorphic forms of *H. timareta* in Ecuador also has a recombinant phenotype – showing hindwing rays, but no red forewing patch (Mallet 1999).

The genome scans described above are exceptionally powerful for localizing functional regions. However, the ability to more finely characterize these regions and to identify functional elements or functional changes using population genomic approaches will ultimately depend on what is causing the strong LD seen across the region. One possibility is that the region contains an inversion that suppresses recombination and locks loci into specific allelic combinations. Inversions have been shown to be important in maintaining a series of major effect color alleles in *H. numata* (Joron et al. 2011). However, we see no evidence for inversions in the red switch locus in *H. erato* and *H. melpomene* pair-end sequence data (Supple et al. 2013), nor were inversions evident in analysis of different color pattern races of *H. melpomene* (Nadeau et al. 2012). Moreover, fine scale analysis of haplotype structure across this region suggests that recombination occurs (Supple et al. 2013). Given the lack of evidence of an inversion and the presence of recombinant individuals, it is most likely that the observed LD is a result of strong natural selection. Strong selection can establish extensive LD among loci, even among unlinked color loci (Mallet et al. 1990). In this case, the presence of recombination raises the possibility that more extensive population and taxon sampling will further refine the boundaries of the functional elements. However, in order to fully understand how variation in this region modulates pattern evolution, population genomic approaches must be coupled with other strategies, including exploring protein-DNA interactions with DNA foot printing (Cai and Huang 2012) and confirming functional mutations with transgenic manipulations (Merlin et al. 2013; Cong et al. 2013).

There has been similar progress in identifying a candidate gene for the forewing melanin shutter. Linkage mapping, gene expression analysis, and


a
Natural selection on color pattern
b
Differential gene expression across the red switch locus
c
Spatial expression of optix
d
Molecular signature of selection

pharmacological treatments all indicate that the *WntA* ligand modulates variation of the forewing band (Martin et al. 2012). The spatial pattern of *WntA* expression corresponds to the black forewing pattern in multiple species across the *Heliconius* radiation. As with *optix*, the *WntA* protein is highly conserved and variation in *cis*-regulatory regions is likely responsible for pattern diversity. *WntA* is a signaling ligand that creates a morphogen gradient across the developing wing tissue and is the type of molecule predicted to underlie pattern formation in theoretical models of wing color pattern development (Kondo and Miura 2010; Nijhout 1991). *WntA* is expressed earlier in pattern formation than *optix* and may act as a negative regulator of *optix*, with the interaction between the two genes being responsible for establishing black versus non-black wing pattern boundaries. This is only the second report of a morphogen involved in pattern generation (see Werner et al. 2010), but it is the first that directly links change in a patterning molecule to the evolution of a highly variable trait with clear adaptive significance (Martin et al. 2012).

The yellow switch and the global melanin shutter have similarly been positionally cloned (Joron et al. 2006; M. Kronforst, pers. comm.). Nonetheless, specific candidate genes and/or functional elements have yet to be identified. The global melanin shutter, in particular, has proven difficult to characterize. It was the first color pattern locus to be positionally cloned and it has the broadest range of phenotypic effects of any of the *Heliconius* color loci. Moreover, this region has recently been shown to underlie pattern change in several Lepidoptera species, including eyespot size in *Bicyclus* (Saenko et al. 2010) and the classic case of industrial melanism in the British peppermoth, *Biston betularia* (van't Hof et al. 2011), underscoring the flexibility and broad evolutionary importance of the *Heliconius* patterning loci. In *Heliconius*, this locus has been hypothesized to be a "supergene" composed of a number of distinct co-adapted protein coding loci. Support for this comes from recent work on *H. numata* that showed that allelic variants at this locus are actually a set of different chromosomal inversions across a region containing at least 18 genes (Joron et al. 2011). However, ongoing expression and genome scan studies in *H. erato* and *H. melpomene* indicate that, similar to the red switch and the forewing melanin shutter, only a single protein coding region at this locus may underlie the global variation in melanin pattern across *Heliconius* wings (C. Jiggins, pers. comm.).

13.4.2 The Origins of Novel Phenotypes

Natural selection can explain why wing patterns of different *Heliconius* species should converge – strong selection against rare color patterns promotes mimicry (Müller 1879). Natural selection can also explain the maintenance of existing wing pattern diversity – strong frequency-dependent selection removes non-mimetic individuals creating sharp transition zones between divergent phenotypes (Mallet et al. 1998). However, natural selection cannot easily explain the origin and spread of new phenotypes in *Heliconius*. This is a complex issue at the core of the mimicry paradox – the frequency – dependent selection that

Fig. 13.5 Linking genotype to phenotype to fitness in *Heliconius*. (**a**) Allele frequency variation at the red switch locus across the Peruvian hybrid zone reflects strong natural selection on color pattern. (**b**) *optix* is the only gene in the red switch locus showing strong and significant differential expression between the red and yellow forewing bands of divergent red phenotypes. (**c**) *In situ* hybridizations show that spatial expression of *optix* exactly prefigures red color pattern elements in pupal wings 60 hours after pupation. (**d**) Sliding window genomic divergence between two major *H. erato* phenotypes ("postman" and "rayed") from three hybrid zones across the red switch locus. There are two peaks of divergence, one near *optix* and one at a 65 kb region 3' of *optix*, indicating potentially functional regions. The 65 kb peak also contains numerous SNPs perfectly associated with phenotype. The divergent peak stands in marked contrast to the lack of differentiation at regions unlinked to color pattern $\left(\widehat{\theta}_x = -0.07\right)$ (Figure modified from Mallet and Barton 1989; Reed et al. 2011; Supple et al. 2013)

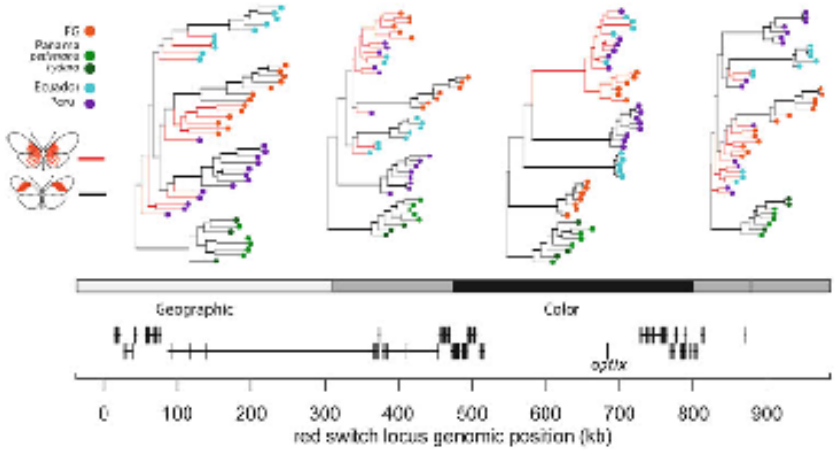


Fig. 13.6 The origins of an adaptive radiation – phylogenetic analysis of the red switch locus in *H. erato*. Phylogenetic analysis of optimal topological partitions highlight a region around the gene *optix* as clustering samples by phenotype, rather than geographic proximity. The tree topologies are shown, with phenotypes represented by branch color and geographic regions by terminal node color. Trees were generated from SNPs determined by aligning short sequencing reads to a reference genome. The grey scale bar is colored by the general history inferred. Around *optix*, samples are clustered by phenotype (*black bar*), while the farthest partition from *optix* clusters by geography (*light grey bar*), and the regions in between are intermediate, clustering by a mix of geography and phenotype (*dark grey bar*). Gene annotations, with the gene *optix* indicated, are shown below (Figure modified from Supple et al. 2013)

stabilizes existing patterns is the same force that eliminates novel forms, yet pattern divergence is rampant (Mallet and Gilbert 1995; Turner and Mallet 1996; Joron and Mallet 1998). To begin to explain this paradox, we first need to understand the evolutionary dynamics of the genomic regions that cause pattern change. An essential first question to ask is whether novel phenotypes arise once and spread within and between species or are there multiple, independent origins of the same phenotype? It has only been with the identification of the regions that modulate phenotypic differences that we can begin to address this question. The answer seems to be a bit of both – phenotypic evolution within races and species with even low levels of hybridization occurs by sharing uniquely derived color pattern alleles; while convergent phenotypes evolve independently in more distantly related co-mimics.

Analyses of the genomic region responsible for color pattern diversity support a single origin for major red color pattern phenotypes within species. For both *H. erato* and *H. melpomene*, variation around the red switch locus sorts by color pattern phenotype (Hines et al. 2011; Supple et al. 2013). In both species, individuals possessing a rayed phenotype cluster together to the exclusion of individuals possessing the postman phenotype (Fig. 13.6). Rayed patterns are found in the Amazon basin and are co-mimetic with several other *Heliconius* species and day flying moths; whereas, the postman phenotypes are largely unique to *H. erato* and *H. melpomene* and are found in multiple disjunct regions around the periphery of the Amazon and in Central America (Fig. 13.2). The pattern of genetic variation around the red switch locus supports the hypothesis that one rayed phenotype evolved in each species and spread quickly, fragmenting the geographic distribution of the older postman phenotypes. This phylogenetic signal is restricted to a region containing the 65 kb divergence peak identified in the hybrid zone comparisons (Fig. 13.5d). As you move away from this region, the phylogenetic signal increasingly reflects a history of recent gene flow, with variation

clustering by geographic proximity and biogeographic boundaries, regardless of color pattern (Fig. 13.6). This pattern of clustering by geography is the same pattern that is observed across regions unlinked to color pattern, which previously led to the incorrect conclusion that similar color pattern phenotypes evolved multiple times within each species (Brower 1994; Flanagan et al. 2004; Quek et al. 2010). This discordance demonstrates how inferences drawn from a specific subset of the genome can be misleading (Hines et al. 2011).

Mimetic convergence between distantly related species, in contrast, likely occurs by independent evolution. For example, population genetic and phylogenetic analyses of the *H. erato* and *H. melpomene* radiations, using variation within color pattern intervals, consistently clusters individuals by species designation, which is similar to the groupings obtained at loci unlinked to color pattern. Thus, although the same genomic region regulates mimetic color pattern variation, the changes responsible for mimetic convergence likely arose independently in the two species. The independent origin of similar color patterns in *H. erato* and *H. melpomene* is perhaps not unexpected, given that the two species diverged from each other over 15 million years ago (Pohl et al. 2009) and do not hybridize.

It is less clear how often more closely related species "borrow" color pattern alleles to acquire a mimetic wing color pattern. A cursory review of a phylogeny of the *Heliconius* radiation shows the high frequency that similar wing patterns are shared by species across the tree (Fig. 13.1). For example, within the melpomene/cydno/silvaniform (MCS) clade, the rayed and postman phenotypes occur within three of the four major lineages (Fig. 13.1). These species are all closely related and many are known to hybridize in the wild (Mallet et al. 2007) and in greenhouses (Gilbert 2003). These observations lead to a proposed model whereby *Heliconius* mimicry evolved by repeated interspecific transfer of color patterning alleles (Gilbert 2003). Adaptive introgression, which is the spread of beneficial variation through interspecific hybridization, may have provided the genetic raw material for both accelerated adaptation and speciation due to the dual role of color patterns in mimicry and mating behavior.

Compelling evidence for hybridization and introgression, particularly around color pattern loci, comes from analyses of closely related *Heliconius* species that share similar mimetic patterns. Phylogenetic analysis of genetic variation from the region identified as crucial to red color pattern differences, clusters populations and species by red pattern phenotype across species boundaries, rather than by phylogenetic relationship (Heliconius Genome Consortium 2012; Pardo-Diaz et al. 2012) (Fig. 13.7). This clustering by phenotype includes *H. elevatus*, which is the only rayed species in the more distantly related silvaniform clade, which usually have orange, black, and yellow tiger patterns and are known to hybridize, albeit rarely, with *H. melpomene*. Additional support comes from genome-wide tests that attempt to distinguish shared ancestral polymorphism from shared polymorphism resulting from recent gene flow (Green et al. 2010; Durand et al. 2011). This analysis shows a statistically significant excess of shared polymorphisms between sympatric co-mimics than would be expected from random sorting of ancestral polymorphisms, with a particularly strong signal at known color pattern loci. Although this test has been shown to be biased by population structure and to be sensitive to genotyping errors (Durand et al. 2011), the pattern of shared polymorphisms combined with the traditional phylogenetic analyses paints a dramatic picture of the adaptive spread of color pattern alleles across species boundaries.

13.4.3 Wing Color Patterns as a "Magic Trait" Promoting Speciation

As we are beginning to understand the role that adaptive introgression plays in the spread of mimetic color pattern alleles, it is becoming equally clear that divergent color pattern alleles in *Heliconius* likewise play a profound role in speciation. Disruptive selection on an ecological trait, such as *Heliconus* wing patterns, imposes

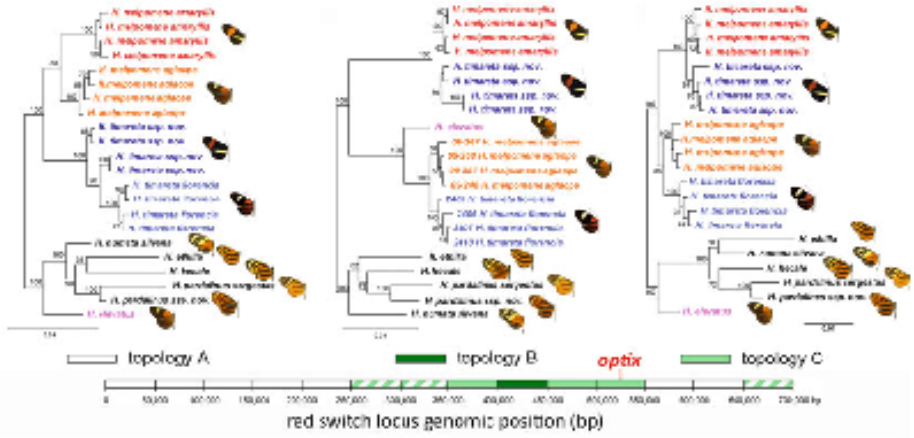


Fig. 13.7 Evolution by adaptive introgression. Phylogenetic analysis across the red switch locus shows introgression between sympatric, mimetic species in the genomic region believed to control red color pattern variation. A single 50 kb region clusters all rayed samples together, including *H. elevatus*, a proposed hybrid species. Windows farthest from this region generate the expected species tree (Figure modified from Heliconius Genome Consortium 2012)

a barrier to gene flow (Schluter 2009; Nosil 2012). We see this clearly in the signature of differentiation across hybrid zones between color pattern races of *H. erato* (Fig. 13.5d). However, in the absence of strong assortative mating, even with this barrier to gene flow, intermediate phenotypes will be continually produced and recombination will prevent speciation. This antagonism is the principal reason why the idea of speciation with gene flow remains extremely controversial (Felsenstein 1981). One way around this difficulty is if the trait under disruptive selection also contributes to nonrandom mating. In this case, there is a clear path to speciation (Dieckmann and Doebeli 1999; Gavrilets 2005) and such traits have become known as "magic traits". Rather than the term "magic trait", which implies a trait encoded by a "magic gene", a better term would be "multiple-effect trait". A multiple-effect trait is simply a trait that has multiple functions – it is under disruptive selection and contributes to nonrandom mating. This definition is of more value because it does not presuppose any particular underlying genetic architecture (c.f. Servedio et al. 2011).

The wing patterns of *Heliconius* provide one of the best experimental systems to study how "magic" or "multiple-effect" traits can generate biodiversity. Research on how these traits can promote speciation is most progressed in *H. melpomene* and *H. cydno*, where experimental manipulation demonstrates the importance of wing color patterns in both natural and sexual selection (Naisbit et al. 2001; Jiggins et al. 2001b; Merrill et al. 2011b, 2012). Recent field and cage experiments demonstrate that wing color patterns are under disruptive natural selection – F_1 hybrids, whose wing color patterns show an intermediate forewing phenotype, were attacked more frequently than either parental species (Merrill et al. 2012). Mate choice experiments demonstrate both color pattern based assortative mating between the two species and disruptive sexual selection against hybrids (Jiggins et al. 2001b; Naisbit et al. 2001; Merrill et al. 2011b). In addition, assortative mating is much higher in populations where *H. melpomene* and *H. cydno* are sympatric, as compared to populations where *H. melpomene* does not encounter *H. cydno*. This is consistent with the expectation of the reinforcement hypothesis – selection against hybrids lead

to the evolution of stronger pre-mating isolation (Jiggins et al. 2001b). It is important to point out that, in addition to color pattern based mating, there are other forms of isolation between the pair, including host plant preferences, microhabitat usage, and sterility barriers. The sterility barriers occur because the F_1 offspring follow Haldane's rule – the homogametic males are fertile and the heterogametic females are sterile. However, the strength of selection against hybrids that results from female sterility is only about as strong as mimicry selection and not nearly as strong as pre-mating isolation due to assortative mating (Naisbit et al. 2002).

In *Heliconius*, ongoing research is beginning to make the connection between multiple-effect traits and the loci that underlie these traits. A series of recent studies have demonstrated that the loci causing color pattern differences and the loci responsible for color pattern based mating preference are physically linked in *Heliconius* (Kronforst et al. 2006a; Merrill et al. 2011b). Physical linkage between a color pattern locus and a male mating preference was demonstrated first in *H. pachinus* and *H. cydno galantus* – mating was strongly assortative by white versus yellow color and variation in male mating preference mapped to the yellow switch locus (Kronforst et al. 2006a). A similar association was demonstrated between *H. cydno* and *H. melpomene*. In this species pair, male approach and courtship behavior was also highly assortative by coloration and strong male preference for red pattern mapped to the red switch locus (Merrill et al. 2011b). This is a very intriguing finding given that the gene that controls the distribution of red on a *Heliconius* wing, *optix*, also plays a role in compound eye development (Seimiya and Gehring 2000). This raises the possibility for a direct link between the perception and transmission of color pattern cues. Ongoing research, including experiments to create introgression lines that differ primarily around the regions responsible for red pattern variation, seeks to better understand the nature of the observed association and its role in promoting the radiation of *Heliconius* butterflies.

In addition to facilitating speciation by divergent natural selection, physical linkage between color pattern traits and the mating preference for those traits can promote the formation of new species through hybridization (Arnold 2006). Hybrid speciation results when hybridization produces novel genotypes that are reproductively isolated from the parental species. In this regard, hybrid genotypes that confer an ecological advantage and influence assortative mating (sensu Smith 1966) could quickly result in the origin of a novel hybrid population that is reproductively isolated from the parental species. This process has been termed hybrid trait speciation (Jiggins et al. 2008). Although hybrid speciation is thought to be rare in animals, hybrid trait speciation may provide a route for hybridization to play a role in animal diversification.

In *Heliconius* there are some of the most compelling *Heliconius* examples of hybrid trait speciation in the animal kingdom (Mavárez et al. 2006; Salazar et al. 2010; Heliconius Genome Consortium 2012). For example, evidence from a number of independent datasets suggest *Heliconius heuripa* arose through hybrid speciation: (i) the observation of regions in Venezuela where hybrids between the proposed parental species commonly occur (Mavárez et al. 2006), (ii) laboratory crosses demonstrating a clear path to the *H. heurippa* phenotype (Mavárez et al. 2006), (iii) molecular genetic analysis showing that the *H. heurippa* genome is a mosaic of pieces from the parental species (Salazar et al. 2010), and (iv) mate choice experiments demonstrating incipient reproductive isolation from the parental species via assortative mating (Mavárez et al. 2006; Melo et al. 2009). *Heliconius elevatus* is another interesting example of a putative hybrid species, as speciation potentially involves both color pattern mimicry and mate choice (Heliconius Genome Consortium 2012). The hypothesis is that hybridization between *H. pardalinus* and *H. melpomene* resulted in adaptive introgression of color pattern alleles, followed by reproductive isolation due to assortative mating on wing color pattern. Genomic data strongly support adaptive introgression of the *H. melpomene* rayed color

pattern allele into a *H. pardalinus* genome (Heliconius Genome Consortium 2012). The prediction of reproductive isolation secondary to adaptive introgression remains untested. It is predicted that the new rayed *H. pardalinus* population became reproductively isolated from other *H. pardalinus* and *H. melpomene* due to assortative mating based on color pattern and perhaps other signals, such as short-range chemical cues (Estrada and Jiggins 2008), resulting in a new species – *H. elevatus*. Although the genomic, ecological, and behavioral evidence for these examples are impressive, further studies are needed as alternative speciation scenarios have been proposed for these species that do not invoke introgression and hybridization (see Brower 2013). A whole genome perspective should help shed light on this debate, but it requires a more fundamental understanding of how genomes change during speciation and what signature hybridization and introgression would leave on expected patterns of genomic divergence.

13.4.4 Genomic Heterogeneity at the Species Boundary

Although hybridization between closely related species is common in nature (Mallet 2005; Rieseberg 2009), the idea that divergence and speciation can occur in the face of ongoing gene flow remains contentious. Nonetheless, over the last decade the debate has largely shifted from questions about the geographic context of speciation towards gaining an understanding of the processes and mechanisms that can generate biodiversity in the face of gene flow (Nosil 2012). More recently, next-generation sequencing technologies have matured to permit a whole-genome perspective on divergence during speciation. These data promise to advance our understanding of the origins of reproductive isolation by moving research towards the processes that shape patterns of divergence across whole genomes (Feder et al. 2012).

With the publication of the first *Heliconius* genome (Heliconius Genome Consortium 2012), a number of studies have used whole genome resequencing to explore the genomic landscape of divergence and speciation along an evolutionary continuum of hybridizing taxa. These studies have focused on the melpomene/cydno/silvaniform clade and use the *H. melpomene* genome as a reference to layer resequence data and to characterize individual variation across the genome (Kronforst et al. 2013; Martin et al. 2013). These studies join several recent studies in other organisms (Hohenlohe et al. 2010; Lawniczak et al. 2010; Ellegren et al. 2012; Gagnaire et al. 2013) to provide the first full genome perspectives on speciation.

Genomic analyses across the *Heliconius* speciation continuum highlight some characteristics that are emerging from these early speciation genomics studies. At the early stages in the speciation continuum, recently diverged populations freely hybridize but show strong divergence at regions of the genome known to be under strong selection. The result is differentiation that is restricted to few areas of the genome. For example, hybridizing races within both *H. erato* and *H. melpomene* showed clear regions of divergence around known mimicry loci, with little divergence evident elsewhere in the genome (Figs. 13.5d and 13.8). It is notable just how restricted differentiation is, even under conditions of very strong natural selection when patterns of divergence are expected to extend well beyond functional sites. Despite strong frequency dependent selection on color pattern, genomic divergence is limited to sharp and narrow peaks tightly linked to color pattern loci. These divergence peaks have long tails that extend about 1 Mb, but differentiation is only slightly above background levels. This observation is interesting given that differences in a number of important ecological traits, including host plant preference and larval survival, map to color pattern regions (Merrill et al. 2013).

As speciation progresses, selection, genetic hitchhiking, and the accumulation of neutral mutations during these latter stages of speciation result in highly heterogeneous patterns of genomic divergence across the genome. This pattern of heterogeneous divergence is evident across the continuum from incipient species with

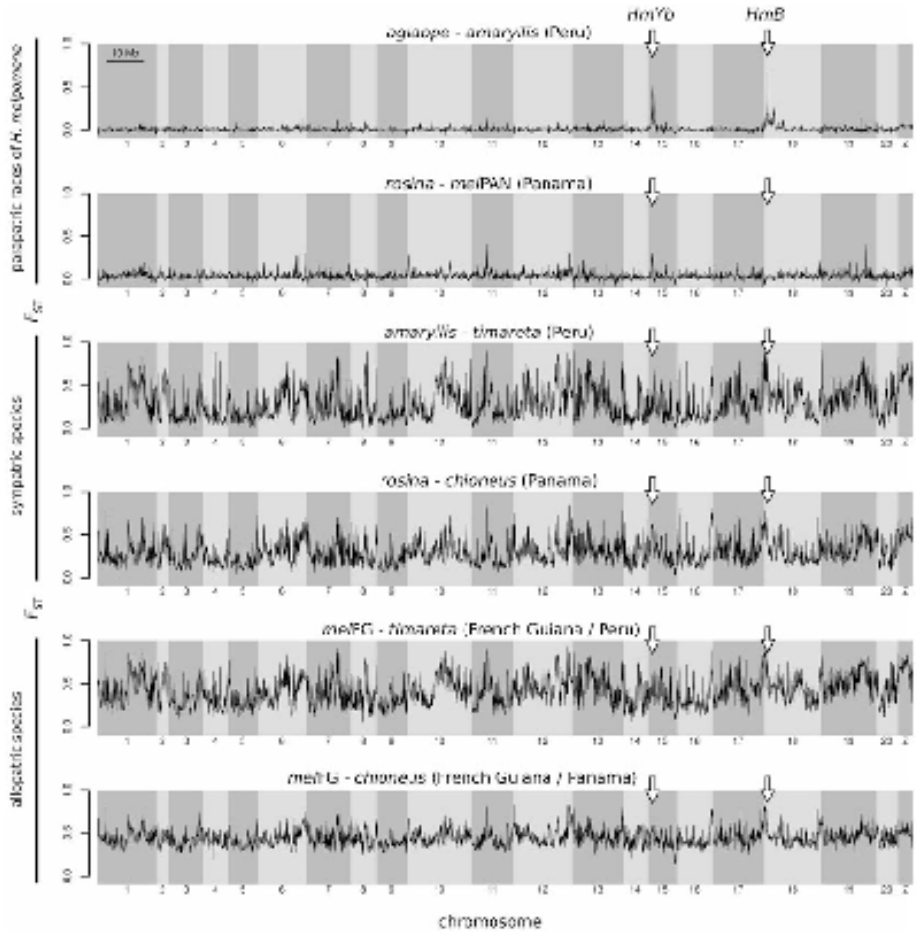


Fig. 13.8 Genomic architecture of speciation – empirical data. The empirical divergence data from the melpomene/cydno clade, with arrows pointing to known color pattern loci. Hybridizing races of *H. melpomene* show islands of divergence at the color pattern loci. Comparisons between closely related, sympatric species, show heterogeneous patterns of divergence across the whole genome. Allopatric species pairs also show high heterogeneity, but with elevated divergence across the genome (Figure modified from Martin et al. 2013)

pre-mating isolation, to species with strong pre-mating and post-mating isolation. For example, *H. cydno* and *H. pachinus* are two closely related species that differ in color pattern and show strong color pattern based assortative mating (Kronforst et al. 2006a), yet hybridize occasionally in narrow regions of overlap in Costa Rica (Kronforst et al. 2006b). In addition to peaks of divergence at known color pattern loci, there are over a dozen regions of the genome that are more divergent than expected by chance and may harbor previously unidentified ecologically important variation. The heterogeneous patterns of divergence persist as the evolutionary distance increases to closely-related species with stronger reproductive isolation, such as *H. melpomene* versus *H. timareta* and *H. cydno*, (Martin et al. 2013) (Fig. 13.8).

Heterogeneity is emerging as a common feature of genomic divergence in a number of recent studies. For example, studies of differentiation in *Anopheles* mosquitoes, *Ficedula* flycatchers, and *Coregonus* whitefish also showed highly heterogeneous patterns of differentiation (Lawniczak et al. 2010; Ellegren et al. 2012; Gagnaire et al. 2013). In *Ficedula* and *Coregonus*, the patterns are thought to be the result of recent admixture following allopatric divergence. In contrast, the *Heliconius* and *Anopheles* patterns are thought to have emerged as a result of speciation without periods of allopatry. Both speciation with gene flow and allopatric divergence with secondary contact can generate genomic heterogeneity. However, finer analysis of the patterns should allow one to distinguish the two scenarios. A commonly used measure of the extent and timing of gene flow is the number and distribution of shared polymorphisms between species. The basic principle is that (i) longer periods of gene flow should result in a greater proportion of shared polymorphic alleles between older species, (ii) recent gene flow will reduce differentiation and increase the proportion of shared alleles among sympatric populations, as compared to allopatric populations, and (iii) recent gene flow should initially result in strong linkage disequilibrium between shared alleles at linked sites, which would breakdown over time if gene flow was ongoing for longer periods. Various methods for comparing the numbers and distribution of shared polymorphisms are being developed and have recently been used to study the role of gene flow during speciation in the handful of organisms with population genomic data available (Kulathinal et al. 2009; Ellegren et al. 2012), including humans and neandertals (Green et al. 2010).

The melpomene/cydno/silvaniform clade provides an ideal opportunity to explore the genomics of speciation with gene flow versus allopatric divergence with secondary contact in a rich comparative framework. This is because the clade includes many allopatric and sympatric/parapatric races and species with varying degrees of known phylogenetic relatedness that can be used to compare patterns of shared polymorphisms and test different speciation models. For example, the observed increase of shared polymorphisms across the genome at increasing phylogenetic depths is suggestive of a long history of gene flow during speciation (Martin et al. 2013). Interestingly, a similar conclusion was reached using a completely different approach that involved modeling introgression rates in a community assessment of genomic differentiation among Costa Rican *Heliconius* species (Kronforst et al. 2013). However, extreme caution in interpreting these patterns is warranted. High heterogeneity in genomic divergence is indicative of the complex interplay between a diverse array of ecological and demographic factors, including selection, gene flow, and population history, as well as intrinsic genomic features such as variation in recombination rate.

With these new data, we are beginning to appreciate the complexity and the challenges of identifying genomic regions responsible for adaptive divergence and reproductive isolation and understanding how they affect genome-wide patterns of divergence throughout the speciation process. This is a serious challenge, yet systems with replicated examples of adaptation or speciation, such as *Heliconius*, sticklebacks, and whitefish, can be extremely powerful for inferring the functional importance of regions of divergence and understanding the history of gene flow between species.

13.5 From Patterns to Process

Moving forward requires a much better understanding of the how genomes diverge. As genomic technologies advance, empirical descriptions of genomic divergence will be layered onto one another to describe how the genomes of species change through space and time. The accumulating genomic data are already revealing extremely heterogeneous patterns of divergence that result from complex interactions between selection and gene flow. The research is quickly transitioning to identifying systems that have the most promise to provide insights into the process of genomic divergence. In this respect, new model systems, such as *Heliconius*, which (i) have replicated examples of adaptation, (ii) are

composed of taxa representing distinct stages of the speciation process, and (iii) have traits that are known to contribute to adaptation and speciation, will provide an important framework to determine the processes that drive ecological divergence and speciation from the patterns of genomic divergence.

Genomic data in *Heliconius* highlight the ability to identify the genomic regions that are known targets of selection and show how divergence around these regions changes when populations are increasingly isolated from each other. However, divergent races and incipient *Heliconius* species differ by more than wing color patterns. They show differences in mating preference (McMillan et al. 1997; Jiggins et al. 2001b; Chamberlain et al. 2009; Merrill et al. 2011a), hybrid sterility (Jiggins et al. 2001a; Naisbit et al. 2002), host plant choice (Brown 1981; Estrada and Jiggins 2002), and physiology (Davison et al. 1999) – all of which may play key roles in speciation. Leveraging the extraordinary radiation for broader insights into the origins of diversity requires that we better utilize genomic datasets. We need to identify regions under divergent selection and understand how they contribute to differences in survival or otherwise cause a reduction in gene flow between incipient forms. This challenge is not unique to *Heliconius* – the overarching goal of ecological and speciation genomic research is to identify and characterize regions of the genome under divergent selection and to understand what role they play in speciation.

To reach this goal, we need new theory that will "transform current predictions concerning genetic divergence into more dynamic recreations of how genomic differentiation unfolds through time during speciation" (Feder et al. 2012). Presently, the analysis of genomic landscapes is largely descriptive. Formal models that explain how selection and genetic hitchhiking can drive patterns of genomic divergence are beginning to emerge (Smadja et al. 2008; Feder and Nosil 2010; Feder et al. 2012), but presently there is no standardized procedure to rigorously delimit the shape, size, and distribution of divergent regions of the genomes, and more importantly, to model how they change through time (Feder et al. 2012). Even among *Heliconius* studies, different strategies were used to identify the location and size of divergent regions and to estimate the degree and timing of gene flow. Without common tools and practices, it becomes very difficult to compare patterns of genomic divergence and to identify general patterns emerging from genome-wide studies of speciation. However, just as new and better datasets will be generated, new and better theories will be developed. The field needs to (i) establish better understandings of the genomic architectures of the traits under divergent selection and influencing reproductive isolation, including the number of loci, their size effect on isolation, and their relative contribution in the speciation process; (ii) investigate how mutation and recombination rates vary locally across the genome and between populations, particularly for those regions of the genome that influence isolation; and (iii) develop increasingly complex models of speciation history and understand how heterogeneous patterns of divergence evolve as species diverge with and without gene flow.

When studying adaptation and speciation, we speculate on the specific historical events that generated the extant genomic patterns that we observe. As such, we have to be very careful to temper our enthusiasm (Nielsen 2009; Barrett and Hoekstra 2011). Molecular "spandrels" (sensu Gould and Lewontin 1979) abound in the genomes of all organisms and establishing direct links between genotype, phenotype, form, and fitness requires integrated datasets. Identifying highly divergent regions of the genome is a starting point for building an integrative understanding of the nature of variation between taxa. For some species, experiments can be designed that measure the success of individuals under specific ecological conditions. In these cases, researchers can actually follow genomic change forward in time. Experimental genomics is moving beyond the laboratory to directly testing hypotheses about how selection causes genome-wide change (Barrett et al. 2008) and provides a powerful approach that can be used in a number of emerging model systems (Barrett and Hoekstra 2011). For other groups, a combination of traditional genetic and functional

genomic approaches, coupled with functional manipulation experiments, remains the best strategy to identify functionally important variation. In either case, the combination of technological and analytical advances ensures that genomic exploration will continue to transform our understanding of the origins of biodiversity.

References

Arias CF, Muñoz AG, Jiggins CD, Mavárez J, Bermingham E, Linares M (2008) A hybrid zone provides evidence for incipient ecological speciation in *Heliconius* butterflies. Mol Ecol 17(21):4699–4712

Arnold ML (2006) Evolution through genetic exchange. Oxford University Press, Oxford

Barrett RD, Hoekstra HE (2011) Molecular spandrels: tests of adaptation at the genetic level. Nat Rev Genet 12(11):767–780

Barrett RD, Rogers SM, Schluter D (2008) Natural selection on a major armor gene in threespine stickleback. Science 322(5899):255–257

Bates HW (1862) Contributions to an insect fauna of the Amazon valley.–Lepidoptera:–Heliconinæ. J Proc Linnean Soc Lond Zool 6(22):73–77

Baxter SW, Nadeau NJ, Maroja LS, Wilkinson P, Counterman BA, Dawson A, Beltran M, Perez-Espona S, Chamberlain N, Ferguson L, Clark R, Davidson C, Glithero R, Mallet J, McMillan WO, Kronforst M, Joron M, ffrench Constant RH, Jiggins CD (2010) Genomic hotspots for adaptation: the population genetics of Müllerian mimicry in the *Heliconius melpomene* clade. PLoS Genet 6(2):e1000794

Beltrán M, Jiggins CD, Brower AV, Bermingham E, Mallet J (2007) Do pollen feeding, pupal-mating and larval gregariousness have a single origin in *Heliconius* butterflies? Inferences from multilocus DNA sequence data. Biol J Linn Soc 92(2):221–239

Benson WW (1972) Natural selection for Müllerian mimicry in *Heliconius erato* in Costa Rica. Science 176(4037):936–939

Bickel RD, Kopp A, Nuzhdin SV (2011) Composite effects of polymorphisms near multiple regulatory elements create a major-effect QTL. PLoS Genet 7(1):e1001275

Boggs CL, Watt WB, Ehrlich PR (2003) Butterflies: ecology and evolution taking flight. University of Chicago Press, Chicago, IL, USA

Brakefield PM, Beldade P, Zwaan BJ (2009) The African butterfly *Bicyclus anynana*: a model for evolutionary genetics and evolutionary developmental biology. Cold Spring Harb Protoc 2009(5): pdb–emo122

Briscoe AD, Bybee SM, Bernard GD, Yuan F, Sison-Mangus MP, Reed RD, Warren AD, Llorente-Bousquets J, Chiao CC (2010) Positive selection of a duplicated UV-sensitive visual pigment coincides with wing pigment evolution in *Heliconius* butterflies. Proc Natl Acad Sci USA 107(8):3628–3633

Brower AV (1994) Rapid morphological radiation and convergence among races of the butterfly *Heliconius erato* inferred from patterns of mitochondrial DNA evolution. Proc Natl Acad Sci USA 91(14):6491–6495

Brower AV (2013) Introgression of wing pattern alleles and speciation via homoploid hybridization in *Heliconius* butterflies: a review of evidence from the genome. Proc R Soc B 280(1752):20122302

Brown KS (1981) The biology of *Heliconius* and related genera. Annu Rev Entomol 26(1):427–457

Brown KS, Benson WW (1974) Adaptive polymorphism associated with multiple Müllerian mimicry in *Heliconius numata* (Lepid. Nymph). Biotropica 6:205–228

Cai YH, Huang H (2012) Advances in the study of protein – DNA interaction. Amino Acids 43:1141–1146

Carroll SB, Grenier J, Weatherbee S (2009) From DNA to diversity: molecular genetics and the evolution of animal design. Wiley-Blackwell, Hoboken, NJ, USA

Chai P (1986) Field observations and feeding experiments on the responses of rufous-tailed jacamars (*Galbula ruficauda*) to free-flying butterflies in a tropical rainforest. Biol J Linn Soc 29(3):161–189

Chamberlain NL, Hill RI, Kapan DD, Gilbert LE, Kronforst MR (2009) Polymorphic butterfly reveals the missing link in ecological speciation. Science 326(5954):847–850

Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini LA et al (2013) Multiplex genome engineering using CRISPR/Cas systems. Science 339(6121):819–823

Counterman BA, Araujo-Perez F, Hines HM, Baxter SW, Morrison CM, Lindstrom DP, Papa R, Ferguson L, Joron M, ffrench Constant RH, Smith CP, Nielsen DM, Chen R, Jiggins CD, Reed RD, Halder G, Mallet J, McMillan WO (2010) Genomic hotspots for adaptation: the population genetics of Müllerian mimicry in *Heliconius erato*. PLoS Genet 6(2):e1000796

Davison A, McMillan WO, Griffin AS, Jiggins CD, Mallet JLB (1999) Behavioral and physiological differences between two parapatric *Heliconius* species. Biotropica 31(4):661–668

Deinert E, Longino J, Gilbert L (1994) Mate competition in butterflies. Nature 370(6484):23–24

Dieckmann U, Doebeli M (1999) On the origin of species by sympatric speciation. Nature 400(6742):354–357

Durand EY, Patterson N, Reich D, Slatkin M (2011) Testing for ancient admixture between closely related populations. Mol Biol Evol 28(8):2239–2252

Eberhard SH, Nemeschkal HL, Krenn HW (2009) Biometrical evidence for adaptations of the salivary glands to pollen feeding in *Heliconius* butterflies (Lepidoptera: nymphalidae). Biol J Linn Soc 97(3):604–612

Ellegren H, Smeds L, Burri R, Olason PI, Backström N, Kawakami T, Künstner A, Mäkinen H, Nadachowska-Brzyska K, Qvarnström A, et al (2012) The genomic landscape of species divergence in *Ficedula* flycatchers. Nature 491(7426):756–760

Engler HS, Spencer KC, Gilbert LE (2000) Insect metabolism: preventing cyanide release from leaves. Nature 406(6792):144–145

Engler-Chaouat HS, Gilbert LE (2007) De novo synthesis vs. sequestration: negatively correlated metabolic traits and the evolution of host plant specialization in cyanogenic butterflies. J Chem Ecol 33(1):25–42

Estrada C, Jiggins CD (2002) Patterns of pollen feeding and habitat preference among *Heliconius* species. Ecol Entomol 27(4):448–456

Estrada C, Jiggins CD (2008) Interspecific sexual attraction because of convergence in warning colouration: is there a conflict between natural and sexual selection in mimetic species? J Evol Biol 21(3):749–760

Feder JL, Nosil P (2010) The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. Evolution 64(6):1729–1747

Feder JL, Egan SP, Nosil P (2012) The genomics of speciation-with-gene-flow. Trends Genet 28:342–350

Felsenstein J (1981) Skepticism towards Santa Rosalia, or why are there so few kinds of animals? Evolution 35(1):124–138

Ferguson L, Lee SF, Chamberlain N, Nadeau N, Joron M, Baxter S, Wilkinson P, Papanicolaou A, Kumar S, KEE TJ, et al (2010) Characterization of a hotspot for mimicry: assembly of a butterfly wing transcriptome to genomic sequence at the HmYb/Sb locus. Mol Ecol 19(Suppl 1):240–254

Flanagan NS, Tobler A, Davison A, Pybus OG, Kapan DD, Planas S, Linares M, Heckel D, McMillan WO (2004) Historical demography of Müllerian mimicry in the neotropical *Heliconius* butterflies. Proc Natl Acad Sci USA 101(26):9704–9709

Frankel N, Erezyilmaz DF, McGregor AP, Wang S, Payre F, Stern DL (2011) Morphological evolution caused by many subtle-effect substitutions in regulatory DNA. Nature 474(7353):598–603

Gagnaire PA, Pavey SA, Normandeau E, Bernatchez L (2013) The genetic architecture of reproductive isolation during speciation-with-gene-flow in lake whitefish species pairs assessed by RAD-sequencing. Evolution

Gavrilets S (2005) "Adaptive speciation" – it is not that easy: reply to Doebeli et al. Evolution 59(3):696–699

Gilbert LE (1972) Pollen feeding and reproductive biology of *Heliconius* butterflies. Proc Natl Acad Sci USA 69(6):1403–1407

Gilbert L (2003) Adaptive novelty through introgression in *Heliconius* wing patterns: evidence for shared genetic "tool box" from synthetic hybrid zones and a theory of diversification. In: Ecology and evolution taking flight: butterflies as model systems. University of Chicago Press, Chicago, pp 281–318

Gompert Z, Lucas LK, Nice CC, Buerkle CA (2012) Genome divergence and the genetic architecture of barriers to gene flow between *Lycaeides idas* and *L. melissa*. Evolution 67:2498–2514

Gould SJ, Lewontin RC (1979) The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. Proc R Soc B 205(1161):581–598

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MHY, et al (2010) A draft sequence of the Neandertal genome. Science 328(5979):710–722

Hanski IA (2011) Eco-evolutionary spatial dynamics in the Glanville fritillary butterfly. Proc Natl Acad Sci USA 108(35):14,397–14,404

Heliconius Genome Consortium (2012) Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature 487:94–98

Hines HM, Counterman BA, Papa R, Albuquerque de Moura P, Cardoso MZ, Linares M, Mallet J, Reed RD, Jiggins CD, Kronforst MR, McMillan WO (2011) Wing patterning gene redefines the mimetic history of *Heliconius* butterflies. Proc Natl Acad Sci USA 108(49):19666–19671

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS Genet 6(2):e1000862

Jiggins CD, Linares M, Naisbit RE, Salazar C, Yang ZH, Mallet J (2001a) Sex-linked hybrid sterility in a butterfly. Evolution 55(8):1631–1638

Jiggins CD, Naisbit RE, Coe RL, Mallet J (2001b) Reproductive isolation caused by colour pattern mimicry. Nature 411:302–305

Jiggins CD, Salazar C, Linares M, Mavarez J (2008) Hybrid trait speciation and *Heliconius* butterflies. Phil Trans R Soc B Biol Sci 363(1506):3047–3054

