Cell Reports
Article

Hybridization Reveals the Evolving Genomic Architecture of Speciation

Marcus R. Kronforst,[1,*] Matthew E.B. Hansen,[2] Nicholas G. Crawford,[3] Jason R. Gallant,[3] Wei Zhang,[1] Rob J. Kulathinal,[2] Durrell D. Kapan,[4,5] and Sean P. Mullen[3,*]
[1]Department of Ecology and Evolution, University of Chicago, Chicago, IL 60637, USA
[2]Department of Biology, Temple University, Philadelphia, PA 19122, USA
[3]Department of Biology, Boston University, Boston, MA 02215, USA
[4]Department of Entomology and Center for Comparative Genomics, California Academy of Sciences, San Francisco, CA 94118, USA
[5]Center for Conservation and Research Training, Pacific Biosciences Research Center, University of Hawaii at Manoa, Honolulu, HI 96822, USA
*Correspondence: mkronforst@uchicago.edu (M.R.K.), smullen@bu.edu (S.P.M.)
http://dx.doi.org/10.1016/j.celrep.2013.09.042

SUMMARY

The rate at which genomes diverge during speciation is unknown, as are the physical dynamics of the process. Here, we compare full genome sequences of 32 butterflies, representing five species from a hybridizing *Heliconius* butterfly community, to examine genome-wide patterns of introgression and infer how divergence evolves during the speciation process. Our analyses reveal that initial divergence is restricted to a small fraction of the genome, largely clustered around known wing-patterning genes. Over time, divergence evolves rapidly, due primarily to the origin of new divergent regions. Furthermore, divergent genomic regions display signatures of both selection and adaptive introgression, demonstrating the link between microevolutionary processes acting within species and the origin of species across macroevolutionary timescales. Our results provide a uniquely comprehensive portrait of the evolving species boundary due to the role that hybridization plays in reducing the background accumulation of divergence at neutral sites.

INTRODUCTION

Gene flow prevents the accumulation of genetic differentiation among populations, and as a result, hybridization is often viewed as an impediment to the speciation process (Mayr, 1963). However, increasing evidence across a variety of plant and animal taxa suggests that speciation with gene flow may be more common than previously recognized (Mallet, 2005). Such examples of divergence with gene flow argue for a critical role of divergent selection in the origin of species (Via, 2009). Importantly, these systems also offer an opportunity to identify the genetic changes that underlie species-level divergence, because background differentiation at neutral sites is reduced by persistent hybridization and interspecific gene flow (Nosil et al., 2009; Via, 2009). This approach circumvents a classic problem in the study of speciation: distinguishing the subset of the genome that plays a critical role in the origin of species from the many changes that accumulate after the evolution of reproductive isolation.

Recent studies have documented genome-wide patterns of divergence between closely related sister taxa (Ellegren et al., 2012; Kulathinal et al., 2009; Lawniczak et al., 2010; Nadeau et al., 2013; Neafsey et al., 2010; Staubach et al., 2012; Turner et al., 2005), but the fundamental question of how divergence evolves throughout the process of speciation remains largely unexplored. Theoretical work suggests that divergent genomic regions protect adjacent, tightly linked neutral polymorphism and enhance genetic hitchhiking locally due to reduced migration (Feder et al., 2012a, 2012b; Feder and Nosil, 2010; Nosil et al., 2009). The expected outcome of this is that as phylogenetic distance increases, divergent genomic regions should increase in physical size, leading to reduced genome-wide patterns of gene flow and increased differentiation. This prediction has not been rigorously investigated using whole-genome sequence data, and it remains unclear whether such islands of divergence increase in size, how quickly they grow, or how the number, density, and chromosomal distribution of divergent regions change over time (Feder et al., 2012a; Nadeau et al., 2013; Nosil et al., 2009).

The butterfly genus *Heliconius* provides a particularly useful system to explore the dynamics of genome evolution during speciation, because this recent radiation has produced a continuum of co-occurring taxa at different stages of speciation. *Heliconius* is a diverse group of 45 species, well known for bold color patterns and widespread wing-pattern mimicry (Brown, 1981; Joron et al., 2006a; Papa et al., 2008; Sheppard et al., 1985). Across the Neotropics, local *Heliconius* communities generally consist of 10 to 15 species, with four or five of these coming from a subclade of closely related species that are known to hybridize (Mallet et al., 2007). In Costa Rica, the hybridizing *Heliconius* community consists of five species (Figure 1A); sister

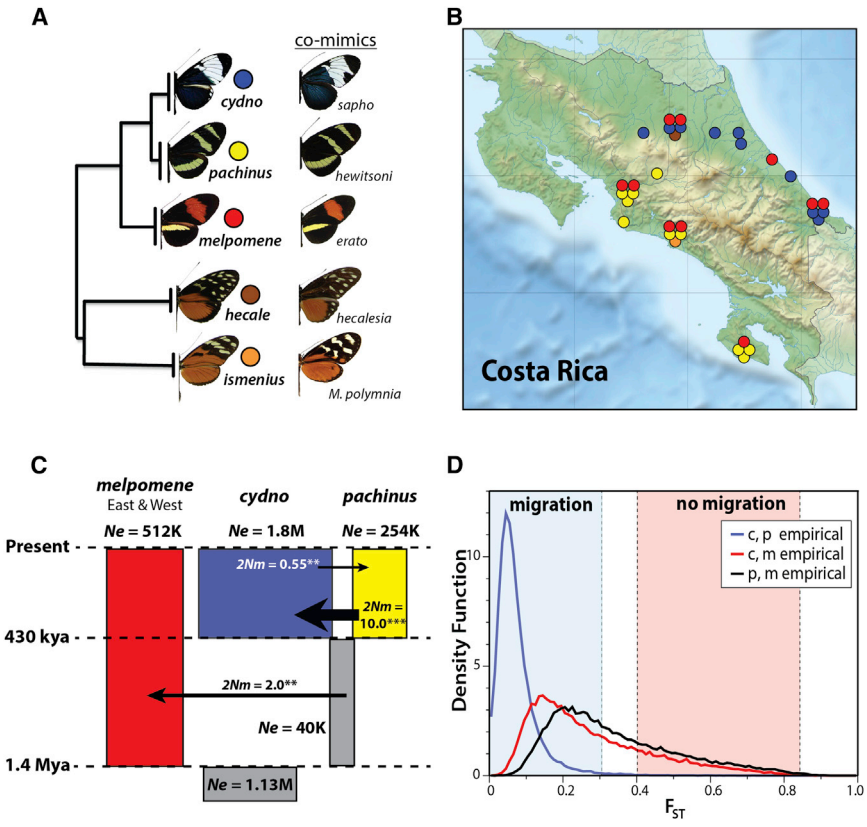

Figure 1. Five Hybridizing Species of *Heliconius* in Costa Rica Demonstrate Varying Levels of Genome-wide Differentiation and Gene Flow

(A) Phylogeny of *H. cydno*, *H. pachinus*, and *H. melpomene*, along with their outgroup species, *H. hecale* and *H. ismenius*, based on genome sequence data. Their distantly related comimics are shown on the right.

(B) Collection sites of individual samples, color-coded according to (A).

(C) History of divergence and gene flow among focal taxa based on analysis of genome-wide data using IMa2 (N_e, effective population size; *2Nm*, population migration rate).

(D) Empirical F_{ST} distributions among *H. cydno*, *H. pachinus*, and *H. melpomene*, with shading indicating F_{ST} distributions based on coalescent simulations with and without interspecific gene flow.

species *H. cydno* and *H. pachinus* are restricted to opposite coastal drainages with a contact zone in the center of the country, while *H. melpomene*, *H. hecale*, and *H. ismenius* are distributed throughout (Figure 1B).