Joron M, Mallet JL (1998) Diversity in mimicry: paradox or paradigm? Trends Ecol Evol 13(11):461–466

Joron M, Wynne IR, Lamas G, Mallet J (1999) Variable selection and the coexistence of multiple mimetic forms of the butterfly *Heliconius numata*. Evol Ecol 13(7):721–754

Joron M, Papa R, Beltrán M, Chamberlain N, Mavárez J, Baxter S, Abanto M, Bermingham E, Humphray SJ, Rogers J, et al (2006) A conserved supergene locus controls colour pattern diversity in *Heliconius* butterflies. PLoS Biol 4(10):e303

Joron M, Frezal L, Jones RT, Chamberlain NL, Lee SF, Haag CR, Whibley A, Becuwe M, Baxter SW, Ferguson L, Wilkinson PA, Salazar C, Davidson C, Clark R, Quail MA, Beasley H, Glithero R, Lloyd C, Sims S, Jones MC, Rogers J, Jiggins CD, ffrench Constant RH (2011) Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. Nature 477(7363):203–206

Kapan DD (2001) Three-butterfly system provides a field test of müllerian mimicry. Nature 409:338–340

Kondo S, Miura T (2010) Reaction-diffusion model as a framework for understanding biological pattern formation. Science 329(5999):1616–1620

Kotiaho JS, Kaitala V, Komonen A, Päivinen J (2005) Predicting the risk of extinction from shared ecological characteristics. Proc Natl Acad Sci USA 102(6):1963–1967

Krenn HW, Penz CM (1998) Mouthparts of *Heliconius* butterflies (Lepidoptera: nymphalidae): a search for anatomical adaptations to pollen-feeding behavior. Int J Insect Morphol Embryol 27(4):301–309

Kronforst MR, Young LG, Kapan DD, McNeely C, O'Neill RJ, Gilbert LE (2006) Linkage of butterfly mate preference and wing color preference cue at the genomic location of *wingless*. Proc Natl Acad Sci USA 103(17):6575–6580

Kronforst MR, Hansen MEB, Crawford NG, Gallant JR, Zhang W, Kulathinal RJ, Kapan DD, Mullen SP (2013) Hybridization reveals the evolving genomic architecture of speciation. Cell Reports

Kulathinal RJ, Stevison LS, Noor MA (2009) The genomics of speciation in *Drosophila*: diversity, divergence, and introgression estimated using low-coverage genome sequencing. PLoS Genet 5(7):e1000550

Lawniczak M, Emrich S, Holloway A, Regier A, Olson M, White B, Redmond S, Fulton L, Appelbaum E, Godfrey J et al (2010) Widespread divergence between incipient *Anopheles gambiae* species revealed by whole genome sequences. Science 330(6003): 512–514

Linares M (1997) Origin of neotropical mimetic biodiversity from a threeway hybrid zone of *Heliconius cydno* butterflies. In: Ulrich H (ed) Tropical biodiversity and systematics. Proceedings of the international symposium on biodiversity and systematics in tropical ecosystems. Zoologisches Forschungsinstitut und Museum Alexander Koenig, Bonn, pp 93–108

Mallet J (1989) The genetics of warning colour in Peruvian hybrid zones of *Heliconius erato* and *H. Melpomene*. Proc R Soc B 236(1283):163–185

Mallet J (1999) Causes and consequences of a lack of coevolution in Müllerian mimicry. Evol Ecol 13(7–8):777–806

Mallet J (2005) Hybridization as an invasion of the genome. Trends Ecol Evol 20(5):229–237

Mallet J (2008) Hybridization, ecological races and the nature of species: empirical evidence for the ease of speciation. Phil Trans R Soc B Biol Sci 363(1506):2971

Mallet J (2009) Rapid speciation, hybridization and adaptive radiation in the *Heliconius melpomene* group. In: Butlin R, Bridle J, Schluter D (eds) Speciation and patterns of diversity. Cambridge University Press, Cambridge, UK, pp 177–194

Mallet J, Barton NH (1989) Strong natural selection in a warning-color hybrid zone. Evolution 43:421–431

Mallet J, Gilbert LE (1995) Why are there so many mimicry rings? Correlations between habitat, behaviour and mimicry in *Heliconius* butterflies. Biol J Linn Soc 55(2):159–180

Mallet J, Barton N, Lamas G, Santisteban J, Muedas M, Feley H (1990) Estimates of selection and gene flow from measures of cline width and linkage disequilibrium in *Heliconius* hybrid zones. Genetics 124(4):921–936

Mallet J, McMillan WO, Jiggins CD (1998) Mimicry and warning color at the boundary between races and species. In: Endless forms: species and speciation. Oxford University Press, New York, pp 390–403

Mallet J, Beltrán M, Neukirchen W, Linares M (2007) Natural hybridization in heliconiine butterflies: the species boundary as a continuum. BMC Evol Biol 7(1):28

Martin A, Papa R, Nadeau NJ, Hill RI, Counterman BA, Halder G, Jiggins CD, Kronforst MR, Long AD, McMillan WO, Reed RD (2012) Diversification of complex butterfly wing patterns by repeated regulatory evolution of a *Wnt* ligand. Proc Natl Acad Sci USA 109:12632–12637

Martin SH, Dasmahapatra KK, Nadeau NJ, Salazar C, Walters JR, Simpson F, Blaxter MD, Manica A, Mallet J, Jiggins CD (2013) Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. Genome Res 23(11)

Mather K (1950) The genetical architecture of heterostyly in *Primula sinensis*. Evolution 4:340–352

Mavárez J, Salazar CA, Bermingham E, Salcedo C, Jiggins CD, Linares M (2006) Speciation by hybridization in *Heliconius* butterflies. Nature 441(7095):868–871

McMillan WO, Jiggins CD, Mallet J (1997) What initiates speciation in passion-vine butterflies? Proc Natl Acad Sci USA 94(16):8628–8633

Melo MC, Salazar C, Jiggins CD, Linares M (2009) Assortative mating preferences among hybrids offers a route to hybrid speciation. Evolution 63(6):1660–1665

Merlin C, Beaver LE, Taylor OR, Wolfe SA, Reppert SM (2013) Efficient targeted mutagenesis in the monarch butterfly using zinc finger nucleases. Genome Res 23:159–168

Merrill RM, Gompert Z, Dembeck LM, Kronforst MR, McMillan WO, Jiggins CD (2011a) Mate preference across the speciation continuum in a clade of mimetic butterflies. Evolution 65(5):1489–1500

Merrill RM, Van Schooten B, Scott JA, Jiggins CD (2011b) Pervasive genetic associations between traits causing reproductive isolation in *Heliconius* butterflies. Proc R Soc B 278(1705):511–518

Merrill RM, Wallbank RW, Bull V, Salazar PC, Mallet J, Stevens M, Jiggins CD (2012) Disruptive ecological selection on a mating cue. Proc R Soc B 279(1749): 4907–4913

Merrill RM, Naisbit RE, Mallet J, Jiggins CD (2013) Ecological and genetic factors influencing the transition between host-use strategies in sympatric *Heliconius* butterflies. J Evol Bio 26(9):1959–1967

Müller F (1879) Ituna and Thyridia: a remarkable case of mimicry in butterflies. Trans Entomol Soc Lond 1879:20–29

Nadeau NJ, Whibley A, Jones RT, Davey JW, Dasmahapatra KK, Baxter SW, Quail MA, Joron M, ffrench Constant RH, Blaxter ML, Mallet J, Jiggins CD (2012) Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-scale targeted sequencing. Phil Trans R Soc B Biol Sci 367(1587):343–353

Naisbit RE (2001) Ecological divergence and speciation in *Heliconius cydno* and *H. melpomene*. PhD thesis, University of London

Naisbit RE, Jiggins CD, Mallet J (2001) Disruptive sexual selection against hybrids contributes to speciation

between *Heliconius cydno* and *Heliconius melpomene*. Proc R Soc B 268(1478):1849–1854

Naisbit RE, Jiggins CD, Linares M, Salazar C, Mallet J (2002) Hybrid sterility, Haldane's rule and speciation in *Heliconius cydno* and *H. Melpomene*. Genetics 161(4):1517–1526

Naisbit RE, Jiggins CD, Mallet J (2003) Mimicry: developmental genes that contribute to speciation. Evol Dev 5(3):269–280

Nielsen R (2009) Adaptionis – 30 years after Gould and Lewontin. Evolution 63(10):2487–2490

Nijhout HF (1991) The development and evolution of butterfly wing patterns. Smithsonian Institution Press, Washington, DC, USA

Nosil P (2012) Ecological speciation. OUP, Oxford

O'Neil ST, Dzurisin JD, Carmichael RD, Lobo NF, Emrich SJ, Hellmann JJ (2010) Population-level transcriptome sequencing of nonmodel organisms *Erynnis propertius* and *Papilio zelicaon*. BMC Genomics 11(1):310

Papa R, Kapan DD, Counterman BA, Maldonado K, Lindstrom DP, Reed R, Nijhout HF, Hrbek T, McMillan WO (2013) Multi-allelic major effect genes interact with minor effect QTLs to control adaptive color pattern variation in *Heliconius erato*. PLoS One. 8(3):e57033

Pardo-Diaz C, Salazar C, Baxter SW, Merot C, Figueiredo-Ready W, Joron M, McMillan WO, Jiggins CD (2012) Adaptive introgression across species boundaries in *Heliconius* butterflies. PLoS Genet 8(6):e1002752

Pohl N, Sison-Mangus M, Yee E, Liswi S, Briscoe A (2009) Impact of duplicate gene copies on phylogenetic analysis and divergence time estimates in butterflies. BMC Evol Biol 9(1):99

Quek SP, Counterman BA, Albuquerque de Moura P, Cardoso MZ, Marshall CR, McMillan WO, Kronforst MR (2010) Dissecting comimetic radiations in *Heliconius* reveals divergent histories of convergent butterflies. Proc Natl Acad Sci USA 107(16):7365–7370

Reed RD, Papa R, Martin A, Hines HM, Counterman BA, Pardo-Diaz C, Jiggins CD, Chamberlain NL, Kronforst MR, Chen R, Halder G, Nijhout HF, McMillan WO (2011) *optix* drives the repeated convergent evolution of butterfly wing pattern mimicry. Science 333(6046):1137–1141

Rieseberg LH (2009) Evolution: replacing genes and traits through hybridization. Curr Biol 19(3):R119–R122

Saenko SV, Brakefield PM, Beldade P (2010) Single locus affects embryonic segment polarityand multiple aspects of an adult evolutionary novelty. BMC Biol 8(1):111

Salazar C, Baxter SW, Pardo-Diaz C, Wu G, Surridge A, Linares M, Bermingham E, Jiggins CD (2010) Genetic evidence for hybrid trait speciation in *Heliconius* butterflies. PLoS Genet 6(4):e1000930

Schluter D (2009) Evidence for ecological speciation and its alternative. Science 323(5915):737–741

Seimiya M, Gehring WJ (2000) The *Drosophila* homeobox gene *optix* is capable of inducing ectopic eyes by an *eyeless*-independent mechanism. Development 127(9):1879–1886

Servedio MR, Doorn G, Kopp M, Frame AM, Nosil P (2011) Magic traits in speciation: 'magic' but not rare? Trends Ecol Evol 26(8):389–397

Sheppard PM, Turner JRG, Brown KS, Benson WW, Singer MC (1985) Genetics and the evolution of Muellerian mimicry in *Heliconius* butterflies. Philos Trans R Soc Lond Ser B Biol Sci 308(1137):433–610

Sivinski J (1989) Mushroom body development in nymphalid butterflies: a correlate of learning? J Insect Behav 2(2):277–283

Smadja C, Galindo J, Butlin R (2008) Hitching a lift on the road to speciation. Mol Ecol 17(19):4177–4180

Smith JM (1966) Sympatric speciation. Am Nat 100:637–650

Spencer KC (1988) Glycosides: the interface between plant secondary and insect primary metabolism. In: ACS symposium series. ACS Publications, Washington, DC, USA, vol 380, pp 403–416

Supple MA, Hines HM, Dasmahapatra KK, Nielsen DM, Lavoie C, Ray DA, Salazar C, McMillan WO, Counterman BA (2013) Genomic architecture of adaptive color pattern divergence and convergence in *Heliconius* butterflies. Genome Res. 23(8):1248–1257

Turner JR, Mallet JL (1996) Did forest islands drive the diversity of warningly coloured butterflies? Biotic drift and the shifting balance. Philos Trans R Soc Lond Ser B Biol Sci 351(1341):835–845

van't Hof AE, Edmonds N, Dalková M, Marec F, Saccheri IJ (2011) Industrial melanism in British peppered moths has a singular and recent mutational origin. Science 332(6032):958–960

Werner T, Koshikawa S, Williams TM, Carroll SB (2010) Generation of a novel wing colour pattern by the Wingless morphogen. Nature 464(7292):break 1143–1148

Zhan S, Merlin C, Boore JL, Reppert SM (2011) The monarch butterfly genome yields insights into long-distance migration. Cell 147(5):1171–1185

CHAPTER 2

Genomic architecture of adaptive color pattern divergence and convergence in *Heliconius* butterflies

Published in *Genome Research*:
Supple MA, Hines HM, Dasmahapatra KK, Lewis JJ, Nielsen DM, Lavoie C, Ray DA, Salazar C, McMillan WO, and Counterman BA. 2013. Genomic architecture of adaptive color pattern divergence and convergence in Heliconius butterflies. *Genome Research* 23:1248-1257.

Research

Genomic architecture of adaptive color pattern divergence and convergence in *Heliconius* butterflies

Megan A. Supple,[1,2] Heather M. Hines,[3,4] Kanchon K. Dasmahapatra,[5,6] James J. Lewis,[7] Dahlia M. Nielsen,[3] Christine Lavoie,[8] David A. Ray,[8] Camilo Salazar,[1,9] W. Owen McMillan,[1,10] and Brian A. Counterman[8,10,11]

[1]*Smithsonian Tropical Research Institute, Panama City, Republic of Panama;* [2]*Biomathematics Program, North Carolina State University, Raleigh, North Carolina 27695, USA;* [3]*Department of Genetics, North Carolina State University, Raleigh, North Carolina 27695, USA;* [4]*Department of Biology, Pennsylvania State University, University Park, Pennsylvania 16802, USA;* [5]*Department of Genetics, Evolution and Environment, University College London, London WC1E 6BT, United Kingdom;* [6]*Department of Biology, University of York, York YO10 5DD, United Kingdom;* [7]*Department of Ecology and Evolutionary Biology, Cornell University, Ithaca, New York 14853, USA;* [8]*Department of Biological Sciences, Mississippi State University, Mississippi State, Mississippi 39762, USA;* [9]*Facultad de Ciencias Naturales y Matemáticas, Universidad del Rosario, Bogotá DC, Colombia*

Identifying the genetic changes driving adaptive variation in natural populations is key to understanding the origins of biodiversity. The mosaic of mimetic wing patterns in *Heliconius* butterflies makes an excellent system for exploring adaptive variation using next-generation sequencing. In this study, we use a combination of techniques to annotate the genomic interval modulating red color pattern variation, identify a narrow region responsible for adaptive divergence and convergence in *Heliconius* wing color patterns, and explore the evolutionary history of these adaptive alleles. We use whole genome resequencing from four hybrid zones between divergent color pattern races of *Heliconius erato* and two hybrid zones of the co-mimic *Heliconius melpomene* to examine genetic variation across 2.2 Mb of a partial reference sequence. In the intergenic region near *optix*, the gene previously shown to be responsible for the complex red pattern variation in *Heliconius*, population genetic analyses identify a shared 65-kb region of divergence that includes several sites perfectly associated with phenotype within each species. This region likely contains multiple *cis*-regulatory elements that control discrete expression domains of *optix*. The parallel signatures of genetic differentiation in *H. erato* and *H. melpomene* support a shared genetic architecture between the two distantly related co-mimics; however, phylogenetic analysis suggests mimetic patterns in each species evolved independently. Using a combination of next-generation sequencing analyses, we have refined our understanding of the genetic architecture of wing pattern variation in *Heliconius* and gained important insights into the evolution of novel adaptive phenotypes in natural populations.

[Supplemental material is available for this article.]

Natural selection acting on heritable genetic variation has generated much of the extraordinary biological diversity we observe in nature. However, in the 150 years since Darwin and Wallace independently posited the theory of natural selection, we still have only a rudimentary understanding of how adaptive variation arises and spreads in natural populations. The molecular basis of adaptive variation in natural populations has been identified in only a handful of traits (Martin and Orgogozo 2013). These examples provide the foundation for our current understanding of the genetic architecture of adaptation and underpin efforts to unite molecular, developmental, and evolutionary biology into a single evolutionary synthesis (Stern and Orgogozo 2009).

The array of adaptive wing color patterns of *Heliconius* butterflies offers an exceptional opportunity to explore the functional changes that drive complex adaptive traits in natural populations. *Heliconius* butterflies display bright wing color patterns that are under strong natural selection (Benson 1972; Mallet and Barton 1989; Mallet et al. 1990; Kapan 2001)—they warn potential predators that the butterflies are unpalatable (Chai 1986). Selection favoring regional mimicry among these and other noxious butterflies drives the remarkable diversity in wing color patterns—characterized by extreme divergence within species and striking convergence among distantly related species (Turner 1975). This pattern of convergence and divergence is best exemplified by the Müllerian mimics, *H. erato* and *H. melpomene*. Since the divergence of the two species 13–26 million years ago (Pohl et al. 2009), both have undergone parallel radiations, such that the range of each species is composed of an identical patchwork of divergent color pattern races stitched together by a series of narrow hybrid zones. This replicate-rich and highly variable system has become a textbook example of evolution by natural selection and provides a remarkable template to explore the repeatability of evolution.

Heliconius show striking variation in complex red color pattern elements across both the forewing and hindwing. Red color patterns in *H. erato* and *H. melpomene* consist of three distinct elements—the color of the forewing band, the presence or absence of the red "dennis" patch on the proximal portion of the forewing, and the presence or absence of red hindwing rays (Sheppard et al. 1985; Papa et al. 2008). These elements comprise two major red phenotypes—"postman" and "rayed" (Fig. 1). Postman races are characterized by a red forewing band and absence of both the red dennis patch and hindwing rays and are found in disjunct pop-

[10]These authors contributed equally to this work.
[11]Corresponding author
E-mail bcounterman@biology.msstate.edu
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.150615.112.

23:1248–1257 © 2013, Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/13; www.genome.org

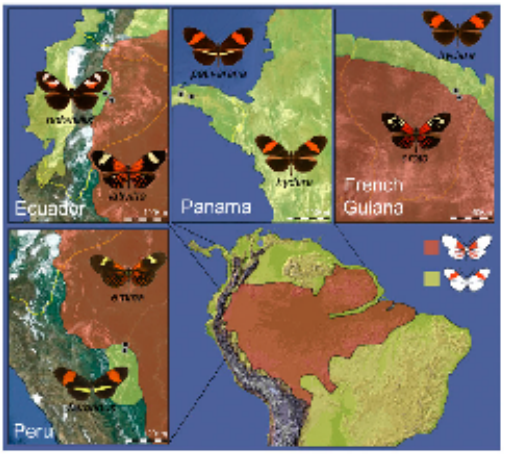


Figure 1. Distribution of the *H. erato* color pattern radiation with approximate sampling locations from four *H. erato* hybrid zones. Shaded areas are distributions of *H. erato* rayed (red) and postman (yellow) races (based on Hines et al. 2011; Rosser et al. 2012).

ulations across Central and South America. Rayed races, in contrast, have a yellow forewing band, red dennis patch, and hindwing rays. The rayed races are found throughout the Amazon Basin but not on the Pacific coast of South America or in Central America (Fig. 1).

Substantial progress has been made in understanding the genetic basis of red phenotypic variation in *Heliconius*, but the causative variant remains elusive. The region was initially described as a complex of three tightly linked loci (Sheppard et al. 1985). However, the region modulating red variation was positionally cloned in both *H. erato* (Counterman et al. 2010) and *H. melpomene* (Baxter et al. 2010) to a single shared 400-kb genomic interval—referred to as the *D* interval in *H. erato* and the *B/D* interval in *H. melpomene*. Targeted resequencing identified an ~150-kb region of divergence between divergent red color pattern races of *H. melpomene* (Nadeau et al. 2012). Gene expression analyses across this interval supported the transcription factor *optix* as the only gene with expression patterns consistent with a role in red pattern formation (Reed et al. 2011). Gene expression differences and the highly conserved amino acid sequence variation in *optix* suggest that variable red patterns are driven by *cis*-regulatory variation (Reed et al. 2011). Phenotypic recombinants between red pattern elements have been occasionally observed (Mallet 1989), implicating the potential involvement of multiple, tightly linked, *cis*-regulatory variants in this region that generate differences in the spatial expression of *optix* and result in diverse red color pattern phenotypes.

Identifying the genomic region responsible for red color pattern variation has allowed reassessment of the history of color pattern diversification in these co-mimics using the region most likely to reflect mimetic history. Phylogenetic reconstruction of the history of color pattern evolution in both *H. erato* and *H. melpomene* found that sequences from the gene *optix* cluster races strongly by color pattern, rather than the clustering by geographic proximity observed with other genomic markers (Hines et al. 2011). This suggests a single origin for geographically disjunct rayed color pattern phenotypes within each species. The evolutionary history between the two co-mimetic species is likely more complex. *Heliconius erato* and *H. melpomene* belong to divergent clades and are not known to hybridize, precluding the hypothesis of adaptive introgression that has been shown between more closely related species within the *melpomene/cydno/silvaniform* clade (The *Heliconius* Genome Consortium 2012; Pardo-Diaz et al. 2012). However, the question remains as to whether or not the convergent color patterns have a common origin through the use of ancestral variation or if they evolved independently through unique de novo mutations. Inferring the evolutionary history of this striking mimetic phenotype requires high resolution sequence data across a more finely defined genomic region regulating color pattern variation in both species.

In this study, we use a combination of next-generation sequencing technologies to (1) explore the genetic variation that modulates spatially complex adaptive red wing color pattern variation in natural populations of *Heliconius erato* butterflies; and (2) compare the genomic regions responsible for parallel mimetic color pattern radiations in *H. erato* and *H. melpomene*. We start by using transcriptome sequencing to annotate protein coding genes across the *H. erato* red color pattern interval. We then examine whole genome resequencing data from divergent color pattern races across four geographically distinct *H. erato* hybrid zones. In these admixture zones, gene flow between divergent races homogenizes the genomes, while strong selection on color pattern creates narrow regions of genetic divergence around these genomic targets of selection. We use these hybrid zone data to localize signatures of selection across 1 Mb of the red color pattern locus and identify a noncoding region likely responsible for regulating expression of *optix* and ultimately driving spatially complex patterns of red across the wings. Finally, we use phylogenetic analyses to demonstrate that mimetic red patterns evolved only once within each species but likely evolved independently in the co-mimics *H. erato* and *H. melpomene*. These two distantly related mimetic species appear to generate their convergent phenotypes through different changes in the same genetic architecture. This study refines the regulatory regions driving adaptive phenotypic variation in a classic mimetic radiation and provides an improved understanding of the repeatability of evolution at the genomic level.

Results

Annotation, synteny, and conservation of the *H. erato* red color pattern (*D*) interval

Using transcriptome alignments, protein homology, and ab initio predictions, we annotated 30 protein coding genes across 1 Mb of the *H. erato D* interval (Supplemental Table S4; GenBank accession KC469894). The genes across the *D* interval are in perfect synteny with the corresponding red pattern region of the co-mimic *H. melpomene* (Supplemental Fig. S1). The annotated region contains a 200-kb gene desert that includes a single gene, *optix*. This gene desert contains the 65-kb peak that shows strong signatures of selection and association (described below). Sequence alignments of the *H. erato D* interval and the orthologous scaffolds from the *H. melpomene* genome identified 182 highly conserved regions (>90% sequence similarity in a 500-bp window), covering a total of 63 kb of sequence, of which 25 kb is located in noncoding regions. The 65-kb peak we have identified contains 20% of the conserved noncoding regions, but transcriptome alignments show

no transcriptional activity. Analysis of potential transcription factor binding sites in the 65-kb region revealed binding sites for numerous transcription factors across the region, but no clear candidates for *optix* regulation (see Supplemental Material section S1; Supplemental Fig. S2).

Summary of resequencing data and genotyping

We examined whole genome data from multiple individuals of divergent color pattern races across four geographically distinct *H. erato* hybrid zones and one *H. melpomene* hybrid zone (Fig. 1; Table 1). We resequenced 45 *H. erato* individuals to a median per base coverage between 15× and 35× per individual and aligned sequence reads across 2.2 Mb of the *H. erato* partial reference genome sequence (1 Mb of the D interval and 1.2 Mb from other regions of the genome). For each individual, we identified, on average, one SNP for every 2.2 bases genotyped (Supplemental Table S6). Additionally, we resequenced six *H. melpomene* individuals to a median per base coverage between 19× and 34× per individual and aligned sequences to the whole *H. melpomene* reference genome (The *Heliconius* Genome Consortium 2012). For each individual, we identified, on average, one SNP for every 48 bases genotyped across the *B/D* interval (Supplemental Table S7).

Genomic divergence and genotype by phenotype association

The sliding window divergence analysis between subpopulations of the two common *H. erato* phenotypes—the postman and the rayed—showed peaks of genetic divergence at two distinctive regions within a 200-kb stretch of the *D* interval (Fig. 2). The first region (the second peak in Fig. 2 spanned ~40 kb [650–690 kb]) had moderate levels of differentiation ($0.3 < \hat{\theta}_s < 0.4$) (Fig. 2) and reduced nucleotide diversity (Fig. 3A). This region was centered immediately 3′ of the transcription factor *optix*, which is the only gene located within the peaks of divergence. The second region spanned ~65 kb of noncoding sequence more distally 3′ of *optix* (515–580 kb). This 65-kb region had very high levels of differentiation ($0.6 < \hat{\theta}_s < 0.7$) between hybridizing races, low levels of nucleotide diversity within races (Fig. 3A), and elevated linkage disequilibrium relative to other regions of the genome (Supplemental Fig. S3), which are all indicative of recent positive selection driving the fixation of haplotypes. Of these two regions that showed signatures of selection in *H. erato*, only the 65-kb peak, which is furthest from *optix*, contained SNPs that were perfectly associated with color pattern phenotype across our entire sampling of individuals (Fig. 2).

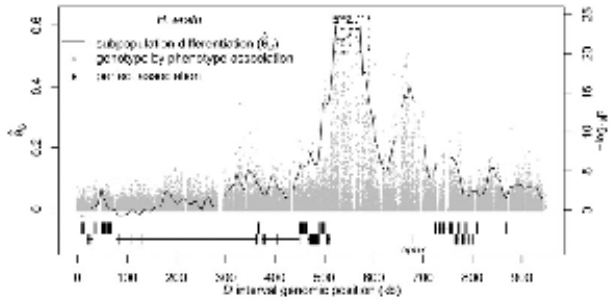


Figure 2. Divergence and association between divergent *H. erato* color pattern races across the *D* interval. The solid line indicates sliding window (15-kb window size, 5-kb step size) subpopulation differentiation ($\hat{\theta}_s$) between the *H. erato* postman and rayed phenotypes for individuals from three hybrid zones (Peru, Ecuador, French Guiana; $n_{postman} = 20$; $n_{rayed} = 17$), requiring a minimum of 75% of individuals genotyped for each phenotype at each position and data for at least 20% of positions in each window. A baseline subpopulation differentiation of $\hat{\theta}_s = -0.07$ was calculated from genomic intervals unlinked to color pattern. The dots indicate genotype by phenotype association calculated for biallelic SNPs using a Fisher's exact test for all four hybrid zones ($n_{postman} = 28$; $n_{rayed} = 17$), requiring a minimum of 75% of individuals genotyped for each phenotype at each SNP. The black dots indicate association for the 76 SNPs perfectly associated with phenotype. The gene annotations are shown *below* for the plus strand (*top*) and minus strand (*bottom*), with the single exon gene *optix* denoted. Wider boxes represent coding exons and the narrower boxes represent introns.

There were striking similarities in patterns of differentiation and association between *H. erato* and *H. melpomene*, two species that show identical shifts in their mimetic wing patterns across their geographic ranges (Turner 1975). The same two peaks of population differentiation and phenotypic association were seen when examining *H. erato* and *H. melpomene* in the Peruvian hybrid zone (Supplemental Fig. S1). The major peak of divergence and association identified in the *H. erato* hybrid zones (Fig. 2) perfectly coincides with the region of peak divergence and association identified in *H. melpomene* hybrid zones (Fig. 4; Supplemental Fig. S1).

Table 1. Hybrid zone sampling

Species	Race	Location	Red phenotype	Sample size
H. erato	*favorinus*	Peru	postman	8
	emma	Peru	rayed	6
	hydara	French Guiana	postman	7
	erato	French Guiana	rayed	6
	notabilis	Ecuador	postman	5
	lativitta	Ecuador	rayed	5
	petiverana	Panama	postman	5
	hydara	Panama	postman	3
H. melpomene	*melpomene*	Colombia	postman	3
	malleti	Colombia	rayed	3

Haplotype structure and recombination under the divergence peak

To investigate what might be driving the high divergence in the 65-kb peak, we looked for evidence of chromosomal rearrange-

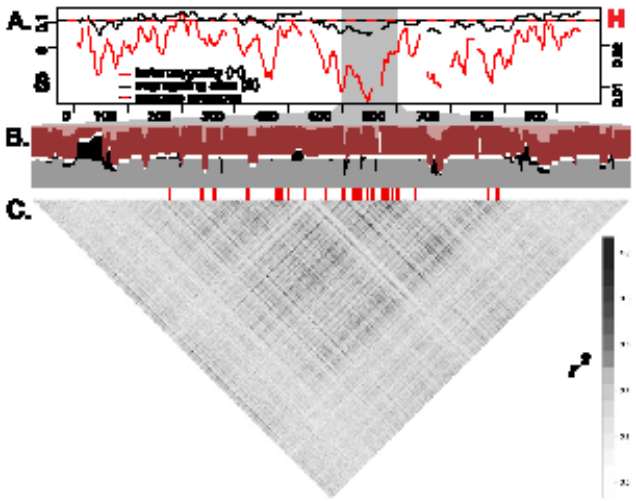


Figure 3. Signatures of selection and recombination across the *D* interval in the *H. erato* Peruvian hybrid zone ($n = 14$). (*A*) Solid lines show sliding window (15-kb window size, 5-kb step size) values for the number of segregating sites (S, black) and heterozygosity (H, red). Horizontal dashed line represents averages of S and H from genomic regions unlinked to color, with axes scaled to match averages. (*B*) Haplotypes clustered from 3227 SNPs from the 500–600 kb region, which includes the peak of divergence. Haplotypes from rayed (shaded dark) and postman (shaded light) races were clustered into two groups ($K = 2$) (colored red and gray). A horizontal white row separates the two clusters. Blocks that have different haplotypes fixed between the two color pattern races are represented by regions with sets of neighboring SNPs in which all rayed individuals are assigned exclusively to the red cluster (red columns shaded entirely dark) and all postman individuals are assigned exclusively to the gray cluster (gray columns shaded entirely light). Blocks where the two color pattern races share haplotypes are represented by regions where individuals of a single color pattern race are assigned to more than one cluster (red columns with light and dark shading; gray columns with light and dark shading). SNPs with fixed allelic differences between color patterns are denoted by red hash marks *below* the haplotype clustering (data from Fig. 2). (*C*) Correlation plot of linkage disequilibrium (r^2) among 3187 biallelic SNPs across the 500–600-kb window on the *D* interval.

ments, which could reduce recombination between color pattern races, resulting in high divergence. Using BreakDancer (Chen et al. 2009) to look for paired-end alignments indicative of inversions, we found no evidence of a chromosomal inversion between color pattern races in the *H. erato* and *H. melpomene* genomic data across this region. These analyses have limited power due to the necessity of having good coverage across inversion breakpoints from read pairs with one pair in the inverted region and the other outside. Despite these limitations, these results, along with the perfect gene synteny between species, suggest that divergent color pattern races are collinear and can recombine across the *D* locus.

We more closely examined haplotypes and linkage disequilibrium (LD) across the 65-kb peak of divergence for patterns of recombination. In genomic regions unlinked to color, LD decayed very rapidly, falling to background levels within a few thousand bases (Supplemental Fig. S3). Across the three hybrid zones between rayed and postman races, LD was consistently higher across the 65-kb peak of divergence than across other regions of the *D* interval (Supplemental Fig. S3).

Although recombination in the 65-kb peak region appears to be substantially reduced, haplotype reconstruction suggests that recombination is present. In the Peruvian hybrid zone, clustering of reconstructed haplotypes into two groups does not support a single large haplotype block distinguishing the two color pattern races. Rather, the clustering revealed several smaller blocks with haplotypes fixed between the color pattern races that were intervened by narrow regions where both races shared the same haplotypes. Similar to the LD analyses, haplotypes clustered more strongly by color pattern in the 65-kb peak than in the regions flanking the peak (Fig. 3B). Further examination of the distribution of SNPs across races revealed that nearly half of the SNPs that differ between hybridizing races were polymorphic in only a single hybrid zone.

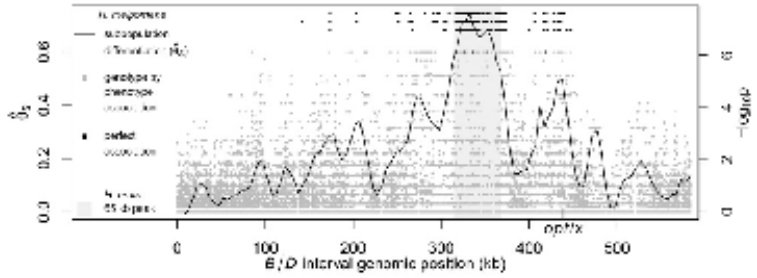


Figure 4. Divergence and association between divergent *H. melpomene* color pattern races across the *B/D* interval. The gray shaded area indicates the region of peak divergence and association identified in the co-mimic *H. erato* (Fig. 2). The solid line indicates sliding window (15-kb window size, 5-kb step size) subpopulation differentiation ($\hat{\theta}_x$) between the *H. melpomene* postman and rayed phenotypes for individuals from two hybrid zones (Peru, Colombia; $n_{postman} = 7$; $n_{rayed} = 7$), requiring a minimum of 75% of individuals genotyped for each phenotype at each position and data for at least 20% of positions in the window. A baseline subpopulation differentiation of $\hat{\theta}_x = -0.03$ was calculated from genomic scaffolds unlinked to color pattern. The dots indicate genotype by phenotype association calculated for biallelic SNPs using a Fisher's exact test for both hybrid zones ($n_{postman} = 7$; $n_{rayed} = 7$), requiring a minimum of 75% of individuals genotyped for each phenotype at each SNP. The black dots indicate association for the 430 SNPs perfectly associated with phenotype.

This large proportion of polymorphic sites private to a specific hybrid zone indicates that alleles and haplotypes shared between hybridizing races across the 65-kb peak likely result from recent gene flow at each hybrid zone, rather than incomplete sorting of ancestral variation among the different races.

The evolutionary history of red wing patterns

We examined phylogenetic relationships across the *D* interval using two methods—a sliding window likelihood tree preference and optimally partitioned phylogenetic topologies (Fig. 5). We identified a major shift in evolutionary history across the 200-kb region supported by the divergence and association analyses. Likelihood scores strongly support a tree that clusters samples by color pattern phenotype, rather than the tree that clusters samples by geographic population (Fig. 5). Additionally, the optimally partitioned phylogenetic topologies showed a partition encompassing the 200-kb region where the tree perfectly clusters by color. Moving away from this 200-kb region, phylogenetic topology shifted to a transitional topology between color-based and geography-based trees before transitioning to a geography-based tree on the far edge of the interval (Fig. 5). The tree in the color partition supports a single origin of the rayed color pattern, clustering rayed phenotypes separate from nonrayed phenotypes, as does a tree inferred from the 65-kb region of highest divergence (Supplemental Fig. S5). Branch lengths across topologies show a signature of reduced gene flow, whereby color pattern alleles have reduced gene flow among races and less variation among individuals relative to markers further from the color pattern locus (Fig. 5; Supplemental Fig. S5).

We compared the evolutionary history of the red color pattern interval between the two co-mimics, *H. erato* and *H. melpomene*, which showed concordant peaks of divergence, indicating that they use the same genetic architecture to generate their convergent phenotypes. We aligned 71 regions showing high conservation between the two reference sequences within a 300-kb window that included both peaks of divergence. Phylogenetic analyses of 69 of the 71 fragments support complete monophyly of each species, rather than the clustering of samples by phenotype that would be expected if the phenotypes shared a common origin. Nonmonophyly in the two remaining fragments was driven by poor alignments of a few samples with extensive missing data. We more closely examined histories of SNPs within the 65-kb peak of divergence. After manual removal of regions with poor alignment and SNPs with missing data for more than 25% of samples, we obtained 1164 SNPs. *Heliconius erato* and *H. melpomene* share allelic variation across 73 of these 1164 SNPs, none of which show patterns of association with phenotype across both species. Phylogenetic analysis of the 1164 SNPs resolved *H. erato* and *H. melpomene* as separate lineages with high support, while clustering races by phenotype within each species (Supplemental Fig. S6). Results from these data support an independent origin of red patterns within each species.

Discussion

The high definition genomic data sets provided by next-generation sequencing techniques can give important insights into how different ecological pressures, complex genetic architectures, and evolutionary histories shape patterns of genomic variation. In this

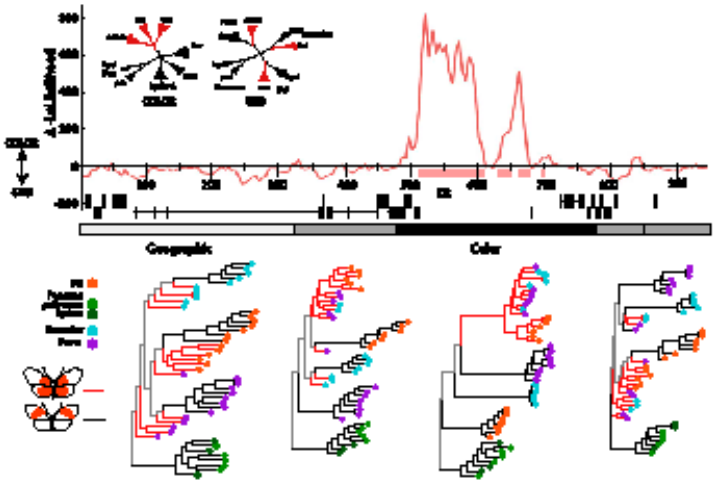


Figure 5. Phylogenetic trees across the *D* interval. The *top* panel is the likelihood preference for color-based versus geography-based tree models (*inset*). Pink bars *beneath* the plot represent the region where neighbor-joining trees in this sliding window have a monophyletic lineage for the rayed phenotype. These peaks are shown relative to the annotated genes in the interval. The gray bar in the *middle* represents the five most optimal topological splits used for phylogenetic tree reconstruction, colored by the general history inferred, with black representing a tree clustering by color pattern, light gray a tree clustering by geographic region, and medium gray for transitional topologies. Tree topologies for these divisions are shown *below*, excluding the tree for the fourth interval, which does not differ substantially from the fifth tree. Nodes supported by a posterior probability >0.95 are represented with bold branches preceding them. Phenotypes are represented by branch color and geographic regions by terminal node color. Internal branches are colored only in cases of clade monophyly of that pattern, otherwise branches are represented in gray. All unrooted tree topologies were arbitrarily rooted at the Panamanian lineage for presentation.

study, we used whole genome resequencing to successfully narrow the region modulating adaptive red color pattern variation in *H. erato* to a 65-kb interval of peak divergence that showed strong signatures of selection and included almost all of the SNPs perfectly associated with wing pattern phenotype (Fig. 2). There were no predicted genes nor any evidence of transcriptional activity across this region, which strongly supports the assertion that red pattern variation is caused by *cis*-regulatory changes controlling gene expression during wing development (Reed et al. 2011). Although this genetic architecture appears to be highly conserved across the *Heliconius* genus, it most likely arose independently in the co-mimics *H. erato* and *H. melpomene*.

Signatures of strong selection and recombination across the 65-kb region

The patterns of genomic variation across the *H. erato D* interval reflect a history of high rates of gene flow coupled with strong selection. Across most of the 2.2 Mb of the genome we analyzed, gene flow has homogenized genomic variation between hybridizing color pattern races of *H. erato*, resulting in levels of differentiation near zero. Within narrow sections of the *D* interval, however, there are sharply defined regions of divergence (Fig. 2), consistent with strong natural selection operating on mimetic color patterns (Benson 1972; Mallet and Barton 1989; Mallet et al. 1990; Kapan 2001). We observed a 65-kb region that showed hallmark signatures of strong selection—high divergence between populations, low nucleotide diversity, and strong LD. This region also contained almost all of the fixed allelic differences between divergent color pattern races.

The 65-kb region showing high divergence is much wider than expected given the rapid decay of LD to background levels in regions unlinked to color pattern, often within just a few thousand bases in *H. erato* (Supplemental Fig. S3; Counterman et al. 2010). Broad and sharp peaks of divergence can be driven by chromosomal rearrangements (e.g., inversion), which are known to suppress recombination (Hartl and Jones 2005), causing elevated patterns of divergence (McGaugh and Noor 2012). Recently, chromosomal inversions were shown to be important in maintaining wing color pattern polymorphisms among sympatric forms of *Heliconius numata* (Joron et al. 2011). However, we found no evidence for large chromosomal inversions between different color pattern races in our analyses of both gene synteny and the orientation of paired-end alignments.

Additional support against chromosomal inversions comes from finer analysis of haplotype structure, which suggests that recombination, while reduced, is occurring in the 65-kb peak of divergence. Across this peak we identified several regions where individuals from both color pattern races in the *H. erato* Peruvian hybrid zone shared similar haplotypes (Fig. 3). These blocks of shared haplotypes likely reflect gene flow between divergent races across these genomic regions, which suggests regions of this peak can recombine and introgress between rayed and postman individuals. The presence of recombination in this region indicates that genotyping of more individuals will provide a finer scale picture of how strong selection and recombination between divergent wing pattern races shape patterns of introgression across this 65-kb region of the genome.

In addition to the 65-kb peak of high divergence, we detected a second peak of moderate divergence located at the gene *optix*. This peak occurred in both *H. erato* and *H. melpomene*; however, the height and exact location of the peak varied between replicate *H. erato* hybrid zones. The peak showed high divergence in all of the western hybrid zones, whereas French Guiana, the only eastern hybrid zone, showed a low peak of divergence. Although the signatures of selection indicate that this region is functionally important, the different patterns of divergence between hybrid zones could be a result of either the unique histories of each hybrid zone or a complex genetic architecture regulating red color pattern variation.

The regulatory architecture of *optix*

We predict that the peaks of divergence identified in our analyses harbor the key regulatory regions of the transcription factor *optix*. Previous work has shown that expression patterns of *optix* are consistent with a critical role in red color pattern variation and the amino acid sequence is conserved across multiple *Heliconius* species, implicating a regulatory mechanism modulating red variation (Reed et al. 2011). We see strong signatures of selection across a 200-kb gene desert that contains only the gene *optix*. We believe that the major 65-kb peak of divergence, which is located in the gene desert ~130 kb 3′ of *optix*, contains multiple *cis*-regulatory elements of *optix*.