These species represent different points on the trajectory of speciation (Mallet et al., 1998; Merrill et al., 2011). For instance, *H. cydno* and *H. pachinus* are closely related, ecologically similar species that are completely interfertile, producing viable, fertile hybrids in captivity (Gilbert, 2003; Kronforst et al., 2006a, 2006c). In nature, however, there is pronounced reproductive isolation between them, mediated by a combination of their largely parapatric distributions, divergent mimicry phenotypes that generate extrinsic postzygotic isolation, and strong assortative mate preferences that generate sexual isolation (Kronforst and Gilbert, 2008; Kronforst et al., 2007a, 2007b, 2006c). *Heliconius melpomene* is sympatric with *H. cydno* on Costa Rica's Caribbean drainage and it is sympatric with *H. pachinus* on the Pacific drainage. Comparison of *H. melpomene* to either *H. cydno* or *H. pachinus* represents a further step in the process of speciation (Mallet et al., 1998, 2011). In addition to divergent mimicry phenotypes (Merrill et al., 2012) and strong sexual isolation (Jiggins et al., 2001), *H. melpomene* and *H. cydno/pachinus* are also ecologically and behaviorally distinct (Benson, 1978; Estrada and Jiggins, 2002; Mallet and Gilbert, 1995; Smiley, 1978), and crosses between them result in Z-linked female sterility (Naisbit et al., 2002) and disruptive sexual selection against hybrids (Naisbit et al., 2001). Yet, despite strong reproductive isolation among species, they are all known to hybridize (Mallet et al., 2007), and previous analyses suggest ongoing gene flow throughout the process of speciation (Beltrán et al., 2002; Bull et al., 2006; Kronforst et al., 2006b, 2008; Martin et al., 2013).

Recent genetic work in this subclade of *Heliconius* has focused on characterizing the molecular basis of wing-pattern mimicry (Baxter et al., 2010; Joron et al., 2006b; Martin et al., 2012; Reed et al., 2011) and then examining signatures of genetic differentiation and introgression around these mimicry genes (Baxter et al., 2010; Chamberlain et al., 2011; Heliconius Genome Consortium, 2012; Nadeau et al., 2012; Pardo-Diaz et al., 2012; Reed et al., 2011). The results of this work indicate that DNA sequence variation around mimicry genes is strongly differentiated between species and subspecies with divergent mimicry phenotypes, and there is evidence that mimicry alleles have introgressed between phenotypically similar species. However, population genomic analyses outside of these mimicry genes have had less resolution because they have utilized small samples sizes and looked at only a small fraction of the genome, using either targeted sequencing of a few regions of the genome (Nadeau et al., 2012), widely spaced molecular markers (Nadeau et al., 2013), or a combination of the two (Heliconius Genome Consortium, 2012).

The recent publication of a reference genome sequence for *H. melpomene* (Heliconius Genome Consortium, 2012) now enables full genome characterization of genetic variation in *Heliconius*, permitting a complete census of genome-wide divergence associated with speciation. Here, we present whole-genome resequencing data for five sympatric hybridizing taxa with divergent mimetic wing patterns to examine how genome divergence is initiated and how it evolves over time during the process of speciation with gene flow. Our results indicate that (1) divergent natural selection acts first on a handful of color-patterning loci, triggering population divergence leading to speciation in *Heliconius*; (2) the species boundary subsequently

evolves very rapidly across the entire genome primarily due to the origin of newly divergent regions; and (3) patterns of molecular variation across the genome reflect a dynamic interplay between selection and gene flow.

RESULTS AND DISCUSSION

Substantial Interspecific Gene Flow Reduces Background Divergence among Species

Hybridization and gene flow among *Heliconius* species is well documented. Sympatric species from across our focal clade hybridize at appreciable frequencies in nature and hybrids that have been collected include both F1 and backcross hybrids (Mallet et al., 2007; Mavárez et al., 2006). Furthermore, advanced generation hybrids are common. Our previous work on the hybridizing community in Costa Rica revealed that a number of field-collected *H. cydno*, *H. pachinus*, and *H. melpomene* individuals had mixed ancestry (Kronforst, 2008; Kronforst et al., 2006b), indicating a relatively recent hybrid ancestor (Figure S1A). This hybridization appears to have resulted in long-term introgression among species as previous studies have routinely documented strong statistical evidence for interspecific gene flow (Bull et al., 2006; Kronforst, 2008; Kronforst et al., 2006b; Martin et al., 2013). In addition, there is good genetic support for (1) hybrid ancestry of field-collected individuals with recombinant wing patterns (Dasmahapatra et al., 2007), (2) at least one instance of hybrid speciation (Jiggins et al., 2008; Mavárez et al., 2006; Salazar et al., 2010), and (3) multiple instances of introgression of wing-patterning alleles across the species boundary (Heliconius Genome Consortium, 2012; Pardo-Diaz et al., 2012; Smith and Kronforst, 2013).

To examine genome-wide patterns of introgression and divergence, we sequenced the genomes of ten wild-caught samples from each of our three focal species, *H. cydno*, *H. pachinus*, and *H. melpomene*, as well as one sample from each of the two closely related outgroup species, *H. hecale* and *H. ismenius*. Each sample was sequenced to an average depth of 16× using an Illumina Hi-Seq 2000 (Tables S1 and S2). We mapped the data for each sample back to the *H. melpomene* reference genome (Heliconius Genome Consortium, 2012) and scored polymorphisms using the GATK (DePristo et al., 2011). Our final data set consisted of approximately 33 million SNPs, covering the entire genome, with over 97% of these covered in each sample (Table S2). Importantly, we selected samples for sequencing that did not show evidence of recent mixed ancestry (Figure S1A) so as to not bias our estimates of interspecific gene flow. We subsequently verified that our sequenced samples showed no recent admixture using our genome-wide SNP data (Figure S1B).

As a first step in characterizing this system, we used the isolation-with-migration model (IMa2), incorporating data from many loci sampled across the genome, to estimate the history of divergence and gene flow among species (Figure 1C; Table S3). The inferred divergence times and migration rates among species are consistent with previous results based on smaller data sets (Bull et al., 2006; Kronforst, 2008; Kronforst et al., 2006b). We further characterized the inferred demographic parameter estimates by simulating genome-scale data, with and without interspecific gene flow. Simulations including persistent interspecific gene flow yielded divergence levels similar to our observed data, whereas simulations without gene flow yielded divergence levels five to six times greater than observed (Figure 1D). Together, these results suggest that rates of gene flow among species are high and sufficient to prevent the strong, neutral genetic differentiation we would expect in the absence of introgression. In other words, interspecific gene flow appears to be partially homogenizing genetic variation in portions of the genome that are free to cross the species boundary, permitting a comprehensive investigation of how species-level divergence is initiated at the genomic level and how it subsequently evolves.

To test this hypothesis, and further document the influence of interspecific gene flow among sympatric species in Costa Rica, we compared measures of genetic divergence and allele sharing between *H. cydno* from Costa Rica and three different populations of *H. melpomene*: sympatric *H. melpomene rosina* from Costa Rica, allopatric *H. melpomene aglaope* from Peru, and allopatric *H. melpomene amaryllis* from Peru (Figure 2). The allopatric *H. melpomene* data consist of approximately 1.8 Mbp of sequence data around two mimicry loci, *B/D* and *Yb*, from four samples of each Peruvian population, which were sequenced as part of the *Heliconius* Genome Project (Heliconius Genome Consortium, 2012). The results reveal that for two different estimates of genetic divergence, F_{ST} and d_{XY}, sympatric *H. melpomene* and *H. cydno* were more similar (Figures 2A–2D). Furthermore, by using Patterson's D statistic (Durand et al., 2011) to compare patterns of derived allele sharing between populations, we found a substantial enrichment of shared derived alleles in sympatric comparisons relative to allopatric comparisons (Figures 2E and 2F), indicative of local introgression. Unlike the adaptive introgression of mimicry documented between other taxa at the *B/D* and *Yb* loci (Heliconius Genome Consortium, 2012; Pardo-Diaz et al., 2012; Smith and Kronforst, 2013), the signatures of gene flow we detected here between *H. melpomene* and *H. cydno* are not related to mimicry introgression because the two species show highly divergent phenotypes at both mimicry loci. It is important to note that these results only hint at the real rates of interspecific gene flow for three reasons. First, this analysis is based on examining sequence variation around mimicry loci, which are under divergent selection between *H. melpomene* and *H. cydno* in Costa Rica and should be (and are) resistant to interspecific gene flow (see below). Hence, the evidence for gene flow we found in these regions is likely to be much more modest than regions of the genome not linked to divergent mimicry loci. Second, we can only document gene flow that has occurred since the subspecies of *H. melpomene* split from one another, which is recent relative to the split between *H. melpomene* and *H. cydno*. Therefore, a longer history of introgression is lost in these analyses. Third, *H. melpomene aglaope* and *amaryllis* have both experienced substantial gene flow with a close relative of *H. cydno*, *H. timareta*, at the *B/D* and *Yb* loci (Heliconius Genome Consortium, 2012; Pardo-Diaz et al., 2012; Smith and Kronforst, 2013). Therefore, our allopatric *melpomene* have potentially experienced the same homogenizing effect with a *cydno*-like genome, which will artificially decrease allopatric F_{ST} and d_{XY} estimates as well as Patterson's D.

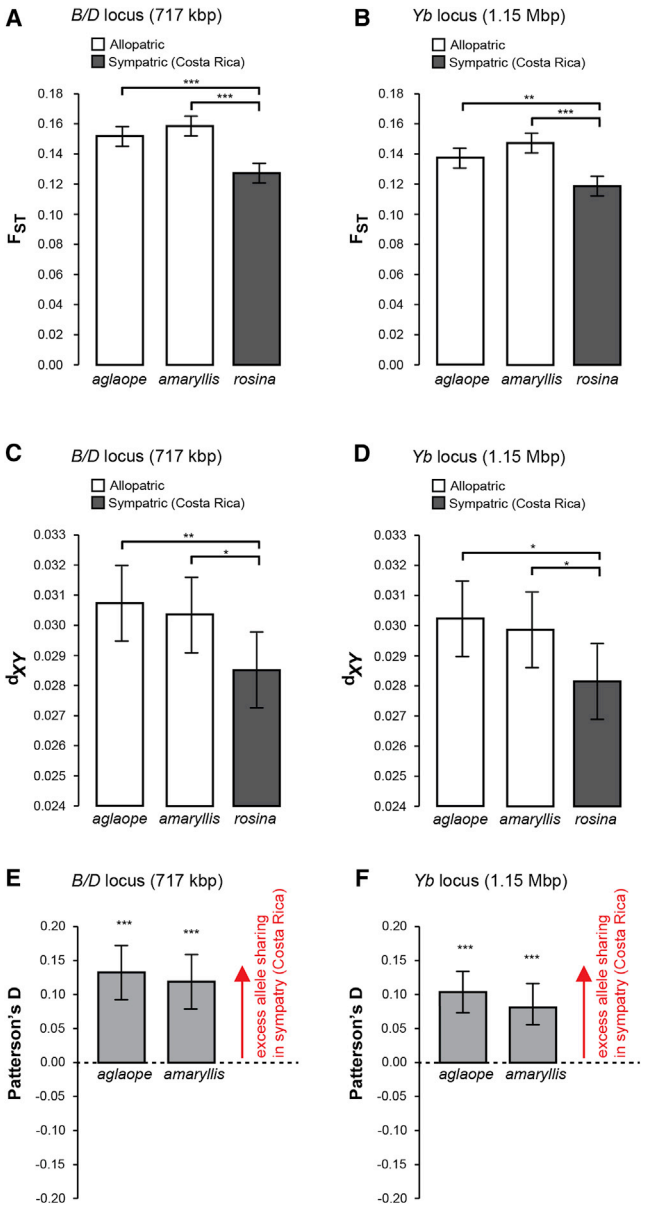

Figure 2. Additional Evidence for Gene Flow among Sympatric Species in Costa Rica

(A–D) Sympatric *H. melpomene* and *H. cydno* show reduced divergence, measured by both F_{ST} and d_{XY}, relative to allopatric comparisons, across two different regions of the genome. Error bars (indicating 95% confidence intervals) and p values are based on bootstrap resampling.

(E and F) Furthermore, Patterson's D statistic is highly elevated in these regions, indicative of biased allele sharing in sympatry due to introgression. Error bars (indicating 95% confidence intervals) and p values are based on bootstrap resampling. *p < 0.05, **p < 0.01, ***p < 0.001.

Genome Divergence at the Earliest Stage of Speciation Centers on Mimicry Genes

We examined the genome-wide distribution of genetic divergence in pairwise comparisons among sympatric *H. cydno*, *H. pachinus*, and *H. melpomene* from Costa Rica. For these analyses, we calculated genetic differentiation, analysis of molecular variance (AMOVA)-based F_{ST} (Excoffier et al., 1992), for 5 kbp windows covering the entire genome and identified outliers using an empirically derived significance threshold (Figure S2). Because adjacent windows showing significant differentiation are not biologically independent (see Experimental Procedures), they were connected into larger divergent segments. Surprisingly, the comparison between the most closely related species, *H. cydno* and *H. pachinus*, revealed only 12 narrow (mean = 14 kbp) divergent regions across the genome, spanning a total of 165 kbp (Figure 3). These regions were so narrow, in fact, that they could have been missed in previous restriction-site-associated DNA (RAD) studies (Heliconius Genome Consortium, 2012; Nadeau et al., 2013), because the average marker spacing of *Heliconius* RADs has been between 27 and 39 kbp (Nadeau et al., 2013).

The distribution of divergent regions between *H. cydno* and *H. pachinus* was highly nonrandom (Fisher's exact test, $p < 0.01$; Figure S3), with eight of them mapping to the locations of known mimicry genes (Baxter et al., 2010; Chamberlain et al., 2011; Kronforst et al., 2006a, 2006c; Martin et al., 2012; Reed et al., 2011). For instance, 4 of the 12 divergent regions sit within 1 Mbp of one another on chromosome 1, in the location of a locus that controls wing color and mate preference in *H. cydno* and *H. pachinus* (Chamberlain et al., 2009; Kronforst et al., 2006c). Similarly, two divergent regions are located on chromosome 10, near the gene *WntA*, which controls melanin patterning on the forewing (Martin et al., 2012). Two additional divergent regions are on chromosome 15, in the location of the mimicry locus that controls melanin patterning on the hindwing (Joron et al., 2006b). There is a signal of enhanced differentiation around the gene *optix*, which controls red patterning in *Heliconius* (Reed et al., 2011), but it did not pass the significance threshold in the comparison between *H. cydno* and *H. pachinus*, both of which lack striking red coloration. However, it is important to note that there was significant divergence in and around *optix* in both comparisons with red-winged *H. melpomene*, which are the comparisons that have radically different alleles at this mimicry locus.