The *D* locus segregates with a simple Mendelian inheritance pattern (Mallet 1989; Kapan et al. 2006); however, phenotypic patterns from *H. erato* hybrid zones suggest that different red color pattern elements are genetically distinct. Phenotypic hybrid forms having a red forewing dennis patch but no red hindwing rays should not exist if the same genetic variant at the *D* locus controls both those red pattern elements. Yet such individuals have been found on very rare occasions in hybrid zones (Mallet 1989), and they exist as naturally occurring color pattern races in both *H. erato* (*amalfreda*) and *H. melpomene* (*meriana*) (Sheppard et al. 1985).

We hypothesize that the 65-kb divergence and association peak actually contains a cluster of tightly linked *cis*-regulatory elements of *optix*, each controlling one of the three distinct red pattern elements. There is increasing evidence in other systems that some loci, originally thought to be single large effect loci, are actually clusters of tightly linked genetic changes (Rebeiz et al. 2009; Frankel et al. 2011; Studer and Doebley 2011; Loehlin and Werren 2012; Linnen et al. 2013). Our haplotype analysis supports this hypothesis because the 65-kb peak is rather broad and contains a number of haplotype blocks that are perfectly associated with phenotype, with haplotype blocks of low association interspersed between them. Although this could be a consequence of the history of recombination alone, it could also be indicative of several conserved subregions that control different aspects of phenotypic variation (Fig. 3).

In *Heliconius*, these clustered regulatory elements may bind different upstream regulatory genes that control the spatially specific expression patterns of *optix* and ultimately determine scale coloration across the wings. The region could thus act as an evolutionary hotspot, with the flexibility to generate diverse phenotypes, but where strong selection and physical linkage ensure that mimetic patterns are inherited as a whole.

Independent origins of a shared genetic architecture between co-mimics

The genetic architecture driving color pattern convergence and divergence appears to be highly conserved across the *Heliconius* genus. The *D* interval showed strikingly consistent differentiation across replicate *H. erato* hybrid zones between divergent color pattern races (Supplemental Fig. S4). Moreover, the 65-kb divergence

and association peak identified in our *H. erato* analysis is nearly identical to the peak of divergence in the co-mimic, *H. melpomene* (Fig. 4; Supplemental Fig. S1). These observations raise some interesting questions about the broader evolutionary history of this region. Did the alleles responsible for mimetic patterns in *H. erato* and *H. melpomene* evolve independently or do they share a common origin, through either adaptive introgression or shared ancestral variation?

The history of color pattern evolution within *H. erato* indicates that races with similar red wing patterns likely share a common origin. Consistent with a previous gene-based study (Hines et al. 2011), sliding window phylogenetic analyses and inferred phylogenetic trees for the 65-kb region strongly support a common origin of the rayed pattern in *H. erato* (Fig. 5; Supplemental Fig. S5).

Adaptive introgression of red color pattern alleles between species could be argued to explain a shared origin of mimetic phenotypes across *Heliconius*, as it does within the *H. melpomene* clade (The *Heliconius* Genome Consortium 2012; Pardo-Diaz et al. 2012). Similarly, ancestral variation, shared between the co-mimics, could also be elicited to explain the shared mimetic phenotypes. Yet, in our comparative analysis across the 65-kb peak that we have identified as responsible for red patterning, we found no evidence for a common origin of shared color pattern variation in these co-mimics. No SNPs were perfectly associated with color pattern in both species, and all genealogies from this region were reciprocally monophyletic for both species. These results are consistent with the deep genetic divergence (Pohl et al. 2009) and lack of hybridization between *H. erato* and *H. melpomene*.

Although color patterns likely have a single origin within each species, patterns of variation suggest similar red patterns evolved independently in the co-mimetic species. Analyses between color pattern races within a species show numerous shared SNPs and strong phylogenetic clustering—indicating a shared origin or the rayed phenotype. Analyses between co-mimetic species show no shared SNPs and no phylogenetic clustering—indicating that it is most likely that the two parallel radiations independently evolved mimetic red patterns through changes in the same regulatory regions.

Although we found no evidence for a shared history of red color patterns between the co-mimics *H. erato* and *H. melpomene*, it is important to point out several caveats. Foremost, we are only beginning to understand the evolutionary dynamics of this region. Patterns of variation, including SNPs and insertions/deletions, across this noncoding genomic interval are complex, and our strategy of mapping short sequencing reads onto a static reference sequence may miss variation that is functionally important. In addition, mapping of these data onto different species-specific references may introduce some bias. The improvement of algorithms for de novo assembly from short read sequence data (Miller et al. 2010; Schatz et al. 2010), coupled with emerging technologies for generating long sequencing reads from single molecules (Schadt et al. 2010), will allow more detailed dissection of the evolutionary dynamics of this region in the future. Resolution of the question of the origin of mimicry in *Heliconius* awaits final discovery of the precise functional changes regulating phenotypic variation in both species.

In conclusion, using genomic sequencing of a small number of individuals from multiple admixed hybrid zones, we were able to confidently localize the noncoding genetic switch modulating an adaptive phenotype. We showed that both distantly related co-mimics used this same genetic architecture, but although regulatory alleles appear to have a single origin within species, mutations were most likely independently acquired in the co-mimetic species to produce the shared phenotypes. This study highlights how next-generation sequencing techniques can be leveraged to identify functional variation and to understand the evolutionary history of adaptive radiations.

Methods

Annotation, synteny, and conservation

We annotated 2.2 Mb of *Heliconius erato petiverana* BAC sequences from the ~400-Mb *H. erato* genome (see Supplemental Material section S1 for further details). Approximately 1 Mb of this reference sequence, called the *D* interval, was previously identified as being involved in red wing color pattern elements (Counterman et al. 2010). We masked repetitive elements in this partial reference genome using RepeatMasker v3-2-9 (Smit et al. 2010) and a *Heliconius* repetitive elements database (The *Heliconius* Genome Consortium 2012).

We generated a partial reference transcriptome for *H. erato* wing tissue using reference-based assembly of Illumina RNA-seq short-read data. We obtained whole transcriptome short-read sequences from hindwing cDNA for 18 individuals, including two divergent *H. erato* color pattern races and three developmental stages. Transcripts were generated from these pooled samples using the Bowtie/TopHat/Cufflinks pipeline. We aligned each sample to the masked reference sequence using TopHat v1.2.0 (Trapnell et al. 2009) and Bowtie v0.12.7.0 (Langmead et al. 2009), with stringent mapping parameters to minimize false alignments. We generated transcripts by analyzing alignments with Cufflinks v1.0.1 (Trapnell et al. 2010).

We produced automated gene annotations across the *H. erato* partial genomic reference sequence using the MAKER pipeline v2.09 (Holt and Yandell 2011). This analysis involved masking repetitive elements in the reference sequence, aligning peptide sequences from the UniRef90 (Suzek et al. 2007) and *Bombyx mori* (Duan et al. 2010) protein databases, and aligning transcripts from the RNA-seq data and previously published EST sequences from wing tissues of *H. erato* and *H. himera* (Papanicolaou et al. 2009). MAKER generated ab initio gene models for both the masked and unmasked reference using AUGUSTUS v2.5.5 (Stanke et al. 2006) trained for *H. melpomene* (The *Heliconius* Genome Consortium 2012) and SNAP v2010-07-28 (Korf 2004) trained for *Bombyx mori*. MAKER generated gene predictions by promoting ab initio models with enough supporting evidence from protein homology or transcriptome alignments. We manually curated the predicted genes within the *D* interval based on the supporting evidence and alignments to annotated *Heliconius* proteins and curated insect proteins. We assigned gene descriptions and putative functions based on homology with known proteins and protein domains. In addition to protein coding genes, we predicted transcription factor binding sites using the Transcription Element Search System (TESS) (Schug and Overton 1997), searching against known *Drosophila* binding sites from the TRANSFAC and JASPAR databases.

We examined gene synteny across the *D* interval between the curated *H. erato* peptide sequences and *H. melpomene* v1.0 peptide sequences (The *Heliconius* Genome Consortium 2012) (see Supplemental Material section S2 for further details). We used Inparanoid v4.0 (Ostlund et al. 2010) to identify one-to-one orthologs and examined genes for consistent order and orientation using OrthoCluster release 2 (Vergara and Chen 2009).

We examined the level of sequence conservation between *H. erato* and *H. melpomene* across the *D* interval (see Supplemental Material section S2 for further details). We used mVista LAGAN

(Brudno et al. 2003) to globally align the *H. erato D* interval sequence and the *H. melpomene* scaffolds containing the orthologous genes identified by the Inparanoid analysis. We examined sequence conservation in 500-bp windows across the interval, identifying regions of >90% similarity.

Sequencing and genotyping

To determine where in the *D* interval different red phenotypes diverge genetically, and therefore where the genetic control of the red phenotype is most likely located, we examined genomic sequence data for divergent red color pattern races of *H. erato* (see Supplemental Material section S3 for further details). We collected 45 individual *H. erato* butterflies from hybrid zones in Peru ($n = 14$), French Guiana ($n = 13$), Ecuador ($n = 10$), and Panama ($n = 8$) (Fig. 1; Supplemental Table S6). We collected phenotypically pure individuals of each color pattern race from admixed populations where the ranges of two color pattern races overlap. For dissecting red color pattern variation, the hybrid zones in Peru, French Guiana, and Ecuador are considered replicate hybrid zones since each involves hybridization between rayed and postman races. The Panamanian hybrid zone serves as a control in that both races are postman phenotypes, showing variation only in the yellow phenotypic elements, which are under independent genetic control from the red elements (Mallet 1986). Additionally, we collected six *H. melpomene* individuals near a hybrid zone in eastern Colombia, three samples representing each of the two major red phenotypes—the postman and the rayed (Supplemental Table S7). We assessed variation across a second *H. melpomene* hybrid zone in Peru, which is also represented by postman and rayed phenotypes, using published targeted resequencing data (Nadeau et al. 2012).

We sequenced the whole genome of each sample on the Illumina platform, producing 100-bp paired-end reads. We aligned the *H. erato* sequencing reads to our unmasked *H. erato* partial reference genome and *H. melpomene* reads to the *H. melpomene* genome v1.1 (The *Heliconius* Genome Consortium 2012) using BWA v0.5.9-r16 (Li and Durbin 2009) with relaxed mapping parameters. We called multisample genotypes across samples for each race using GATK's (DePristo et al. 2011; McKenna et al. 2010) UnifiedGenotyper with heterozygosity set to 0.025 and filtered genotype calls for quality using GATK's VariantFiltration, applying both site and individual sample filters to remove low quality genotypes, low coverage regions, and hypercoverage regions indicative of repetitive elements. We used BreakDancer v1.2.6 (Chen et al. 2009) to identify regions of the reference sequence that showed paired end alignments with incorrect orientations and unexpected distances between pairs, indicating possible structural rearrangement.

Divergence and association analyses

We examined signatures of selection and genotype by phenotype association between divergent color pattern races (see Supplemental Material section S4 for further details). We calculated sliding window genomic divergence between pairs of *H. erato* color pattern races at each hybrid zone independently and across all hybrid zones combined, with samples classified as either postman or rayed phenotypes. We used a model for diploid data with populations as random effects ($\hat{\theta}$) (Weir 1996) and no simplifying assumptions regarding sample sizes or number of populations (Weir and Cockerham 1984). For analyzing all hybrid zones combined, we incorporated the geographic structure of the populations by using a three-level hierarchy method ($\hat{\theta}_{IT}$) (Weir 1996). For all comparisons, we calculated divergence at a position only if at least 75% of the individuals were genotyped for each phenotype. We evaluated 15-kb sliding windows at 5-kb steps across the genomic intervals and required a window to have divergence calculated for at least 20% of the positions in the window. We calculated a baseline level of divergence for each comparison as the level of divergence observed across intervals unlinked to color pattern (*H. erato*: three unlinked BACs; *H. melpomene*: 38 unlinked scaffolds).

We estimated genotype by phenotype association at each *H. erato* hybrid zone independently and across all four *H. erato* hybrid zones combined. We examined each biallelic SNP using a two-tailed Fisher's exact test based on allele counts. Positions were excluded if <75% of individuals were genotyped for each phenotype.

To look for signatures of selection, we calculated sliding window values for the proportion of segregating sites and heterozygosity for the *H. erato* Peruvian hybrid zone. We calculated estimates of these parameters for a genomic position only if at least 75% of individuals were genotyped and then looked at 15-kb windows with a 5-kb step size, for windows with at least 20% of positions with parameter estimates. To explore linkage disequilibrium (LD), we examined all biallelic SNPs with at least 75% of individuals genotyped. We calculated correlations (r^2) between all pairwise SNPs using PLINK (Purcell et al. 2007), which for unphased data is based on genotype allele counts.

We assessed divergence and association in the *H. melpomene* hybrid zones as described above. Additionally, for each *H. erato* fixed SNP, we attempted to identify an orthologous SNP in *H. melpomene* and determined if the SNP was associated with phenotype in both species.

Linkage disequilibrium and haplotype clustering

We explored haplotype structure in the Peruvian hybrid zone by estimating haplotypes across a 100-kb window of the *D* interval (500–600 kb) containing the 65-kb peak of divergence and flanking regions using fastPHASE v1.2 (Scheet and Stephens 2006) (see Supplemental Material section S6 for further details). We filtered biallelic SNPs across this 100-kb region to remove sites that had genotypes from <75% of the individuals of each race, resulting in 3227 SNPs. Haplotypes were clustered during phase estimation to two clusters ($K = 2$) and the proportion of rayed and postman individuals assigned to each cluster at each SNP was determined. We used Haploscope (San Lucas et al. 2012) to visualize regions where the two races had fixed haplotype block differences and where individuals from both races shared the same haplotypes. Using the haplotype estimations from fastPHASE, for each SNP, Haploscope visualizes the portion of individuals from a race (light vs. dark) assigned to each cluster (red vs. gray) across the 100-kb region (San Lucas et al. 2012).

Phylogenetic analyses

We constructed phylogenetic trees across sliding windows in the *D* interval, sampling 15 kb of sequence every 5 kb (see Supplemental Material section S5 for further details). For each window, we tested the log likelihood of the data with two alternative trees: The geographic tree assumes samples cluster by geographic hybrid zone and the color based tree groups races with a similar color pattern (rayed or postman) (Fig. 5). Likelihood values were calculated for each interval and tree topology using scripts in PAUP* 4b10 (Swofford 2002), using a GTR + G model inferred for the interval as a whole using Modeltest v3.7 (Posada and Crandall 1998). Neighbor-joining trees across these sliding windows constructed in PAUP* were used to infer regions of monophyly by color phenotype.

To summarize variation in phylogenetic topology across the interval, we constrained division of the interval into the five most distinct topologies using the MDL method, raising the likelihood

score penalty until five clusters of SNP blocks were reached (Ané 2011). Tree topologies for each of these five regions of the interval were constructed using MrBayes v3.1.2 (Ronquist and Huelsenbeck 2003). Analyses involved three runs for 3 million generations each, sampling every 500 generations and removing 33% burn-in and runs that did not converge (as assessed in MrBayes and Tracer v1.5 [Rambaut and Drummond 2007]). Models were assigned using MrModeltest v2.3 (Nylander 2004) and included the GTR model for the second and third regions of the interval and GTR + G for the remaining regions. In addition to these phylogenies, phylogenetic and network-based trees were constructed for both the 65-kb peak region of population differentiation and unlinked genetic regions.

To test whether shared color patterns between the mimics could result from a common origin, we also performed phylogenetic analyses combining *H. erato* and *H. melpomene* sequences along this interval. We focused on regions of high conservation between *H. erato* and *H. melpomene* (>80% conservation in a 500-bp window) from our mVista alignment within the 515–580 kb region of peak divergence. We used ClustalW2 (Larkin et al. 2007) and manual edits to align sequences from all 45 *H. erato* individuals (Supplemental Table S6) and 14 *H. melpomene* individuals (Supplemental Table S7). After filtering, 1134 SNPs were concatenated and used for a Bayesian analysis using the same parameters above and a GTR model.

Data access

Annotated gene models are available on the *Heliconius* Genome Project website (http://butterflygenome.org/) through the Genome Browser (Data Source: Hera_D_Jan2012) and as a downloadable gff file. Reference sequences are available on Genbank (http://www.ncbi.nlm.nih.gov/genbank) under accession numbers KC469892–KC469895 and AC208805–AC208806. Aligned sequencing reads are available at the NCBI Sequence Read Archive (SRA) (http://www.ncbi.nlm.nih.gov/sra) under accession numbers SRA059512 (resequencing) and SRA060220 (RNA-seq). SNP data, scripts, and tree files for phylogenetic analyses are available at Dryad (datadryad.org) under doi: 10.5061/dryad.rn65n.

Acknowledgments

We wish to thank Claudia Rosales for the tremendous amount of work she did in the laboratory; Jamie Walters and Chris Jiggins for contributing sequencing data; and Chris Smith for computational support. We thank Arnaud Martin, the editor, and three reviewers for their insightful comments on previous versions of the manuscript. We also thank the following permitting agencies for permission to collect butterflies: Peruvian Ministerio de Agricultura and Instituto Nacional de Recursos Naturales (004-2008-INRENA-IFFS-DCB and 011756-AG-INRENA); Ecuadorian Ministerio del Ambiente Ecuadorian (013-09 IC-FAU-DNB/MA); French Guiana Ministere de L'Ecologie, de L'Energie, du Developpemet Durable et de la Mar (BIODAD-2010-0433); Panamanian Autoridad Nacional del Ambiente (SC/A-7-11); and Colombian Ministerio de Ambiente, Vivienda y Desarrollo Territorial (RGE0027-LAM3483). This work was funded by the following awards: CNRS Nouraugues (B.A.C.); NIH F32 GM88994 2 (H.M.H.) and T32 HD060555 (J.J.L.); NSF DEB-1257839 (B.A.C.), DEB-0844244 (W.O.M.), DEB-0715096 (W.O.M.), and IOS-1305686 (J.J.L.); and the Smithsonian Institution.

Author contributions: Experimental design is credited to W.O.M., B.A.C., H.M.H., D.M.N., and M.A.S.; data collection was carried out by H.M.H., B.A.C., M.A.S., W.O.M., and C.S.; data analysis was done by M.A.S., H.M.H., B.A.C., K.K.D., J.J.L., and W.O.M. with assistance from D.A.R. and C.L.; and the manuscript was prepared by M.A.S., B.A.C., W.O.M., and H.M.H.

References

Ané C. 2011. Detecting phylogenetic breakpoints and discordance from genome-wide alignments for species tree reconstruction. *Genome Biol Evol* **3:** 246–258.

Baxter SW, Nadeau NJ, Maroja LS, Wilkinson P, Counterman BA, Dawson A, Beltran M, Perez-Espona S, Chamberlain N, Ferguson L, et al. 2010. Genomic hotspots for adaptation: The population genetics of Müllerian mimicry in the *Heliconius melpomene* clade. *PLoS Genet* **6:** e1000794.

Benson WW. 1972. Natural selection for Müllerian mimicry in *Heliconius erato* in Costa Rica. *Science* **176:** 936–939.

Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Program NCS, Green ED, Sidow A, Batzoglou S. 2003. LAGAN and Multi-LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res* **13:** 721–731.

Chai P. 1986. Field observations and feeding experiments on the responses of rufous-tailed jacamars (*Galbula ruficauda*) to free-flying butterflies in a tropical rainforest. *Biol J Linn Soc Lond* **29:** 161–189.

Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, et al. 2009. BreakDancer: An algorithm for high-resolution mapping of genomic structural variation. *Nat Methods* **6:** 677–681.

Counterman BA, Araujo-Perez F, Hines HM, Baxter SW, Morrison CM, Lindstrom DP, Papa R, Ferguson L, Joron M, ffrench-Constant RH, et al. 2010. Genomic hotspots for adaptation: The population genetics of Müllerian mimicry in *Heliconius erato*. *PLoS Genet* **6:** e1000796.

DePristo M, Banks E, Poplin R, Garimella K, Maguire J, Hartl C, Philippakis A, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43:** 491–498.

Duan J, Li R, Cheng D, Fan W, Zha X, Cheng T, Wu Y, Wang J, Mita K, Xiang Z, et al. 2010. SilkDB v2.0: A platform for silkworm (*Bombyx mori*) genome biology. *Nucleic Acids Res* **38:** D453–D456.

Frankel N, Erezyilmaz DF, McGregor AP, Wang S, Payre F, Stern DL. 2011. Morphological evolution caused by many subtle-effect substitutions in regulatory DNA. *Nature* **474:** 598–603.

Hartl DL, Jones EW. 2005. *Genetics: Analysis of genes and genomes*, 6th ed. Jones & Bartlett, Sudbury, MA.

The *Heliconius* Genome Consortium. 2012. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* **487:** 94–98.

Hines HM, Counterman BA, Papa R, Albuquerque de Moura P, Cardoso MZ, Linares M, Mallet J, Reed RD, Jiggins CD, Kronforst MR, et al. 2011. Wing patterning gene redefines the mimetic history of *Heliconius* butterflies. *Proc Natl Acad Sci* **108:** 19666–19671.

Holt C, Yandell M. 2011. MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12:** 491.

Joron M, Frezal L, Jones RT, Chamberlain NL, Lee SF, Haag CR, Whibley A, Becuwe M, Baxter SW, Ferguson L, et al. 2011. Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature* **477:** 203–206.

Kapan DD. 2001. Three-butterfly system provides a field test of Müllerian mimicry. *Nature* **409:** 338–340.

Kapan DD, Flanagan NS, Tobler A, Papa R, Reed RD, Acevedo Gonzalez J, Ramirez Restrepo M, Martinez L, Maldonado K, Ritschoff C, et al. 2006. Localization of Müllerian mimicry genes on a dense linkage map of *Heliconius erato*. *Genetics* **173:** 735–757.

Korf I. 2004. Gene finding in novel genomes. *BMC Bioinformatics* **5:** 59.

Langmead B, Trapnell C, Pop M, Salzberg S. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10:** R25.

Larkin M, Blackshields G, Brown N, Chenna R, McGettigan P, McWilliam H, Valentin F, Wallace I, Wilm A, Lopez R, et al. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* **23:** 2947–2948.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25:** 1754–1760.

Linnen CR, Poh YP, Peterson BK, Barrett RD, Larson JG, Jensen JD, Hoekstra HE. 2013. Adaptive evolution of multiple traits through multiple mutations at a single gene. *Science* **339:** 1312–1316.

Loehlin DW, Werren JH. 2012. Evolution of shape by multiple regulatory changes to a growth gene. *Science* **335:** 943–947.

Mallet J. 1986. Hybrid zones of *Heliconius* butterflies in Panama and the stability and movement of warning colour clines. *Heredity* **56:** 191–202.

Mallet J. 1989. The genetics of warning colour in Peruvian hybrid zones of *Heliconius erato* and *H. melpomene*. *Proc R Soc Lond B Biol Sci* **236:** 163–185.

Mallet J, Barton NH. 1989. Strong natural selection in a warning-color hybrid zone. *Evolution* **43:** 421–431.

Mallet J, Barton N, Lamas G, Santisteban J, Muedas M, Eeley H. 1990. Estimates of selection and gene flow from measures of cline width and

linkage disequilibrium in *Heliconius* hybrid zones. *Genetics* **124:** 921–936.
Martin A, Orgogozo V. 2013. The loci of repeated evolution: A catalog of genetic hotspots of phenotypic variation. *Evolution* **67:** 1235–1250.
McGaugh SE, Noor MAF. 2012. Genomic impacts of chromosomal inversions in parapatric *Drosophila* species. *Philos Trans R Soc Lond B Biol Sci* **367:** 422–429.
McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20:** 1297–1303.
Miller JR, Koren S, Sutton G. 2010. Assembly algorithms for next-generation sequencing data. *Genomics* **95:** 315–327.
Nadeau NJ, Whibley A, Jones RT, Davey JW, Dasmahapatra KK, Baxter SW, Quail MA, Joron M, ffrench-Constant RH, Blaxter ML, et al. 2012. Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-scale targeted sequencing. *Philos Trans R Soc Lond B Biol Sci* **367:** 343–353.
Nylander JAA. 2004. MrModeltest v2. (Program distributed by the author.) Evolutionary Biology Centre, Uppsala University, Sweden.
Ostlund G, Schmitt T, Forslund K, Köstler T, Messina DN, Roopra S, Frings O, Sonnhammer ELL. 2010. Inparanoid 7: New algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res* **38:** D196–D203.
Papa R, Martin A, Reed RD. 2008. Genomic hotspots of adaptation in butterfly wing pattern evolution. *Curr Opin Genet Dev* **18:** 559–564.
Papanicolaou A, Stevens JR, ffrench-Constant R, Heckel D. 2009. Next generation transcriptomes for next generation genomes using *est2assembly*. *BMC Bioinformatics* **10:** 447.
Pardo-Diaz C, Salazar C, Baxter SW, Merot C, Figueiredo-Ready W, Joron M, McMillan WO, Jiggins CD. 2012. Adaptive introgression across species boundaries in *Heliconius* butterflies. *PLoS Genet* **8:** e1002752.
Pohl N, Sison-Mangus M, Yee E, Iiswi S, Briscoe A. 2009. Impact of duplicate gene copies on phylogenetic analysis and divergence time estimates in butterflies. *BMC Evol Biol* **9:** 99.
Posada D, Crandall KA. 1998. MODELTEST: Testing the model of DNA substitution. *Bioinformatics* **14:** 817–818.
Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, et al. 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81:** 559–575.
Rambaut A, Drummond AJ. 2007. Tracer v1.4. http://beast.bio.ed.ac.uk/Tracer.
Rebeiz M, Pool JE, Kassner VA, Aquadro CF, Carroll SB. 2009. Stepwise modification of a modular enhancer underlies adaptation in a *Drosophila* population. *Science* **326:** 1663–1667.
Reed RD, Papa R, Martin A, Hines HM, Counterman BA, Pardo-Diaz C, Jiggins CD, Chamberlain NL, Kronforst MR, Chen R, et al. 2011. *optix* drives the repeated convergent evolution of butterfly wing pattern mimicry. *Science* **333:** 1137–1141.
Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19:** 1572–1574.
Rosser N, Phillimore AB, Huertas B, Willmott KR, Mallet J. 2012. Testing historical explanations for gradients in species richness in heliconiine butterflies of tropical America. *Biol J Linn Soc Lond* **105:** 479–497.
San Lucas FA, Rosenberg NA, Scheet P. 2012. Haploscope: A tool for the graphical display of haplotype structure in populations. *Genet Epidemiol* **36:** 17–21.
Schadt EE, Turner S, Kasarskis A. 2010. A window into third-generation sequencing. *Hum Mol Genet* **19:** R227–R240.
Schatz MC, Delcher AL, Salzberg SL. 2010. Assembly of large genomes using second-generation sequencing. *Genome Res* **20:** 1165–1173.
Scheet P, Stephens M. 2006. A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* **78:** 629–644.
Schug J, Overton GC. 1997. TESS: Transcription element search software on the www. In *Technical Report CBIL-TR-1997-1001-v0.0*. Computational Biology and Informatics Laboratory, School of Medicine, University of Pennsylvania. http://www.cbil.upenn.edu/node/30.
Sheppard PM, Turner JRG, Brown KS, Benson WW, Singer MC. 1985. Genetics and the evolution of Muellerian mimicry in *Heliconius* butterflies. *Philos Trans R Soc Lond B Biol Sci* **308:** 433–610.
Smit AFA, Hubley R, Green P. 2010. RepeatMasker Open-3.0. http://www.repeatmasker.org.
Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. 2006. AUGUSTUS: Ab initio prediction of alternative transcripts. *Nucleic Acids Res* **34:** W435–W439.
Stern DL, Orgogozo V. 2009. Is genetic evolution predictable? *Science* **323:** 746–751.
Studer AJ, Doebley JF. 2011. Do large effect QTL fractionate? A case study at the maize domestication QTL *teosinte branched1*. *Genetics* **188:** 673–681.
Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. 2007. UniRef: Comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23:** 1282–1288.
Swofford DL. 2002. *PAUP*. Phylogenetic analysis using parsimony (*and other methods)*, 4th ed. Sinauer Associates, Sunderland, MA.
Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* **25:** 1105–1111.
Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28:** 511–515.
Turner JRG. 1975. A tale of two butterflies. *Nat Hist* **84:** 28–37.
Vergara IA, Chen N. 2009. Using OrthoCluster for the detection of synteny blocks among multiple genomes. *Curr Protoc Bioinformatics* **27:** 6.10.1–6.10.18.
Weir BS. 1996. *Genetic data analysis II*. Sinauer Associates, Sunderland, MA.
Weir BS, Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. *Evolution* **38:** 1358–1370.

Received October 15, 2012; accepted in revised form May 7, 2013.

Supplementary Information—Supple et al. 2013

Table of Contents

S1. Annotation of the *H. erato* Red Color-Pattern (*D*) Interval

S1.1 Annotation Methods

We annotated genes across the red "*D*" color pattern interval of *H. erato*. To provide supporting evidence for gene models, we sequenced and aligned short-read transcriptome data from several races and stages of *H. erato* to the available partial genomic reference sequence. Additional supporting evidence was provided by aligning available EST and protein databases. We manually curated the predicted genes and compared them to predicted genes in the *H. melpomene* genome v1.1 (Heliconius Genome Consortium 2012).

Reference sequence

We examined 2.2 Mb of the approximately 400 Mb *H. erato* genome. A 1 Mb genomic region involved in red wing color pattern (the *D* interval) was sequenced from an *H. e. petiverana* BAC library (Counterman et al. 2010). In addition, we sequenced approximately 1.2 Mb from BAC clones unlinked to red color pattern. We compiled all sequences into a single *H. erato* "reference" genome (Table S1). We masked repetitive elements in this reference using RepeatMasker v3-2-9 (Smit et al. 2010) and a *Heliconius* repetitive elements database (Heliconius Genome Consortium 2012).

Table S1: *H. erato* reference sequence contigs

NCBI accession	size (bp)	description
KC469892	144495	unlinked to red color pattern
KC469893	743824	unlinked to red color pattern
KC469894	948009	red color-pattern (*D*) interval
KC469895	100927	unlinked to red color pattern
AC208805	70206	unlinked to red color pattern
AC208806	134000	unlinked to red color pattern

Transcriptome assembly

We generated a partial reference transcriptome for *H. erato* wing tissue using reference based assembly of Illumina RNA-seq short-read data from hind wing cDNA from 18 individuals, representing two divergent color pattern races (*H. e. favorinus* and *H. e. emma*) that meet in the Peruvian hybrid zone. Each race was sampled in three biological replicates across three developmental stages (5th instar, day 1 pupae, day 3 pupae). These stages are most relevant to phenotypic differentiation, as they precede the physical manifestation of the color phenotype, which occurs around 5 days after pupation, and include the stage of initial differential expression in *optix*, which occurs at day 3 (Reed et al. 2011).

For each individual, we obtained cDNA from whole hindwing tissues and prepared libraries for sequencing using a slightly-modified Illumina protocol (Illumina 2008, outlined in Supplementary Protocols). Briefly, this involved RNA extraction and isolation of mRNA using poly-A tail binding. Transcripts were chemically fragmented and converted to cDNA using random primers. Adapters were ligated and the resulting fragments were size selected from 100-250 bp using gel extraction, and amplified using a 15-cycle PCR.

Each sample was run in a single Illumina lane and paired end sequenced at either 36, 66, or 75 bp lengths on an Illumina GAIIx at either the UNC-CH Genome Analysis Facility or NCSU

Genome Sequence Laboratories. The number of reads ranged from 19.2 to 55.9 million per sample. Sample quality was assessed using FastQC v0.8.0 (Andrews 2011) and a few low quality samples were rerun. To standardize read length and remove adapter contamination due to read-through of short fragments, all reads were trimmed to 36 bp using FASTX-Toolkit's fastx_trimmer v0.0.13 (Gordon 2010). We obtained empirical estimates of the distribution of sequenced fragment lengths for each sample by aligning the reads to the unmasked *H. erato* reference sequence using BWA v0.5.9-r16 (Li and Durbin 2009) with default parameters.

We generated transcripts with a reference-based assembly method using the Bowtie/TopHat/Cufflinks pipeline. Each sample was aligned, using the empirically estimated fragment lengths, to the masked reference sequence using the closure search option in TopHat v1.2.0 (Trapnell et al. 2009) and utilizing Bowtie v0.12.7.0 (Langmead et al. 2009). We used stringent mapping parameters to minimize false alignments (see Table S2). No more than a single mismatch was allowed per 36 bp read, each aligned read could only map to a single location, and each splice junction had to be supported by at least one read with at least 12 bases on each side with no mismatches. Intron and exon size parameters were determined by examining the distribution of sizes in *Bombyx mori* (Duan et al. 2010) and *Drosophila melanogaster* (McQuilton et al. 2012). We used a first round of TopHat alignments to create a library of potential splice junctions across all samples. This splice junction library was used when each sample was realigned using TopHat with all other parameters unchanged. To generate transcripts, alignments for all samples were first merged using SamTools v0.1.9.0 (Li et al. 2009) and then analyzed with Cufflinks v1.0.1 (Trapnell et al. 2010), using default parameters, except for decreasing the minimum and maximum exon lengths.

Table S2: TopHat and Cufflink parameters

TopHat parameter	**value**	**description**
max-multihits	1	number of alignments to allow per read
segment-mismatches	1	number of mismatches allowed per segment
min-intron-length	20	minimum intron length
max-intron-length	4000	maximum intron length
min-anchor	12	minimum number of aligned bases on each side to report a splice junction
allow-indels	true	allows indels
no-coverage-search	true	disables coverage search
closure-search	true	enables closure search
min-closure-exon	3	minimum exon length for closure search
min-closure-intron	20	minimum intron length for closure search
min-segment-intron	20	minimum intron length for split segment
max-segment-intron	20000	maximum intron length for split segment
Cufflink parameter	**value**	**description**
min-intron-length	20	minimum intron length
max-intron-length	20000	maximum intron length

Automated gene annotation

We produced automated gene annotations for the *H. erato* partial genomic reference sequence using the MAKER pipeline v2.09 (Holt and Yandell 2011) with modified parameters (Table S3).

This analysis begins with masking repetitive elements in the reference sequence using RepeatMasker v3-2-9 (Smit et al. 2010) with the *Heliconius* repetitive elements database (Heliconius Genome Consortium 2012). MAKER next aligned peptide sequences from the Uniref90 (Suzek et al. 2007) and *Bombyx mori* (Duan et al. 2010) protein databases using NCBI BLASTX v2.2.24 (Altschul et al. 1997) and polished these alignments with Exonerate v2.2.0 (Slater and Birney 2005) to ensure that multiple hits within a single protein are ordered properly and utilize consensus splice sites. MAKER then aligned *H. erato* ESTs from a previous *de novo* assembly built from Sanger and 454 sequences from wing tissues of several H. erato races and H. himera (Papanicolaou et al. 2009). These ESTs were aligned using NCBI BLASTN v2.2.24 (Altschul et al. 1997) and further polished with Exonerate v2.2.0 (Slater and Birney 2005). We included the set of aligned RNA-seq transcripts as additional EST evidence in the MAKER pipeline. MAKER next generated *ab initio* gene predictions for both the masked and unmasked reference using Augustus v2.5.5 (Stanke et al. 2006) trained for *H. melpomene* (Heliconius Genome Consortium 2012) and SNAP v2010-07-28 (Korf 2004) trained for *Bombyx mori*. Finally, MAKER determined which *ab initio* gene models had enough supporting evidence from aligned peptide sequences, ESTs, and RNA-seq to be promoted to predicted genes. We required promoted models to produce a protein with at least 30 amino acids and have an annotation edit distance (AED) no greater than 0.5, which is a measure of the difference between the model and the supporting evidence (Holt and Yandell 2011). In the event of overlapping models, only the model with the lowest AED was promoted.

Table S3: MAKER behavior options

MAKER parameter	value	description
pred_flank	500	extent of surrounding evidence to pass to gene predictors
AED_threshold	0.5	maximum annotation edit distance
min_protein	30	minimum number of amino acids in a predicted protein
alt_splice	1 [yes]	take additional steps to find alternative splicing?
always_complete	0 [no]	force start and stop codons for every gene?
keep_preds	0 [no]	include unsupported gene predictions to final gene set?
split_hit	20000	expected max intron size for alignments
single_exon	1 [yes]	include single exon EST evidence?
single_length	250	minimum length of single exon ESTs

Manual curation and functional annotation of predicted genes

We manually curated all predicted genes in the *D* interval using Apollo and following the BeeBase protocols, section IV (Munoz-Torres et al. 2011). Curation involved manually examining each predicted gene to see how well it matched the supporting evidence, adding or removing exons based on supporting evidence and shifted exon boundaries to match RNA-seq models. We blasted the resulting peptide sequences against NCBI's non-redundant (nr) protein database (http://www.ncbi.nlm.nih.gov/). Using ClustalW2 (Larkin et al. 2007), we examined alignments between top *Heliconius* hits and non-*Heliconius* hits, focusing on insect proteins from NCBI's Reference Sequence (RefSeq) collection (Pruitt et al. 2007), which is reviewed and curated. We examined alignments for major gaps or differences and attempted to modify the predicted gene to better match the top blast hits.

We assigned gene descriptions based on the top BLAST hit of the curated proteins to the SwissProt protein database (Boeckmann et al. 2003) with an e-value of at least 0.001 and Blast2Go functional annotation (Conesa et al. 2005) of the curated coding sequences blasted against NCBI's non-redundant (nr) protein database (http://www.ncbi.nlm.nih.gov/). We assigned putative functions to genes based on gene ontology terms from the Blast2Go analysis and known functions of domains identified with InterProScan domain recognition analysis (Hunter et al. 2012).

Transcription factor binding site prediction

In silico transcription factor binding site prediction was performed with the Transcription Element Search System (TESS) (Schug and Overton 1997) using the default settings, searching against known *Drosophila* binding sites from the TRANSFAC and JASPAR databases. Custom scripts were used to divide the genomic region downstream of *optix* into non-overlapping 2 kb fasta sequences for upload to TESS. Additional custom post-processing scripts were used to reassemble TESS output files for the full locus and to parse out transcription factor binding site predictions with p-values less than 0.05. The remaining predictions were uploaded to a private UCSC genome browser track for visual inspection.

S1.2 Annotation Results & Discussion

Using the MAKER pipeline, we annotated 30 protein coding genes in the *D* interval based on *ab initio* models with supporting evidence from homology to known proteins and regions of active transcription identified from ESTs and RNA-Seq data (GenBank accession KC469894). The annotated genes cover a wide variety of functions (Table S4). The distribution of genes across the *D* interval showed a 250 kb "gene desert", which contains only a single gene, *optix* (Figure 2), which has been shown to be involved in the red phenotype (Reed et al. 2011).

We used the RNA-seq alignments to identify potential non-coding transcriptional activity in the gene desert. There is transcriptional activity immediately 3' of *optix*. There are a few additional regions throughout the gene desert where RNA-seq reads aligned, but visual inspection revealed these to be artifacts due to repetitive sequences.

TESS transcription factor binding site prediction within the region of peak divergence revealed over 12,000 putative binding sites, 6,927 of which had a p-value less than 0.05. Filtering of these results for potential transcription factors associated with *optix* was unsuccessful due to a lack of known candidate binding sites. The modMine, through the *Drosophila* modEncode Project (Celniker et al. 2009), identifies *eyeless* (*ey*) as the only gene known to bind and regulate *optix*, and identifies a 1435 bp regulatory region that putatively contains the *ey* binding region. A BLAST search did not show a region of significant sequence similarity to this candidate regulatory region in our *H. erato* reference sequences. It is important to note that there is no known evidence of *optix* expression in *Drosophila* wings and it is unknown if similar genes bind and regulate *optix* in developing eyes and wings. The limited data of gene interactions with *optix* hinders our ability to identify which of the thousands of putative transcription factor binding sites across the 65 kb region may be involved in regulating *optix* expression during wing pattern development.

Table S4: Annotated coding genes, their putative functions, and *H. melpomene* orthologs

gene	description	putative function	translation start	translation stop	*H. melpomene* ortholog
HERA000001	hypothetical protein	unknown	7853	14570	HMEL003289
HERA000002	sideroflexin-2-like	cation transmembrane transporter activity	30936	19952	HMEL003292
HERA000003	PWWP domain-containing	transcription factor regulating a developmental process	34342	37275	HMEL003293
HERA000004	haspin-like	protein phosphorylation/serine threonine-protein kinase activity	54012	63492	HMEL003294
HERA000005	hypothetical protein	unknown	66376	69249	HMEL003296
HERA000006	max dimerization-like	regulation of transcription	360347	84307	HMEL001000
HERA000007	DARL anticodon-binding domain-containing	tRNA ligase activity	365880	361488	HMEL001004
HERA000008	DnaJ domain-containing	heat shock protein binding	366972	367988	HMEL001006
HERA000009	blood vessel epicardial substance-like	cell motility and cell adhesion	449332	374194	HMEL001009
HERA000010	ashwin-like	involved in embryonic morphogenesis	450061	451140	HMEL001022
HERA000011	phosphodiesterase 10a-like	involved in signal transduction	454498	464434	HMEL001021
HERA000012	sorting nexin 12-like	phosphatidylinositol binding	470177	466059	HMEL001020
HERA000013	step ii splicing factor slu7-like	Pre-mRNA splicing factor	477368	471255	HMEL001019
HERA000014	kinesin-like	microtubule motor activity	488674	480075	HMEL001018
HERA000015	G protein-coupled receptor-like	transmembrane signaling receptor activity	489421	500099	HMEL001017
HERA000016	epoxide hydrolase 4-like	hydrolase and catalytic activity	511162	505139	HMEL001014

Table S4 (cont.)

gene	description	putative function	translation start	translation stop	*H. melpomene* ortholog
HERA000017	six sine homebox transcription factor (*optix*)	regulation of transcription	680834	680031	HMEL001028
HERA000018	integrator complex subunit 7-like	subunit of the integrator complex which mediates snRNA processing	723974	754207	HMEL001044
HERA000019	leucine repeat-rich	protein binding	754592	756603	HMEL001043
HERA000020	leucine repeat-rich	protein binding	757308	760265	HMEL001042
HERA000021	strabismus/van gogh-like	involved in development	770295	764754	HMEL001039
HERA000022	monocarboxylate transporter-like	transport across membranes	771067	774473	HMEL001038
HERA000023	SCY1-like protein 2-like	protein phosphorylation/serine threonine-protein kinase activity	783021	777647	HMEL001037
HERA000024	TM2 domain-containing protein CG11103-like	unknown	783509	784039	HMEL001036
HERA000025	40S ribosomal protein S13-like	structural constituent of ribosome	786109	785302	HMEL001035
HERA000026	nadh:ubiquinone dehydrogenase-like	NADH dehydrogenase (ubiquinone) activity	786716	787140	HMEL001034
HERA000027	trafficking protein particle complex subunit 5-like	involved in vesicular transport from endoplasmic reticulum to Golgi	794553	792627	HMEL001033
HERA000028	ras-related protein rab-39b-like	involved in small GTPase mediated signal transduction	801187	798560	HMEL001031
HERA000029	THAP domain-containing	nucleic acid binding	808019	810386	HMEL001029
HERA000030	hypothetical protein	unknown	868108	868452	HMEL002053

S2. Synteny and Conservation between Co-Mimics

S2.1 Synteny and Conservation Methods

We examined gene synteny across the *D* interval between *H. erato* and *H. melpomene* genomic reference sequences. We compared the thirty curated peptide sequences from the *H. erato D* interval to the *H. melpomene* v1.0 gene set peptide sequences (Heliconius Genome Consortium 2012) using Inparanoid v4.0 (Ostlund et al. 2010) to identify one-to-one orthologs. Only matches with bootstrap support of >95% and a score of >50 were retained for analysis. We examined gene rearrangements using OrthoCluster release 2 (Vergara and Chen 2009) in rs mode. We estimated the expected number of rearrangements per Mb between *H. erato* and *H. melpomene* using a divergence time of 13.5-26.1 million years (Pohl et al. 2009) and a rearrangement rate of 0.04-0.29, which are the minimum and maximum estimates from comparisons of *H. melpomene*, *Danaus plexipus* (monarch) and *Bombyx mori* (silkworm) genome assemblies (Heliconius Genome Consortium 2012).