These results suggest a central role for mimicry evolution in promoting the earliest stages of speciation in *Heliconius*. This finding matches well with previous research on *Heliconius* showing that mimetic wing patterns experience strong divergent natural selection (Kapan, 2001; Mallet et al., 1990; Mallet and Barton, 1989) and that shifts in wing pattern generate reproductive isolation, both premating and extrinsic postzygotic (Chamberlain et al., 2009; Jiggins et al., 2001; Kronforst et al., 2006c; Merrill et al., 2011, 2012; Naisbit et al., 2001). The extent to which our genome-scan results overlap with previous ecological and behavioral research as well as recent positional cloning of mimicry loci is remarkable, and the intersection of these various forms of data provide compelling evidence for ecological speciation in *Heliconius* butterflies. While previous work has documented divergence around mimicry genes in *Heliconius* (Nadeau et al., 2012), our unbiased survey of the entire genome allows us to show that these loci do genuinely stand out from the rest of the genome as the initial targets of selection that then precipitate speciation.

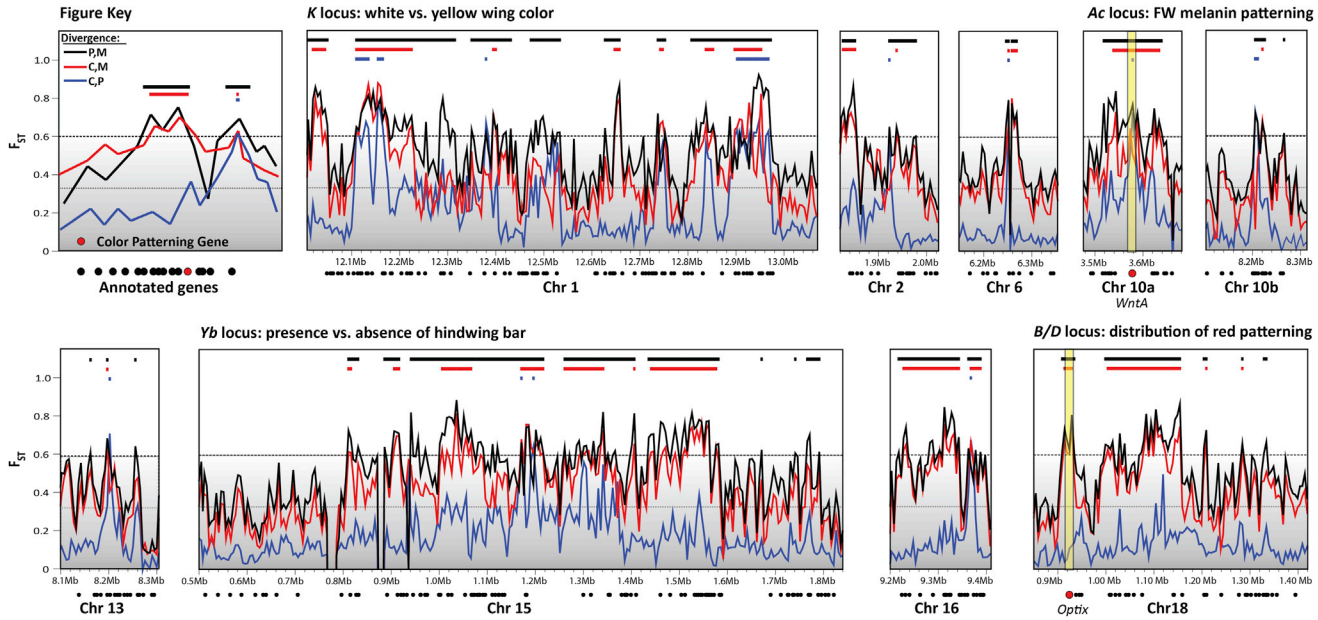

Figure 3. Signatures of Genomic Differentiation, Focusing on the 12 Regions that Are Divergent between *H. cydno* and *H. pachinus*

Known wing color patterning loci (*K*, *Ac*, *Yb*, *B/D*) are listed, as are genes *WntA* and *Optix*. F_{ST} plots and divergent segment markers are color coded by pairwise comparison.

The few highly divergent regions not linked to mimicry loci suggest additional genes that are likely to play an important role in the early stages of speciation. These four regions contain only six genes: the fatty acid synthase gene *p260* on chromosome 2, *abl-interactor 2* on chromosome 6, a fatty acid elongase gene on chromosome 13, and three clustered genes on chromosome 16 (a cytoplasmic dynein 1 intermediate chain gene similar to *short wing* in *Drosophila*, a peptide deformylase gene, and *3-hydroxyisobutyryl-coenzyme A hydrolase*). Interestingly, chromosomal inversions and the Z (sex) chromosome do not appear to play a role in maintaining this young species boundary (Table S4; Figure 4), suggesting that these factors emerge later in *Heliconius* speciation, following initial ecological divergence.

Genome-wide Divergence Grows Rapidly, Primarily due to the Origin of Newly Divergent Regions

We next examined how genome-wide divergence evolves over time. Pairwise comparisons between *H. melpomene* and either *H. cydno* or *H. pachinus* revealed 100 to 200 times more divergence, with the cumulative portion of the genome showing significant differentiation increasing from 165 kbp in the *cydno/pachinus* comparison to 19 Mbp and 33 Mbp in the two comparisons with *H. melpomene* (Table 1). The two comparisons with *H. melpomene* are not phylogenetically independent, but the comparison between *H. cydno* and *H. pachinus* is independent of the comparison between their common ancestor and *H. melpomene*. Given that only approximately 1 million years separates these divergence events, the sizeable divergence in comparisons with *H. melpomene* appears to be much more than that predicted by the modest divergence between *H. cydno* and *H. pachinus*. This result suggests a nonlinear relationship between time since speciation and the accumulation of genome-wide divergence.

To examine the evolution of divergence further, we separated our *H. melpomene* samples into two populations: one from the Caribbean drainage (east) and one from the Pacific drainage (west), and we compared them to estimate the amount of genome divergence for a within-species comparison. This intraspecific comparison yielded a single, 10 kbp divergent region that distinguished Caribbean *H. melpomene* from Pacific *H. melpomene*. We also estimated DNA sequence divergence in all comparisons as mean d_{XY}. We then plotted the aggregate portion of the genome contained in highly divergent regions, as a function of time since divergence, for the following comparisons: *melpomene* east versus *melpomene* west, *cydno* versus *pachinus*, *and melpomene* versus the common ancestor of *cydno* and pachinus (estimated as the subset of highly divergent regions shared between *melpomene* versus *cydno* and *melpomene* versus *pachinus* comparisons). This yielded three phylogenetically independent comparisons. We also plotted mean d_{XY} for the following comparisons: *melpomene* east versus *melpomene* west, *cydno* versus *pachinus*, *melpomene* versus *cydno*, and *melpomene* versus *pachinus*. Given the divergence time estimates, this analysis indicates that genome-wide divergence accumulates slowly then rapidly rises, despite a constant substitution rate (Figure 5A). The observed relationship hinges on how genome-wide differentiation occurs during the earliest stages of speciation when phenotypic and behavioral differences are apparent but most of the genome has not yet diverged. Our data suggest that an exponential model is more likely than a linear one (Akaike information criterion [AIC] = 9.06 versus 61.7, 2 df). We explored this same phenomenon using a separate approach, counting the number of fixed differences in pairwise