To determine if a genomic inversion might be present in the regions of high divergence (see below), we examined a 200 kb region with the greatest divergence between races in the *H. erato D* interval and the *H. melpomene B/D* interval. We used BreakDancer v1.2.6 (Chen et al. 2009), with default parameters, to identify regions of the reference sequence that showed paired end alignments (see below) with incorrect orientations and unexpected distances between pairs.

To examine the level of sequence conservation between *H. erato* and *H. melpomene* across the *D* interval, we used mVista LAGAN (Brudno et al. 2003) to globally align the *H. erato D* interval sequence and the *H. melpomene* scaffolds containing the orthologous genes identified by the Inparanoid analysis. We examined sequence conservation in 500 bp windows across the interval, identifying regions of greater than 90% similarity.

S2.2 Synteny and Conservation Results

Each gene in the *H. erato D* interval had an *H. melpomene* ortholog (Table S4). These orthologs identify two *H. melpomene* scaffolds (HE671887 and HE670865) orthologous to the *H. erato D* interval. The HE670865 scaffold was previously identified as the *H. melpomene B/D* interval, which is responsible for the red color phenotypes, and all genes on this scaffold had been manually curated (Heliconius Genome Consortium 2012). HE671887 was not previously identified as being adjacent to the *B/D* interval and gene HMEL003292 required manual curation. We manually curated the gene based on *H. melpomene* evidence (Heliconius Genome Consortium 2012). For the 30 *H. erato D* interval genes, gene order and orientation were completely conserved relative to *H. melpomene* (Figure S1) and all protein coding *H. melpomene* genes in the homologous regions were present in the *H. erato* annotations. Additionally, the BreakDancer analysis of read pair orientation did not highlight any inversions that could be driving elevated divergence between divergent races. Despite an expected 0.5–8.0 rearrangements/Mb between *H. erato* and *H. melpomene*, we did not detect any gene rearrangements across nearly 1 Mb of sequence across the *D* interval.

We aligned the *H. erato D* interval to the two orthologous *H. melpomene* scaffolds and examined sequence conservation across the alignment. There were 182 highly conserved regions (>90% sequence similarity in a 500 bp window), covering a total of 63 kb of sequence.

Most of the highly conserved regions (82%) fell in coding exons, covering 38 kb of sequence. Of the conserved regions not located in exons, the gene desert near *optix* contained a higher proportion—8% of the gene desert was highly conserved, while only 3% of the rest of the non-exon sequence for the entire interval was highly conserved. Several of these highly conserved regions in the gene desert contain SNPs that show perfect associations with color pattern phenotype.

S3. Sampling and Genotyping Across Replicate Hybrid Zones

S3.1 Sampling and Genotyping Methods

To determine where different red phenotypes diverge genetically, and therefore, where the genetic control of the red phenotype is most likely located, we examined genomic sequence data for multiple individuals from eight different races of *H. erato* and four races of *H. melpomene*, representing two major red phenotypes. These samples were from multiple hybrid zones, where whole genome sequencing of individuals from regions of admixture between divergent color pattern races allows fine dissection of genomic regions driving phenotypic divergence. In hybrid zones between divergent red color pattern races of *Heliconius*, the free exchange of genes will homogenize the genomes, while strong selection on the red color pattern phenotype will create peaks of genetic divergence around the genomic targets of selection.

Sampling

We collected 45 individual *H. erato* butterflies from hybrid zones in Peru, French Guiana, Ecuador, and Panama (Figure 1). Adult individuals were preserved for DNA extraction or transported live to insectaries in Gamboa, Panama to establish phenotypically pure stocks. For each of the four hybrid zones, we collected phenotypically pure samples from admixed populations where the ranges of two color pattern races overlap. In these regions of admixture, gene flow homogenizes the genomes of the two races, while strong selection on color pattern phenotype drives divergence at genomic regions responsible for color pattern phenotypes. For dissecting red color pattern variation, the hybrid zones in Peru, French Guiana, and Ecuador are considered replicate hybrid zones since each involves hybridization between rayed and postman races. The Panamanian hybrid zone serves as a control in that both races are postman phenotypes, showing variation only in the yellow phenotypic elements, which are under independent genetic control from the red elements (Mallet 1986). For each of the eight color pattern races, we collected three to eight phenotypically pure individuals.

Additionally, we collected six *H. melpomene* individuals near a hybrid zone in eastern Colombia, three samples representing each of the two major red phenotypes—the postman (*H. m. melpomene*) and the rayed (*H. m. malleti*). We assessed history across a second *H. melpomene* hybrid zone in Peru—including postman (*H. m. amaryllis*) and rayed (*H. m. aglaope*) phenotypes—using published genome resequencing data (Nadeau et al. 2012).

Sequencing and genotyping

For each sample, we extracted genomic DNA from a partial thorax or whole pupae. We prepared whole genome Illumina libraries (outlined in Supplementary Protocols). Briefly this involved shearing the DNA with a Covaris machine, followed by bead purification, and then

standard Illumina library preparation. We assessed library quality using a fluorimeter and qPCR. Whole genomes of each individual were sequenced on either an Illumina GAIIx or HiSeq at Baylor College of Medicine, producing 100 bp paired end reads. We examined sequence quality for each pair of each sample separately using FastQC v0.8.0 (Andrews 2011) and hard trimmed all reads in a set using FASTX-Toolkit's fastx_trimmer v0.0.13 (Gordon 2010) where the 25th percentile base quality score dropped below 20.

We aligned the sequencing reads to our unmasked *H. erato* reference genome using BWA v0.5.9-r16 (Li and Durbin 2009) with relaxed mapping parameters (Table S5). We assessed the quality and coverage of alignments using FlagStat and DepthOfCoverage from GATK v1.2-4 (McKenna et al. 2010, DePristo et al. 2011). We used Picard v1.53 (Broad Institute 2009) and GATK to refine the alignments by marking duplicate reads using Picard's MarkDuplicates and realigning around potential indels using GATK's RealignerTargetCreator and IndelRealigner.

We called multi-sample genotypes across samples for each race using GATK's UnifiedGenotyper with default parameters, except heterozygosity set to 0.025, and filtered genotype calls for quality using GATK's VariantFiltration, applying both site and individual sample filters (Table S5) to remove low confidence genotypes. If a site did not pass the site filtering criteria, we assigned all individuals of that race a genotype of N/N. If an individual's genotype did not pass the individual sample filtering criteria, we assigned that individual a genotype of N/N. Hypercoverage regions are indicative of repetitive elements, so based on the distribution of coverage per site for each individual, we empirically choose a hypercoverage threshold of 100x per sample.

We used the same pipeline and parameters for the *H. melpomene* Colombia data, aligning to the *H. melpomene* genome v1.1 (Heliconius Genome Consortium 2012). Additionally, we obtained unfiltered genotype calls for four individuals of each race from the *H. melpomene* hybrid zones in Peru (Nadeau et al. 2012). We filtered the genotypes using the same criteria above, with the exception of a hypercoverage cutoff of 150 due to the higher overall coverage of these samples.

Genotyping samples by aligning short sequence reads to a reference genome has inherent errors associated with it that result in incorrect genotypes. We introduced an additional source of error when we aligned whole genome reads to just a small portion of the genome. We estimated this additional error rate by aligning a single *H. timareta* sample to two *H. melpomene* reference sequences—the whole genome (v1.1) and a 2 Mb partial genome comprised of the color pattern regions. An *H. timareta* sample was used in the analysis because the amount of genetic diversity in *H. melpomene* is reduced relative to *H. erato* (Flanagan et al. 2004), while the amount of genetic diversity between *H. timareta* and *H. melpomene* is similar to that within *H. erato* (see Figure 3 in Beltrán et al. 2007). We aligned reads to the reference sequences using BWA and called genotypes with the GATK pipeline (Heliconius Genome Consortium 2012). We assumed that likely erroneous genotypes were ones that disagreed between the two methods or that were called for the reduced reference, but not the whole genome reference. We estimated the error rate as the number of likely erroneous genotypes divided by the total number of genotypes called from the partial genome alignment.

Table S5: Genotype calling parameters

BWA parameter	value	description
l	35	seed length
k	2	maximum edit in seed
n	8	maximum edits per 100 bp
o	2	maximum number of gap opens
e	3	maximum number of gap extensions
GATK parameter	**value**	**description**
heterozygosity	0.025	estimated heterozygosity
GATK filter	**value to filter out**	**description**
stand_call_conf	<30	standard minimum confidence threshold for the position, which equates to a probability of a misidentified segregating SNP of less than 0.001
DP	>100 * number of samples	hypercoverage per race
genotype GQ	<30	genotype quality for the sample, which equates to a probability of greater than 0.001 that the genotype called is incorrect
genotype DP	<10	low coverage per sample
genotype DP	>100	hypercoverage per sample
QD	<5.0	quality by depth
FS	>200	strand bias
HRun	>5	homopolymer run

S3.2 Sampling and Genotyping Results

The alignments of our *H. erato* Illumina reads to the partial genomic reference produced, on average, 75% properly paired reads—both pairs mapped in the correct orientation to each other and within the expected distance distribution. For each individual, on average, we called genotypes at 50% of the positions in our intervals overall and 56% of positions across the *D* interval (Table S6). Alignment of *H. melpomene* reads to the full *H. melpomene* reference genome produced, on average, 93% properly paired reads and 74% of genotypes called across the *B/D* interval per sample (Table S7).

Genotyping samples by aligning short sequence reads to a reference genome has inherent errors from a number of sources, including the sequencing error, alignment errors, and genotyping calling errors. These sources of error have been discussed elsewhere (Pool et al. 2010) and are affected by multiple factors, including depth of coverage. To determine the impact on error rate of mapping whole genome sequence data to only a partial genomic reference, we compared genotype calls of alignments of a single *H. timareta* individual to two different reference genomes—i) the entire *H. melpomene* reference genome (approximately 269 Mb) and ii) a 2 Mb portion of the *H. melpomene* reference genome (Table S7). This analysis suggests an additional 2.5% genotyping error rate is introduced when aligning whole genome reads to a partial genomic reference.

Table S6: Samples and sequencing data for *H. erato*

hybrid zone	race (phenotype)	sample	geolocation	number of paired end reads	mapped reads (%)	properly mapped pairs (% of mapped)	median coverage	positions genotyped (%)		SNPs* per genotyped position (%)	
								all reference	*D* interval	all reference	*D* interval
Peru	favorinus (postman)	GS012	06°27'41"S 76°20'31"W	69055119	8.5	75.5	20	45.7	52.2	4.6	4.5
		NCS0471	06°28'27"S 76°00'37"W	52625225	8.1	75.2	17	43.7	50.4	4.4	4.2
		NCS0473	06°28'27"S 76°00'37"W	49703856	8.3	75.3	17	42.4	48.7	4.4	4.3
		NCS0476	06°28'27"S 76°00'37"W	59869367	7.5	75.0	21	48.5	55.1	4.8	4.7
		NCS0478	06°28'27"S 76°00'37"W	70514138	8.3	74.3	24	50.6	57.3	5.0	4.9
		NCS0479	06°28'27"S 76°00'37"W	57097304	7.7	73.3	20	48.1	54.4	4.7	4.7
		NCS2554	06°27'41"S 76°20'31"W	51391495	7.5	75.1	18	46.6	52.6	4.6	4.4
		NCS2555	06°27'41"S 76°20'31"W	66232416	7.8	74.7	23	49.7	56.4	4.8	4.7
	emma (rayed)	GS020	06°10'55"S 76°14'50"W	62389463	8.4	75.7	18	44.4	50.3	4.4	4.4
		NCS1671	06°10'55"S 76°14'50"W	67708573	7.5	74.1	24	50.5	56.3	5.0	5.0
		NCS1672	06°10'55"S 76°14'50"W	60859534	7.8	73.9	21	49.0	54.8	4.9	4.8
		NCS1673	06°10'55"S 76°14'50"W	65914675	7.6	74.4	23	50.0	55.8	5.0	5.0
		NCS1674	06°10'55"S 76°14'50"W	60470606	7.9	74.2	22	49.3	55.5	5.0	4.9
		NCS1675	06°10'55"S 76°14'50"W	43527879	8.3	75.5	15	39.1	44.1	4.3	4.2
French Guiana	hydara (postman)	NCS1179	04°42'13"N 52°18'13"W	47188142	8.3	75.2	17	44.9	52.2	4.2	4.0
		NCS1979	04°34'18"N 52°13'24"W	53857631	8.4	75.4	19	48.0	55.7	4.4	4.2
		NCS2080	04°36'28"N 52°16'21"W	52696592	8.3	75.4	20	48.2	56.1	4.4	4.0
		NCS2211	04°32'50"N 52°10'13"W	61935440	8.2	74.8	22	51.2	59.0	4.6	4.2
		NCS2217	04°32'40"N 52°09'09"W	69615030	8.3	75.0	25	52.4	60.1	4.7	4.3
		NCS2581	04°47'48"N 52°19'28"W	87489610	8.0	76.0	29	53.2	61.0	4.8	4.5
		NCS2609	04°47'48"N 52°19'28"W	72210232	8.5	77.2	25	51.4	59.0	4.6	4.3

*SNPs are variation relative to the reference genome

Table S6 (cont.)

hybrid zone	race (phenotype)	sample	geolocation	number of paired end reads	mapped reads (%)	properly mapped pairs (% of mapped)	median coverage	positions genotyped (%)		SNPs* per genotyped position (%)	
								all reference	*D* interval	all reference	*D* interval
French Guiana (cont.)	erato (rayed)	NCS2005	04°38'19"N 52°18'06"W	64612926	8.6	75.1	21	48.8	54.9	4.6	4.5
		NCS2012	04°38'19"N 52°18'06"W	73271811	8.7	76.4	22	49.4	55.9	4.5	4.4
		NCS2020	04°35'06"N 52°14'44"W	97083432	8.0	75.8	32	53.8	60.2	4.9	4.8
		NCS2023	04°38'19"N 52°18'06"W	76874048	8.3	76.5	23	50.9	57.0	4.6	4.5
		NCS2025	04°35'06"N 52°14'44"W	64208302	8.6	76.0	20	48.1	54.2	4.5	4.4
		NCS2556	04°37'19"N 52°22'34"W	107961988	8.0	72.0	35	55.6	62.2	5.5	5.4
Ecuador	notabilis (postman)	BC0410	01°23'57"S 78°10'52"W	78516169	7.7	75.7	27	54.0	60.8	4.6	4.4
		NOT01	01°23'57"S 78°10'52"W	56434329	8.3	73.9	18	47.8	54.4	4.1	3.8
		NOT02	01°23'57"S 78°10'52"W	58901620	7.8	73.6	18	48.9	55.8	4.3	4.0
		NOT03	01°23'57"S 78°10'52"W	64868484	8.3	73.5	21	50.9	58.1	4.3	4.0
		NOT04	01°23'57"S 78°10'52"W	59804065	8.3	73.7	20	49.9	56.9	4.3	4.1
	lattivita (rayed)	BC0411	01°05'54"S 77°35'02"W	82234945	7.5	76.0	27	51.9	58.4	4.9	4.8
		LAT01	01°05'54"S 77°35'02"W	55007275	7.8	75.4	18	44.3	50.0	4.4	4.4
		LAT02	01°05'54"S 77°35'02"W	70156062	8.1	75.9	22	48.2	54.2	4.6	4.6
		LAT03	01°05'54"S 77°35'02"W	80058495	8.5	76.5	26	50.4	56.7	4.8	4.8
		LAT04	00°42'45"S 77°44'26"W	84018273	8.4	75.5	25	51.3	57.5	4.8	4.7
Panama	petiverana (postman)	ED3	09°07'46"N 79°42'55"W	79848146	8.5	77.3	30	55.6	60.1	3.6	3.9
		ED4	09°07'46"N 79°42'55"W	77997401	8.4	77.2	29	54.5	59.2	3.5	3.7
		ED5	09°07'46"N 79°42'55"W	60922100	8.9	76.8	23	50.8	55.6	3.4	3.6
		ED6	09°07'46"N 79°42'55"W	72039988	8.7	77.0	28	54.1	58.5	3.5	3.7
		STRI0033	09°09'09"N 78°41'23"W	50606981	9.9	67.4	21	52.5	56.8	4.5	3.8
	hydara (postman)	STRI0039	09°09'09"N 78°41'23"W	53723260	8.8	76.5	21	53.3	59.1	3.7	3.8
		STRI0040	09°09'09"N 78°41'23"W	54985081	8.7	76.6	22	54.9	60.3	3.7	3.7
		STRI0042	09°09'09"N 78°41'23"W	55879081	9.1	76.5	21	53.6	59.3	3.8	3.8

Table S7: Samples and sequencing for *H. melpomene* and *H. timareta*

hybrid zone	species (phenotype)	sample ID	geolocation	number of paired end reads	mapped reads (%)	properly mapped pairs (% of mapped)	median coverage	*B/D* positions genotyped (%)	*B/D* SNPs* per genotyped position (%)
Colombia	*H. melpomene melpomene* (postman)	HMCS25	4°12'48"N 73°47'70"W	42161297	79.5	93.9	22	74.6	1.8
		HMCS27	5°37'01"N 72°18'00"W	66272922	79.4	94.9	34	88.4	1.8
		STRI006	5°37'01"N 72°18'00"W	63418043	76.6	93.5	26	81.2	1.7
	H. melpomene malleti (rayed)	HMCS21	1°48'49"N 75°40'07"W	58085997	75.0	92.1	25	73.7	2.6
		HMCS22	1°36'35"N 75°40'01"W	52027258	75.1	91.9	23	68.1	2.5
		HMCS24	1°45'02"N 75°37'55"W	44144209	75.6	91.7	19	57.0	2.3
Peru	*H. melpomene amaryllis* (postman)	09-332	see Nadeau et al. 2012					78.6	2.5
		09-333	see Nadeau et al. 2012					79.0	2.5
		09-79	see Nadeau et al. 2012					79.6	2.5
		09-75	see Nadeau et al. 2012					75.3	2.4
	H. melpomene aglaope (rayed)	09-246	see Nadeau et al. 2012					68.2	2.6
		09-267	see Nadeau et al. 2012					76.8	2.8
		09-268	see Nadeau et al. 2012					76.4	2.8
		09-357	see Nadeau et al. 2012					73.1	2.6
	H. timareta (aligned full reference)	09-313	6°27'11"S 76°17'19"W	59607967	67.4	92.0	22	84.7	2.2
	(aligned partial reference)				4.9	71.7	28	76.2	2.3

*SNPs are variation relative to the reference genome

S4. Population Genetic Analyses between Divergent Races

S4.1 Population Genetic Methods

We used a number of population genetic analyses to identify putative functional regions. We examined signatures of selection, including increased genomic divergence between divergent color pattern races, and genotype by phenotype association to highlight regions showing patterns consistent with strong selection acting on functional variation.

Signatures of selection

We examined genomic divergence between pairs of *H. erato* color pattern races at each of four hybrid zone independently and across all three postman/rayed hybrid zones combined. To analyze each hybrid zone independently, we calculated sliding window population differentiation using a method that uses diploid data and models populations as random effects, to account for both statistical and genetic sampling processes ($\hat{\theta}$, Weir 1996). The model makes no simplifying assumptions regarding sample sizes or number of populations (Weir and Cockerham 1984). Calculations were done using a custom Perl script that implemented the Bio::PopGen::PopStats module from BioPerl (www.bioperl.org). To examine genomic divergence between the red phenotypes across the three postman/rayed *H. erato* hybrid zones combined, while accounting for the geographic structure of the populations, we estimated differentiation in a three-level hierarchy method ($\hat{\theta}_S$, Weir 1996). For level one, the populations, we examined the three hybrid zones that showed variation in the red phenotype—Peru, French Guiana, and Ecuador. For level two, the subpopulations, we examined the two color pattern races at each hybrid zone—the postman and the rayed. For level three, the individuals, we examined five to eight individuals per subpopulation. We calculated the sliding window subpopulation differentiation ($\hat{\theta}_S$) using a custom BioPerl module. For all comparisons, we calculated divergence at a position only if at least 75% of the individuals were genotyped for each phenotype. We evaluated 15 kb sliding windows at 5 kb steps across the genomic intervals and required a window to have divergence calculated for at least 20% of the positions in the window. We calculated a baseline level of divergence for each comparison as the level of divergence observed across intervals unlinked to color pattern (*H. erato*—three unlinked BACs, *H. melpomene*—38 unlinked scaffolds).

We calculated sliding window values for the proportion of segregating sites and heterozygosity to look for signatures of selection in *H. erato*. A segregating site was defined as having more than one allele in the population. The proportion of segregating sites was the total number of segregating sites per window divided by the total number of sites examined in the window. We obtained the proportion of heterozygotes by summing the number of heterozygote individuals at each position in the window and dividing that by the sum of the number of individuals genotyped at that position. We calculated baseline values from the three contigs unlinked to color pattern. We calculated estimates of these parameters for a genomic position only if at least 75% of individuals were genotyped and then examined 15 kb windows with a 5 kb step size. We required a window to have parameters estimated for at least 20% of the positions in the window.

Genotype by phenotype analyses

We estimated genotype by phenotype association at each *H. erato* hybrid zone independently, comparing the two color pattern phenotypes that occur in each hybrid zone. We also examined association with red phenotype across all four *H. erato* hybrid zones combined, by assigning all individuals to one of the two major red phenotypes—the postman or the rayed. We estimated association at each biallelic SNP using a two tailed Fisher's exact test, based on allele counts. Positions were excluded if less than 75% of individuals were genotyped for each phenotype.

Population genetic analyses in *H. melpomene*

We also assessed divergence and association in the *H. melpomene* hybrid zones in Peru and Colombia, which both consists of the two major red phenotypes—the postman and the rayed. We calculated sliding window subpopulation differentiation and genotype by phenotype association as described above. Additionally, we compared the positions of fixed SNPs between *H. erato* and *H. melpomene* to determine if any shared fixed SNPs existed. For each fixed SNP in *H. erato*, we attempted to identify an orthologous SNP in *H. melpomene* by manually inspecting the mVista LAGAN alignment between the reference sequences for the two species. If we were able to identify an orthologous SNP, we then compared the genotype calls for *H. erato* and *H. melpomene* individuals to determine if the SNP was associated with phenotype in both species.

S4.2 Population Genetic Results

See main text for population genetic results.

S5. Linkage Disequilibrium (LD) and Haplotype Structure

S5.1 LD and Haplotype Methods

We explored linkage disequilibrium (LD) and haplotype structure across the *D* interval, and regions unlinked to color pattern, in the Peruvian hybrid zone. We focused on a single hybrid zone because we wanted to remove the influence of geography and we chose Peru because it had the largest sample size. The data included all biallelic SNPs with at least 75% of individuals genotyped. We calculated correlations (r^2) between all pairwise SNPs using PLINK (Purcell et al. 2007), which for unphased data is based on genotype allele counts. To understand how LD decays with the distance between SNPs, we averaged the correlations for all pairwise SNPs from 100 bp bins of distance.

We estimated haplotypes from the Peruvian hybrid zone across a 100 kb window of the *D* interval (500-600 kb) containing the 65 kb peak of divergence and flanking regions using fastPHASE v1.2 (Scheet and Stephens 2006). We filtered bialleleic SNPs across this 100 kb region to remove sites that had genotypes from less than 75% of the individuals of each race, resulting in 3227 SNPs. Haplotypes were clustered during phase estimation into two clusters (K=2) and the proportion of rayed and postman individuals assigned to each cluster at each SNP was determined. We used Haploscope (San Lucas et al. 2012) to visualize regions where the two races had fixed haplotype block differences and where individuals from both races shared the same haplotypes. Using the haplotype estimations from fastPHASE, for each SNP Haploscope

visualizes the portion of individuals from a race (light vs. dark) assigned to each cluster (red vs. grey) across the 100 kb region (San Lucas et al. 2012).

S5.2 LD and Haplotype Results

See main text for LD and haplotype results.

S6. Phylogenetic Analyses of Evolutionary History

S6.1 Phylogenetic Methods

We constructed phylogenetic trees across sliding windows in the *D* interval, sampling 15 kb of sequence every 5 kb. For each window, we tested the log likelihood of the data with two alternative trees: the geographic tree assumes samples cluster by geographic hybrid zone and the color based tree groups races with a similar color pattern (rayed or postman) in a monophyletic clade (Figure 5). In each case, races are assumed to be monophyletic so that hypotheses of racial structure are equivalent. Neither geographic regions nor similarly colored races were resolved relative to one another, to avoid the influence of other topological hypotheses on the results. Likelihood values were calculated for each interval and tree topology using scripts in PAUP* 4b10 (Swofford 2002), using a GTR + G model inferred for the interval as a whole using Modeltest v3.7 (Posada and Crandall 1998). In addition to calculating likelihoods, we constructed neighbor-joining trees across these sliding windows in PAUP* to infer where in the interval lineages were monophyletic by color phenotype.

To summarize variation in phylogenetic topology across the interval we constrained division of the interval into the five most distinct topologies using the MDL method (Ané 2011). Default likelihood penalties for this method support a different topology for every block of 500 consecutive SNPs assessed. To divide the region more broadly, we raised the likelihood score penalty until five clusters of SNP blocks were reached. Tree topologies for each of these five regions of the interval were constructed using MrBayes v3.1.2 (Ronquist and Huelsenbeck 2003) run on CIPRES Science Gateway (Miller et al. 2010). Analyses involved 3 runs for 3 million generations each, sampling every 500 generations and removing 33% burn-in and runs that did not converge (as assessed in MrBayes and Tracer v1.5 (Rambaut and Drummond 2007)). Models were assigned using MrModeltest v2.3 (Nylander 2004) and included the GTR model for the 2nd and 3rd regions of the interval and GTR+G for the remaining regions.

In addition to these phylogenies, to infer a "best" tree of color pattern history, a phylogenetic tree was constructed for the 515-580 kb region across the peak of population differentiation. This tree was constructed under the same parameters in MrBayes (model = GTR) using only SNPs with low missing data, including at least 75% coverage of individuals within each red phenotype (1419 SNPs). A general phylogeny was also constructed across three genomic regions unlinked to color pattern using variable sites with at least 75% coverage of individuals under the same Bayesian methods (model = GTR + G, 3534 SNPs). For these "best" color-linked and color-unlinked datasets, we also reconstructed unrooted neighbor-net splits tree networks using SplitsTree v4.8 (Huson and Bryant 2006) and pairwise distances. Unlike most other phylogenetic analyses, which treat polymorphisms as ambiguities ("W" as "A or T"), in this analysis we were able to treat characters additively as "averages" ("W" as "A and T").

When treated as averages, sites where all individuals are heterozygotes are informative, thus more sites were retained as variable for this analysis (3440). Treating sites as averages should more accurately reflect the history of these characters which are by nature additive: two heterozygotes are more similar to each other than they are to homozygotes of either allele and in the additive model, heterozygotes are treated with 50% similarity to homozygotes, rather than as identical. These phylogenetic networks have the additional advantage of graphically representing areas and degrees of character conflict in phylogenetic construction, brought on through hybridization and recombination, ancestral sorting, or homoplasy.

To test whether shared color patterns between the mimics could result from a common origin, we also performed phylogenetic analyses combining *H. erato* and *H. melpomene* sequences along this interval. We focused on regions of high conservation between *H. erato* and *H. melpomene* (>80% conservation in a 500 bp window) from our mVista alignment. Across the 450 to 750 kb interval, we found 71 highly conserved regions that were relatively evenly distributed across the region and ranged in size from 430 to 3857 bp. For each conserved region, we used ClustalW2 (Larkin et al. 2007) to align sequences from all 45 *H. erato* individuals (Table S6) and 14 *H. melpomene* individuals (Table S7). We constructed neighbor-joining trees from pairwise distances of taxa in each these fragments and examined the resulting trees for species monophyly.

To infer a "best" *D* locus tree of *H. melpomene* and *H. erato* combined, we inferred the history in the peak of association from 515-580 kb in *H. erato* after further filtering SNPs from the regions of high conservation. This included first manually editing the alignments by removing regions of highly ambiguous alignment and correcting obvious misalignments. We then removed invariant sites and sites with more than 25% missing data. The resulting 1134 SNPs were concatenated and used for a Bayesian analysis using all the same parameters as listed above, including a GTR model inferred independently for this dataset in MrModeltest. We characterized SNPs across the interval by their patterns of fixation with respect to species and phenotype.

S6.2 Phylogenetic Results

To infer the optimal history of color pattern diversification in *H. erato* we constructed a Bayesian tree and a network-based tree of the 65 kb region that showed the strongest divergence and color pattern association. These trees support a single origin of the rayed color pattern, clustering rayed phenotypes separate from non-rayed phenotypes (Figure S6A). Trees based on SNPs from color-pattern unlinked regions clustered largely by hybrid zone (Figure S6B). Branch lengths across topologies show a signature of reduced gene flow, whereby color pattern alleles have reduced gene flow among races and less variation among individuals relative to markers unlinked to color pattern loci.

Comparing the history of this region between the two co-mimics, *H. erato* and *H. melpomene*, is difficult, as aligning non-coding regions is problematic; however, we were able to align regions of conservation between the two species. Of the 71 aligned fragments within the 300 KB window including the association peaks, 69 resulted in complete monophyly of *H. melpomene* with respect to *H. erato*. The remaining two trees had a few individuals admixing between the two in a manner unrelated to phenotype. Examination of the sequence files for

these fragments revealed problems with the automated alignment due to extensive missing data for these taxa.

We focused further analyses on the 65 KB region of highest association. After manual alignment and removal of sites with greater than 25% missing data in this narrowed region, 1164 SNPs were retained. Among these SNPs, 360 were fixed by species, 591 varied only in *H. erato*, 140 varied only in *H. melpomene*, and 73 shared allelic variation between the two species. Of the SNPs with alleles shared between the two species, none of the alleles that were fixed by phenotype within *H. melpomene* (n=18) or *H. erato* (n=1) had a signal that was associated with color phenotype in the opposite species. A phylogenetic analysis of the 1164 SNPs resolves *H. erato* and *H. melpomene* as separate lineages with high support, while resolving races by phenotype within each species (Figure S5). Results from these data thus support an independent origin of red patterns within each species.

References

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, and Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25: 3389–3402.

Andrews S. 2011. FastQC v0.8.0. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Ané C. 2011. Detecting phylogenetic breakpoints and discordance from genome-wide alignments for species tree reconstruction. *Genome Biology and Evolution* 3: 246–258.

Beltrán M, Jiggins CD, Brower AV, Bermingham E, and Mallet J. 2007. Do pollen feeding, pupal-mating and larval gregariousness have a single origin in *Heliconius* butterflies? inferences from multilocus DNA sequence data. *Biological Journal of the Linnean Society* 92: 221–239.

Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, et al. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research* 31: 365–370.

Broad Institute. 2009. Picard. http://picard.sourceforge.net.

Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Program NCS, Green ED, Sidow A, and Batzoglou S. 2003. LAGAN and Multi-LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA. *Genome Research* 13: 721–731.

Celniker SE, Dillon LA, Gerstein MB, Gunsalus KC, Henikoff S, Karpen GH, Kellis M, Lai EC, Lieb JD, MacAlpine DM, et al. 2009. Unlocking the secrets of the genome. *Nature* 459: 927–930.

Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, et al. 2009. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature Methods* 6: 677–681.

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, and Robles M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21: 3674–3676.

Counterman BA, Araujo-Perez F, Hines HM, Baxter SW, Morrison CM, Lindstrom DP, Papa R, Ferguson L, Joron M, ffrench Constant RH, et al. 2010. Genomic hotspots for adaptation: The population genetics of Müllerian mimicry in *Heliconius erato*. *PLoS Genetics* 6: e1000796. doi:10.1371/journal.pgen.1000796.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* 43: 491–498.

Duan J, Li R, Cheng D, Fan W, Zha X, Cheng T, Wu Y, Wang J, Mita K, Xiang Z, et al. 2010. SilkDB v2.0: a platform for silkworm (*Bombyx mori*) genome biology. *Nucleic Acids Research* 38: D453–D456.

Flanagan NS, Tobler A, Davison A, Pybus OG, Kapan DD, Planas S, Linares M, Heckel D, and McMillan WO. 2004. Historical demography of Müllerian mimicry in the neotropical *Heliconius* butterflies. *Proceedings of the National Academy of Sciences of the United States of America* **101**: 9704–9709.

Gordon A. 2010. FASTX Toolkit v0.0.13.1. http://hannonlab.cshl.edu/fastx_toolkit/index.html.

Heliconius Genome Consortium. 2012. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* **487**: 94–98.

Holt C and Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**: 491. doi:10.1186/1471–2105–12–491.

Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, Bernard T, Binns D, Bork P, Burge S, et al. 2012. InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Research* **40**: D306–D312.

Huson DH and Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* **23**: 254–267.

Illumina. 2008. *Preparing Samples for Sequencing of mRNA*. Illumina.

Korf I. 2004. Gene finding in novel genomes. *BMC Bioinformatics* **5**: 59. doi:10.1186/1471–2105–5–59.

Langmead B, Trapnell C, Pop M, and Salzberg S. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**: R25. doi:10.1186/gb–2009–10–3–r25.

Larkin M, Blackshields G, Brown N, Chenna R, McGettigan P, McWilliam H, Valentin F, Wallace I, Wilm A, Lopez R, et al. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* **23**: 2947–2948.

Li H and Durbin R. 2009. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* **25**: 1754–60.

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

Mallet J. 1986. Hybrid zones of *Heliconius* butterflies in Panama and the stability and movement of warning colour clines. *Heredity* **56**: 191–202.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**: 1297–1303.

McQuilton P, St Pierre SE, Thurmond J, and the FlyBase Consortium. 2012. FlyBase 101—the basics of navigating FlyBase. *Nucleic Acids Research* **40**: D706–D714.

Miller JR, Koren S, and Sutton G. 2010. Assembly algorithms for next-generation sequencing data. *Genomics* **95**: 315–327.

Munoz-Torres MC, Reese JT, and Sundaram JP. 2011. *Bee gene model annotation using Apollo*. Elsik Computational Genomics Laboratory.

Nadeau NJ, Whibley A, Jones RT, Davey JW, Dasmahapatra KK, Baxter SW, Quail MA, Joron M, ffrench Constant RH, Blaxter ML, et al. 2012. Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-scale targeted sequencing. *Philosophical Transactions of the Royal Society B: Biological Sciences* **367**: 343–353.

Nylander JAA. 2004. Mrmodeltest v2. Program distributed by the author.

Ostlund G, Schmitt T, Forslund K, Köstler T, Messina DN, Roopra S, Frings O, and Sonnhammer ELL. 2010. Inparanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Research* **38**: D196–D203.

Papanicolaou A, Stierli R, ffrench Constant R, and Heckel D. 2009. Next generation transcriptomes for next generation genomes using est2assembly. *BMC Bioinformatics* **10**: 447. doi:10.1186/1471–2105–10–447.

Pohl N, Sison-Mangus M, Yee E, Liswi S, and Briscoe A. 2009. Impact of duplicate gene copies on phylogenetic analysis and divergence time estimates in butterflies. *BMC Evolutionary Biology* **9**: 99. doi:10.1186/1471–2148–9–99.

Pool JE, Hellmann I, Jensen JD, and Nielsen R. 2010. Population genetic inference from genomic sequence variation. *Genome Research* **20**: 291–300.

Posada D and Crandall KA. 1998. MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**: 817–818.

Pruitt KD, Tatusova T, and Maglott DR. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research* **35**: D61–D65.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, et al. 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* **81**: 559 – 575.

Rambaut A and Drummond AJ. 2007. Tracer v1.4. http://beast.bio.ed.ac.uk/Tracer.

Reed RD, Papa R, Martin A, Hines HM, Counterman BA, Pardo-Diaz C, Jiggins CD, Chamberlain NL, Kronforst MR, Chen R, et al. 2011. optix drives the repeated convergent evolution of butterfly wing pattern mimicry. *Science* **333**: 1137–1141.

Ronquist F and Huelsenbeck JP. 2003. Mrbayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**: 1572–1574.

San Lucas FA, Rosenberg NA, and P S. 2012. Haploscope: a tool for the graphical display of haplotype structure in populations. *Genetic Epidemiology* **36**: 17–21.

Scheet P and Stephens M. 2006. A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *The American Journal of Human Genetics* **78**: 629 – 644.

Schug J and Overton GC. 1997. TESS: Transcription element search software on the WWW. In *Technical Report CBIL-TR-1997-1001-v0.0*. Computational Biology and Informatics Laboratory, School of Medicine, University of Pennsylvania.

Slater G and Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**: 31. doi:10.1186/1471–2105–6–31.

Smit AFA, Hubley R, and Green P. 2010. RepeatMasker Open-3.0. http://www.repeatmasker.org.

Stanke M, Keller O, Gunduz I, Hayes A, Waack S, and Morgenstern B. 2006. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research* **34**: W435–W439.

Suzek BE, Huang H, McGarvey P, Mazumder R, and Wu CH. 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23**: 1282–1288.

Swofford DL. 2002. *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods)*, 4th edition. Sinauer Associates, Sunderland, Massachusetts.

Trapnell C, Pachter L, and Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**: 1105–1111.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, and Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* **28**: 511–515.

Vergara IA and Chen N. 2009. Using OrthoCluster for the detection of synteny blocks among multiple genomes. *Current Protocols in Bioinformatics* pp. 1–18.

Weir BS. 1996. *Genetic Data Analysis II*. Sinauer Associates, Sunderland, Massachusetts.

Weir BS and Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. *Evolution* **38**: 1358–1370.

Supplementary Protocols—Supple et al. 2013

Supple MA, Hines HM, Dasmahapatra KK, Lewis JJ, Nielsen DM, Lavoie C, Ray DA, Salazar C, McMillan WO, and Counterman BA. 2013. Genomic architecture of adaptive color pattern divergence and convergence in *Heliconius* butterflies. *Genome Research*. doi: 10.1101/gr.150615.112

RNA-Seq Illumina Paired-End Sequencing Sample Preparation Protocol

This protocol is derived from two protocols. The steps from mRNA isolation to cDNA synthesis were modified from protocol RNASeq_v2.2.doc from the Yoav Gilad Lab. The subsequent steps were based on the protocol "Preparing Samples for Sequencing ChIP-seq DNA – Illumina GA II" from the HTSF UNC at Chapel Hill, Version 1.3 by Piotr Mieczkowski.

I. Total RNA isolation and QC

1. Store dissected tissues in RNAlater at -20°C.
2. Remove wings from RNAlater with forceps, removing as much liquid as possible, and weigh in a pre-weighed 1.5 mL tube.
3. Follow RNAqueous kit (Invitrogen) extraction protocol: the RNAqueous micro kit for 5th instar wings and the RNAqueous Kit for pupal wings.

LYSIS: Add the appropriate volume of lysis buffer given tissue weight. Homogenize tissue using handheld Tissueruptor for 10-15 seconds. Wash the tip of the homogenizer between samples by running it briefly through 5 washes: 1. water + soap, 2. ethanol + water, 3. distilled water, 4. RNase Zap (Ambion), 5. distilled water.

ELUTION: 40µL first wash (50µL for Day 3 or larger wings), 10µL second wash.

4. Run TURBO DNA-free (Invitrogen) protocol. 1µL TURBO DNAase for smaller samples (5th and Day 1) and 2µL for larger samples (>Day 3).
5. Measure RNA concentration (typically 200-1000 ng/µL) using Nanodrop and assess quality using Bioanalyzer.

II. mRNA isolation

Isolation with Dynabeads mRNA purification kit (Invitrogen) with adjustments to accommodate smaller RNA concentration. This kit should remove all RNA without a long poly-A region, thus isolating mostly mRNA from other types of RNA.

1. Dilute 10µg of total RNA with nuclease-free H_2O to 50µL in a 1.5 mL RNase-free non-stick tube.
2. Heat the sample at 65°C for 5 minutes and place the tube on ice.
3. Aliquot 100µL of Dynal oligo(dT) beads (Invitrogen) into a 1.5mL RNase-free non-stick tube.
4. Wash the beads twice with 100µL of Binding Buffer and remove the supernatant.
5. Resuspend the beads in 50µL of Binding Buffer, and add the 50µL of total RNA; rotate the tube at room temperature for 5 minutes, and remove the supernatant.
6. Wash the beads once with 100µL of Washing Buffer B and a second time in 30µL, remove supernatant.
7. Aliquoting 80µL of Binding Buffer to a new 1.5mL RNase-free non-sticky Eppendorf tube.
8. Add 20µL of 10mM Tris-HCl to the tube with beads and heat the beads at 80°C for 2 minutes to elute mRNA. Immediately put the beads on the magnet stand and transfer the supernatant (mRNA) to the tube from step 7.
9. Heat the mRNA sample from step 8 at 65°C for 5 minutes.

10. Wash the beads from step 8 twice with 30µL of Washing Buffer B, and remove the supernatant.
11. Add 100µL of mRNA sample from step 9 to the beads; rotate the tube at room temperature for 5 minutes.
12. Remove the supernatant and wash the beads once with 100µL of Washing Buffer B and a second time with 30µL.
13. Remove the supernatant from the beads, add 10µL of 10mM Tris-HCl and heat the beads at 80°C for 2 minutes to elute mRNA. Put the beads on the magnet stand and transfer the supernatant (mRNA) to a fresh 200µL thin wall PCR tube.

III. mRNA fragmentation

1. Add 1µL 10x Fragmentation Buffer (Ambion) to 9µL mRNA at 100ng.
2. Incubate at 70°C for 5 minutes.
3. Add 1µL of Stop Buffer included with the fragmentation buffer kit and put the tube on ice.
4. Transfer the solution to a 1.5 ml microcentrifuge tube. Add 1µL of 3M NaOAC, pH 5.2, 2µL of glycogen (5ug/µL), and 30µL of 100% EtOH. Incubate the tube at -80°C for 30 minutes.
5. Spin the tube at 14000 rpm for 25 minutes at 4°C in a microcentrifuge
6. Wash the pellet with 70% EtOH and air-dry the pellet (20 min).
7. Resuspend the RNA in 5.5µL of RNase free water.

IV. cDNA synthesis – 1st strand

1. In a 200ml thin wall PCR tube put 6µL random hexamer primer (500ng/µL, Promega) with 5.5µL RNA.
2. Incubate at 65°C for 5 min., then put tube on ice.
3. Mix for each sample:

5X first strand buffer (Invitrogen)	4 µL
100mM DTT (Invitrogen)	2 µL
dNTP mix 10mM	1 µL
RNase OUT (20U/uL) (Invitrogen)	1 µL

4. Add 7.5µL of the mix to each tube, vortex, and heat at 25°C for 2 min in thermocycler.
5. Add 1uL Superscript II (Invitrogen) to each sample then put in thermocycler at:

25°C	10min
42°C	50min
70°C	15min
4°C	HOLD

Put tubes on ice.

V. cDNA synthesis – 2nd strand

1. Add 61µL water to each sample.
2. To each sample add 10X second strand buffer (500mM Tris-HCl pH7.8, 50mM MgCl2, 10mM DTT) + 3µL dNTP mix 10mM.
3. Mix and incubate on ice for 5 min.
4. Add 1µL RNaseH 2U/µL (Invitrogen) and 5 uL DNAPol I 10U/µL (Invitrogen).
5. Mix and incubate at 16°C for 2.5 hrs.
6. Purify the DNA with Qiaquick PCR purification kit (Qiagen) and elute in 31µL elution buffer.

VI. End Repair

1. Prepare the following reaction mix, in order:

Eluted DNA	30µL
H_2O	10µL
T4 DNA ligase buffer with 10mM ATP (NEB)	5µL
dNTP mix (10mM)	2µL
T4 DNA polymerase (3U/µL) (NEB)	1.2µL
Klenow DNA polymerase (5U/µL) (NEB), diluted 1:4 with water	0.8µL
T4 Polynucleotide kinase (10U/µL) (NEB)	1µL

2. Mix well using pipettor.
3. Incubate the sample at 20°C for 30 min.
4. Purify with a QIAquick PCR spin column (Qiagen), and elute in 35µL of EB.

VII. Adding A's to DNA

1. Prepare the following reaction mix:

Eluted DNA	34µL
Klenow buffer (NEB buffer 2)	5µL
dATP (1 mM)	10µL
Klenow 3' to 5' exo- (5U/µL) (NEB)	1.2µL

2. Mix well using pipettor.
3. Incubate at 37°C in for 30 min.
4. Purify with a QIAquick MinElute column (Qiagen) and elute in 11µL of EB.

VIII. Ligating Adaptors

1. Prepare the following reaction mix (Total 30ul) :

Eluted DNA	10µL
H_2O	to 30 uL
2x Quick Ligation buffer (NEB, Quick Ligation Kit)	15µL
Adaptor oligo mix	Xµl
Quick DNA Ligase (NEB)	1.5µL

Adaptor: Add 1µL of 1:9 µL adaptor dilution for <100ng. Typically we had 150-200ng so added 2µL.

2. Mix well using pipettor.
3. Incubate the sample at room temperature for 15min.
4. Purify the DNA with QIAquick MinElute column (Qiagen) and elute in 10µL of EB.

IX. Purify Ligation Products

1. Prepare a 60mL gel – 2% with GenePure HiRes Agarose + 1X TAE. After cooling add 2.5uL EtBr to gel (10mg/ml).
2. Pour gel, add 3ul 6X orange loading dye to each 10uL sample.
3. Add 8 uL of 0'GeneRuler Ladder.
4. Run at 120V for 60 min.
5. Cut bands from 150-275 bp.
6. Perform DNA purification using Qiagen Gel Extraction Kit using 6X volume QG to 1X volume gel. Incubate at room temperature and add 2 gel volumes isopropanol after gel has dissolved.

X. PCR enrichment

1. Set up PCR master mix:

DNA	36ul
5 × Phusion Buffer HF	10µL
PCR primer 1.1	1µL
PCR primer 2.1	1µL
10 mM dNTP mix	1.5µL
Phusion polymerase	0.5µL
Total volume 50ul	

2. Run following PCR cycle:
   - 98°C 30 sec
   - 18 cycles of: 98°C 10 sec; 65°C 30 sec; 72°C 30 sec
   - 72°C 5 min
   - 4°C hold
3. Purify with QIAquick MinElute column (Qiagen) and elute in 14µL of EB.

Paired End DNA oligonucleotide sequences

PE Adapters

5' /Phos/-GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
5' ACACTCTTTCCCTACACGACGCTCTTCCGATCT

PE PCR Primers

5'AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
5'CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT

Whole Genome Illumina Paired-End Sequencing Sample Preparation Protocol

I. SHEAR GENOMIC DNA

In 100uL volume, shear 1ug of genomic DNA to 300-500bp using covaris machine.

- 10% Duty cycle
- Intensity → 4
- Cycle burst → 200

] 50sec

Make a 1.2% agarose gel to check fragment size.
Purify using 1.8X volume of **Agencourt AMPure beads**. Elute with 50uL EB.

II. XP BEADS DNA PURIFICATION

1. Add appropriate amount (1.0X or 1.8X) of AgentCourt AMPure XP beads to each DNA sample.
2. Vortex. Incubate at room temperature for 10 minutes on a rotator.
3. Quick spin. Place in magnetic separation stand.
4. After solution clears, carefully remove the supernatant without disturbing the pellet (be careful not to remove any beads).
5. Without disturbing the pellet, add 500uL (or enough to cover the beads) of freshly made 70% Ethanol. Wait for 2 or 3 minutes.
6. Carefully remove ethanol without disturbing the pellet or taking any beads.
   NOTE: Do not remove tubes from the magnetic separation stand while doing the last two steps.
7. Repeat 70% ethanol wash. Try to remove ethanol as much as possible in one pipetting.
8. Remove from magnetic separation stand and let dry on a thermomixer at 37°C for 2 or 3 minutes.
9. Elute the DNA from beads with EB buffer. The amount of EB buffer is dependent on your downstream reaction but usually more than 30uL.
10. Vortex. Allow DNA to elute for a few minutes.
11. Quick spin. Place in magnetic separation stand.
12. After the solution clears, collect supernatant in a new tube.

III. LARGE SCALE SOLEXA LIBRARY.

1. End repair

Reagents	uL
DNA (>500ng)	50
10X T4 DNA ligation buffer (NEB)	10
dNTP mix (10mM)	4
ATP (100mM)	1
T4 DNA Polymerase (3u/uL)	5
Klenow DNA Polymerase (5u/uL)	1
T4 PNK (10u/uL)	5
H_20 (NF)	24
TOTAL	**100**

Incubate at 20°C for 30 min. on a thermomixer.
Purify using 1.8X volume of Agencourt AMPure beads. Elute with 32uL EB.

2. 3' Adenylation (Poly A tail)

Reagents	uL
NEB buffer 2 (≈ Klenow buffer)	5
dATP (1mM)	10
DNA	32
Klenow 3'-5' Exo minus (5u/uL)	3
TOTAL	**50**

Incubate at 37°C for 30 min. on a thermomixer.
Purify using 1.8X volume of Agencourt AMPure beads. Elute with 35uL EB.

3. Ligation of adaptors to DNA fragments

Reagents	uL
DNA	35
Adaptor mix (15uM)	5
2X Quick ligase reaction buffer	45
Quick ligase	5
TOTAL	**90**

Incubate at room temperature for 30 min.
Purify using 1.0X volume of AgentCourt AMPure beads. Elute with 30uL EB.

4. Enrich the adaptor-modified DNA fragment by PCR

Here we use Phusion High-Fidelity DNA Polymerase kit by New England BioLab.

Reagents	uL
DNA	3
5X HF buffer	10
PCR primer 1.1 (25uM)	0.5
PCR primer 2.1 (25uM)	0.5
dNTP mix (10mM)	1
Phusion enzyme	0.5
H_20 (NF)	34.5
TOTAL	**50**

Cycle

- 98°C → 30 sec
- 8 cycles
  - 98°C → 10 sec
  - 65°C → 30 sec
  - 72°C → 30 sec
- 72°C → 5 min
- Hold at 4°

PCR primer 2.1 is different for each sample and must be added separately.
Purify using 1.0X volume of AgentCourt AMPure beads. Elute with 30uL EB.

Supplemental Figure S1: Comparison of population differentiation, association, and gene synteny between *H. erato* and *H. melpomene* in the Peruvian hybrid zone.
The top and bottom panels show sliding window (15 kb window size, 5 kb step size) population differentiation and SNPs perfectly associated with phenotype between postman and rayed phenotypes within *H. erato* and *H. melpomene*, respectively, for the Peruvian hybrid zone ($n_{erato\_postman}$=8, n_{erato_rayed}=6, $n_{melpomene\_postman}$=4, $n_{melpomene\_rayed}$=4) requiring a minimum of 75% of individuals genotyped for each phenotype at each position and data for at least 20% of positions in the window. The grey shaded regions highlight the 65 kb region of peak divergence across multiple *H. erato* hybrid zones (Figure 2). Baseline population differentiations of $\hat{\theta}_{erato}$=0.03 and $\hat{\theta}_{melpomene}$=0.02 were calculated from genomic intervals unlinked to color pattern. The middle panel shows gene synteny between the *H. erato D* interval genes (top row) and their orthologous *H. melpomene* genes (bottom row), with selected genes highlighted in red. The larger boxes represent coding exons and the narrower boxes represent introns. The synteny lines connect the beginnings and endings of orthologous genes.

Supplemental Figure S2: Annotation and analyses across the peak of divergence
Analysis across the region of peak divergence showing regions of high conservation between *H. erato* and *H. melpomene* (top panel, mVista alignment with 500 bp windows, red indicates >90% sequence identity), candidate SNPs perfectly associated with phenotype ("fixed SNPs"), probable repetitive regions based on hypercoverage of whole genome alignments ("hypercoverage"), regions with substantial missing data ("missing data", less than 20% of the positions in a 15 kb window have at least 75% of samples genotyped), and the number of predicted transcription factor (TF) binding sites (bottom panel, non-overlapping 100 bp windows across the 520-570 kb region).

Supplemental Figure S3: Decay of LD in *H. erato* across the 65 kb peak and color unlinked regions.
Correlation between SNPs as a function of distance between SNPs (colored symbols—65 kb peak of divergence; grey symbols—color unlinked regions). SNPs were binned into 100 bp windows based on the distance between SNPs. For each bin, the average correlation was calculated from all pairwise biallelic SNPs with at least 75% of individuals genotyped in the hybrid zone.

Supplemental Figure S4: *H. erato* genomic differentiation and association by hybrid zone.
Each panel indicates sliding window (15 kb window size, 5 kb step size) population differentiation and association values between two color pattern races for individuals from a single *H. erato* hybrid zone ($n_{peru\_postman}$=8, n_{peru_rayed}=6, $n_{ecuador\_postman}$=5, $n_{ecuador\_rayed}$=5, $n_{frenchguiana\_postman}$=7, $n_{frenchguiana\_rayed}$=6, $n_{panama\_postman1}$=5, $n_{panama\_postman2}$=3), requiring a minimum of 75% of individuals genotyped for each phenotype at each position and data for at least 20% of positions in the window. The green shaded area indicates the region of peak divergence across the *H. erato* hybrid zones combined (Figure 2). Baseline population differentiations of $\hat{\theta}_{peru}$=0.03, $\hat{\theta}_{ecuador}$=0.05, $\hat{\theta}_{frenchguiana}$=0.02, and $\hat{\theta}_{panama}$=0.01 were calculated from BACs unlinked to color pattern. The races from the hybrid zones in French Guiana, Ecuador, and Peru show phenotypic

variation in the red color pattern elements, while the races from the Panamanian hybrid zone do not.

Supplemental Figure S5: Phylogenetic relationships in the 65 kb peak of divergence and color unlinked regions.
Phylogenetic relationships among individuals from A) the region of strongest association in the *D* interval (515-580 kb) and B) from BACs unlinked to color pattern. Trees were constructed from datasets filtered to remove SNPs with more than 75% missing data and include neighbor-net phylogenetic networks (left) and Bayesian tree topologies (right). Terminal nodes in the networks and sample names in the Bayesian trees are color-coded respective to race. Dashed lines in the networks indicate rayed vs. non-rayed patterns for the *D* interval and hybrid zones for the unlinked data. Red branches in the Bayesian tree indicate rayed lineages. Specimen names follow a naming system including voucher number, hybrid zone, and race. Nodal values are posterior probabilities of clade support. Unrooted Bayesian trees were arbitrarily rooted for presentation.

Supplemental Figure S6: Phylogeny of *H. erato* and *H. melpomene* across the 65 kb peak of divergence.
Bayesian phylogeny using reliably aligned SNPs across the 65 kb region of highest divergence and association. Lineages with rayed patterns are colored in red. Posterior probability support values are displayed for major lineages. Degree of support for all lineages is indicated by branch thickness, with thicker branches for nodes supported by a posterior probability >90.

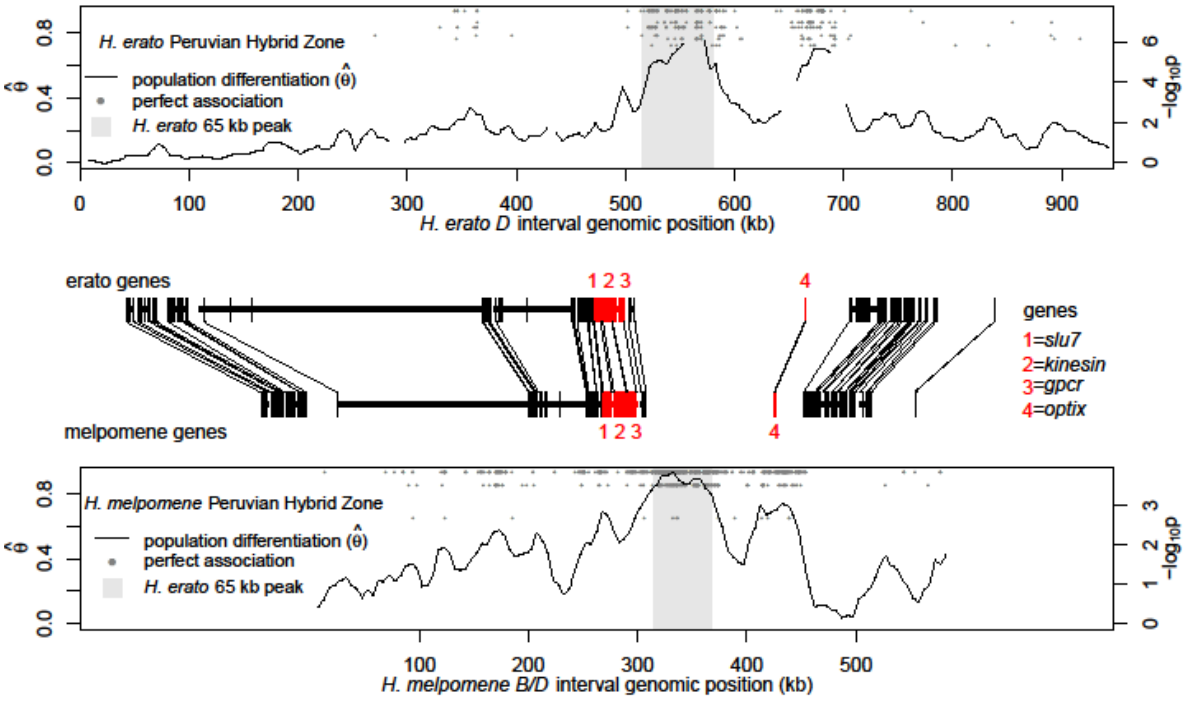


Figure S1

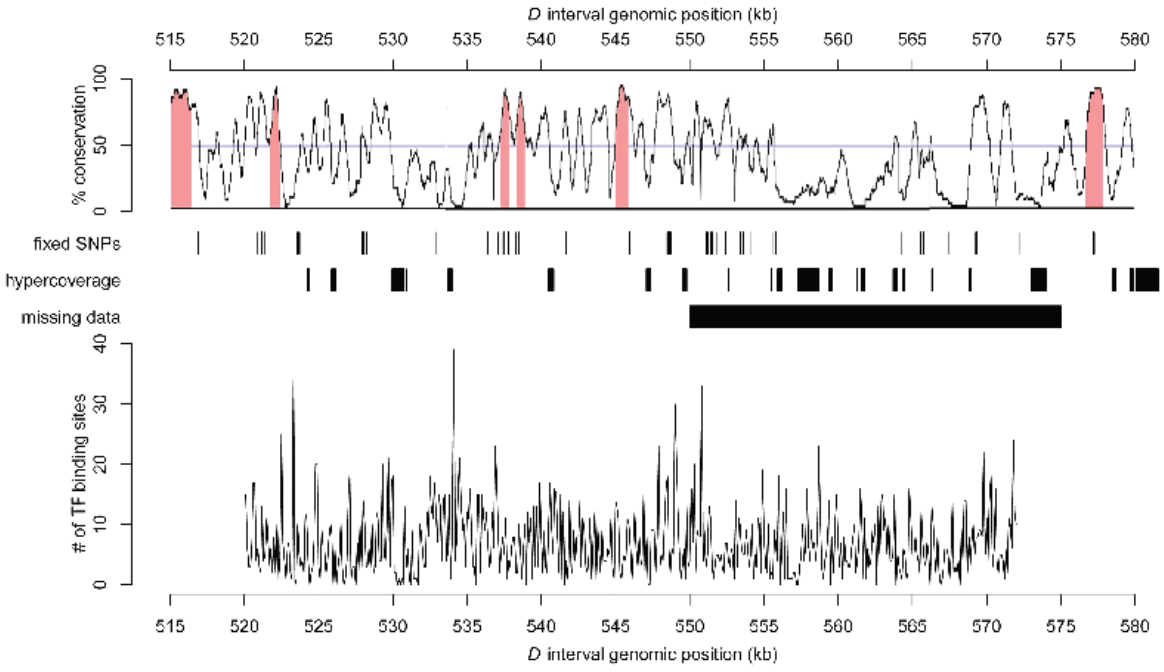


Figure S2

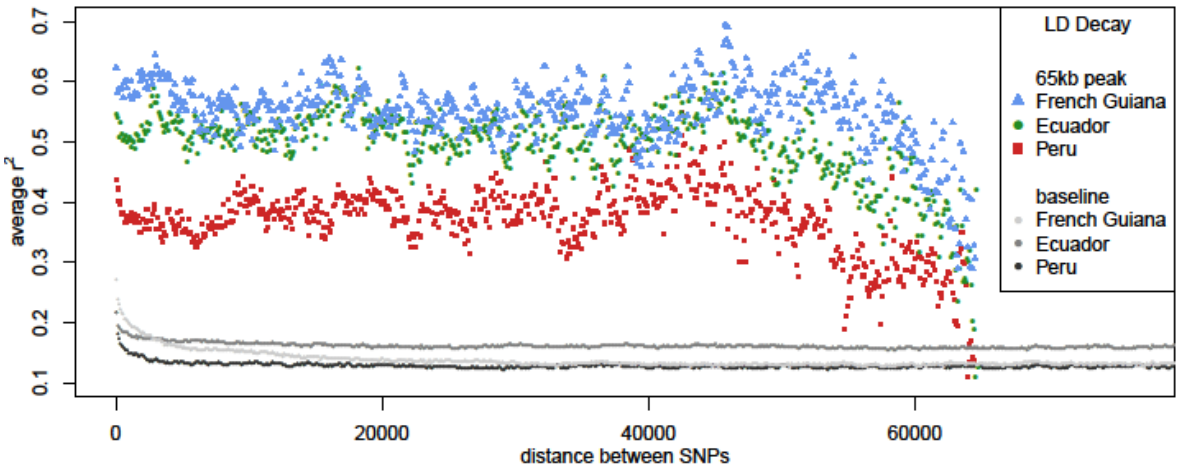

LD Decay
65kb peak
French Guiana
Ecuador
Peru
baseline
French Guiana
Ecuador
Peru
average r2
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
20000
40000
60000
distance between SNPs

Figure S3

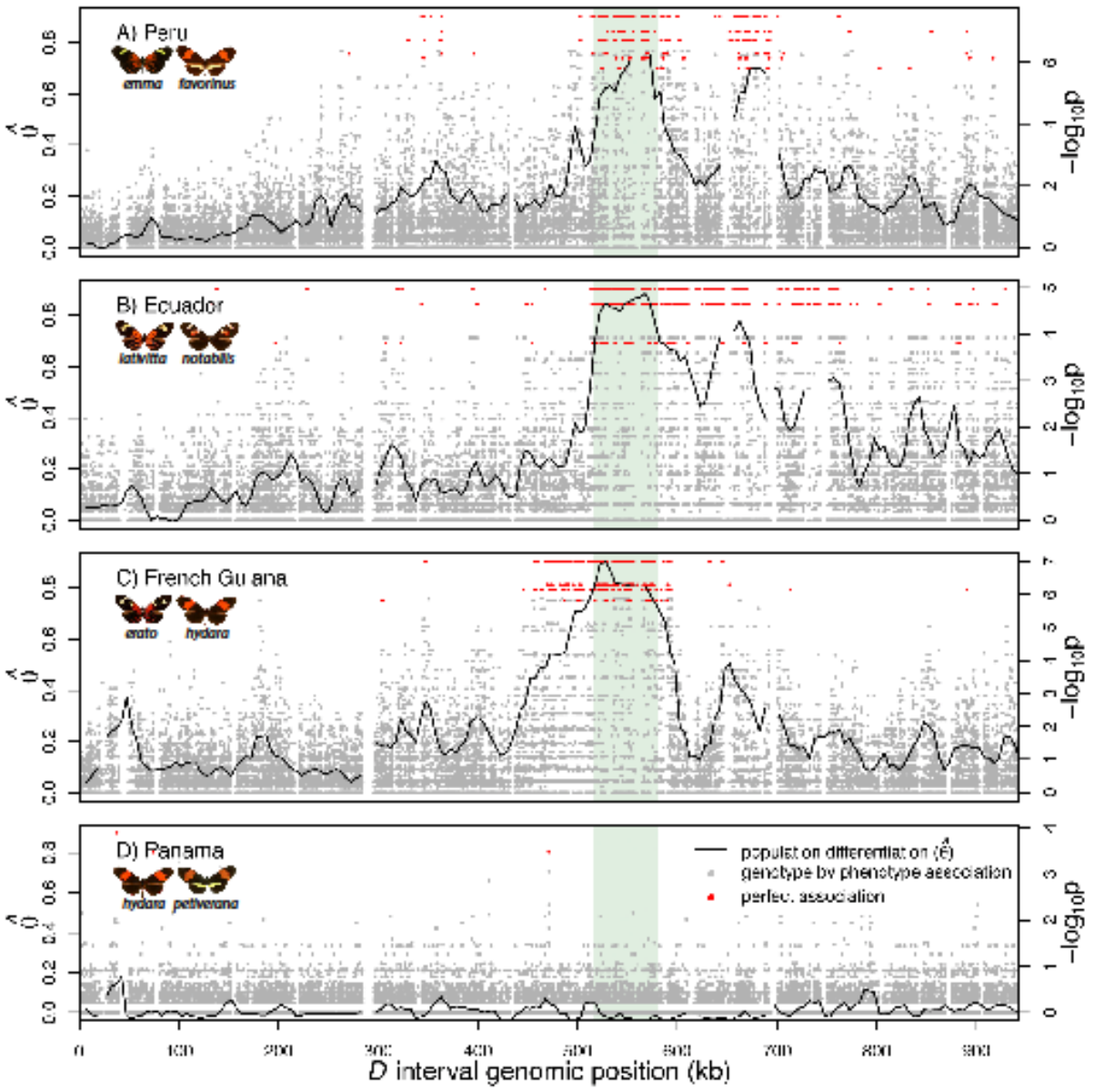

A) Peru
emma
favorinus
B) Ecuador
lativitta
notabilis
C) French Guiana
erato
hydara
D) Panama
hydara
petiverana
population differentiation ($\hat{F}_{ST}$)
genotype by phenotype association
perfect association
$-\log_{10}p$
0
100
200
300
400
500
600
700
800
900
D interval genomic position (kb)

Figure S4

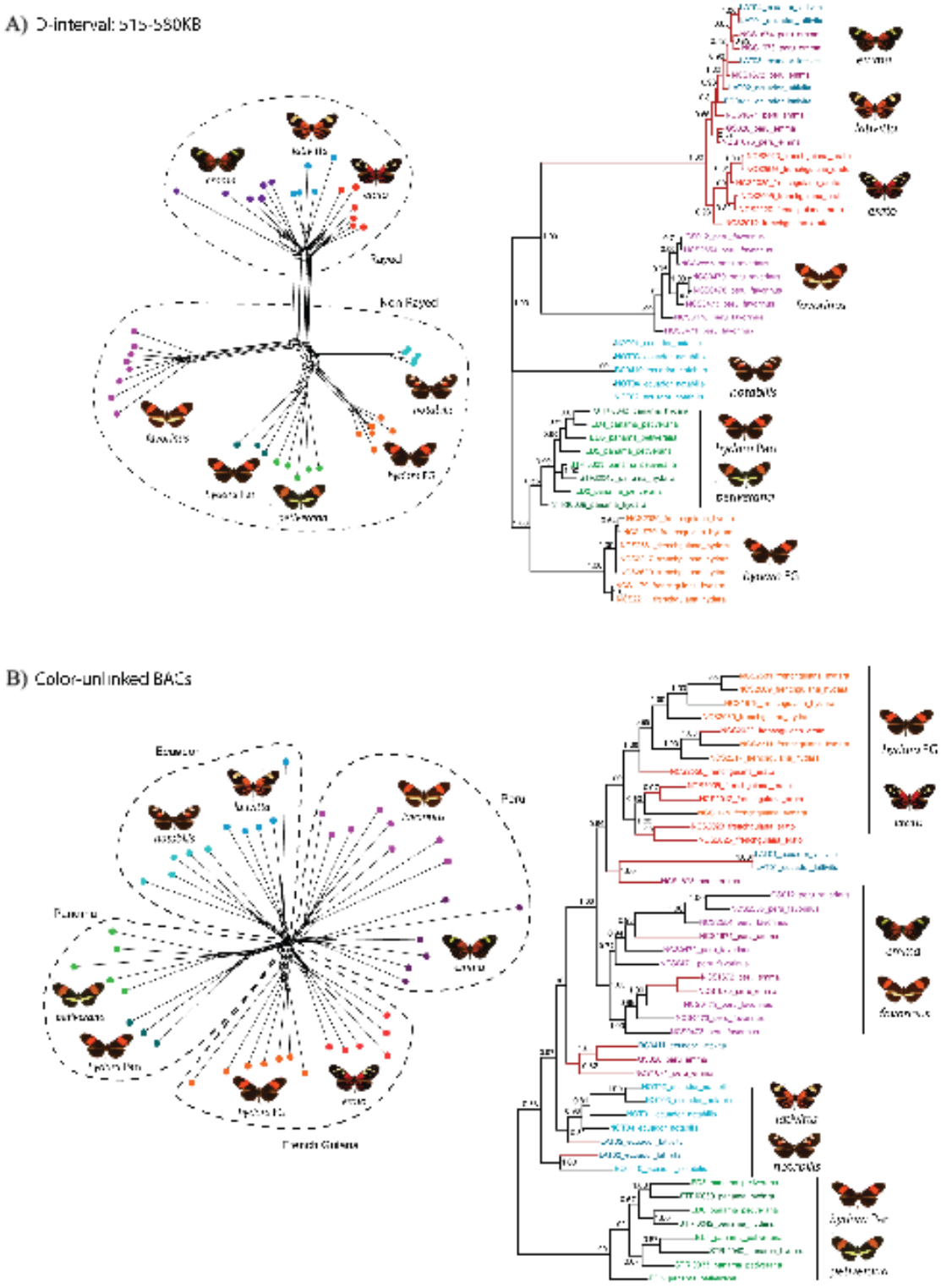


Figure S5

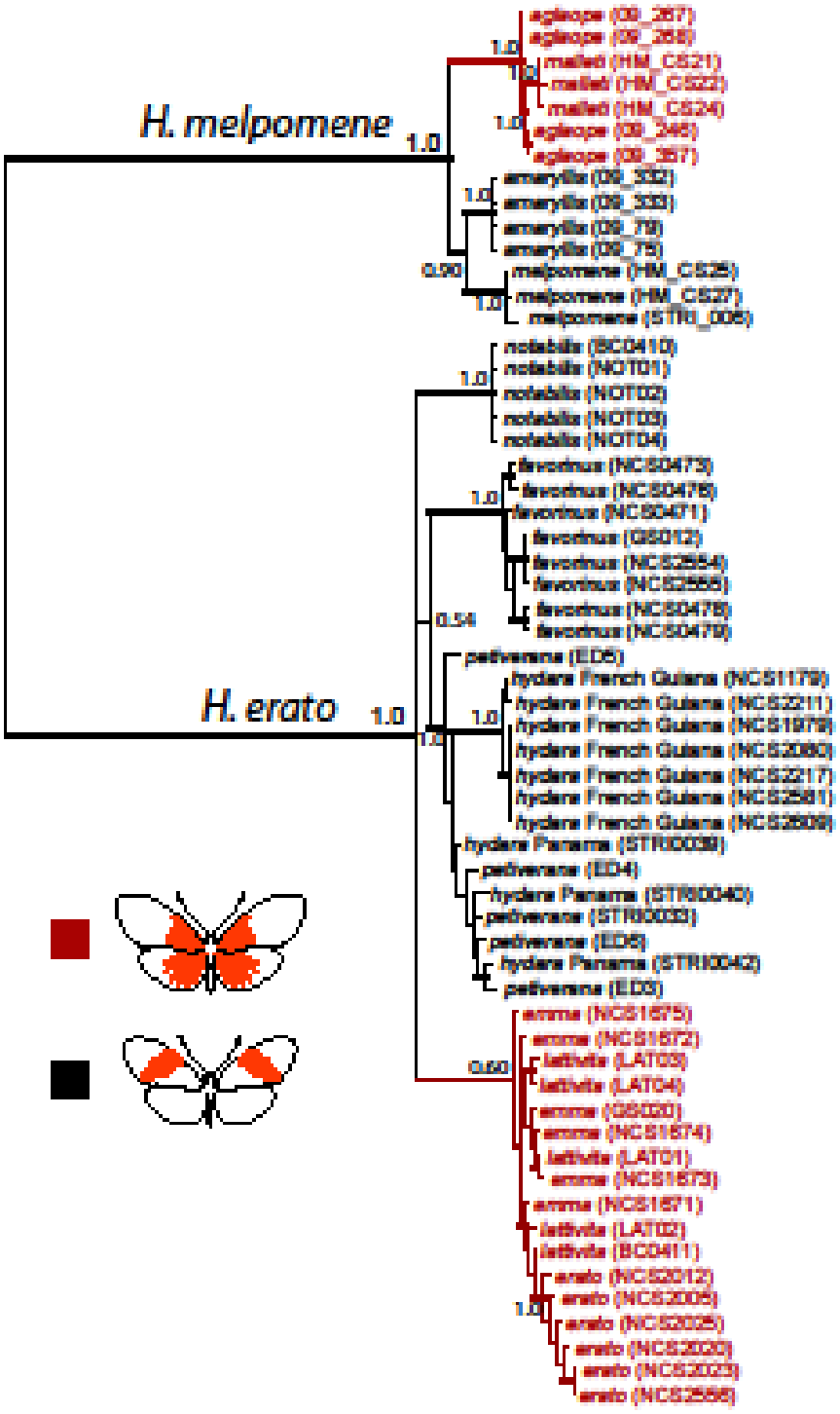

H. melpomene
1.0
H. erato
1.0
1.0
0.90
0.54
0.60
aglaope (09_267)
aglaope (09_268)
malleti (HM_CS21)
malleti (HM_CS22)
malleti (HM_CS24)
aglaope (09_246)
aglaope (09_357)
amaryllis (09_332)
amaryllis (09_333)
amaryllis (09_79)
amaryllis (09_75)
melpomene (HM_CS25)
melpomene (HM_CS27)
melpomene (STRI_006)
notabilis (BCD410)
notabilis (NOT01)
notabilis (NOT02)
notabilis (NOT03)
notabilis (NOT04)
favorinus (NCS0473)
favorinus (NCS0476)
favorinus (NCS0471)
favorinus (GS012)
favorinus (NCS2554)
favorinus (NCS2555)
favorinus (NCS0478)
favorinus (NCS0479)
petiverana (ED5)
hydara French Guiana (NCS1179)
hydara French Guiana (NCS2211)
hydara French Guiana (NCS1979)
hydara French Guiana (NCS2080)
hydara French Guiana (NCS2217)
hydara French Guiana (NCS2581)
hydara French Guiana (NCS2609)
hydara Panama (STRI0039)
petiverana (ED4)
hydara Panama (STRI0040)
petiverana (STRI0033)
petiverana (ED6)
hydara Panama (STRI0042)
petiverana (ED3)
emma (NCS1575)
emma (NCS1572)
lativitta (LAT03)
lativitta (LAT04)
emma (GS020)
emma (NCS1574)
lativitta (LAT01)
emma (NCS1573)
emma (NCS1571)
lativitta (LAT02)
lativitta (BCD411)
erato (NCS2012)
erato (NCS2005)
erato (NCS2025)
erato (NCS2020)
erato (NCS2023)
erato (NCS2556)

Figure S6

CHAPTER 3

Rapid phenotypic evolution through modular enhancer shuffling

In preparation for submission:
Supple MA, Counterman BA, McMillan WO, and Hines HM. Rapid phenotypic evolution through modular enhancer shuffling.

Abstract

Identifying the mechanisms that generate phenotypic variation is key to understanding the origins of biodiversity. Cis-regulatory evolution has been proposed as a main driver of the tremendous amount of phenotypic variation found in nature. The extensive diversity of wing color patterns across the adaptive radiation of *Heliconius* butterflies provides a natural laboratory to explore the origins of novel phenotypes. Using whole genome sequencing across the spectrum of natural variation in *Heliconius*, we identify putative modular enhancer regions controlling red color pattern variation. Each of these enhancers drives the expression of the gene *optix* in a spatially specific pattern across the wing. Within the enhancer controlling the presence or absence of hindwing rays, we use a combination of population genomic and transcription factor binding analyses to identify candidate transcription factors. These candidates regulate *optix* by binding to the enhancer region, with differential binding driving phenotypic diversity. We further demonstrate that novel phenotypes evolved through the shuffling of these enhancer regions. Our results suggest that cis-regulatory evolution has played a key role in generating the diversity of wing color patterns in *Heliconius*.

Introduction

Phenotypic variation abounds in nature, yet we still know little about the mechanisms by which this extraordinary diversity arises and proliferates. Among the evolutionary processes that can generate phenotypic variation, there is a growing body of evidence that morphological diversity is often associated with changes in gene expression, particularly

through modifications of enhancers in cis-regulatory regions (Carroll 2000, 2008; Wray 2007; Wittkopp and Kalay 2012). Enhancers consist of non-coding DNA, including transcription factor binding sites, that act to coordinate the fine scale temporal and spatial regulation of genes. A single gene can have multiple enhancers that each control the transcription of the gene in a different tissue or during a different developmental stage. New enhancers with novel expression domains, and hence novel functions, can evolve without affecting existing expression patterns. This modularity in gene regulation can minimize negative downstream pleiotropic effects (see Blackwood and Kadonaga 1998; Carroll 2000; Wittkopp and Kalay 2012 for reviews).

Despite the compelling logic, we actually know relatively little about how changes in regulatory regions drive morphological variation. There are just a handful of studies, mostly in *Drosophila* and other model organisms, which have started to pinpoint the molecular changes in enhancers that underlie morphological variation. The potentially modular nature of cis-regulation has been demonstrated in studies of melanism in *Drosophila*. Reporter assays have revealed an array of cis-regulatory elements of pigmentation genes that drive gene expression in different regions of the body and enable species-specific patterns of pigmentation (Gompel et al. 2005; Rebeiz et al. 2009; Kalay and Wittkopp 2010). The modularity of these enhancers was elegantly demonstrated for the *ebony* gene, where functional substitutions in an abdominal specific module did not result in changes in gene expression in other spatial domains (Rebeiz et al. 2009).

These early studies are also uniting the theoretical prediction that evolution proceeds by many, small effect changes with the empirical evidence suggesting fewer, large effect

changes (Orr 2005). Evidence is emerging that you can essentially assemble a large effect allele from multiple small effect mutations. This aggregation has been demonstrated in loci regulating cuticle pigmentation and trichome variation in *Drosophlia* (Rebeiz et al. 2009; Bickel et al. 2011; Frankel et al. 2011) and coat color in mice (Linnen et al. 2013). Overall, these studies are providing the crucial insights into how enhancer evolution can generate phenotypic diversity. The universality of such cis-regulatory modules as critical material for phenotypic evolution, however, awaits the further dissection of the genetic basis of phenotypic variation across a diverse set of systems.

Here we explore the evolution of novel phenotypic variation in *Heliconius*, a group of Neotropical butterflies renowned for their extraordinary diversity of wing color patterns (Turner 1975). Specifically, we assess the modularity of the evolution of red pattern elements within the *H. erato* radiation. The extensive geographic range of *H. erato* is composed of a mosaic of over 25 different color pattern races. Many of these races differ with respect to three distinct red color pattern elements on the forewing and hindwing (Sheppard et al. 1985). Previous mapping and expression studies showed that a single gene, the transcription factor *optix*, controls red pattern variation in *H. erato* and its distantly related co-mimic, *H. melpomene* (Reed et al. 2011). Its differential expression and the conserved amino acid sequence of *optix* across the genus lead to the hypothesis that it is regulation of *optix* expression that ultimately drives variation in red phenotypic elements (Hines et al. 2011; Reed et al. 2011).

More recent population genomic analyses have narrowed the location of this cis-regulatory region to a 65-kb genomic region about 100-kb down stream of *optix*. This region

shows both strong genotype by phenotype association and high genomic divergence between divergent red phenotypes (Supple et al. 2013). We believe that this 65-kb region harbors key cis-regulatory elements that drive variation in the spatial expression patterns of *optix* across the wing surface. Our proposed model for how *optix* might regulate color pattern variation is one where different *cis*-regulatory elements of *optix* read the inputs of various aspects of a conserved pre-pattern laid down during wing development (Figure 1). The pre-pattern, often called the nymphalid groundplan, is a set of conserved wing development elements that occur in spatially specific regions of the wing (Nijhout 2001). Butterflies can build their diverse wing color patterns using this pre-pattern when pigmentation genes acquire the ability to respond to these elements. We hypothesize that *optix* has acquired novel binding sites for a select group of these elements, resulting in a spatially specific pattern of *optix* across the wing. *optix* then coordinates the expression of other downstream scale-cell maturation and pigment pathways. In this respect, *optix* acts as a classic "input-output" gene (see Stern and Orgogozo 2009).

We leverage whole genome sequence data from across the spectrum of natural phenotypic variants in the *H. erato* color pattern radiation and closely related species to define the functional regions driving differences in the spatial expression of *optix* and to explore the evolutionary histories of the adaptive alleles. We compare individuals of rayed or postman phenotypes across multiple hybrid zones examined in previous studies (Supple et al. 2013, *in prep*) and to this we add additional phenotypic sampling, including individuals from races and species that are phenotypic recombinants of the two major color patterns (Figure 2). This phenotypic sampling enables us to identify distinct enhancers in the 65-kb

regulatory region that likely drive variation of each phenotypic element. Within the putative enhancer controlling hindwing rays, we combine population genomic analyses, transcription factor binding site prediction, and gene expression profiles to generate strong candidates for upstream regulation of *optix*. Finally, we show how novel phenotypic variation can be created by shuffling existing enhancer alleles. Collectively our results highlight how rapid phenotypic variation can be driven by modification and shuffling of multiple enhancers of a single gene.

Methods

Sampling and sequencing

We examined sequence variation of 19 taxa (n=69) across the broader *Heliconius erato*/*sara*/*sapho* clade, representing both convergent and divergent phenotypes (Figure 2). We used previously published data for 12 taxa (n=58) (Supple et al. 2013, *in prep*) and we collected and sequenced 7 new taxa (n=11) (Table 1). Our overall dataset consisted of multiple races of *H. erato*, representing the traditional postman phenotype (6 races, n=32) and the traditional rayed phenotype (4 races, n=18). We added distantly related species within the clade with the traditional postman phenotype (1 species, n=2) and the traditional rayed phenotype (1 species, n=1). We sampled *H. e. amalfreda* (n=5), which like the rayed phenotype has a red patch in the basal part of the forewing and a yellow forewing band, but like the postman phenotype lacks red hindwing rays. We sampled, an incipient species from the *H. erato* radiation, *H. himera* (n=5), which like the rayed phenotype has red on the

hindwing, but like the postman phenotype lacks red in proximal part of the forewing. We also sampled additional species with color patterns similar to *H. himera* (3 races, n=4). Finally, we sampled distantly related species within the clade with a yellow forewing band and no red color elements (2 species, n=2). Samples were sequenced to at least 30x coverage using whole genome sequencing of 100-bp paired end sequencing reads on an Illumina HiSeq platform and genotyped as described in Supple et al. (2013).

Identifying modular enhancers

We performed multiple population genomic comparisons to begin to localize functionally important regions within the 65-kb block of strong divergence between postman and rayed *H. erato* races. We examined genotype by phenotype association of sequences derived from alignment to a partial reference sequence as described in Supple et al. (2013). Briefly, this involved aligning sequencing reads to a partial genomic reference sequence using BWA (Li and Durbin 2009) with relaxed mapping parameters and calling genotypes with GATK (McKenna et al. 2010). We calculated a per position genotype by phenotype association using a two-tailed Fisher's exact test based on allele counts and identified SNPs showing perfect genotype by phenotype association. We filtered out positions if less than 75% of individuals were genotyped for each phenotype. We identified putative enhancer regions by looking for regions of high genotype by phenotype association, including the presence of perfectly associated SNPs. To ensure we kept potentially important flanking sequences, these regions were extended on either side to include the next called SNP that was identified as fixed in the initial analysis of *H. erato* hybrid zones (Supple et al. 2013).

We identified the functional region modulating the presence of hindwing rays by examining association between three races of *H. erato* from neighboring populations: *H. e. hydara* (postman), *H. e. erato* (rayed), and *H. e. amalfreda* (recombinant). *H. e. amalfreda* has a yellow forewing band and red dennis patch that is characteristic of the rayed phenotype, but similar to the postman phenotype it lacks the hindwing rays. We examined genotype by phenotype association of rayed samples (n=6) versus non-rayed samples (n=12). Additionally, to identify a narrower, high priority region within this area, we examined genotype by phenotype association across the broader *erato*/*sara*/*sapho* clade (n_{ray}=19, $n_{noHWred}$=41).

Similarily, we identified the functional region modulating the presence or absence of the red forewing basal "dennis" patch by examining association across the *erato* clade. In addition to the classic postman (no dennis patch) and rayed (dennis patch) in *H. erato,* this extended sampling included 3 species, including *H. himera,* that have a yellow forewing band, a red hindwing bar, but lack the dennis patch. We examined genotype by phenotype association between dennis samples (n=23) and non-dennis samples (n=42). We did an additional association analysis across the broader *erato*/*sara*/*sapho* clade (n_{dennis}=24, $n_{no\text{-}dennis}$=45).

Finally, we localized the functional region modulating the color of the forewing band by extending our analysis across the *erato*/*sara*/*sapho* clade. This extended sampling includes additional species that have a yellow forewing band and no red color elements. The yellow forewing band appears to be the ancestral phenotype and it is shared with the

otherwise derived rayed phenotype. We examined genotype by phenotype association between the yellow forewing band (n=35) and the red forewing band (n=34).

Identifying functional regions in the rays enhancer

To more finely dissect functional regions within the putative rays enhancer, we examined variation across the enhancer region in five races. Two races are postman phenotypes (*H. e. hydara* from French Guiana and *H. e. favorinus* from Peru), two races are rayed phenotypes (*H. e. erato* from French Guiana and *H. e. emma* from Peru), and the fifth race is a recombinant phenotype (*H. e. amalfreda* from Suriname) that has two of the rayed phenotypic elements, but lacks the hindwing rays. Genomic regions where *H. e. amalfreda* alleles are similar to the other non-rayed races and divergent from the rayed races are candidate regions controlling the hindwing rays.