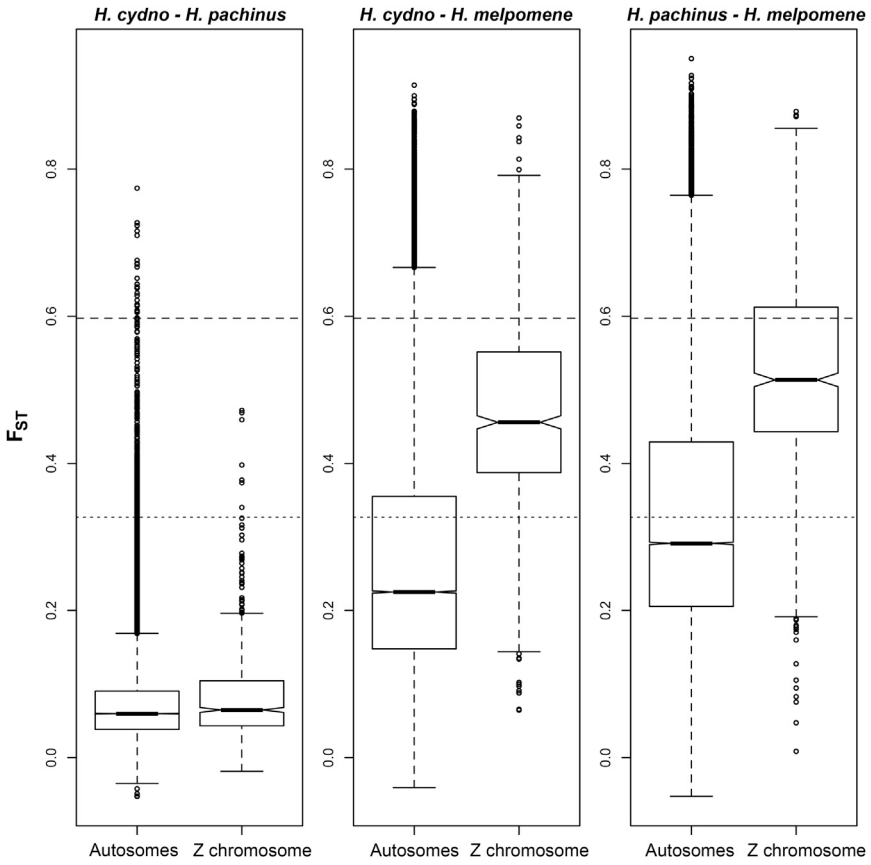

Figure 4. Z Chromosome and Autosome Divergence in Pairwise Comparisons between Species

Pairwise F_{ST} represented as boxplots with whiskers between (1) *cydno-pachnius* (left), (2) *cydno-melpomene* (middle), and (3) *pachinus-melpomene* (right) for autosomes versus the Z chromosome, highlighting elevated divergence on the Z chromosome in comparisons with *H. melpomene*. Similar distributions, separated out by chromosome, are shown in Figure S7.

comparisons. Here too, we see evidence for a nonlinear accumulation of genetic differentiation (Table S5). Our results are also consistent with a step change, whereby divergence shifts rapidly from low to high levels, but more data points will be required to determine the exact shape of this function.

Why do the rates of accumulation for fixed differences and highly differentiated portions of the genome increase over evolutionary time? We suspect that this is a direct consequence of the interspecific gene flow we have documented and how this parameter changes over time. Specifically, our results suggest that rates of hybridization and introgression decrease with time during the speciation process, as expected. The patterns we observe suggest that there is a tipping point in the rate of interspecific gene flow, below which its homogenizing effect is overwhelmed by other evolutionary processes. Hence, much of the genome remains quite similar for an extended period of time following initial divergence due to gene flow, but then genome-wide differentiation grows explosively later in the speciation process. Interestingly, the apparent exponential growth of genome-wide divergence found here reflects what has been shown for at least one byproduct of genome divergence: the accumulation of intrinsic postzygotic incompatibilities (Matute et al., 2010; Moyle and Nakazato, 2010).

Traditionally, the snowball effect for hybrid incompatibilities has been interpreted as a product of the nonlinear accumulation of epistatic interactions that are expected to result from a linear gene substitution process. While tentative, our results raise the intriguing possibility that a second phenomenon, the nonlinear rate of genome divergence, may also contribute to this snowball effect. It remains to be seen whether our observation of exponential growth holds up as additional data points are added, whether this is a general phenomenon or one that only applies to systems experiencing divergence with gene flow, and what is ultimately responsible for the phenomenon.

Our results revealed a high degree of overlap in the divergent regions across all comparisons (Figure 5B). While these comparisons are not independent, the fact that almost all of the divergent regions between closely related *H. cydno* and *H. pachinus* are also divergent in comparisons with *H. melpomene* suggests that the process of divergence is repeatable. Furthermore, while islands of divergence do grow over time, they remain quite narrow, such that the vast majority of increased genomic divergence in comparisons with *H. melpomene* results from the origin of new divergent regions (Table 1). This result is in contrast to a divergence hitchhiking model of speciation with gene flow whereby genome-wide divergence is achieved by expansion in the physical size of initial islands of divergence. The rapid origin of new divergent regions appears to be partially driven by selection (see below), but it also may be influenced by genomic hitchhiking, whereby genome-wide divergence is facilitated by reductions in gene flow resulting from divergent selection. This conclusion remains to be tested further but, intriguingly, while we found that divergent regions were distributed nonrandomly in the genome when comparing *H. cydno* and *H. pachinus*, comparisons with *H. melpomene* revealed no clustering of divergent regions among chromosomes ($p > 0.61$ in both comparisons), except on the Z chromosome, which exhibited enhanced divergence in comparisons with *H. melpomene* (Figure 4). Enhanced divergence on the Z chromosome is consistent with both a neutral process, whereby this chromosome diverges faster as a result of its reduced effective population size and the fact that an important component of reproductive isolation, hybrid female sterility, is Z linked in crosses between *H. melpomene* and *H. cydno* (Naisbit et al., 2002). Finally, we found that gene content across all divergent regions was enriched for a variety of Gene Ontology (GO) terms,

Table 1. Dynamics of Genome Divergence across the *Heliconius* Phylogeny

Species Pairing	No. of Divergent Regions	Cumulative Region Size (bp)	Average Region Size (bp)
cydno, *pachinus*	12	165,000	13,750
cydno, *melpomene*	688	18,949,219	27,542
pachinus, *melpomene*	933	32,615,794	34,958

including categories that are likely to be important in the evolutionary history of *Heliconius*, such as vision, learning, and morphogenesis (Table S6).

Genome Divergence Associated with Speciation Is Fueled by Selection and Adaptive Introgression

Given the history of interspecific gene flow among species, what is responsible for observed divergence between species? One possibility is that F_{ST} outliers are driven primarily by linked selection, including processes such as genetic hitchhiking and background selection, which will reduce intraspecific diversity and elevate F_{ST}. However, this predicts that regions of high F_{ST} should localize to regions of the genome with reduced recombination. In contrast to this prediction, our previous genetic mapping results (Kronforst et al., 2006a, 2006c) reveal that mimicry loci, which are the first regions to diverge during speciation, are not in regions of low recombination (Figure S4). Rather, we hypothesize that observed genome divergence exists because of natural (Kapan, 2001; Mallet et al., 1990; Mallet and Barton, 1989; Merrill et al., 2012) and sexual selection (Chamberlain et al., 2009; Jiggins et al., 2001; Kronforst et al., 2006c; Naisbit et al., 2001). Furthermore, the evolution of mimicry proceeds by initial, strong divergent selection followed by long-term purifying selection. If divergent genome regions generally behave like the mimicry loci, we might expect to see the combined actions of both divergent and purifying selection.