The high level of variation between races of *H. erato*, in particular insertions and deletions, results in substantial missing data when aligning sequencing reads to our *H. e. petiverana* reference sequence from Panama. To improve sequence coverage within the putative rays enhancer region, we generated race-specific reference sequences by tiling PCR sequences for a single individual of each race across the region of interest. The race-specific references were aligned to each other, as well as the *H. e. petiverana* genomic reference for that region. Completely ambiguous positions (Ns) in each race-specific reference were filled in with the consensus sequence from all of the race-specific references. Each race-specific reference was exported as two fasta files, one including gap positions and one excluding

gaps. The non-N IUPAC ambiguous base codes in the references were replaced with a random representative base using a custom PERL script.

To capture within race nucleotide variation, we aligned the whole genome sequencing reads for each race to its ungapped race-specific reference using BWA v0.5.9-r16 (Li and Durbin 2009) with relaxed mapping parameters. We disabled insert size estimation and Smith-Waterman alignment due to the short reference length with multiple repetitive regions present. We called genotypes on each race as described in Supple et al. (2013). We used a custom PERL script to generate an aligned, multi-sample fasta file from the vcf of genotype calls. The fasta entries for each race were then aligned to each other by aligning to the race-specific gapped reference sequence using ClustalW v2.1 (Larkin et al. 2007). Examining all the races combined, we assessed genotype by phenotype association, as described above and including gaps as a called genotype.

Using both the association analyses and manual examination of aligned genotype calls, we identified regions potentially modulating phenotypic variation. Candidate regions were areas where genotypic variation (SNPs, insertions, and deletions) sorted by phenotype. Regions with substantial amounts of variation between races that did not sort by phenotype were ruled out. Given the large number of insertions in postman races, we only considered large regions with high conservation across all postman phenotypes. We considered all insertions, large and small, consistent across the rayed races.

For each region identified, we predicted transcription factor binding sites for each phenotype. Two methods were used to predict binding sites, both using the JASPAR CORE-insecta matrix profiles for DNA-binding sites (Mathelier et al. 2013). The first method used

the JASPAR server with default setting, except a relative profile score threshold of 95% (Mathelier et al. 2013). The second method used the RSAT server (van Helden 2003; Turatsinze et al. 2008). A background model was generated using RSAT's oligo-analysis of the entire 65kb peak region, with a markov order of 5. RSAT's matrix-scan was run using this background model and filtering for p-values less than 3e-3 and scores greater than 4. From each analysis, we generated a list of transcription factors that showed differential binding between the two phenotypes. To reduce the high rate of false positives in the prediction of transcription factor binding sites, we only considered transcription factors that were identified in both analyses. We then manually compared candidate binding sites to the JASPAR matrix to ensure that high conservation sites from the matrix were present in the candidate site. From this we generated a list of candidate transcription factors that are potentially involved in phenotypic variation.

For each candidate transcription factor, the protein sequence was downloaded from NCBI's GenBank. To determine if these transcription factors were expressed on the *H. erato* wing, these protein sequences were blasted, using tblastn, to two *H. erato* transcriptome databases. One transcriptome was assembled from Sanger/454 data (Papanicolaou et al. 2009) and one was assembled from existing RNA-seq reads from whole hindwings (Supple et al., 2013, assembly unpublished). We determined possible functions and expression patterns across *Heliconius* wings based on known functions and patterns in *Drosophila*.

For the four main candidate transcription factors identified using the methods above (*ct, Deaf1, vvl, ara*), we examined the distribution of binding sites across the rayed module for four races—two postman and two rayed. We excluded *H. e. favorinus* from this analysis

due to excessive missing data across the PCR reference sequence. For each race, we used Geneious v7.1.3 (Biomatters 2014) to generate a strict consensus sequence (only ATCGN) from the original PCR sequence and the sequences for each sample generated from the alignment of the raw sequencing reads. We generated predicted binding sites using RSAT with a p-value cutoff of 3e-3 and estimated the background model from the input sequence, except for binding sites of less than six basepairs we used the background model described above. We also generated predictions from the JASPAR server with a relative profile score threshold of 95%. We retained predicted binding sites that were present in both analyses. For each retained binding site, we determined how it associated with phenotype, in particular if it was present in both rayed races and no postman race, present in both postman races and no rayed races, or present in all four races.

We also examined regions enriched for predicted binding sites of the primary candidate transcription factors. We used RSAT's cis-regulatory element enriched regions analysis for each candidate separately. We estimated the background model from the input sequence, except for binding sites of less than six basepairs we used the background model described above. We used a site p-value=3e-3, region size=100-1000, significance=1, and a region p-value=0.05. For overlapping predictions, we choose the prediction with the lowest p-value. We identified enriched regions present in all races and perfectly associated with phenotype.

Phylogenetic analyses

We used phylogenetic analyses to examine the history of *H. e. amalfreda and H. himera* relative to the larger *H. erato* radiation. First, we examined where *H. e. amalfreda* and *H. himera* fell within the *H. erato* radiation across the 65-kb regulatory region by generating non-overlapping neighbor joining trees for 5-kb windows from all 19 taxa (n=69) using PAUP* (Swofford 2002). We then used a reduced dataset to test the log likelihood of the data under alternative trees. Each tree had four *H. erato* races (*hydara, erato, favorinus, emma)* plus the race or species of interest. The five taxa were assumed to be monophyletic and all samples within the five race/species were unresolved relative to each other. For each comparison, the two trees only differed in their placement of the taxon of interest. For *H. e. amalfreda*, we tested whether it clustered with the rayed or the non-rayed samples. For *H. himera*, we tested whether it clustered with the dennis or the non-dennis samples. In both cases, we selected a nucleotide substitution model and estimated model parameters for the 65-kb region as a whole using a hierarchical likelihood ratio test implemented in PAUP* v4 (Swofford 2002) and MrModeltest v2 (Nylander 2004). We examined sliding windows across the region, with 5-kb windows and 1-kb slide. For each window, we used PAUP* LScores to determine the negative log likelihood of the data under each of the two hypothesis trees.

Results

Identifying modular enhancers

Our entire genomic dataset consisted of 69 individuals representing 9 species, including 11 different geographic races of *H. erato* (Figure 2). Through examination of the genotype by phenotype association based on the alignments of resequencing data to the *H. erato petiverana* BAC reference, we identified regions with genetic associations that suggested that they contain the functional regions modulating the three color pattern elements (Figure 3).

For the rayed region, we analyzed three parapatric races (n=18), which are known to hybridize in nature, and identified just 12 SNPs that were perfectly associated with the presence or absence of the hindwing rays. Of these 12 sites, 11 are within a single 12-kb region (Figure 3A) and the remaining site was located just downstream of the gene *optix*. Additionally, two of these SNPs that are 4 bp apart remained perfectly associated when the analysis was extended to include all taxa across the broader radiation.

For the dennis region, analysis of 15 taxa, including *H. erato* and *H. himera*, identified 10 SNPs perfectly associated with the presence or absence of the forewing dennis patch (Figure 3B). Of these 10 SNPs, 7 are located in a 7-kb region. The other 3 are within the broader 65-kb regulatory region. Based on these additional SNPs we have highlighted two additional regions for further examination. The first additional region is 3.5-kb and surrounds the first SNP outside of the 7-kb core region. It is the only SNP that remained perfectly associated across the broader radiation. The second region is 1-kb and surrounds the second SNP outside the core region. The third SNP outside the core region was

discarded because the association is largely being driven by a general postman versus rayed comparison due to too much missing data in the most informative races.

For the forewing color region, we examined 19 taxa across the broader *Heliconius erato*/*sara*/*sapho* clade. The association identified 4 SNPs in a 5-kb region perfectly associated with the color of the forewing band (Figure 3C). Three of these SNPs clustered in in a 34-bp window.

Identifying functional regions within the rays enhancer

We examined variation across the 12-kb region associated with the presence of hindwing rays by aligning the resequencing reads to the race specific references we generated though a series of tiled PCR reactions. Across this region, we identified 3 regions with good sequence across all five races with SNP variation sorting by phenotype and predicted differential binding affinities between the phenotypes. We identified 2 regions with fixed deletions in the rayed races, relative to the postman races, that showed consistent binding sites across the postman races. We identified 3 insertions in the rayed races that showed consistent binding sites across the rayed races (Figure 4, top).

From these regions, 19 candidate transcription factors were identified (Table 2). Of these, 5 primary candidates were identified based on the higher quality of the binding sites and segregating genomic variation (Figure S1). One of these candidates was eliminated because it was not present in either *H. erato* transcriptome database and it has no known function in *Drosophila*. The remaining 4 candidates are all expressed in the hindwing of *H. erato* and are known to function in wing development in *Drosophila*. The candidate

transcription factors are *cut* (*ct*), *Deformed epidermal autoregulatory factor 1* (*Deaf1*), *ventral veins lacking* (*vvl*), and *araucan* (*ara*) (Figure S1).

We examined the distribution of binding sites for the candidate transcription factors (Figure 4, bottom). Three candidates (*vvl, ara, Deaf1*) had regions with a significantly larger than expected number of binding sites (cis-regulatory enriched regions, CRERs) that segregated with phenotype. The fourth candidate (*ct*) had a region of enrichment that was present in both phenotypes, but a nearby binding site showed differential binding affinity between the two phenotypes and contained a perfectly associated SNP.

Evolutionary history of recombinant phenotypes

We examined phylogenetic trees across the 65-kb regulatory region to understand the evolutionary history of *H. e. amalfreda*, which has a phenotype that has color pattern elements from both the traditional postman and rayed phenotypes. These analyses reveal that across most of the regulatory region, *H. e. amalfreda* is more closely related to the rayed races than to the postman races (Figure 5). There is a single region where *H. e. amalfreda* is more closely related to the postman races. This region is about 10-15-kb and corresponds to the ray enhancer region we have identified. The neighbor joining 5-kb trees show *H. e. amalfreda* clustering with the rayed races, except for two adjoining 5-kb windows where *H. e. amalfreda* switches to cluster with the postman races. The likelihood analysis favors *H. e. amalfreda* clustering with the rayed across most of the region, except a 16-kb region, where the alternative tree, with *H. e. amalfreda* clustering with the postman, is favored (Figure 5).

This suggests that *H. e. amalfreda*'s recombinant phenotype is a result of genomic recombination between postman and rayed alleles.

Phylogenetic analyses involving the recombinant phenotype *H. himera* show that *H. himera* is associated with the rayed taxa across much of the regulatory region (Figure 6). There is one primary region where *H. himera* clusters with the postman races. This region corresponds to our primary dennis enhancer region predicted from genotype by phenotype association. Both *H. himera* and the postman phenotypes lack the dennis patch, therefore this adds additional support to our identification of the dennis enhancer region. There were two additional dennis regions we identified through our population genomic analyses, each region based on a single perfectly associated SNP. This phylogenetic analysis show that *H. himera* is more associated with the rayed phenotype in these regions, making these regions less likely candidates for the dennis enhancer (Figure 6).

Discussion

Leveraging natural phenotypic variation for functional identification

Using naturally occurring phenotypic variation, we were able to narrow down the functional genomic regions that modulate phenotypic variation. Previously, using two major phenotypic classes within *H. erato*, we identified a 65-kb regulatory region. Here, using additional phenotypic variation, including recombinant phenotypes across the broader *H. erato* radiation, we further characterized functional variation within the regulatory region. We identified putative enhancer regions modulating three distinct phenotypic elements, as

well as identifying potentially functional transcription factor binding sites with differential predicted binding affinities that is associated with phenotypic variation. The ability to identify functional genomic variation, particularly in non-coding regulatory regions, from naturally occurring variation is key to promoting the continued transition to adaptive genomics in non-model organisms. Until recently, these types of studies were restricted to classic model organisms with extensive genomic resources. We can now begin to understand the complex connection between genotype and phenotype in a vast array of traits and organisms.

Candidate pre-patterning genes

We identified five primary candidate transcription factors that likely produce a pre-pattern across the wing that interacts with *optix*, resulting in a complex spatial pattern of red across the wing. These candidates were selected based on genomic analyses showing binding sites with phenotypically differential binding affinities that were in genomic regions with variation that perfectly associates with phenotype. These analyses were completely independent of functional and expression data, but a number of additional factors add support to the selected candidates.

Four of the top five candidates modulating the presence of hindwing rays are known to be expressed in *Heliconius* hindwings (Wallbank et al. *in prep*; Papanicolaou et al. 2009; Hines et al. 2012; Supple et al. 2013). Three have expression patterns in *Drosophila* wings that make them excellent candidates for generating hindwing rays. The *H. erato* hindwing rays occur between wing veins and do not extend to the wing margin. The candidates

araucan (*ara*) and *ventral veins lacking* (*vvl*) are expressed along wing veins and provide positional information for developing wing veins (de Celis et al. 1995; Gómez-Skarmeta et al. 1996; de Celis 2003). The candidate *cut* (*ct*) is expressed at the wing margins. The spatial expression of *Deaf-1* across the pupal wing is unknown, but its overexpression in imaginal discs can disrupt eye and wing development (Veraksa et al. 2002).

In addition to expression patterns, the four top genes are known to interact directly with genes that interact with *optix* (de Celis et al. 1995; Gómez-Skarmeta and Modolell 1996; Gross and McGinnis 1996; Wang and Sun 2012). *optix* is the gene at the center of red color pattern variation across *Heliconius*, yet even in *Drosophila* relatively little is known about its function and interactions. A handful of genes have been identified as interacting with *optix* in *Drosophila*, making them, and the genes that interact with them, excellent candidates for our upstream genes. Of particular interest is *ct* because it plays a crucial role in determination of cells fated to become the eye, which is the same developmental pathway that that *optix* is known to be a part of (Wang and Sun 2012). The genomic analysis, together with known functions and patterns of expression, make these strong candidate pre-patterning genes for determining the spatial variation of red across the wings of *Heliconius*.

The genomic architecture of cis-regulation

Cis-regulatory variation has the ability to modulate fine scale gene expression in space and time, providing a powerful mechanism for generating phenotypic variation. In particular, the ability of distinct enhancer elements to control expression in different domains enables a wide array of variation, even with a limited toolkit of genes. The wing color patterns of

Heliconius butterflies are made up of a number of distinct color pattern elements. Variation in red color pattern in *H. erato* is comprised of three elements, which are all controlled by variation in a 65-kb regulatory region modulating the expression of a single gene, *optix* (Reed et al. 2011; Supple et al. 2013). Using natural phenotypic variation we have identified three distinct regions of the 65-kb regulatory region, each controlling a different color pattern element. As new genomic techniques are enabling fine scale dissection of functional variation, this modularity of cis-regulation is emerging as a major theme enabling phenotypic variation.

In addition to modularity, another key aspect of the architecture of cis-regulation that is emerging is that large effect loci can often be broken down into multiple, small effect changes. This is enabling resolution between conflicting results—the theoretical results, which show evolution proceeds by many small effect changes, and the empirical results, which suggests that fewer large effect changes are the norm. We show that the major shift in red color pattern is controlled by a single locus that can be broken down into distinct enhancer regions, each controlling one element of red color pattern. Further, the size of the identified region and the fact that we identified multiple candidate regions within the enhancer suggest that the enhancer itself is composed of multiple changes, each potentially having a small effect on phenotype.

Evolution of novel phenotypes through enhancer shuffling

The regulatory regions we have identified are made up of multiple changes held together to create single alleles of major effect. We have no evidence that genomic inversions are

maintaining the cohesion of the alleles, as has been shown in other major effect alleles controlling wing color patterns in butterflies (Joron et al. 2011; Kunte et al. 2014). Rather, it seems that strong selection acting on the allele as a whole, through the resulting phenotype, is causing multiple changes to aggregate into a single, large effect allele.

This regulatory architecture consisting of modular enhancers held together by selection has the potential to promote rapid phenotypic evolution through the shuffling of these enhancers in the event that selection is relaxed. Our phylogenetic analyses suggest that enhancer shuffling has generated novel phenotypes in *Heliconius*. We show that two taxa that are a phenotypic recombinants between the traditional postman and rayed phenotypes are in fact genotypic recombinants. Each taxa generated its novel phenotype using an assortment of enhancers from the divergent postman and rayed taxa. Overall, our study suggests that the genetic architecture of cis-regulation enables rapid phenotypic evolution.

References

Bickel, R. D., A. Kopp, and S. V Nuzhdin. 2011. Composite effects of polymorphisms near multiple regulatory elements create a major-effect QTL. PLoS Genet. 7:e1001275.

Biomatters. 2014. Geneious.

Blackwood, E. M., and J. T. Kadonaga. 1998. Going the Distance: A Current View of Enhancer Action. Science 281:60–63.

Carroll, S. 2000. Endless forms: the evolution of gene regulation and morphological diversity. Cell 101:577–80.

Carroll, S. B. 2008. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell 134:25–36.

De Celis, J. F. 2003. Pattern formation in the *Drosophila* wing: The development of the veins. BioEssays 25:443–51.

De Celis, J. F., M. Llimargas, and J. Casanova. 1995. *ventral veinless*, the gene encoding the Cf1a transcription factor, links positional information and cell differentiation during embryonic and imaginal development in *Drosophila melanogaster*. Development 121:3405–16.

Frankel, N., D. F. Erezyilmaz, A. P. McGregor, S. Wang, F. Payre, and D. L. Stern. 2011. Morphological evolution caused by many subtle-effect substitutions in regulatory DNA. Nature 474:598–603.

Gómez-Skarmeta, J. L., R. Diez del Corral, E. de la Calle-Mustienes, D. Ferré-Marcó, and J. Modolell. 1996. *araucan* and *caupolican*, two members of the novel iroquois complex, encode homeoproteins that control proneural and vein-forming genes. Cell 85:95–105.

Gómez-Skarmeta, J. L., and J. Modolell. 1996. *araucan* and *caupolican* provide a link between compartment subdivisions and patterning of sensory organs and veins in the *Drosophila* wing. Genes Dev. 10:2935–2945.

Gompel, N., B. Prud'homme, P. J. Wittkopp, V. a Kassner, and S. B. Carroll. 2005. Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in *Drosophila*. Nature 433:481–7.

Gross, C., and W. McGinnis. 1996. DEAF-1, a novel protein that binds an essential region in a Deformed response element. EMBO J. 15:1961–1970.

Hines, H. M., B. A. Counterman, R. Papa, P. Albuquerque de Moura, M. Z. Cardoso, M. Linares, J. Mallet, R. D. Reed, C. D. Jiggins, M. R. Kronforst, and W. O. McMillan. 2011. Wing patterning gene redefines the mimetic history of *Heliconius* butterflies. Proc. Natl. Acad. Sci. U. S. A. 108:19666–19671.

Hines, H. M., R. Papa, M. Ruiz, A. Papanicolaou, C. Wang, H. F. Nijhout, W. O. McMillan, and R. D. Reed. 2012. Transcriptome analysis reveals novel patterning and pigmentation genes underlying *Heliconius* butterfly wing pattern variation.

Joron, M., L. Frezal, R. T. Jones, N. L. Chamberlain, S. F. Lee, C. R. Haag, A. Whibley, M. Becuwe, S. W. Baxter, L. Ferguson, P. A. Wilkinson, C. Salazar, C. Davidson, R. Clark, M. A. Quail, H. Beasley, R. Glithero, C. Lloyd, S. Sims, M. C. Jones, J. Rogers, C. D. Jiggins, and R. H. ffrench-Constant. 2011. Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. Nature 477:203–6. Nature Publishing Group.

Kalay, G., and P. J. Wittkopp. 2010. Nomadic enhancers: tissue-specific *cis*-regulatory elements of *yellow* have divergent genomic positions among *Drosophila* species. PLoS Genet. 6:e1001222.

Kunte, K., W. Zhang, A. Tenger-Trolander, D. H. Palmer, A. Martin, R. D. Reed, S. P. Mullen, and M. R. Kronforst. 2014. *doublesex* is a mimicry supergene. Nature 507:229–32. Nature Publishing Group.

Larkin, M., G. Blackshields, N. Brown, R. Chenna, P. McGettigan, H. McWilliam, F. Valentin, I. Wallace, A. Wilm, R. Lopez, J. Thompson, T. Gibson, and D. Higgins. 2007. ClustalW and ClustalX version 2. Bioinformatics 23:2947–2948.

Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760.

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

Mathelier, A., X. Zhao, A. W. Zhang, F. Parcy, R. Worsley-Hunt, D. J. Arenillas, S. Buchman, C. Chen, A. Chou, H. Ienasescu, J. Lim, C. Shyr, G. Tan, M. Zhou, B. Lenhard, A. Sandelin, and W. W. Wasserman. 2013. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Res. 42:D142–7.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. 2010. The Genome

Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297–1303.

Nijhout, H. F. 2001. Elements of butterfly wing patterns. J. Exp. Zool. 291:213–25.

Nylander, J. A. A. 2004. MrModeltest v2. Evolutionary Biology Centre, Uppsala University.

Orr, H. A. 2005. The genetic theory of adaptation: a brief history. Nat. Rev. Genet. 6:119–27.

Papanicolaou, A., R. Stierli, R. H. Ffrench-Constant, and D. G. Heckel. 2009. Next generation transcriptomes for next generation genomes using est2assembly. BMC Bioinformatics 10:447.

Rebeiz, M., J. E. Pool, V. a Kassner, C. F. Aquadro, and S. B. Carroll. 2009. Stepwise modification of a modular enhancer underlies adaptation in a *Drosophila* population. Science 326:1663–7.

Reed, R. D., R. Papa, A. Martin, H. M. Hines, B. A. Counterman, C. Pardo-Diaz, C. D. Jiggins, N. L. Chamberlain, M. R. Kronforst, R. Chen, G. Halder, H. F. Nijhout, and W. O. McMillan. 2011. *optix* drives the repeated convergent evolution of butterfly wing pattern mimicry. Science 333:1137–41.

Sheppard, P. M., J. R. G. Turner, K. S. Brown, W. W. Benson, and M. C. Singer. 1985. Genetics and the evolution of Mullerian mimicry in *Heliconius* butterflies. Philos. Trans. R. Soc. London. Ser. B, Biol. 308:433–610.

Stern, D., and V. Orgogozo. 2009. Is genetic evolution predictable? Science 323:746–751.

Supple, M. A., H. M. Hines, K. K. Dasmahapatra, J. J. Lewis, D. M. Nielsen, C. Lavoie, D. A. Ray, C. Salazar, W. O. McMillan, and B. A. Counterman. 2013. Genomic architecture of adaptive color pattern divergence and convergence in *Heliconius* butterflies. Genome Res. 23:1248–1257.

Supple, M. A., R. Papa, H. M. Hines, W. O. McMillan, and B. A. Counterman. n.d. Selection drives genomic divergence during speciation in Heliconius butterfliese.

Swofford, D. 2002. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, MA.

Turatsinze, J.-V., M. Thomas-Chollier, M. Defrance, and J. van Helden. 2008. Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. Nat. Protoc. 3:1578–88.

Turner, J. R. G. 1975. A Tale of Two Butterflies. Nat. Hist. 84:24–37.

Van Helden, J. 2003. Regulatory Sequence Analysis Tools. Nucleic Acids Res. 31:3593–3596.

Veraksa, A., J. Kennison, and W. McGinnis. 2002. DEAF-1 function is essential for the early embryonic development of *Drosophila*. Genesis 33:67–76.

Wallbank, R. W., S. W. Baxter, C. Pardo-Diaz, J. Hanly, S. Martin, J. Mallet, K. Dasmahapatra, M. Joron, N. Nadeau, O. McMillan, and C. D. Jiggins. n.d. The origins of an evolutionary novelty through modular regulation of an input-output gene.

Wang, C. W., and Y. H. Sun. 2012. Segregation of eye and antenna fates maintained by mutual antagonism in *Drosophila*. Development 139:3413–21.

Wittkopp, P. J., and G. Kalay. 2012. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nat. Rev. Genet. 13:59–69.

Wray, G. a. 2007. The evolutionary significance of cis-regulatory mutations. Nat. Rev. Genet. 8:206–16.

Figures

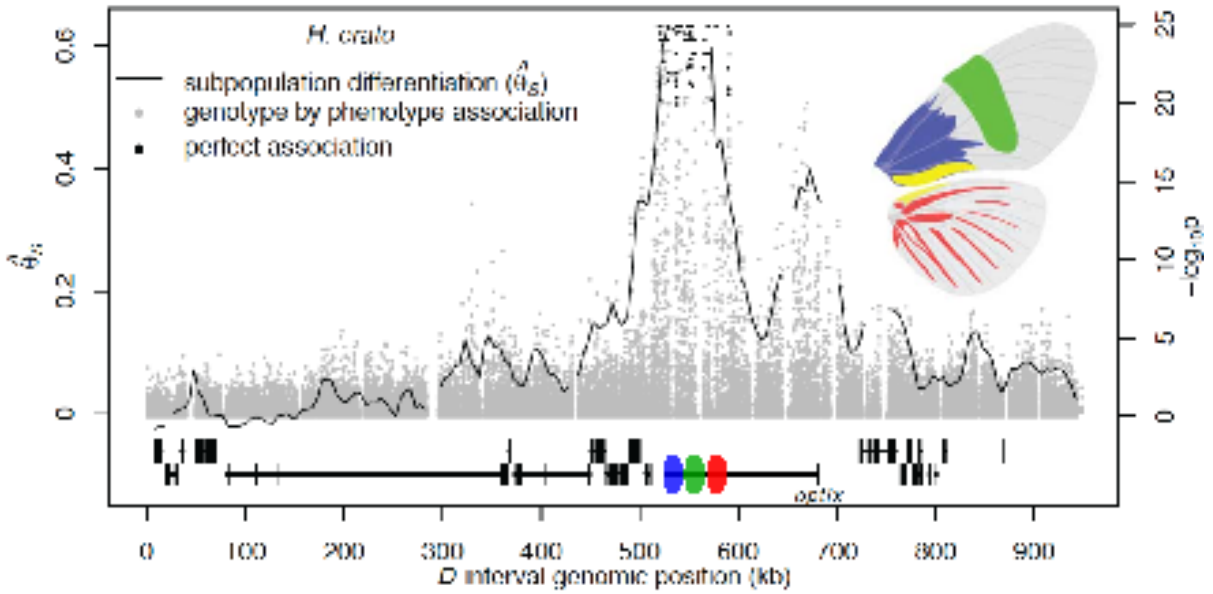


Figure 1: The modular enhancer hypothesis

The black and white background figure is previous divergence and association analysis (Supple et al. 2013) that identified a putative 65-kb regulatory region of the gene *optix*, which is shown in the gene annotation at the bottom. The colored bars represent hypothetical regulatory regions controlling spatially specific red color pattern variation, shown in the same color on the inset butterfly wing.

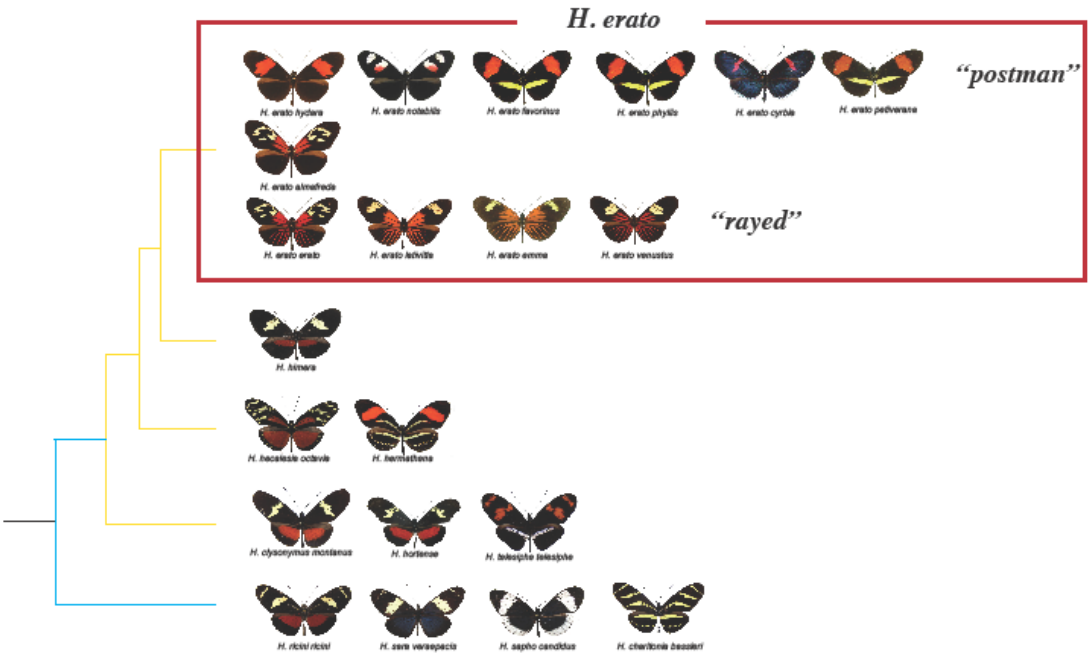
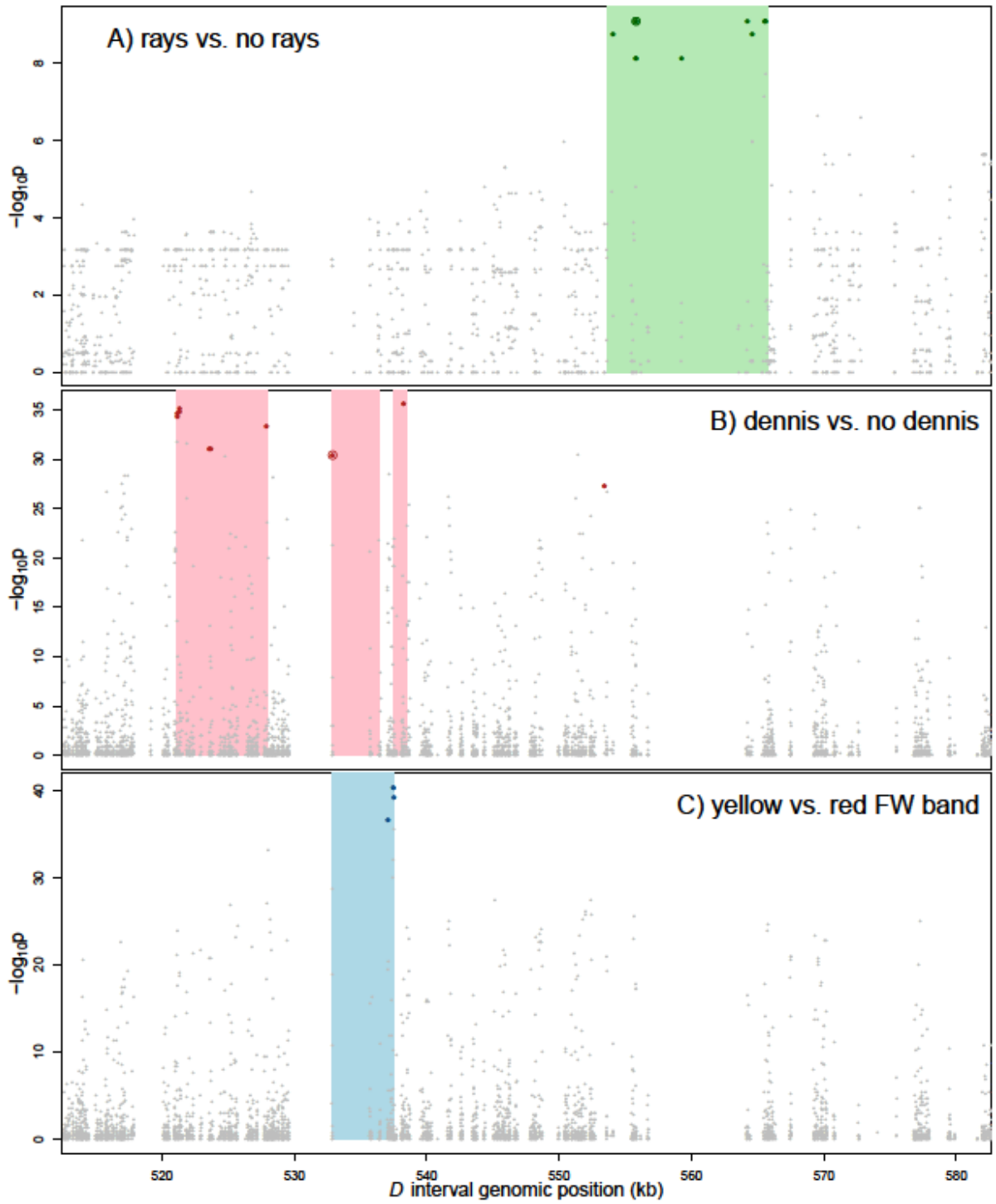


Figure 2: Taxon sampling

The tree representing phylogenetic and phenotypic sampling used to identify modular enhancers that control distinct red color pattern elements.

Figure 3: Modular enhancers for distinct color pattern elements

Genotype by phenotype association was used to identify putative enhancer regions for each phenotypic element—presence or absence of hindwing rays (A), presence or absence of basal dennis patch (B), and the color of the forewing band (C). The colored boxes highlight the identified regions. Dots represent genotype by phenotype calculated for biallelic SNPs using a Fisher's exact test, with colored dots representing sites perfectly associated with phenotype. Colored dots surrounded by circles in panels (A) and (B) indicate sites that are perfectly associated across the broader *H. erato* radiation.


A) rays vs. no rays
B) dennis vs. no dennis
C) yellow vs. red FW band
−log10P
D interval genomic position (kb)
520
530
540
550
560
570
580

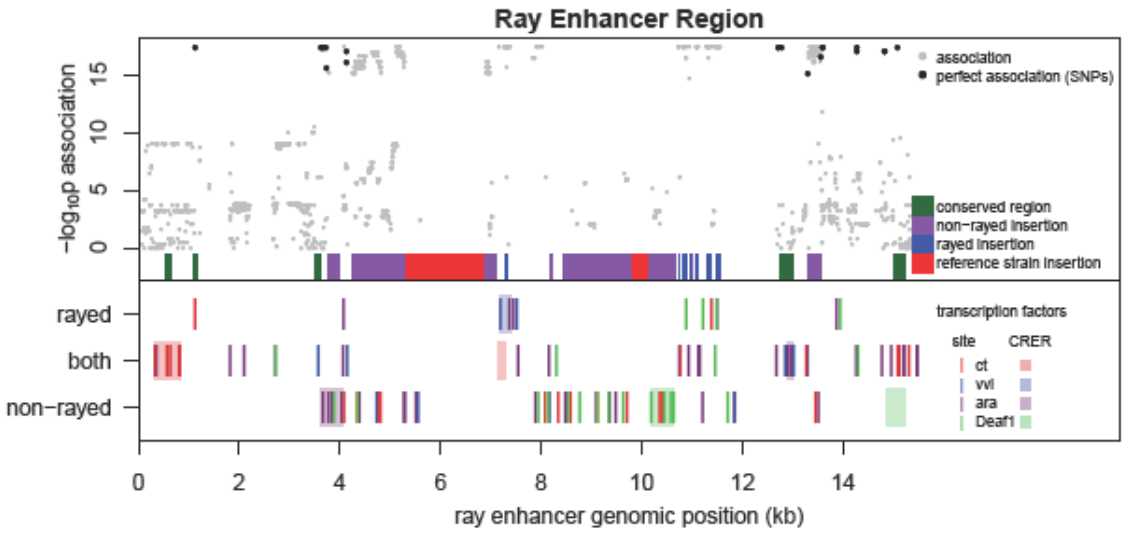


Figure 4: Features of the ray enhancer region

The top panel shows genotype by phenotype association (grey dots) based on alignment to the PCR generated reference sequences across the ray enhancer region, with perfectly associated SNPs indicated (black dots). The genomic structure is indicated with color boxes highlighting regions of high sequence conservation (green), phenotypic specific insertions/deletions (purple, blue), as well as major insertions in the *H. erato petiverana* reference sequence that results in the appearance of missing data in other races. The bottom panel shows predicted binding sites (hash marks) and cis-regulatory enhanced regions (CRER; shaded boxes) for the four candidate transcription factors. These elements are shown when they are present in both rayed races only, in all four races, or in both non-rayed races only. CRERs with no associated binding site indicated are due to the more stringent filtering during the site analysis than the CRER analysis.

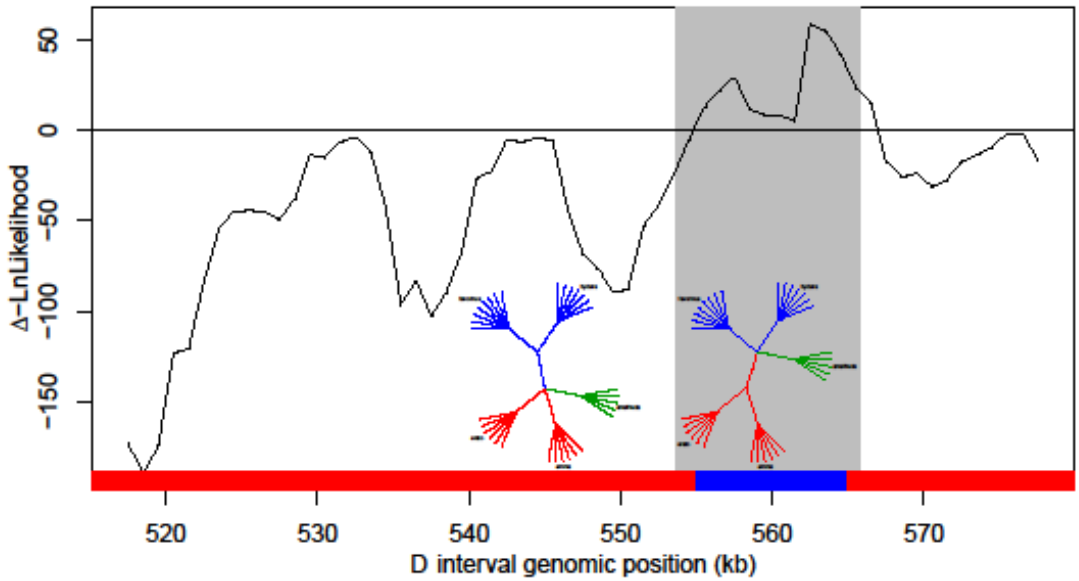


Figure 5: Evolutionary history of *H. e. amalfreda,* a recombinant phenotype

The recombinant phenotype of *H. e. amalfreda* is the result of a recombinant genotype across the region regulating red color variation. The curve shows the difference in negative log likelihoods of the data under two alternative trees (inset). Regions where the curve is negative indicate where *H. e. amalfreda* clusters with the rayed phenotypes (left tree), while positive regions indicate where *H. e. amalfreda* clusters with the postman phenotypes (right tree). The grey shaded box highlights the putative ray enhancer region identified by genotype by phenotype association (Figure 3A). The colored bars at the bottom show where the non-overlapping, 5-kb neighbor joining trees clustered *H. e. amalfreda* with the rayed races (red) or postman races (blue).

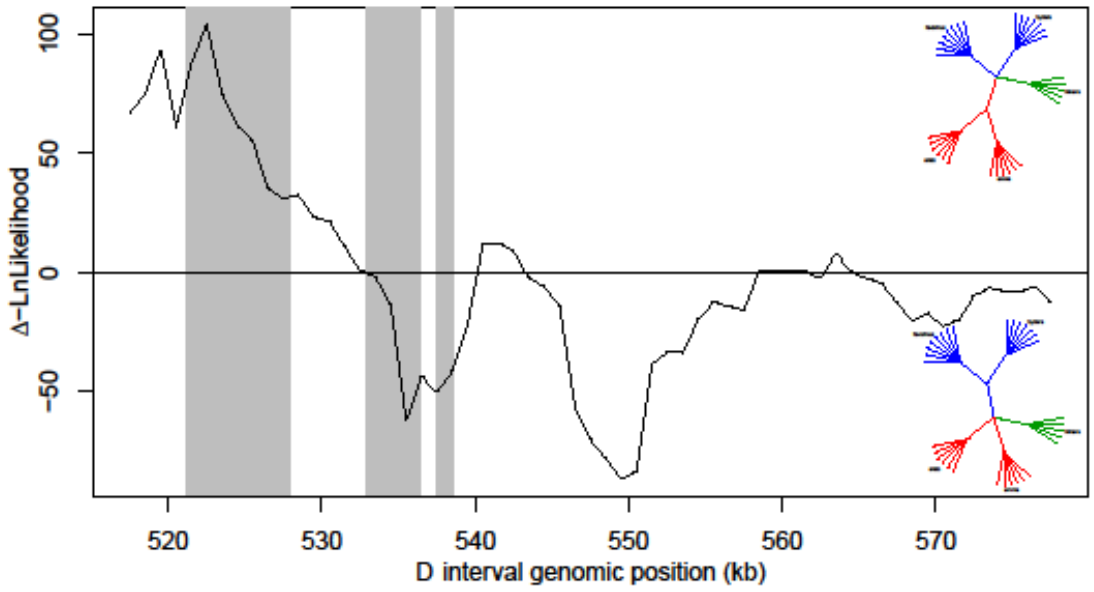


Figure 6: Evolutionary history of *H. himera,* a recombinant phenotype

The phenotype of *H. himera* contains elements of both the postman and the rayed phenotypes. The curve shows the difference in negative log likelihoods of the data under two alternative trees (inset). Regions where the curve is negative indicate where *H. himera* clusters with the rayed phenotypes (bottom tree), while positive regions indicate where *H. himera* clusters with the postman phenotypes (top tree). The grey shaded boxes highlight the putative dennis enhancer region identified by genotype by phenotype association (Figure 3B).

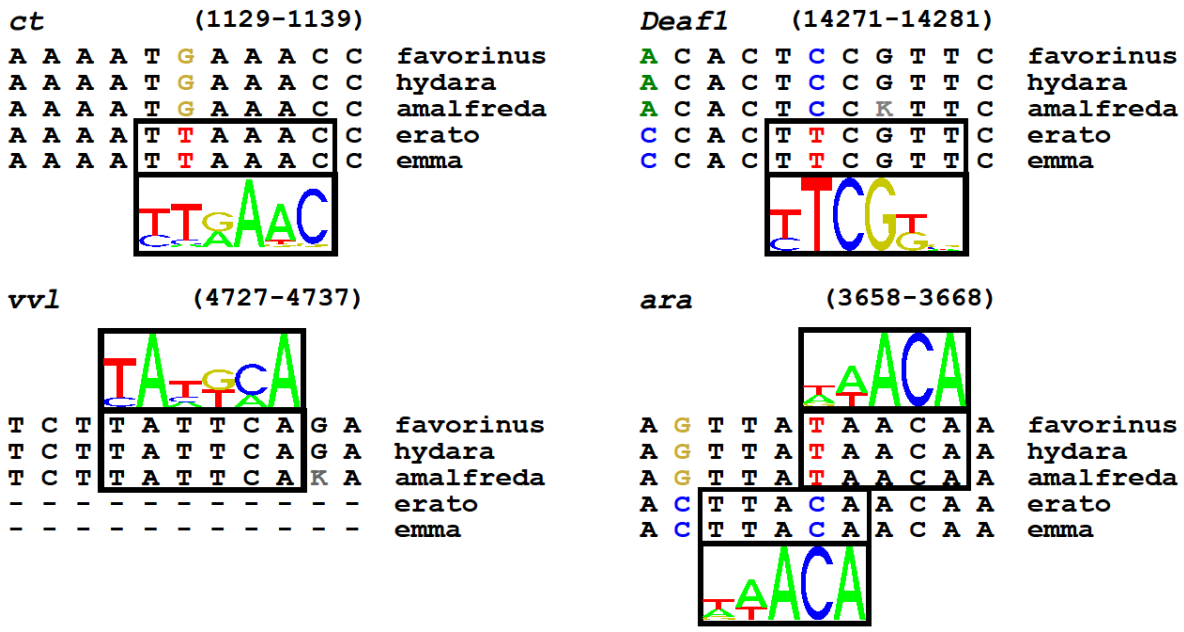


Figure S1: Binding sites for candidate transcription factors

Examples of high quality differential binding sites for the four top candidates (*ct*, *Deaf1*, *vvl*, *ara*). The transcription factor abbreviation is noted at the top left and the position within the rayed enhancer is noted at the top right. Sequences are shown for five races. Boxes indicate good binding sequences and the sequence logo for the candidate from JASPAR.