To test these hypotheses, we scanned the genome with multiple population genetic statistics and then compared divergent regions to the rest of the genome. This analysis revealed multiple, classic signatures of divergent selection as well as evidence for long-term purifying selection. For instance, divergent regions displayed (1) reduced polymorphism (Figures 6A and 6B), (2) increased derived allele frequency (Figure 6C), (3) increased linkage disequilibrium (Figure 6D), and (4) negative Tajima's D values (Figure 6E). Furthermore, consistent with a history of selective constraint following initial divergent selection, divergent regions were highly enriched for fixed differences between species (Figure 6F) yet showed reduced total sequence divergence (d_{XY}) between species (Figure 6G), the latter being a classic signature of purifying selection (Haddrill et al., 2005; Halligan and Keightley, 2006; Marais et al., 2005; Parsch, 2003).

Finally, we wanted to determine the source of genetic variation contributing to divergence. Previous work has shown a signature of shared ancestry among *Heliconius* species around wing-patterning loci (Heliconius Genome Consortium, 2012; Pardo-Diaz et al., 2012; Smith and Kronforst, 2013), suggestive of a role for introgression in the evolution of mimicry. Given the amount of hybridization among these taxa, it is possible that interspecific gene flow may have played a more general role in facilitating adaptation. To test this possibility, we scanned the genome using Patterson's D (Durand et al., 2011), a measure of shared ancestry, and then compared divergent regions to the rest of the genome. We found that divergent genome regions had more extreme values of *D*, compared to the rest of the genome (Figure 6H), and this pattern remained even after excluding divergence associated with mimicry loci (permutation test, $p < 0.001$). This introgression is likely to be adaptive because the signal of shared ancestry is enriched in these highly differentiated regions of the genome that also have multiple signatures of selection. Hence, adaptive introgression appears to be pervasive among hybridizing *Heliconius* species, potentially influencing many aspects of their biology.

Conclusions

The study of speciation is inherently challenging because it generally involves inferring a piecemeal process of divergence after reproductive isolation is complete. Systems such as *Heliconius* permit direct investigation of the genetic changes associated with speciation because species that are phenotypically well differentiated, and often sympatric, continue to hybridize, reducing divergence at neutral sites. We validated this basic expectation of divergence with gene flow and then used the resultant heterogeneity in genomic divergence to characterize the shape and depth of the species boundary as a function of divergent selection, phylogenetic distance, and hybridization. Our results provide unique insights into (1) what defines genomic regions of divergence associated with speciation, (2) how divergence evolves over time, (3) what the targets of selection are at the genetic level, and (4) the repeatability of this process. Beyond that, our work reveals important, creative roles for both selection and introgression in the origin of species. It is quite possible that this combined action of gene flow and selection may have a more general role in driving instances of rapid diversification (Seehausen, 2004). In addition, these results help elucidate the relative roles of divergent selection, divergence hitchhiking, and genome hitchhiking during the process of speciation with gene flow. Specifically, our data point to an essential role for divergent selection in initiating speciation, and we also see signs consistent with genome hitchhiking later in the process. In contrast, the role of divergence hitchhiking appears to be modest relative to these other two processes. These empirical results agree well with recent simulations in which all three processes are allowed to operate (Feder et al., 2012b; Flaxman et al., 2013). Ongoing work in this and a variety of other biological systems (Hendry et al., 2009; Kitano et al., 2009; Martin et al., 2013; McKinnon and Rundle, 2002; Michel et al., 2010; Nosil et al., 2012a, 2012b) will help expand on the generality of these results.

EXPERIMENTAL PROCEDURES

For more information, see Supplemental Experimental Procedures.

Samples

We collected 32 samples from 13 locations across Costa Rica (Table S1) and sequenced each to an average depth of 16× coverage using an Illumina HiSeq 2000 (2 × 100 paired-end sequencing). These data were aligned to the

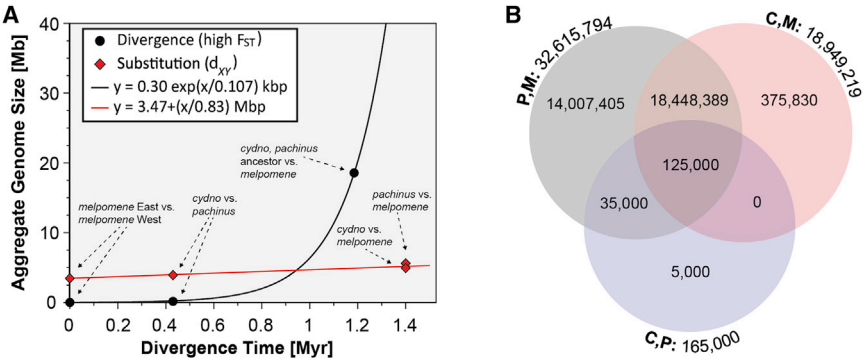

Figure 5. Dynamics of Genome-wide Divergence during Speciation

(A) Exponential growth in genome-wide divergence compared to linear substitutions as a function of divergence time. Note that d_{XY} is expressed as the total number of nucleotide substitutions across the genome, rather than a proportion, so the same y axis applies to both the divergence and substitution lines.

(B) Venn diagram of the total base-pair overlap between divergent regions in pairwise comparisons.

Hmel 1.1 reference genome (Heliconius Genome Consortium, 2012) using Stampy (Lunter and Goodson, 2011) and SNPs were called simultaneously for all samples using the multiallelic calling function in GATK version 1.5 (DePristo et al., 2011; McKenna et al., 2010). The final data set consisted of 33,061,085 SNPs, with 97% of these sites covered in each sample (Table S2).

Genome-wide Demographic Inference

Coalescent simulations, implemented in IMa2 (Hey, 2010; Nielsen and Wakeley, 2001), were used to generate neutral estimates of migration (*2Nm*), effective population size (θ), and divergence times ($t\mu$;TMRCA). Ten 10 kbp windows were drawn randomly from each chromosome, and each window was phased using BEAGLE version 3.3.2 (Browning and Browning, 2007). The phased SNPs were converted to FASTA formatted haplotypes, and the longest nonrecombining block within each window was identified with IMgc (Woerner et al., 2007). Each of the resulting ten, 21 locus (representing each chromosome) data sets was analyzed in IMa2. Results are summarized across the ten data sets in Figure 1C, Table S3, and Table S7.

Simulations

Gene trees were simulated under a neutral model using Hudson's program *ms* (Hudson, 2002). The full migration model, with population size changes, was modeled as follows: ms 60 10000 -t 34.6 -I 3 20 20 20 -ma x 11.53 11.53 0 × 12.56 0 4.89 x -n 1 0.35 –n 2 1.59 –n 3 0.22 -ej 0.761 3 2 –en 0.761 2 0.035 -ej 2.48 2 1 –en 2.48 1 1. Coalescent trees without migration were simulated using the following command line: ms 60 10000 -t 34.6 -I 3 20 20 20 -n 1 0.35 –n 2 1.59 –n 3 0.22 -ej 0.761 3 2 –en 0.761 2 0.035 -ej 2.48 2 1 –en 2.48 1 1. Sixty 5 kbp DNA segments were then generated for each of the coalescent gene trees using Seq-Gen (Rambaut and Grassly, 1997) and used to determine the neutral distribution of F_{ST} for each comparison using Arlequin 3.5.1.3 (Excoffier and Lischer, 2010). F_{ST} distributions under models with and without migration were then compared to our empirical distributions (Figure S1).