Table 1: Taxon Sampling

species	race	phenotype			sample size	reference
		fw band color	dennis	hindwing rays		
H. erato	*amalfreda*	yellow	yes	no	5	new
	venustus	yellow	yes	yes	1	new
	hydara (FG)	red	no	no	7	Supple et al. 2013
	erato	yellow	yes	yes	6	Supple et al. 2013
	favorinus	red	no	no	8	Supple et al. 2013
	emma	yellow	yes	yes	6	Supple et al. 2013
	notabilis	red	no	no	5	Supple et al. 2013
	lativitta	yellow	yes	yes	5	Supple et al. 2013
	petiverana	red	no	no	5	Supple et al. 2013
	hydara (PAN)	red	no	no	3	Supple et al. 2013
	cyrbia	red	no	no	4	Supple et al. *in prep*
H. himera	-	yellow	no	no	5	Supple et al. *in prep*
H. clysonymus	-	yellow	no	no	2	Supple et al. *in prep*
H. telesiphe	-	red	no	no	2	Supple et al. *in prep*
H. hortense	-	yellow	no	no	1	new
H. sara	-	yellow	no	no	1	new
H. demeter	-	yellow	yes	yes	1	new
H. ricini	-	yellow	no	no	1	new
H. chaithonia	-	yellow	no	no	1	new

Table 2: Candidate transcription factors

transcription factor	symbol	expressed in *Heliconoius* wing?	genbank accession
cut	ct	yes	gi|123334|sp|P10180.1
Deformed epidermal autoregulatory factor 1	Deaf1	yes	gi|22256750|sp|Q24180.1
Ventral veins lacking	vvl	yes	gi|27735164|sp|P16241.5
araucan	ara	yes	gi|19863700|sp|Q24248.2
CG4328	CG4328	no	gi|75027533|sp|Q9VTW3
BarH1	B-H1	yes	gi|33112231|sp|Q24255.2
BarH2	B-H2	yes	gi|33112232|sp|Q24256.3
onecut	onecut	yes	gi|13632097|sp|Q9NJB5.2
caupolican	caup	yes	gi|19860750|sp|P54269.2
putative homeodomain protein	PHDP	yes	gi|17986113|ref|NP_523834.1
abdominal-B	Abd-B	no	gi|1708231|sp|P09087.3
CG42234	CG42234	no	gi|74871980|sp|Q9W064
H2.0	H2.0	no	gi|66774180|sp|P10035.2
C15	C15	no	gi|74876548|sp|Q7KS72
bicoid	bcd	no	gi|47117827|sp|P09081.3
goosecoid	Gsc	no	gi|1708053|sp|P54366.1
Pituitary homeobox homolog	Ptx1	no	gi|38258871|sp|O18400.2
ocelliless/orthodenticle	oc	no	gi|226693532|sp|P22810.2
CG7056	CG7056	no	gi|74947796|sp|Q9VDF0

CHAPTER 4

Selection drives genomic divergence during speciation in *Heliconius* butterflies

Submitted for review:
Supple MA, Papa R, Hines HM, McMillan WO, and Counterman BA. Selection drives genomic divergence during speciation in *Heliconius* butterflies.

Abstract

A key to understanding the origins of species is determining the evolutionary processes that drive the patterns of genomic divergence during speciation. Using whole genome sequencing, we examine patterns of divergence between parapatric and allopatric populations of *Heliconius* butterflies with varying levels of reproductive isolation. We examine these patterns around a locus responsible for a major phenotypic switch in red wing color patterns. This genomic region is known to be under divergent selection and to drive reproductive isolation through assortative mating. As predicted, we find that genomic divergence increases with the degree of reproductive isolation. However, we observed unexpectedly high levels of genomic divergence between the incipient species *H. erato* and *H. himera*, given that reproductive isolation is incomplete and hybridization is common. This divergence between the incipient species is substantially higher than levels of divergence between parapatric hybridizing races and between allopatric races. This result indicates that reduced gene flow and selection on color pattern alone cannot explain the high levels of divergence between the incipient species. Rather, our results suggest that selection on multiple loci drives genome-wide divergence to accumulate early during speciation with gene flow.

Introduction

The ability to study high-resolution genomic patterns of divergence in natural populations has catalyzed a fundamental shift in our understanding of the origins of species. Historically,

speciation research considered geographic isolation (allopatry) as the main barrier that allowed populations to diverge and reproductive incompatibilities to evolve. However, more recent theoretical and empirical studies have demonstrated that natural selection can be an important driver of divergence and reproductive isolation, even when populations overlap geographically (parapatry or sympatry) (see Nosil 2012 for a review). As a result, there has been a shift towards understanding how populations diverge as reproductive isolation increases, but gene flow continues to occur (Harrison 2012). Genome scans between incompletely reproductively isolated populations and species have revealed that patterns of divergence can be highly heterogeneous across the genome (Lawniczak et al. 2010; Ellegren et al. 2012; Gompert et al. 2012; Roesti et al. 2012; Andrew and Rieseberg 2013; Gagnaire et al. 2013). There are now snapshots of genomic divergence across a broad range of organisms, yet little is known about how or why these patterns evolve.

Initially, metaphorical models invoking selection and differential gene flow were proposed to explain the heterogeneous patterns of genomic divergence that occur as speciation progresses (Nosil et al. 2009). In these models, selection drives "islands" of high genomic divergence, while homogenizing gene flow reduces the surrounding areas to a lower "sea level" divergence. Two types of genetic hitchhiking, divergence hitchhiking and genomic hitchhiking, were introduced to explain how selection could affect levels of gene flow locally and globally across the genome (Feder et al. 2012a). Divergence hitchhiking is the result of selection acting on loci responsible for ecological divergence. This causes the local reduction of gene flow in physically linked regions, which leads to the further accumulation of allelic differences around the targets of selection. In genomic hitchhiking,

the reduction of gene flow due to divergent selection reduces genome-wide rates of gene flow. This results in the global accumulation of genetic differences. While these two processes are similar, they reflect two fundamentally different ways that genomes may evolve during speciation—through the local accumulation of divergence around targets of selection or through the genome-wide accumulation of divergence (Feder et al. 2012a).

Heliconius butterflies offer an exceptional opportunity to examine the processes that drive heterogeneous patterns of divergence during speciation with gene flow using empirical genomic data from natural populations. The radiation is composed of a continuum of divergent races and species that provides an exceptional opportunity to study taxa at different stages of speciation. At one end of this continuum are divergent color pattern races that frequently hybridize and are only weakly reproductively isolated from each other. These racial boundaries are maintained primarily by selection on wing color patterns (Mallet 1986; Mallet and Barton 1989; Mallet et al. 1990; Blum 2002, 2008). Speciation is more progressed in other pairs where hybridization occurs, but where taxa show additional forms of reproductive isolation. This isolation is due to divergence in mating preference, hybrid sterility and inviability, differences in ecology and physiology, or a combination of all these factors (Descimon and Mast de Maeght 1984; Jiggins et al. 1996; McMillan et al. 1997; Mallet et al. 1998a; Naisbit et al. 2002; Arias et al. 2008; Jiggins 2008; Merrill et al. 2011a, 2014). An example of taxa that are further progressed are the incipient species *H. e. cyrbia* and *H. himera*. These taxa have a very narrow hybrid zone across which there is evidence of strong pre-mating isolation based on color pattern, but no evidence for post-mating isolation in the form of hybrid inviability or sterility (Descimon and Mast de Maeght 1984; Jiggins et

al. 1996; McMillan et al. 1997; Mallet et al. 1998a; Merrill et al. 2014). The parental species are associated with very different habitats: *H. himera* is found in very dry habitats at altitudes above 1000 meters; whereas, *H. e. cyrbia* is found in wetter forests at elevations below 1200 meters (Jiggins et al. 1996). Despite the strong pre-mating isolation and clear habitat differences between these two species, hybrids comprise 5-10% of the population in the narrow zone of overlap (Descimon and Mast de Maeght 1984; Jiggins et al. 1996; Mallet et al. 1998a).

There has been substantial progress identifying the loci that underlie phenotypic variation in the vivid wing patterns that define the group (Baxter et al. 2010; Counterman et al. 2010; Ferguson et al. 2010; Joron et al. 2011; Reed et al. 2011; Martin et al. 2012; Papa et al. 2013). These loci are under very strong natural selection and enable us to examine the build up of genomic divergence around a target of selection during speciation. In addition to being an important signal to predators, the color pattern elements also act as a reproductive barrier due to their role in assortative mating (Jiggins et al. 2001; Kronforst et al. 2006; Merrill et al. 2011b, 2014). The best characterized of these loci is the *D* or "red" locus, which controls the presence or absence of various red elements on the wings. A combination of linkage analysis, gene expression, and genotype by phenotype association strongly implicate cis-regulatory changes in the transcription factor *optix* as causing variation in red pattern elements (Reed et al. 2011). More recently, genome scans suggest that a 65-kb region about 100-kb upstream of *optix* likely contains the functional variation driving phenotypic differences (Supple et al. 2013).

Here we leverage the radiation and recent progress identifying functional variation to sample taxa at various levels of reproductive isolation and to examine patterns of divergence around loci known to be involved in both divergent selection and reproductive isolation. We focus specifically on *H. erato*—a radiation that has produced over 25 geographic races with distinct wing color patterns across Central and South America (Mallet 1993). We sample variation across two major red color pattern phenotypes—the postman, which has a red forewing band, and the rayed, which has yellow forewing band, red proximal forewing patch, and red hindwing rays (Figure 1). We sampled this variation across five geographically distinct hybrid zones. These hybrid zones were chosen to include taxa pairs where 1) hybrids are common and parental races show no differences in red color pattern elements, 2) hybrids are common but where parental taxa possess the divergent red color pattern phenotypes, and 3) hybrids are rare and parental types also have divergent color pattern phenotypes. Importantly, individual hybrid zones are 1000's of kilometers apart from each other and are often separated by major topographic features that impede gene flow.

Thus, our experimental design allows us to explore genomic divergence between functionally important and functionally neutral regions and examine how it varies relative to geographic proximity (parapatry and allopatry), wing color pattern (postman and rayed), and degree of reproductive isolation (freely hybridizing versus rarely hybridizing). Our results yield unexpected patterns whereby selection creates islands of differentiation at color pattern loci in allopatric taxa that do not differ by color pattern. This result suggests caution should be used when interpreting peaks of divergence to assess phenotypic regulation. Furthermore, we demonstrate a rapid accumulation of whole-genome divergence along the speciation

continuum and show that this divergence is likely driven by selection on multiple loci across the genome.

Methods

Sampling the *H. erato* speciation continuum

We examined phenotypically pure samples from five hybrid zones across the *H. erato* radiation, with varying degrees of reproductive isolation. The first four hybrid zones are all between hybridizing races of *H. erato* that show no evidence of pre-mating or post-mating isolation (Mallet 1986, 1989; Mallet et al. 1998b). One of these hybrid zones is between two different postman phenotypes (*H. e. hydara* x *H. e. petiverana* in Panama, n=8). The other three are between divergent postman and rayed phenotypes (*H. e. erato* x *H. e. hydara* in French Guiana, n=13; *H. e. emma* x *H. e. favorinus* in Peru, n=14; and *H. e. lativitta* x *H. e. notabilis* in Ecuador, n=10). For these four within species hybrid zones, we utilized previously published data (NCBI SRA accession SRA059512) (Supple et al. 2013). We sampled a fifth hybrid zone between incipient species with pre-mating isolation, but no post-mating isolation. We collected phenotypically pure samples of *H. himera* (n=5) and *H. e. cyrbia* (n=4) on either side of a narrow region of hybridization in Loja province, Ecuador. We also collected samples from two outgroup species, *H. clysonymus* (n=2) and *H. telesiphe* (n=2) in Peru. See Supplemental Table S1 for exact sampling locations.

Within all of these intraspecific and interspecific hybrid zones, gene flow is known to occur based on the presence of hybrid individuals with wing patterns that result from backcross color pattern genotypes (e.g. individuals homozygous at one color pattern locus, but heterozygous at another) (Descimon and Mast de Maeght 1984; Mallet 1989). However, these hybrid zones are geographically isolated from each other, with some on opposite sides of South America and separated by major topographic features. This isolation severely restricts gene flow between races from the different hybrid zones.

Sequencing and genotyping

The samples were sequenced and genotyped as described in Supple et al. (2013). Briefly, this involved whole genome sequencing 100-bp paired end reads on the Illumina HiSeq platform and aligning reads to a partial genomic reference (2.2-Mb, 0.5% of the *H. erato* genome) (GenBank accessions KC469892-KC469895, AC208805-AC208806) with BWA v0.59-r16 (Li and Durbin 2009) using relaxed mapping parameters. Each sample was sequenced to a realized median per base coverage between 8x and 21x (Supplemental Table S1). Multi-sample genotypes were then called and filtered with GATK v1.2 (McKenna et al. 2010; DePristo et al. 2011). Aligned sequencing reads are available at NCBI SRA (accessions XXX-XXX) and genotype VCF files are available at Dryad (doi:XXX). For comparison to the incipient species pair, we also examined within species patterns of divergence in four previously published parapatric *H. erato* racial hybrid zones (NCBI SRA accession SRA059512) (Supple et al. 2013). We examined sequence variation for these 58

samples across a 1-Mb region containing the locus controlling red color pattern variation and 350-kb of additional genomic regions unlinked to color pattern loci.

Phylogenetic analysis

We performed phylogenetic analyses of all the genomic data to i) determine the phylogenetic relationships of newly sequenced taxa, especially with regard to the position of *H. himera* relative to *H. erato*, which has been disputed (Brower 1994; Flanagan et al. 2004), and ii) develop expectations for relative genetic divergence of the lineages based on the their phylogenetic placement. We generated phylogenetic consensus trees from 5,399 SNPs across the previously identified 65-kb region that controls red color pattern variation (Supple et al. 2013) and 15,714 SNPs across the 350-kb of genomic regions unlinked to color pattern. To reduce the size of the unlinked dataset, we removed invariant sites.

We selected nucleotide substitution models for both datasets using a hierarchical likelihood ratio test implemented in PAUP* v4 (Swofford 2002) and MrModeltest v2 (Nylander 2004). We used MrBayes v3.2.2 (Ronquist et al. 2012) to generate consensus trees with the selected models—GTR+I+G for the 65-kb functional region and GTR+G for the unlinked loci. For each tree, we ran 5 runs for 5 million generations each, sampling every 500 generations and removing 25% burnin. We assessed burnin and convergence in MrBayes and Tracer v1.6 (Rambaut et al. 2013). For the unlinked tree, we removed two runs that converged to a slightly lower likelihood than the other three runs. We generated a consensus tree from converged runs and rooted the trees with the *H. clysonymus* and *H. telesiphe* samples. The consensus trees are available at TreeBASE (accession XXX).

We performed an additional assessment to test the robustness of the position of *H. himera* relative to the broader *H. erato* radiation at genomic regions unlinked to color pattern loci. We accounted for gene flow and phylogenetic conflict in inferring relationships by constructing an unrooted neighbor-net splits tree network on unlinked data using SplitsTree v4.13.1 (Huson and Bryant 2006). This analysis utilized pairwise distances and treated polymorphisms additively, rather than as ambiguities.

Genomic divergence analysis

We compared patterns of divergence between the incipient species, *H. himera* and *H. e. cyrbia*, to patterns of divergence within species using both parapatric and allopatric pairs of *H. erato* races. The within species parapatric comparisons consisted of three hybrid zones between postman and rayed phenotypes (Peru, Ecuador, and French Guiana) and one hybrid zone between two different postman phenotypes (Panama). The within species allopatric comparisons consisted of all remaining pairwise comparisons between these same races, with the two Panamanian postman races being treated as a single race. These comparisons can be divided into three categories (postman versus rayed, n=9 comparisons; postman versus postman, n=6 comparisons; rayed versus rayed, n=3 comparisons). Note that *H. erato cyrbia* is only included in comparison to *H. himera*. It is not included in the within *H. erato* comparisons, due to its interspecies hybridization likely making it more divergent from other *H. erato* races.

For each comparison, we calculated divergence per genomic position using a model for diploid data with populations as random effects ($\hat{\theta}$) (Weir 1996), with details described in

Supple et al. (2013). We filtered data to remove positions where less than 75% of individuals were genotyped for each of the two taxa being compared. To combine estimates across loci and across comparisons, the numerators and denominators are each summed across the pairwise estimates and then divided to produce the final combined estimate. We calculated the genomic background level of divergence from three genomic intervals (0.35-Mb) unlinked to color pattern, using 1000 bootstrap replicates to determine 95% confidence intervals (CIs). For sliding window analyses, we examined 15-kb windows, with 5-kb steps, requiring estimates for at least 20% of the positions in the window. To examine the decay of divergence with recombination distance, we defined the causative locus as the center of the 65-kb region previously identified to modulate red color pattern variation in *H. erato* (Supple et al. 2013). We converted genomic distance (bp) to recombination distance (cM) assuming a constant recombination rate using the *H. erato* linkage map size (1430 cM) (Kapan et al. 2006) and the estimated size of the *H. erato* genome (400 Mb) (Tobler et al. 2005). Loci were binned by recombination distance from the causative locus in bins of size 0.01 cM. Decay data were loess smoothed for presentation.

Results & Discussion

Heliconius himera—an incipient species from within the H. erato radiation

Our phylogenetic analysis indicates that *H. himera* evolved from within the broader *H. erato* color pattern radiation, rather than predating it. *Heliconius himera* is a geographic replacement of *H. erato* in mid-elevation forests of southern Ecuador and northern Peru

(Jiggins et al. 1996). Previous research has been equivocal about *H. himera's* relationship to *H. erato*. Early phylogenetic analysis using mtDNA placed *H. himera* as a sister species to the *H. erato* radiation (Brower 1994); however, later analysis of rapidly evolving nuclear introns placed *H. himera* within the broader *H. erato* radiation, although with low support (Flanagan et al. 2004). Our phylogenetic analysis based on more extensive genomic data places *H. himera* within the *H. erato* radiation, both at the color pattern locus and at loci unlinked to color pattern (Figure S1 & S2).

The MrBayes phylogenetic tree of *H. erato* races across loci unlinked to color pattern shows a strong geographic signal, with no clustering by color pattern (Figure S1). Similar to previous research, there is a strong phylogenetic break across the Andes (Quek et al. 2010; Hines et al. 2011). All five *H. himera* individuals fall on a well-differentiated lineage that clusters with *H. erato* races from the eastern slope of the Andes and thus within the *H. erato* radiation (Figure S1). Given the long branch length and weak support values leading to the *H. himera* lineage, we used an additional methods to assess the support for *H. himera* evolving from within the *H. erato* radiation versus outside it as a sister lineage. The splits tree neighbor-net network shows *H. himera* as a separate lineage nested within the *H. erato* radiation, with closer affinity for the *H. erato* taxa from the eastern slopes of the Andes than the clade from the western slopes (Figure S3). Collectively, our analyses consistently support *H. himera* as part of the *H. erato* radiation, with a close association to *H. erato* samples from east of the Andes.

A very different phylogenetic history was observed across the genomic region that controls red color pattern variation, reflecting the evolution of the color pattern alleles. This

tree strongly clusters taxa by color pattern, showing two reciprocally monophyletic clades sorting perfectly by red phenotype. The "postman" clade contains all of the *H. erato* postman races, including *H. erato cyrbia*, while the "rayed" clade contains all of the *H. erato* rayed races and *H. himera*. Even though *H. himera* does not have the characteristic rays, it can be phenotypically classified as a rayed based on the occurrence of red on the hindwing, a classification supported by this phylogenetic analysis.

While the analyses at loci unlinked to color pattern and across the genomic region containing the functional variants show different patterns of clustering (by geography and phenotype, respectively), both place *H. himera* within the *H. erato* radiation, supporting *H. himera* as an incipient species derived from within *H. erato*. For the incipient species pair, each taxa falls into a separate, monophyletic lineage at both color pattern linked and unlinked loci. From an experimental perspective, the presence of two well-supported monophyletic lineages suggests that taxa pairs distributed between these lineages should have similar divergence times. This enables us to make comparisons between levels of genomic divergence for the incipient species pair and other taxa pairs across the two lineages without having to control for divergence times.

Divergence accumulates rapidly during speciation

The phylogenetic placement of *H. himera* within the broader *H. erato* radiation makes it an important reference point along the speciation continuum with which to examine how genomes evolve during speciation. Is there a local reduction in gene flow around genomic targets of selection, causing islands of divergence to grow, as is expected with divergence

hitchhiking? Or is gene flow reduced across the genome, causing increased divergence genome wide, as is expected with genomic hitchhiking?

Our divergence analysis shows distinct peaks of divergence at the functional regions that modulate adaptive differences in wing color patterns in comparisons between *H. erato* parapatric races with divergent phenotypes (Figure 2A). As previously demonstrated, when comparing parapatric postman and rayed races, there are two high peaks of divergence that show steep drop-offs almost all the way to the near zero genomic baseline divergence (Supple et al. 2013). These peaks are in sharp contrast to the parapatric postman versus postman comparison, which shows near zero divergence across the entire red color pattern interval and at intervals unlinked to color pattern. Somewhat unexpectedly, comparisons between allopatric postman phenotypes also show peaks of divergence around the functional regions (see discussion below). In these comparisons, the levels of divergence between identical but allopatric races are very similar to the pattern seen in parapatric comparisons between divergent color pattern races (Figure 2B). Comparisons between allopatric rayed races reveal relatively low levels of divergence except for very small peaks around the functional regions (Figure 2B).

There was no compelling evidence that the genomic islands that distinguished divergent *H. erato* color pattern races expand with increased reproductive isolation. To determine if genomic islands grew, we compared the distances where the curve representing the decay of divergence from the target of selection intersected the line representing the background genomic divergence at loci unlinked to color pattern. In both the within species (within *H. erato*) and between species (between *H. himera* and *H. erato*) parapatric

comparisons, divergence decays from the peak to background genomic levels at approximately the same recombination distance (Figure 3).

In fact, there is very little evidence for any distinct islands between *H. himera* and *H. erato* across the color pattern region, as we see a high divergence genome wide (Figure 2A). The level of divergence between the incipient species *H. himera* and *H. e. cyrbia* is substantially higher than the divergence seen in all other comparisons. This is true at every region we examined, except at the narrow region containing the functional variation where most comparisons have very high divergence. Unlike parapatric races, the examination of the level of divergence across the color pattern locus between the incipient species reveals strong divergence extending across the entire 1-Mb interval (Figure 2). The unlinked loci show similarly high divergence between the incipient species (average=0.519), which is substantially higher than any of the within species comparisons (maximum average=0.191) (Table S2). This high level of divergence is somewhat unexpected given that *H. himera* is a recent incipient species and shows the same type of phenotypic transition at hybrid zones as seen between divergent *H. erato* races (Figure S2).

The observations that islands of divergence do not grow and that the incipient species show high genome wide divergence both support a more prevalent role of genome hitchhiking during early speciation, as opposed to divergence hitchhiking. These results contrast with a study in a different *Heliconius* clade study that suggested divergence hitchhiking may play a prominent role during the early stages of speciation (Nadeau et al. 2012). Between incipient species, they observed additional peaks of divergence near the color pattern loci, however regions unlinked to color pattern did not show a similar increase.

Although the incipient species in that study and the incipient species we examine here have similar estimated times of divergence (Kozak et al. *in review*), we observed increased divergence across the genome relative to within species comparisons. These differences may be an artifact of stochasticity in genomic divergence and limited sample size for the incipient species in the previous study (i.e. n=1). Alternatively, the incipient species *H. himera* and *H. e. cyrbia* may have diverged ecologically and genetically more rapidly than the incipient species in the *melpomene* clade. Overall, our results are consistent with several theoretical and empirical studies that report high genome wide divergence and a limited role for divergence hitchhiking in promoting speciation (Feder and Nosil 2010; Lawniczak et al. 2010; Michel et al. 2010; Feder et al. 2012b; Andrew and Rieseberg 2013; Flaxman et al. 2013; Kronforst et al. 2013; Powell et al. 2013). Collectively, these studies suggest that genome-wide divergence can occur in the earliest stages of speciation.

Selection within populations drives divergence between populations

The patterns of divergence within *H. erato* are most consistent with strong selection acting within populations. The comparisons between divergent phenotypes in parapatry show a strong peak of divergence (Figure 2A), as expected given the strong divergent selection on the color pattern locus operating in these contact zones. This pattern stands in marked contrast to divergence observed across the postman versus postman hybrid zone (Figure 2A). We have previously determined that the genomic region containing these peaks of divergence also shows signatures of natural selection, including reduced heterozygosity (Supple et al. 2013).

Further evidence that natural selection is driving divergence between populations comes from our comparisons of allopatric races. Unexpectedly, all of the allopatric comparisons show peaks of divergence around the locus modulating red color pattern variation, including the comparisons between races with similar red phenotypes (Figure 2B). The pattern of divergence among allopatric postman versus postman is particularly noteworthy, as it shows strong divergence peaks similar to those observed between parapatric populations with divergent (postman versus rayed) phenotypes. This is surprising given that these geographically isolated races share a similar red color pattern phenotype, which has a common origin across the *H. erato* radiation (Figure S2). The fact that the peaks of divergence are not as strong in the allopatric rayed versus rayed comparisons likely reflects the younger age of the rayed phenotype coupled with higher likelihood of gene flow among rayed populations. The rayed phenotypes form a contiguous population of races across the Amazon basin that differ in forewing band shape, which is controlled by a separate independent locus (Martin et al. 2012).

In the absence of selection, any divergence among allopatric races should accumulate at a rate similar to the baseline level seen across regions not under selection; however, the background divergence at unlinked loci is much lower than across the color pattern locus (Figure 2B). The observed pattern likely reflects unique histories of mutation and selection at each hybrid zone. Geographic isolation will lead to the accumulation of population specific mutations across the genome over time. Recurrent selection against foreign color patterns on the edges of hybrid zones can result in the rapid fixation of population specific mutations that are tightly linked to the color pattern loci. We believe this selection leads to

the accelerated accumulation of population specific alleles at the locus that controls red color pattern variation. Local selection can reduce within population diversity, driving peaks of divergence, even in the absence of gene flow (Charlesworth et al. 1997; Charlesworth 1998). This resulted in the observed peak of divergence between allopatric races, even when they share a similar red color pattern. Collectively, these results provide evidence that selection *within* populations can leave the same signature of selection as ecological divergence *between* populations.

The result that local selection within populations can drive global patterns of divergence between populations in complex ways has important ramifications for the interpretation of population genomic data. Divergence peaks are often interpreted as a result of divergent selection acting directly on the two taxa being compared. However, we demonstrate that peaks of divergence can be driven by strong selection acting independently within each population. Given the complex patterns of divergence resulting from the interactions between various evolutionary forces, peaks of divergence should be interpreted with caution, especially when additional supporting data is not available.

Selection drives high divergence between incipient species

Given our result that selection drives divergence within species, it is natural to hypothesize that selection also drives divergence between species. However, patterns of divergence could be driven by other evolutionary forces, including gene flow. Our experimental design allows us to begin to disentangle the effects of selection and gene flow on patterns of genomic divergence during speciation. By comparing allopatric and parapatric races, we can see the

effect of gene flow. Consistent with reduced gene flow, the allopatric within species comparisons reveal higher levels of background genomic divergence than the levels observed in the within species parapatric comparisons (Figure 3, blue shaded area versus green shaded area).

Similarly, by comparing divergence levels among allopatric races and the divergence between the incipient species, *H. e. cyrbia* and *H. himera*, we can determine the effect restricted gene flow has on species divergence. Given that both the red locus and unlinked phylogenetic trees show two distinct lineages, one which contains *H. himera* and one which contains *H. erato cyrbia* (Figures S1 and S2), the times of divergence between taxa pairs that are similarly distributed across the same distinct lineages are expected to be similar. As such, allopatric pairs across these lineages provide an expectation of neutral variation in the absence of gene flow. If the patterns of divergence between species were driven by the accumulation of neutral variation resulting from reduced gene flow, we expect that the incipient species *H. himera* and *H. erato cyrbia* would have a similar pattern of divergence as the allopatric within species comparisons. In contrast to this expectation, the genomic divergence between the incipient species is much higher at genomic regions both linked and unlinked to the color pattern locus than the allopatric comparisons (Figure 3, red curve versus blue curve; red horizontal line versus blue shaded area). The fact that genomic divergence is so much higher than our neutral expectation suggests that the neutral accumulation of differences due to limited gene flow is insufficient to explain the high divergence observed between *H. himera* and *H. erato cyrbia*. Rather, these patterns argue that natural selection is rapidly driving genomic divergence in this incipient species pair.

Our data suggest that selection on color pattern alone is insufficient to explain the high divergence across the genome. We see no evidence that the size of the island of divergence around the color pattern locus grows with increasing reproductive isolation. Rather the build-up occurs genome-wide and is likely the result of natural selection acting on many loci. Unlike geographic races of the *H. erato*, *H. himera* and *H. e. cyrbia* show strong divergence in characteristics other than color pattern, including differences in mating preference, larval developmental time, adult physiology, and habitat preferences (Descimon and Mast de Maeght 1984; Jiggins et al. 1996; McMillan et al. 1997; Mallet et al. 1998a; Davison et al. 1999; Pardo-Diaz et al. 2012; Merrill et al. 2014). Divergent selection on multiple traits, including habitat preference, is likely driving the globally high divergence between the incipient species *H. himera* and *H. e. cyrbia*.

These findings add to a building body of evidence, from both theoretical models and empirical data, that support the hypothesis that speciation is driven by selection at multiple loci across the genome and that natural selection can generate genome wide divergence in the face of ongoing gene flow (Feder and Nosil 2010; Michel et al. 2010; Roesti et al. 2012; Flaxman et al. 2013; Powell et al. 2013).

The *Heliconius* speciation continuum

Heliconius butterflies provide a full continuum of taxa pairs at varying stages of reproductive isolation—from freely hybridizing color pattern races to completely reproductively isolated species. By sampling at various stages across the *Heliconius* speciation continuum, we can observe how genomes diverge through the speciation process and begin to disentangle the

evolutionary forces that drive species divergence. Our results add to a growing body of research that is uncovering how genomes diverge as speciation proceeds through time.

A common feature of these studies is that divergence can be extremely heterogeneous between incipient species (Lawniczak et al. 2010; Ellegren et al. 2012; Gagnaire et al. 2013). Consistent with previous work in *Heliconius*, we show that within species, isolated regions of the genome show exceptionally high divergence, while the rest of the genome shows little to no divergence (Nadeau et al. 2012; Martin et al. 2013). As speciation progresses, there is an increase in divergence, coupled with an increase in heterogeneity (Nadeau et al. 2012; Kronforst et al. 2013; Martin et al. 2013). Several studies in *Heliconius* have shown that this increase in genomic divergence can occur rapidly in the early stages of speciation; however, as discussed previously, Nadeau et al. (2012) suggested a more gradual build up of divergence through speciation. Our results show that genomic divergence occurs quite rapidly, with incipient species pairs in the earliest stages of speciation already showing very high divergence genome-wide.

Further, our results show that reduced gene flow is insufficient to explain the rapid divergence observed between incipient species, which suggests selection may be major of driver of genomic divergence during speciation. Our results are consistent with theoretical models (Gavrilets 2000) and empirical studies (see references in Nosil 2012) that show a primary role for selection in speciation. Our results differ, however, from a recent study of the *Heliconius melpomene* clade that showed similar levels of genomic divergence between sympatric species (*H. melpomene*, *H. cydno* and *H. timareta*) and allopatric within species comparisons (allopatric *H. melpomene* color pattern races) (Martin et al. 2013), suggesting

that selection may not play a large role during the early stages speciation in the *melpomene* clade. Although the incipient species pairs in our study and theirs have similar divergence times (Kozak et al. *in review*), they have important differences. The species compared in the *melpomene* clade are sympatric, as opposed to our parapatric comparisons. The much larger geographic region of overlap likely allows for more hybridization events, resulting in higher realized gene flow between species. Additionally, *H. erato* and *H. himera* show stronger evidence of habitat based ecological divergence than the incipient species in the *melpomene* clade (Mallet et al. 1998b). We suggest our incipient species pair is more ecologically isolated and further along in the speciation continuum than those studied in *H. melpomene*, thus explaining the higher levels of divergence observed across the genome.

The interpretation that the observed high levels of genomic divergence between the incipient species are associated with greater ecological divergence (i.e. divergence not only in warning color, but also habitat) supports a growing number of studies that suggest increases in genome-wide levels of divergence are caused by selection acting on many loci across the genome (Feder et al. 2012b; Flaxman et al. 2013). Collectively, these studies and the *Heliconius* speciation continuum offer a powerful empirical framework to test specific predictions of how various ecological factors and evolutionary forces affect patterns of genomic divergence.

Acknowledgements

Samples were collected with permission from Ecuador's Ministerio del Ambiente (006-2012-IC-FAU-DPL-MA, 005-13 IC-FAU-DNB/MA, 013-09 IC-FAU-DNB/MA); Peru's Ministerio de Agricultura and Instituto Nacional de Recursos Naturales (004-2008-INRENAIFFS-DCB, 011756-AG-INRENA); French Guiana's Ministere de L'Ecologie, de L'Energie, du Developpemet Durable et de laMar (BIODAD-2010-0433); Panama's Autoridad Nacional del Ambiente (SC/A-7-11). This work was funded by the following awards: Hanne and Torkel Weis-Fogh Fund (sample collection, awarded to Nicola Nadeau and Richard Merrill); CNRS Nouraugues (B.A.C.); NSF DEB-1257839 (B.A.C.), DEB-1257689 (W.O.M.), DEB-1027019 (W.O.M.); and the Smithsonian Institution.

Author contributions: Experimental design is credited to B.A.C., R.P., M.A.S., W.O.M., H.M.H.; data collection was carried out by M.A.S., B.A.C., W.O.M.; data analysis was done by M.A.S., B.A.C., H.M.H.; the manuscript was prepared by M.A.S., B.A.C., with input from all authors. The authors declare no conflict of interest.

References

Andrew, R. L., and L. H. Rieseberg. 2013. Divergence is focused on few genomic regions early in speciation: incipient speciation of sunflower ecotypes. Evolution 67:2468–2482.

Arias, C. F., A. G. Muñoz, C. D. Jiggins, J. Mavárez, E. Bermingham, and M. Linares. 2008. A hybrid zone provides evidence for incipient ecological speciation in *Heliconius* butterflies. Mol. Ecol. 17:4699–4712.

Baxter, S. W., N. J. Nadeau, L. S. Maroja, P. Wilkinson, B. A. Counterman, A. Dawson, M. Beltran, S. Perez-Espona, N. Chamberlain, L. Ferguson, R. Clark, C. Davidson, R. Glithero, J. Mallet, W. O. McMillan, M. Kronforst, M. Joron, R. H. ffrench-Constant, and C. D. Jiggins. 2010. Genomic hotspots for adaptation: the population genetics of Müllerian mimicry in the *Heliconius melpomene* clade. PLoS Genet. 6:e1000794.

Blum, M. J. 2008. Ecological and genetic associations across a *Heliconius* hybrid zone. J. Evol. Biol. 21:330–341.

Blum, M. J. 2002. Rapid movement of a *Heliconius* hybrid zone: evidence for phase III of Wright's shifting balance theory? Evolution 56:1992–1998.

Brower, A. V. Z. 1994. Rapid morphological radiation and convergence among races of the butterfly *Heliconius erato* inferred from patterns of mitochondrial DNA evolution. Proc. Natl. Acad. Sci. U. S. A. 91:6491–5.

Charlesworth, B. 1998. Measures of divergence between populations and the effect of forces that reduce variability. Mol. Biol. Evol. 15:538–543.

Charlesworth, B., M. Nordborg, and D. Charlesworth. 1997. The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. Genet. Res. 70:155–74.

Counterman, B. A., F. Araujo-Perez, H. M. Hines, S. W. Baxter, C. M. Morrison, D. P. Lindstrom, R. Papa, L. Ferguson, M. Joron, R. H. ffrench-Constant, C. P. Smith, D. M. Nielsen, R. Chen, C. D. Jiggins, R. D. Reed, G. Halder, J. Mallet, and W. O. McMillan. 2010. Genomic hotspots for adaptation: the population genetics of Müllerian mimicry in *Heliconius erato*. PLoS Genet. 6:e1000796.

Davison, A., W. O. McMillan, A. S. Griffin, C. D. Jiggins, and J. L. B. Mallet. 1999. Behavioral and physiological differences between two parapatric *Heliconius* species. Biotropica 31:661–668.

DePristo, M. A., E. Banks, R. Poplin, K. V Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, and M. J. Daly. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43:491–498.

Descimon, H., and J. Mast de Maeght. 1984. Semispecies relationships between *Heliconius erato cyrbia* Godt. and *H. himera* Hew. in Southwestern Ecuador. J Res Lepid 22:229–237.

Ellegren, H., L. Smeds, R. Burri, P. I. Olason, N. Backström, T. Kawakami, A. Künstner, H. Mäkinen, K. Nadachowska-Brzyska, A. Qvarnström, S. Uebbing, and J. B. W. Wolf. 2012. The genomic landscape of species divergence in *Ficedula* flycatchers. Nature 491:756–760.

Feder, J. L., S. P. Egan, and P. Nosil. 2012a. The genomics of speciation-with-gene-flow. Trends Genet. 28:342–350.

Feder, J. L., R. Gejji, S. Yeaman, and P. Nosil. 2012b. Establishment of new mutations under divergence and genome hitchhiking. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 367:461–474.

Feder, J. L., and P. Nosil. 2010. The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. Evolution 64:1729–1747.

Ferguson, L., S. F. Lee, N. Chamberlain, N. Nadeau, M. Joron, S. Baxter, P. Wilkinson, A. Papanicolaou, S. Kumar, T. J. Kee, R. Clark, C. Davidson, R. Glithero, H. Beasley, H. Vogel, R. ffrench-Constant, and C. Jiggins. 2010. Characterization of a hotspot for mimicry: assembly of a butterfly wing transcriptome to genomic sequence at the *HmYb/Sb* locus. Mol. Ecol. 19 Suppl 1:240–54.

Flanagan, N. S., A. Tobler, A. Davison, O. G. Pybus, D. D. Kapan, S. Planas, M. Linares, D. Heckel, and W. O. McMillan. 2004. Historical demography of Müllerian mimicry in the neotropical *Heliconius* butterflies. Proc. Natl. Acad. Sci. U. S. A. 101:9704–9.

Flaxman, S. M., J. L. Feder, and P. Nosil. 2013. Genetic hitchhiking and the dynamic buildup of genomic divergence during speciation with gene flow. Evolution 67:2577–2591.

Gagnaire, P., S. Pavey, E. Normandeau, and L. Bernatchez. 2013. The genetic architecture of reproductive isolation during speciation-with-gene-flow in lake whitefish species pairs assessed by RAD sequencing. Evolution 67:2483–2497.

Gavrilets, S. 2000. Waiting time to parapatric speciation. Proc. Biol. Sci. 267:2483–2492.

Gompert, Z., L. K. Lucas, C. C. Nice, and C. A. Buerkle. 2012. Genome divergence and the genetic architecture of barriers to gene flow between *Lycaeides idas* and *L. melissa*. Evolution 67:2498–2514.

Harrison, R. G. 2012. The language of speciation. Evolution 66:3643–3657.

Hines, H. M., B. A. Counterman, R. Papa, P. Albuquerque de Moura, M. Z. Cardoso, M. Linares, J. Mallet, R. D. Reed, C. D. Jiggins, M. R. Kronforst, and W. O. McMillan. 2011. Wing patterning gene redefines the mimetic history of *Heliconius* butterflies. Proc. Natl. Acad. Sci. U. S. A. 108:19666–19671.

Huson, D. H., and D. Bryant. 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23:254–67.

Jiggins, C. D. 2008. Ecological Speciation in Mimetic Butterflies. Bioscience 58:541–548.

Jiggins, C. D., W. O. McMillan, W. Neukirchen, and J. Mallet. 1996. What can hybrid zones tell us about speciation? The case of *Heliconius erato* and *H. himera* (Lepidoptera: Nymphalidae). Biol. J. Linn. Soc. 59:221–242.

Jiggins, C. D., R. E. Naisbit, R. L. Coe, and J. Mallet. 2001. Reproductive isolation caused by colour pattern mimicry. Nature 411:302–305.

Joron, M., L. Frezal, R. T. Jones, N. L. Chamberlain, S. F. Lee, C. R. Haag, A. Whibley, M. Becuwe, S. W. Baxter, L. Ferguson, P. A. Wilkinson, C. Salazar, C. Davidson, R. Clark, M. A. Quail, H. Beasley, R. Glithero, C. Lloyd, S. Sims, M. C. Jones, J. Rogers, C. D. Jiggins, and R. H. ffrench-Constant. 2011. Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. Nature 477:203–6. Nature Publishing Group.

Kapan, D. D., N. S. Flanagan, A. Tobler, R. Papa, R. D. Reed, J. A. Gonzalez, M. R. Restrepo, L. Martinez, K. Maldonado, C. Ritschoff, D. G. Heckel, and W. O. McMillan. 2006. Localization of Müllerian mimicry genes on a dense linkage map of *Heliconius erato*. Genetics 173:735–57.

Kozak, K. M., N. Wahlberg, A. Neild, K. K. Dasmahapatra, J. Mallet, and C. D. Jiggins. 2014. Multilocus species trees show the recent adaptive radiation of the mimetic *Heliconius* butterflies. bioRxiv, doi: 10.1101/003749.

Kronforst, M. R., M. E. B. Hansen, N. G. Crawford, J. R. Gallant, W. Zhang, R. J. Kulathinal, D. D. Kapan, and S. P. Mullen. 2013. Hybridization reveals the evolving genomic architecture of speciation. Cell Rep. 5:666–677.

Kronforst, M. R., L. G. Young, D. D. Kapan, C. McNeely, R. J. O'Neill, and L. E. Gilbert. 2006. Linkage of butterfly mate preference and wing color preference cue at the genomic location of *wingless*. Proc. Natl. Acad. Sci. U. S. A. 103:6575–6580.

Lawniczak, M. K. N., S. J. Emrich, A. K. Holloway, A. P. Regier, M. Olson, B. White, S. Redmond, L. Fulton, E. Appelbaum, J. Godfrey, C. Farmer, A. Chinwalla, S.-P. Yang, P. Minx, J. Nelson, K. Kyung, B. P. Walenz, E. Garcia-Hernandez, M. Aguiar, L. D. Viswanathan, Y.-H. Rogers, R. L. Strausberg, C. A. Saski, D. Lawson, F. H. Collins, F. C. Kafatos, G. K. Christophides, S. W. Clifton, E. F. Kirkness, and N. J. Besansky. 2010. Widespread divergence between incipient *Anopheles gambiae* species revealed by whole genome sequences. Science 330:512–514.

Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760.

Mallet, J. 1986. Hybrid zones of *Heliconius* butterflies in Panama and the stability and movement of warning colour clines. Heredity 56:191–202.