Identifying Divergent Genomic Regions

Every scaffold was divided into 5 kbp windows and F_{ST} values were calculated for each window in three pairwise comparisons: *H. cydno-H. pachinus*, *H. cydno-H. melpomene*, and *H. pachinus-H. melpomene*. To identify a common scale across which to compare genomic divergence, and to reduce the statistical nonindependence of F_{ST} comparisons for 5 kbp windows, we estimated empirical significance thresholds and linked adjacent windows that exhibited elevated differentiation (Figure S2). Windows with F_{ST} values greater than the 95th percentile ($F_{ST} \geq 0.598$) were treated as highly divergent windows. For each pair of consecutive, though not necessarily adjacent, highly divergent windows, all the enclosed windows were classified as divergent if none of their F_{ST} values fell below the 75th percentile ($F_{ST} \geq 0.325$).

Population Genomics

For most of our analyses, we grouped samples by species, *H. cydno*, *H. pachinus*, and *H. melpomene*, except for those presented in Figure 5A, for which we separated *H. melpomene* samples into east and west collecting locations. We took the union of all divergent regions between the species pairs *H. cydno-H. pachinus*, *H. cydno-H. melpomene*, and *H. pachinus-H. melpomene* as a combined set, which was then compared to the remaining portion of the genome for a variety of population genetic statistics (Figure 6). This set consisted of 941 genomic regions, containing 6,637 windows, spanning 32,983,224 bp of the genome (14.6% of the mapped chromosomes). The 97.5 and 2.5 percentile confidence intervals around the mean values were computed by bootstrap resampling from the entire set of windows 10,000 times. p values were estimated by bootstrap resampling and were adjusted to control for multiple tests (Benjamini and Hochberg, 1995). Pairwise linkage disequilibrium (LD) was calculated as the squared correlation coefficient (r^2) between allele counts observed at two SNPs using the VCFtools software package (Danecek et al., 2011). This approach is computationally feasible for large data sets since it does not require haplotype reconstruction, but it provides only an approximation of the true LD (Rogers and Huff, 2009). Derived allele frequency and Patterson's D both require identifying ancestral and derived alleles, which we did using *H. ismenius* and *H. hecale* as a combined outgroup.

Clustering Analysis

To test if the counts of divergent regions were overrepresented or underrepresented on any chromosome in the *H. cydno-H. pachinus* comparison, we used a Monte-Carlo-simulated nonparametric paired Wilcoxon test ($Z = -1.949$, $p = 0.05$). The probability of observing regions of high divergence between *H. cydno* and *H. pachinus* on a chromosome containing a known color-pattern locus (chr1, chr10, chr15, chr18) was estimated using Fisher's exact test ($p < 0.01$). Equivalent tests for *H. cydno-H. melpomene* and *H. pachinus-H. melpomene* were performed using the nonparametric simulated paired Wilcoxon test, as above (all $Z \leq -5.06$, all $p > 0.61$). To test for enrichment of divergent regions on color-pattern chromosomes, we tested a contingency table of regions on color pattern chromosomes versus not on these chromosomes, normalized by chromosome length (Fisher's exact tests, $p > 0.538$ in both cases).

GO Term Enrichment Analysis

Gene sequences were extracted from Hmel1.1 and annotated using FlyBase and GO Elite. We combined permuted probabilities from the merged GO Elite analysis for the three interspecific comparisons using Fisher's method and then adjusted the tests for multiple comparisons based on the total number of genes in the comparison set, multiplied by 3 to further correct for the three nonindependent comparisons (Table S6).

ACCESSION NUMBERS

The NCBI SRA ID number for the sequence data reported in this paper is SRA106228.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, seven figures, and seven tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2013.09.042.

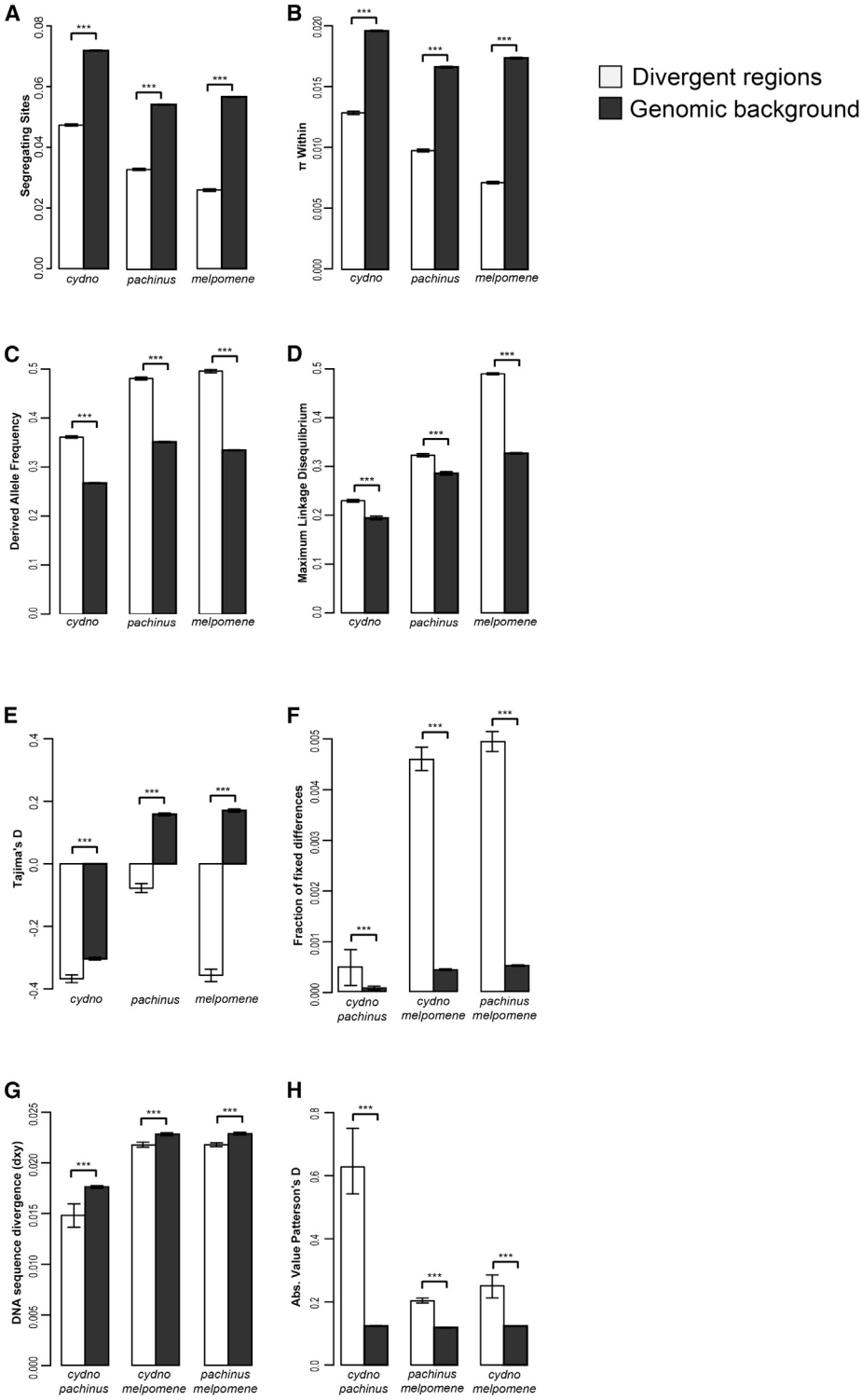

Figure 6. Divergent Regions of the Genome Exhibit Signatures of Selection and Adaptive Introgression

Each panel shows the mean values of population genetic statistics inside divergent regions (white bars) versus the genomic background (gray bars). Segregating site density (A), π within species (B), derived allele frequency (C), maximum linkage disequilibrium (D), Tajima's D (E), fraction of fixed differences between species (F), mean pairwise sequence divergence between species (d_{XY}) (G), and absolute value of Patterson's D statistic for the four taxon ordering: *H. cydno*, *H. pachinus*, *H. melpomene*, outgroup (*H. hecale* and *H. ismenius*) (H). Error bars (indicating 95% confidence intervals) and p values are based on bootstrap resampling. ***$p < 0.0001$.