Mallet, J. 1993. Speciation, raciation, and color pattern evolution in *Heliconius* butterflies: evidence from hybrid zones. Pp. 226–260 *in* R. G. Harrison, ed. Hybrid zones and the evolutionary process. Oxford University Press, New York, NY.

Mallet, J. 1989. The genetics of warning colour in Peruvian hybrid zones of *Heliconius erato* and *H . melpomene*. Proc. R. Soc. London. Ser. B, Biol. Sci. 236:163–185.

Mallet, J., and N. Barton. 1989. Strong natural selection in a warning-color hybrid zone. Evolution 43:421–431.

Mallet, J., N. Barton, G. Lamas, J. Santisteban, M. Muedas, and H. Eeley. 1990. Estimates of selection and gene flow from measures of cline width and linkage disequilibrium in *Heliconius* hybrid zones. Genetics 124:921–936.

Mallet, J., W. McMillan, and C. Jiggins. 1998a. Estimating the mating behavior of a pair of hybridizing *Heliconius* species in the wild. Evolution 52:503–510.

Mallet, J., W. McMillan, and C. Jiggins. 1998b. Mimicry and warning color at the boundary between races and species. Pp. 390–403 *in* D. J. Howard and S. H. Berlocher, eds. Endless forms: species and speciation. Oxford University Press.

Martin, A., R. Papa, N. J. Nadeau, R. I. Hill, B. A. Counterman, G. Halder, C. D. Jiggins, M. R. Kronforst, A. D. Long, W. O. McMillan, and R. D. Reed. 2012. Diversification of complex butterfly wing patterns by repeated regulatory evolution of a *Wnt* ligand. Proc. Natl. Acad. Sci. U. S. A. 109:12632–7.

Martin, S. H., K. K. Dasmahapatra, N. J. Nadeau, C. Salazar, J. R. Walters, F. Simpson, M. Blaxter, A. Manica, J. Mallet, and C. D. Jiggins. 2013. Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. Genome Res. 23:1817–1828.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297–1303.

McMillan, W. O., C. D. Jiggins, and J. Mallet. 1997. What initiates speciation in passion-vine butterflies? Proc. Natl. Acad. Sci. U. S. A. 94:8628–8633.

Merrill, R. M., A. Chia, and N. J. Nadeau. 2014. Divergent warning patterns contribute to assortative mating between incipient *Heliconius* species. Ecol. Evol. 4:911–7.

Merrill, R. M., Z. Gompert, L. M. Dembeck, M. R. Kronforst, W. O. McMillan, and C. D. Jiggins. 2011a. Mate preference across the speciation continuum in a clade of mimetic butterflies. Evolution 65:1489–500.

Merrill, R. M., B. Van Schooten, J. A. Scott, and C. D. Jiggins. 2011b. Pervasive genetic associations between traits causing reproductive isolation in *Heliconius* butterflies. Proc. Biol. Sci. 278:511–518.

Michel, A., S. Sim, T. H. Q. Powell, M. S. Taylor, P. Nosil, and J. L. Feder. 2010. Widespread genomic divergence during sympatric speciation. Proc. Natl. Acad. Sci. U. S. A. 107:9724–9729.

Nadeau, N. J., A. Whibley, R. T. Jones, J. W. Davey, K. K. Dasmahapatra, S. W. Baxter, M. A. Quail, M. Joron, R. H. ffrench-Constant, M. L. Blaxter, J. Mallet, and C. D. Jiggins. 2012. Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-scale targeted sequencing. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 367:343–353.

Naisbit, R. E., C. D. Jiggins, M. Linares, C. Salazar, and J. Mallet. 2002. Hybrid Sterility, Haldane's Rule and Speciation in *Heliconius cydno* and *H. melpomene*. Genetics 161:1517–1526.

Nosil, P. 2012. Ecological Speciation. Oxford University Press.

Nosil, P., D. J. Funk, and D. Ortiz-Barrientos. 2009. Divergent selection and heterogeneous genomic divergence. Mol. Ecol. 18:375–402.

Nylander, J. A. A. 2004. MrModeltest v2. Evolutionary Biology Centre, Uppsala University.

Papa, R., D. D. Kapan, B. A. Counterman, K. Maldonado, D. P. Lindstrom, R. D. Reed, H. F. Nijhout, T. Hrbek, and W. O. McMillan. 2013. Multi-allelic major effect genes interact with minor effect QTLs to control adaptive color pattern variation in *Heliconius erato*. PLoS One 8:e57033.

Pardo-Diaz, C., C. Salazar, S. W. Baxter, C. Merot, W. Figueiredo-Ready, M. Joron, W. O. McMillan, and C. D. Jiggins. 2012. Adaptive introgression across species boundaries in *Heliconius* butterflies. PLoS Genet. 8:e1002752.

Powell, T. H. Q., G. R. Hood, M. O. Murphy, J. S. Heilveil, S. H. Berlocher, P. Nosil, and J. L. Feder. 2013. Genetic divergence along the speciation continuum: the transition from host race to species in *Rhagoletis* (Diptera: tephritidae). Evolution 67:2561–2576.

Quek, S. P., B. A. Counterman, P. Albuquerque de Moura, M. Z. Cardoso, C. R. Marshall, W. O. McMillan, and M. R. Kronforst. 2010. Dissecting comimetic radiations in *Heliconius* reveals divergent histories of convergent butterflies. Proc. Natl. Acad. Sci. U. S. A. 107:7365–70.

Rambaut, A., M. Suchard, and A. Drummond. 2013. Tracer v1.6.

Reed, R. D., R. Papa, A. Martin, H. M. Hines, B. A. Counterman, C. Pardo-Diaz, C. D. Jiggins, N. L. Chamberlain, M. R. Kronforst, R. Chen, G. Halder, H. F. Nijhout, and W. O. McMillan. 2011. *optix* drives the repeated convergent evolution of butterfly wing pattern mimicry. Science 333:1137–41.

Roesti, M., A. P. Hendry, W. Salzburger, and D. Berner. 2012. Genome divergence during evolutionary diversification as revealed in replicate lake-stream stickleback population pairs. Mol. Ecol. 21:2852–2862.

Ronquist, F., M. Teslenko, P. van der Mark, D. L. Ayres, A. Darling, S. Höhna, B. Larget, L. Liu, M. A. Suchard, and J. P. Huelsenbeck. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61:539–42.

Supple, M. A., H. M. Hines, K. K. Dasmahapatra, J. J. Lewis, D. M. Nielsen, C. Lavoie, D. A. Ray, C. Salazar, W. O. McMillan, and B. A. Counterman. 2013. Genomic architecture of adaptive color pattern divergence and convergence in *Heliconius* butterflies. Genome Res. 23:1248–1257.

Swofford, D. 2002. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, MA.

Tobler, A., D. Kapan, N. S. Flanagan, C. Gonzalez, E. Peterson, C. D. Jiggins, J. S. Johntson, D. G. Heckel, and W. O. McMillan. 2005. First-generation linkage map of the warningly colored butterfly *Heliconius erato*. Heredity 94:408–417.

Weir, B. S. 1996. Genetic data analysis II. Sinauer Associates, Sunderland, MA.

Figures

Figure 1: Expected phylogenetic relationships at the red color pattern (*D*) locus. Genomic sequences from *H. himera* and *H. erato* races are expected to cluster samples based on color pattern phenotype at the genomic region responsible for red color pattern variation. The incipient, parapatric species, *H. himera* and *H. e. cyrbia*, are expected to cluster in separate clades. Based on the red forewing band, *H. e. cyrbia* is expected to cluster with the postman phenotypes and based on the presence of red on the hindwing, *H. himera* is expected to clusters with the rayed phenotypes. The taxa are listed by hybrid zone pairs, noting the country where the hybrid zone is located. The hybrid zone in Panama is between two postman phenotypes that vary in the presence or absence of the yellow hindwing bar, which is controlled by a separate, unlinked locus. *H. erato* within species hybrid zone races are Ecuador: *H. e. lativitta* (rayed), *H. e. notabilis* (postman); Peru: *H. e. emma* (rayed), *H. e. favorinus* (postman); French Guiana: *H. e. erato* (rayed), *H. e. hydara* (postman); Panama: *H. e. petiverana* (postman), *H. e. hydara* (postman). The interspecific hybrid zone is between the incipient species *H. himera* and *H. e. cyrbia* in Ecuador. The topology of this expected tree matches the tree inferred from empirical data at the red color pattern locus (Figure S2), but both differ fundamentally from the tree at loci unlinked to color pattern (Figure S1).

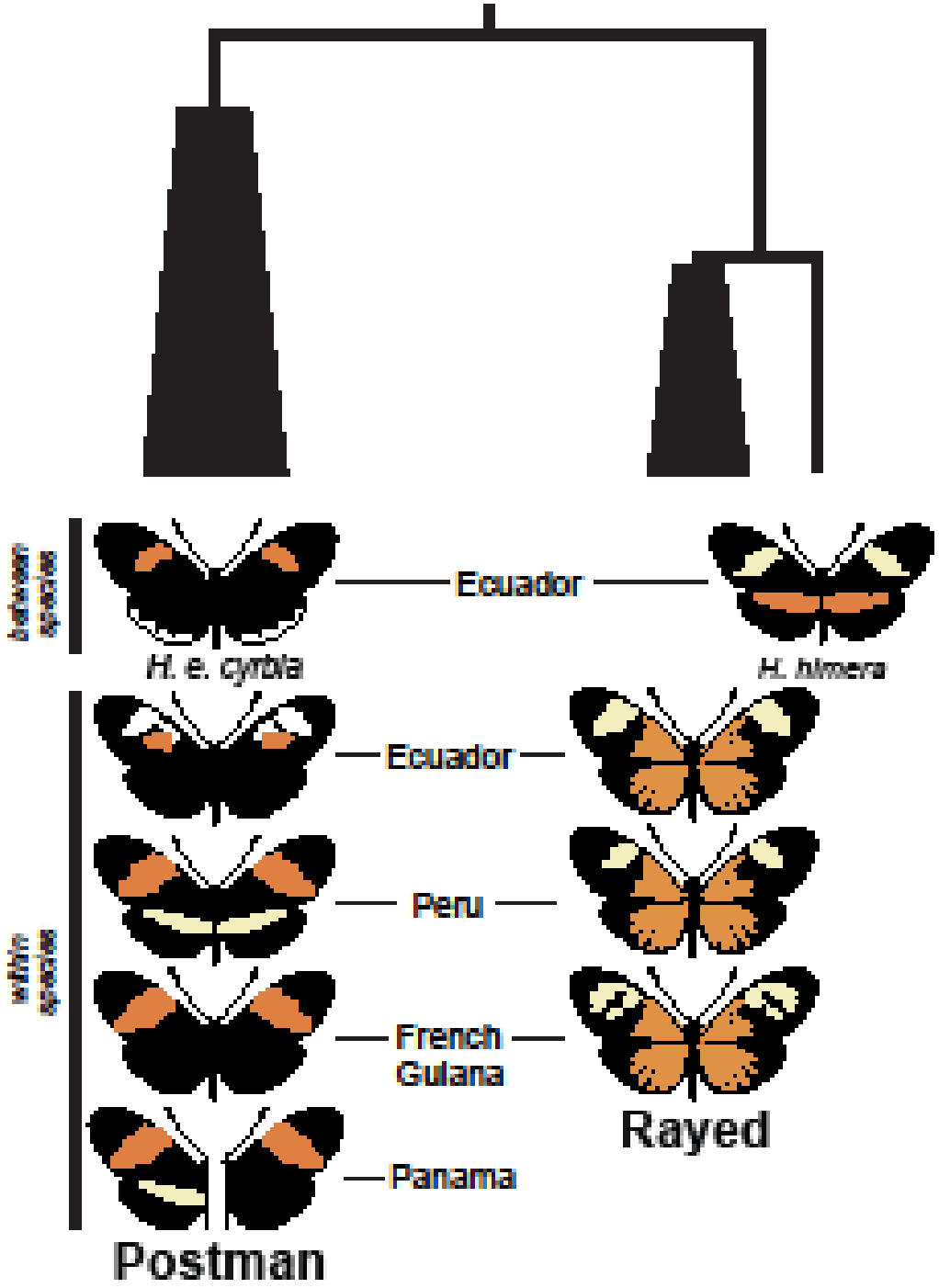
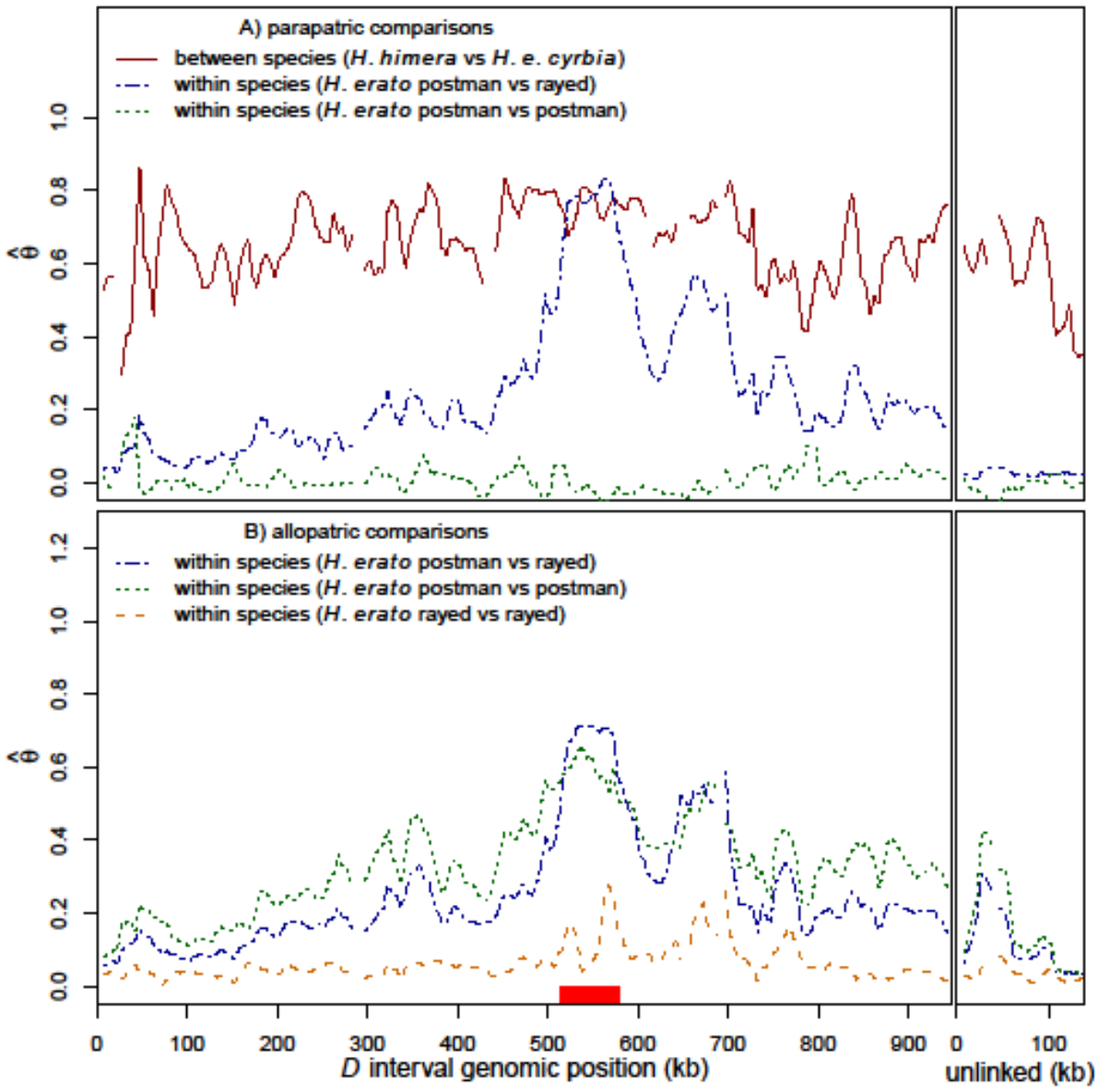

between species
Ecuador
H. e. cyrbia
H. hilmera
within species
Ecuador
Peru
French Guiana
Panama
Postman
Rayed

Figure 2: Genomic divergence across the red color pattern (*D*) interval and unlinked loci.

Sliding window (15-kb window size, 5-kb step size) genomic differentiation across the red (*D*) interval and genomic regions unlinked to color pattern. The box on the x-axis indicates the previously identified 65-kb functional region modulating red color pattern variation (Supple et al. 2013). See Table S2 for the taxa pairs included in each comparison, samples sizes, and estimates of baseline divergence from intervals unlinked to color pattern. **(A)** Parapatric comparisons show elevated divergence between the incipient species (solid line) at the color pattern locus and loci unlinked to color pattern. The within species comparisons between divergent taxa show a large peak of divergence at the functional region and low baseline divergence (dotted-dashed line). The within species comparisons between taxa with identical red phenotypes show low divergence throughout, with no discernable peaks of divergence (dotted line). **(B)** The allopatric within species comparisons show peaks of divergence at the functional region between both divergent taxa and mimetic taxa, with the rayed versus rayed comparison (dashed line) showing much smaller peaks.


A) parapatric comparisons
between species (H. himera vs H. e. cyrbia)
within species (H. erato postman vs rayed)
within species (H. erato postman vs postman)
B) allopatric comparisons
within species (H. erato postman vs rayed)
within species (H. erato postman vs postman)
within species (H. erato rayed vs rayed)
θ̂
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
100
200
300
400
500
600
700
800
900
D interval genomic position (kb)
unlinked (kb)

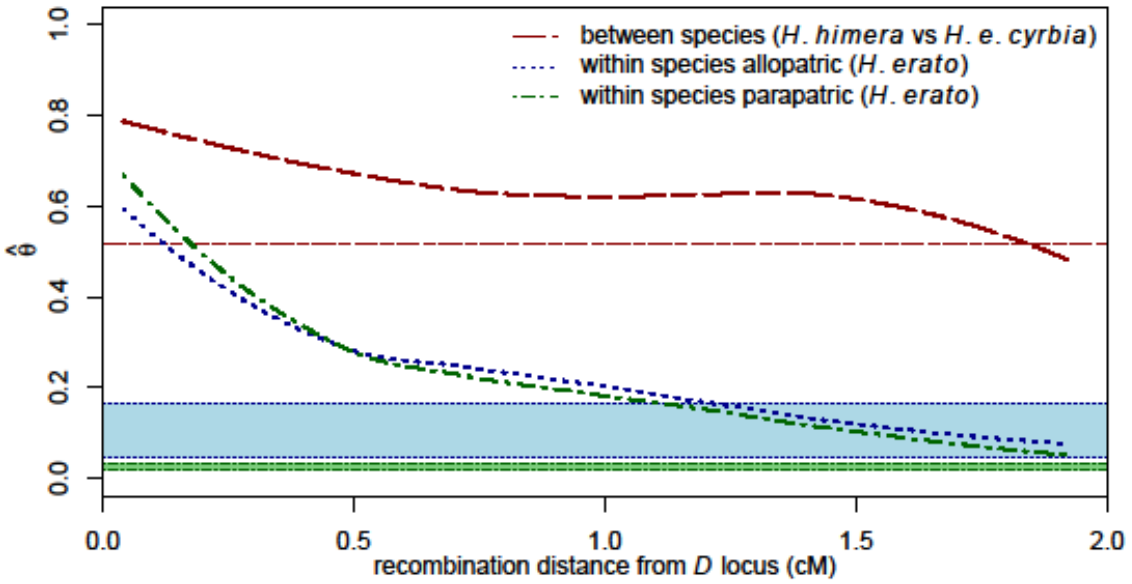


Figure 3: Decay of divergence with recombination distance from the causative locus between taxa pairs with divergent phenotypes.

The curves represent the decay of genomic divergence with distance from the functional locus. See Table S2 for the taxa pairs included in each comparison, samples sizes, and estimates of baseline divergence from intervals unlinked to color pattern. The comparisons are between incipient species (*H. himera* versus *H. erato cyrbia*; red dashed), average of within species parapatric postman versus rayed pairs (green dotted-dashed), and average of within species allopatric postman versus rayed pairs (blue dotted). The horizontal red dashed line represents the background genomic divergence between the incipient species at loci unlinked to color pattern. The blue and green shaded boxes are the range of baseline divergences at unlinked loci for within species allopatric and parapatric comparisons, respectively. The divergence at both the color pattern locus and unlinked genomic regions is substantially higher between species than it is within species.

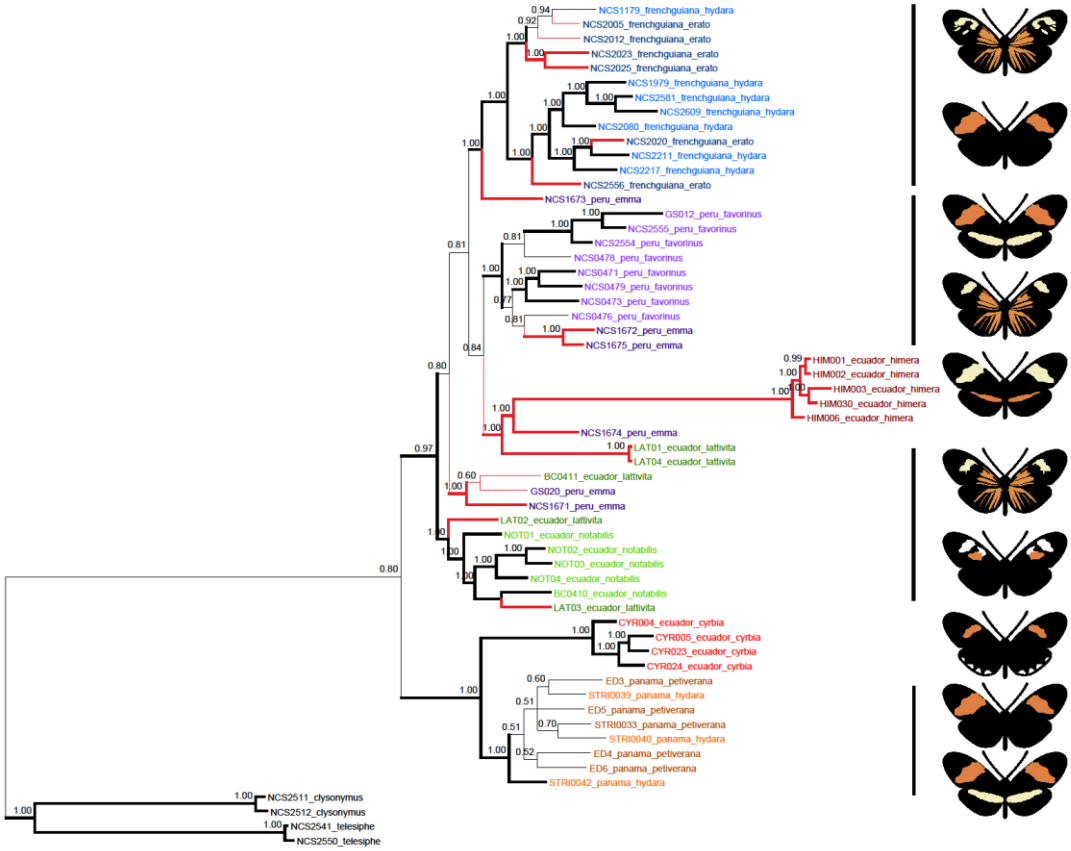
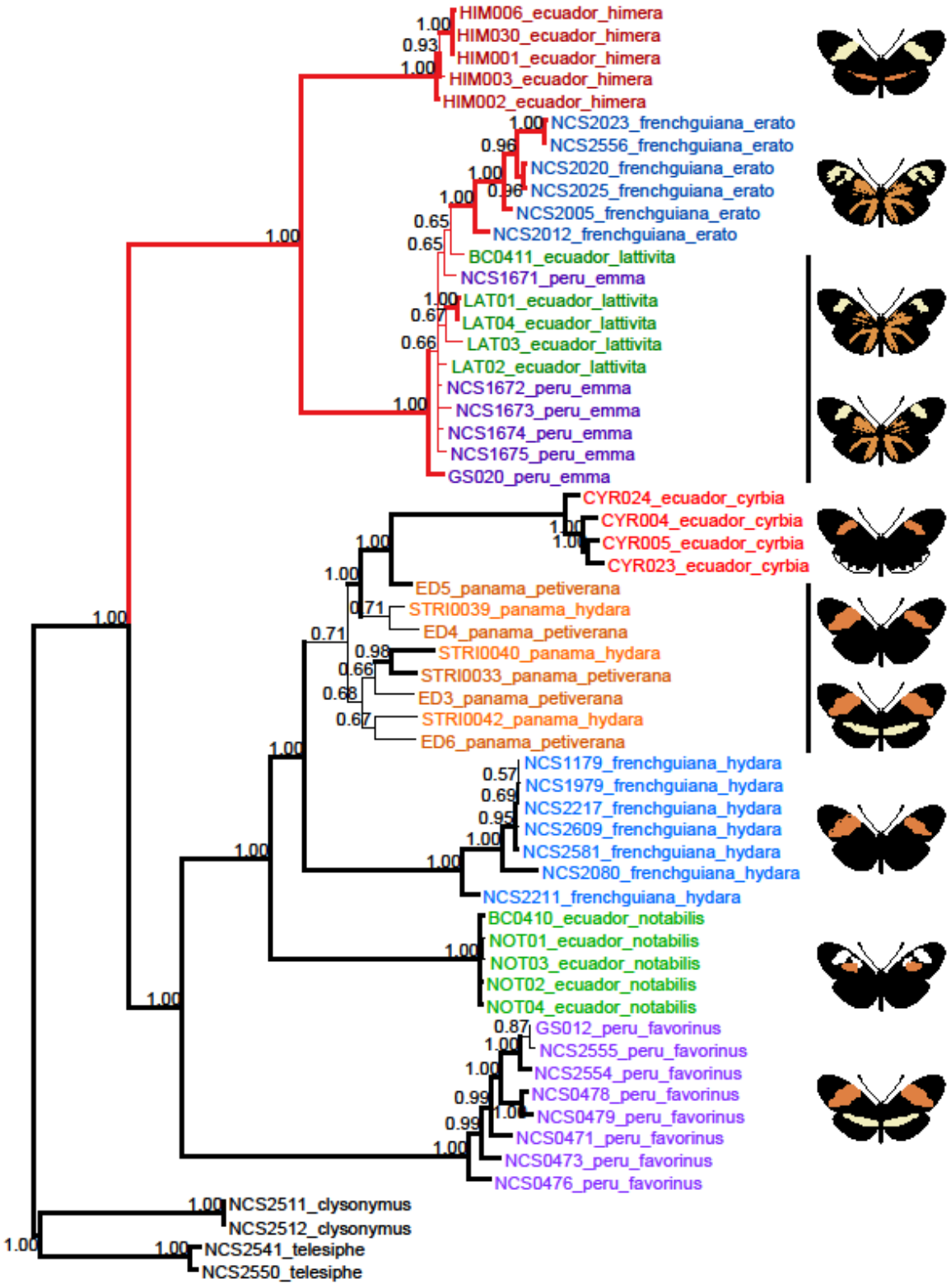


Figure S1: Bayesian phylogenetic relationships at loci unlinked to color pattern.

Phylogenetic clustering across loci unlinked to color pattern variation. Red branches indicate rayed lineages. Nodal values are posterior probabilities of clade support, with bold braches indicating a node with >95% support. Sample names (ID, hybrid zone, race) are indicated at the terminal nodes, color coded by race. Taxa sort strongly by geographic location. *Heliconius himera* clusters within *H. erato*, rather than sister to the radiation.

Figure S2: Bayesian phylogenetic relationships at the red color pattern (*D*) locus. Phylogenetic clustering across a 65-kb region that modulates red color pattern variation. Red branches indicate rayed lineages. Nodal values are posterior probabilities of clade support, with bold braches indicating a node with >95% support. Sample names (ID, hybrid zone, race) are indicated at the terminal nodes, color coded by race. The postman and the rayed clades form reciprocally monophyletic lineages, with *H. himera* clustering with the rayed races and *H. e. cyrbia* clustering with the postman races. Within each phenotype, samples cluster by geography. *Heliconius himera* clusters within *H. erato,* rather than sister to the radiation.


HIM006_ecuador_himera
HIM030_ecuador_himera
HIM001_ecuador_himera
HIM003_ecuador_himera
HIM002_ecuador_himera
NCS2023_frenchguiana_erato
NCS2556_frenchguiana_erato
NCS2020_frenchguiana_erato
NCS2025_frenchguiana_erato
NCS2005_frenchguiana_erato
NCS2012_frenchguiana_erato
BC0411_ecuador_lattivita
NCS1671_peru_emma
LAT01_ecuador_lattivita
LAT04_ecuador_lattivita
LAT03_ecuador_lattivita
LAT02_ecuador_lattivita
NCS1672_peru_emma
NCS1673_peru_emma
NCS1674_peru_emma
NCS1675_peru_emma
GS020_peru_emma
CYR024_ecuador_cyrbia
CYR004_ecuador_cyrbia
CYR005_ecuador_cyrbia
CYR023_ecuador_cyrbia
ED5_panama_petiverana
STRI0039_panama_hydara
ED4_panama_petiverana
STRI0040_panama_hydara
STRI0033_panama_petiverana
ED3_panama_petiverana
STRI0042_panama_hydara
ED6_panama_petiverana
NCS1179_frenchguiana_hydara
NCS1979_frenchguiana_hydara
NCS2217_frenchguiana_hydara
NCS2609_frenchguiana_hydara
NCS2581_frenchguiana_hydara
NCS2080_frenchguiana_hydara
NCS2211_frenchguiana_hydara
BC0410_ecuador_notabilis
NOT01_ecuador_notabilis
NOT03_ecuador_notabilis
NOT02_ecuador_notabilis
NOT04_ecuador_notabilis
GS012_peru_favorinus
NCS2555_peru_favorinus
NCS2554_peru_favorinus
NCS0478_peru_favorinus
NCS0479_peru_favorinus
NCS0471_peru_favorinus
NCS0473_peru_favorinus
NCS0476_peru_favorinus
NCS2511_clysonymus
NCS2512_clysonymus
NCS2541_telesiphe
NCS2550_telesiphe

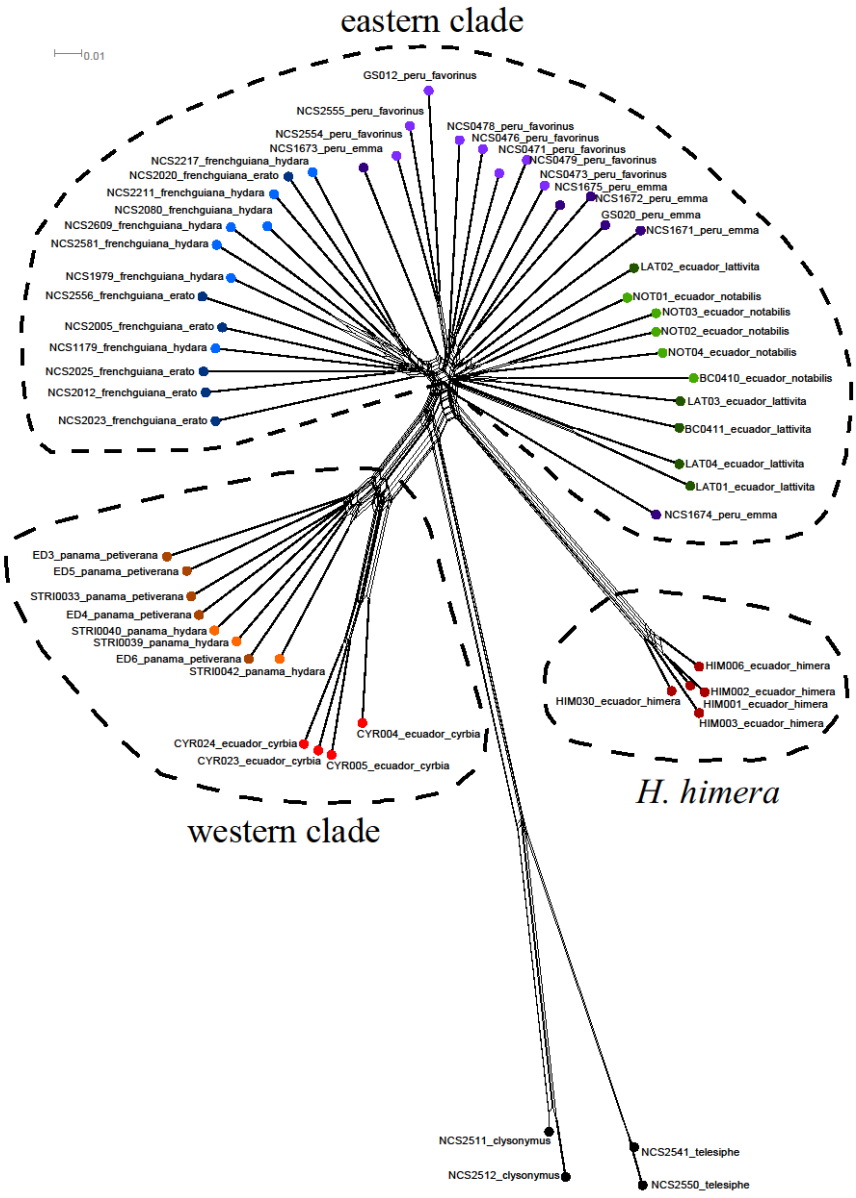


Figure S3: Phylogenetic network at loci unlinked to color pattern.

Neighbor-net splits tree network at loci unlinked to color pattern. Terminal nodes indicate voucher number, hybrid zone, and race of each sample. *Heliconius himera* shows a stronger affinity to taxa east of the Andes than the taxa from west of the Andes, which cluster separately.

Table S1: Samples and sequencing data

taxa (location)	sample ID	geolocation	number of paired end reads	mapped reads (%)	properly mapped pairs (% of mapped)	median coverage	positions genotyped (%)		SNPs* per genotyped position (%)	
							all reference	*D* interval	all reference	*D* interval
H. himera (Ecuador)	HIM001	04°16'34''S 79°11'45''W	47953654	8.8	71.7	17	45.8	51.7	4.3	3.6
	HIM002	04°16'34''S 79°11'45''W	52114768	8.8	71.4	19	47.8	54.2	4.4	3.7
	HIM003	04°16'34''S 79°11'45''W	54102534	8.5	71.7	20	49.3	55.5	4.9	3.8
	HIM006	04°16'34''S 79°11'45''W	38797900	8.8	70.8	14	41.8	47.5	4.1	3.5
	HIM030	04°16'34''S 79°11'45''W	36866545	8.8	70.2	13	39.7	45.3	4.1	3.3
H. erato cyrbia (Ecuador)	CYR004	03°43'35''S 79°50'12''W	58110947	10.8	74.2	21	52.0	59.4	4.8	3.7
	CYR005	03°43'35''S 79°50'12''W	53098890	10.7	74.0	20	50.2	57.3	4.5	3.5
	CYR023	03°43'35''S 79°50'12''W	52694366	9.9	72.1	20	52.4	59.2	4.7	3.6
	CYR024	03°43'35''S 79°50'12''W	52080755	10.9	74.3	19	49.3	56.6	4.6	3.5
H. clysonymus (Peru)	NCS2511	00°42'46''S 77°44'27''W	45312108	7.1	63.5	8	25.0	28.4	5.0	4.1
	NCS2512	00°42'46''S 77°44'27''W	54844792	7.6	65.3	9	28.1	32.2	5.5	4.6
H. telesiphe (Peru)	NCS2541	00°43'04''S 77°40'56''W	53997408	7.3	66.4	10	28.9	33.6	5.3	4.4
	NCS2550	00°42'46''S 77°44'27''W	59716322	7.1	65.0	12	32.2	37.0	5.7	4.8

Table S2: Taxa pairs for divergence analyses

geographic relationship	phenotypic comparison	taxon 1			taxon 2			baseline differentiation (95% CI)
		name (location)	phenotype	sample size	name (location)	phenotype	sample size	
parapatric	himera vs cyrbia	*H. himera* (Ecuador)	rayed	5	*H. e. cyrbia* (Ecuador)	postman	4	0.519 (0.507, 0.531)
	postman vs rayed	*H. e. emma* (Peru)	rayed	6	*H. e. favorinus* (Peru)	postman	8	0.022 (0.019, 0.025)
		H. e. lativitta (Ecuador)	rayed	5	*H. e. notabilis* (Ecuador)	postman	5	0.031 (0.028, 0.035)
		H. e. erato (French Guiana)	rayed	6	*H. e. hydara* (French Guiana)	postman	7	0.020 (0.017,0.023)
	postman vs postman	*H. e. petiverana* (Panama)	postman	5	*H. e. hydara* (Panama)	postman	3	0.001 (0.000, 0.006)
allopatric	postman vs postman	*H. e. favorinus* (Peru)	postman	8	*H. e. notabilis* (Ecuador)	postman	5	0.061 (0.057, 0.066)
		H. e. favorinus (Peru)	postman	8	*H. e. hydara* (French Guiana)	postman	7	0.079 (0.074, 0.084)
		H. e. favorinus (Peru)	postman	8	*H. e. petiverana* & *H. e. hydara* (Panama)	postman	8	0.191 (0.182, 0.200)
		H. e. notabilis (Ecuador)	postman	5	*H. e. hydara* (French Guiana)	postman	7	0.092 (0.086, 0.098)
		H. e. notabilis (Ecuador)	postman	5	*H. e. petiverana* & *H. e. hydara* (Panama)	postman	8	0.161 (0.152, 0.169)
		H. e. hydara (French Guiana)	postman	7	*H. e. petiverana* & *H. e. hydara* (Panama)	postman	8	0.174 (0.165, 0.182)
	rayed vs rayed	*H. e. emma* (Peru)	rayed	6	*H. e. lativitta* (Ecuador)	rayed	5	0.027 (0.024, 0.031)
		H. e. emma (Peru)	rayed	6	*H. e. erato* (French Guiana)	rayed	6	0.048 (0.044, 0.052)
		H. e. lativitta (Ecuador)	rayed	5	*H. e. erato* (French Guiana)	rayed	6	0.053 (0.049, 0.058)

Table S2 continued

geographic relationship	phenotypic comparison	taxon 1			taxon 2			baseline differentiation (95% CI)
		name (location)	phenotype	sample size	name (location)	phenotype	sample size	
allopatric	postman vs rayed	*H. e. emma* (Peru)	rayed	6	*H. e. notabilis* (Ecuador)	postman	5	0.047 (0.042, 0.051)
		H. e. emma (Peru)	rayed	6	*H. e. hydara* (French Guiana)	postman	7	0.072 (0.067, 0.077)
		H. e. emma (Peru)	rayed	6	*H. e. petiverana* & *H. e. hydara* (Panama)	postman	8	0.161 (0.152, 0.169)
		H. e. lativitta (Ecuador)	rayed	5	*H. e. favorinus* (Peru)	postman	8	0.052 (0.048, 0.056)
		H. e. lativitta (Ecuador)	rayed	5	*H. e. hydara* (French Guiana)	postman	7	0.079 (0.075, 0.085)
		H. e. lativitta (Ecuador)	rayed	5	*H. e. petiverana* & *H. e. hydara* (Panama)	postman	8	0.159 (0.150, 0.168)
		H. e. erato (French Guiana)	rayed	6	*H. e. favorinus* (Peru)	postman	8	0.056 (0.052, 0.060)
		H. e. erato (French Guiana)	rayed	6	*H. e. notabilis* (Ecuador)	postman	5	0.068 (0.063, 0.073)
		H. e. erato (French Guiana)	rayed	6	*H. e. petiverana* & *H. e. hydara* (Panama)	postman	8	0.164 (0.156, 0.173)

CONCLUSIONS

In this dissertation, I examined genomic and phenotypic variation across the adaptive radiation of *Heliconius* butterflies to provide insights into the origins of biodiversity. I used natural variation to identify a regulatory region that modulates a major phenotypic switch in red color pattern variation and demonstrated that distinct modular enhancers within this region control different red color pattern elements. I elucidated the path from genotype to phenotype by identifying candidate upstream transcription factors that differentially bind to this regulatory region, driving color pattern variation. I explored the evolution of novel phenotypes and showed that within a closely related clade, novel phenotypes can arise through shuffling of modular enhancers. I examined the origins of mimicry by demonstrating that two distantly related co-mimetic species in the genus use the same genomic region to generate their mimetic phenotypes, but that the allele evolved independently in each species. I examine the evolutionary processes that can drive genomic divergence and showed that selection on multiple loci likely drove rapid genomic divergence early in speciation.

Evaluation of candidate transcription factors

I have identified four strong candidate transcription factors modulating the presence or absence of hindwing rays. Additional work needs to be done to confirm their role in color pattern variation by further linking changes in DNA sequence to changes in the spatial distribution of red across the wing. These transcription factors are hypothesized to lay down

a pre-pattern across the wing surfaces. Races with differences in red color pattern elements will have differential binding of these transcription factors to the regulatory region of *optix*, driving differential expression of *optix* and ultimately variation in red color pattern elements. The first aspect that needs further investigation is to determine if our *in silico* predicted transcription factor binding sites occur *in vivo*. ATAC-seq can be used to infer DNA binding sites. By running parallel ATAC-seq experiments on closely related taxa that vary in the presence or absence of hindwing rays, predicted transcription factor binding sites with differential binding affinities between phenotypes can be confirmed. The next aspect that needs to be investigated further is whether the pattern of spatial expression of the candidates across the *Heliconius* wing is consistent with the expected pre-pattern. Three of these candidates (*ara*, *vvl*, *ct*) have known expression across the *Drosophila* wing. In situ hybridizations will enable confirmation of this pattern of expression in *Heliconius* and determine the expression pattern in the other candidate. Finally, the ultimate test to connect genotype to phenotype is transgenic manipulation. This has proven challenging in butterflies, but new techniques are providing hope that we will soon be able to manipulate DNA in a controlled manner to demonstrate the effect of specific genomic variants on phenotype.

Candidates for other color pattern elements

The candidate transcription factors described above modulate the presence or absence of hindwing rays. There are additional red color pattern elements for which enhancer regions have been determined, but candidates have not been identified. Additionally, many races

vary in the presence or absence of a yellow hindwing bar, which is controlled by an unlinked locus. The genetic basis of these elements can be dissected using the methods presented here. Additionally, RNA-seq can be used to identify candidates as well. Candidate transcription factors that establish the pre-pattern can be identified by dissecting pupal wings using persistent landmarks to separate parts of the wing that will mature to different color pattern elements. This can identify genes that are differentially expressed between regions of the wing that are fated to become different colors. RNA-seq can also be used to identify a primary candidate gene for the yellow hindwing bar by both comparing the black and yellow portions of the hindwing and by comparing hindwing gene expression in the yellow bar region between hybridizing races that differ in the presence or absence of the bar.

Hotspots of mimetic variation

Mimicry offers a powerful way to test the repeatability of evolution. The loci modulating wing color patterns in *Heliconius* have been shown to be hotspots of adaptation. However, the question remains at what level are they hotspots? I have shown that two mimetic species independently evolved their matching phenotypes using the same broad genomic region. The identification of the functional sites in each species will enable determination of whether it is the same SNP, the same upstream transcription factor, or the same gene network that produces the mimetic phenotypes.

Concluding thoughts

By examining the relationship between genomic variation and phenotypic diversity at multiple loci from multiple species across the *Heliconius* speciation continuum, we can begin to understand the generality of our results within *Heliconius*. These results can then be added to a growing body of empirical studies in other systems to provide a broad perspective on the connection between genotype and phenotype. Collectively, these studies will help us understand the evolution of biodiversity.