AUTHOR CONTRIBUTIONS

M.R.K. and S.P.M. conceived the study and M.R.K., D.D.K., and S.P.M. developed the experimental design. N.G.C. and D.D.K. oversaw the generation of the sequence data, performed quality filtering, and handled read-mapping and SNP discovery. M.E.B.H. and R.J.K. identified divergent regions, calculated population genetic summary statistics, and tested for evidence of isolation by distance. N.G.C. implemented sliding-window phylogenetic analyses and performed the GO enrichment analysis and the inversions analysis. N.G.C., J.R.G., and S.P.M. estimated genome-wide demographic parameters. W.Z. performed allopatric/sympatric comparisons. M.R.K. and S.P.M. implemented neutral coalescent simulations. D.D.K., N.G.C., and M.E.B.H. examined clustering. J.R.G. calculated LD statistics. M.E.B.H. and D.D.K. calculated and interpreted the ABBA-BABA statistics. All authors contributed to, read, and approved the manuscript. The manuscript was primarily written by M.R.K., with extensive input from all coauthors.

ACKNOWLEDGMENTS

We thank the government of Costa Rica for permission to collect butterflies, BGI for sequencing, R. Hudson for assistance with *ms*, and reviewers for comments on the manuscript. Computational infrastructure, data storage, and resources were provided to individual laboratories by Boston University, Temple University, University of Chicago, the University of Hawaii, and the California Academy of Sciences. Additional high-performance computing was facilitated by an NSF XCEDE start-up allocation (TG-MCB120130) to J.R.G. and funding from NSF EPSCoR (0554657 PI J. Gaines) at the University of Hawaii. Funding for this project was provided by National Science Foundation grants DEB-1316037 (to M.R.K.) and DEB-1021036 (to S.P.M.).

Received: May 23, 2013
Revised: August 9, 2013
Accepted: September 25, 2013
Published: October 31, 2013

REFERENCES

Baxter, S.W., Nadeau, N.J., Maroja, L.S., Wilkinson, P., Counterman, B.A., 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.

Beltrán, M., Jiggins, C.D., Bull, V., Linares, M., Mallet, J., McMillan, W.O., and Bermingham, E. (2002). Phylogenetic discordance at the species boundary: comparative gene genealogies among rapidly radiating Heliconius butterflies. Mol. Biol. Evol. *19*, 2176–2190.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc., B *57*, 289–300.

Benson, W.W. (1978). Resource partitioning in passion vine butterflies. Evolution *32*, 493–518.

Brown, K.S. (1981). The biology of *Heliconius* and related genera. Annu. Rev. Entomol. *26*, 427–456.

Browning, S.R., and Browning, B.L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. *81*, 1084–1097.

Bull, V., Beltrán, M., Jiggins, C.D., McMillan, W.O., Bermingham, E., and Mallet, J. (2006). Polyphyly and gene flow between non-sibling Heliconius species. BMC Biol. *4*, 11.

Chamberlain, N.L., Hill, R.I., Kapan, D.D., Gilbert, L.E., and Kronforst, M.R. (2009). Polymorphic butterfly reveals the missing link in ecological speciation. Science *326*, 847–850.

Chamberlain, N.L., Hill, R.I., Baxter, S.W., Jiggins, C.D., and Kronforst, M.R. (2011). Comparative population genetics of a mimicry locus among hybridizing *Heliconius* butterfly species. Heredity (Edinb). *107*, 200–204.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al.; 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158.

Dasmahapatra, K.K., Silva-Vásquez, A., Chung, J.W., and Mallet, J. (2007). Genetic analysis of a wild-caught hybrid between non-sister *Heliconius* butterfly species. Biol. Lett. *3*, 660–663.

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

Durand, E.Y., Patterson, N., Reich, D., and Slatkin, M. (2011). Testing for ancient admixture between closely related populations. Mol. Biol. Evol. *28*, 2239–2252.

Ellegren, H., Smeds, L., Burri, R., Olason, P.I., 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*, 756–760.

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

Excoffier, L., and Lischer, H.E.L. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol. Ecol. Resour. *10*, 564–567.

Excoffier, L., Smouse, P.E., and Quattro, J.M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics *131*, 479–491.

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

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

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

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

Gilbert, L.E. (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, C.L. Boggs, W.B. Watt, and P.R. Ehrlich, eds. (Chicago, IL: University of Chicago Press), pp. 281–318.

Haddrill, P.R., Charlesworth, B., Halligan, D.L., and Andolfatto, P. (2005). Patterns of intron sequence evolution in *Drosophila* are dependent upon length and GC content. Genome Biol. *6*, R67.

Halligan, D.L., and Keightley, P.D. (2006). Ubiquitous selective constraints in the *Drosophila* genome revealed by a genome-wide interspecies comparison. Genome Res. *16*, 875–884.

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

Hendry, A.P., Bolnick, D.I., Berner, D., and Peichel, C.L. (2009). Along the speciation continuum in sticklebacks. J. Fish Biol. *75*, 2000–2036.

Hey, J. (2010). Isolation with migration models for more than two populations. Mol. Biol. Evol. *27*, 905–920.

Hudson, R.R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics *18*, 337–338.

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

Jiggins, C.D., Salazar, C., Linares, M., and Mavarez, J. (2008). Review. Hybrid trait speciation and *Heliconius* butterflies. Philos. Trans. R. Soc. Lond. B Biol. Sci. *363*, 3047–3054.

Joron, M., Jiggins, C.D., Papanicolaou, A., and McMillan, W.O. (2006a). *Heliconius* wing patterns: an evo-devo model for understanding phenotypic diversity. Heredity (Edinb). *97*, 157–167.

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

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

Kitano, J., Ross, J.A., Mori, S., Kume, M., Jones, F.C., Chan, Y.F., Absher, D.M., Grimwood, J., Schmutz, J., Myers, R.M., et al. (2009). A role for a neo-sex chromosome in stickleback speciation. Nature *461*, 1079–1083.

Kronforst, M.R. (2008). Gene flow persists millions of years after speciation in *Heliconius* butterflies. BMC Evol. Biol. *8*, 98.

Kronforst, M.R., and Gilbert, L.E. (2008). The population genetics of mimetic diversity in *Heliconius* butterflies. Proc. Biol. Sci. *275*, 493–500.

Kronforst, M.R., Kapan, D.D., and Gilbert, L.E. (2006a). Parallel genetic architecture of parallel adaptive radiations in mimetic *Heliconius* butterflies. Genetics *174*, 535–539.

Kronforst, M.R., Young, L.G., Blume, L.M., and Gilbert, L.E. (2006b). Multilocus analyses of admixture and introgression among hybridizing *Heliconius* butterflies. Evolution *60*, 1254–1268.

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

Kronforst, M.R., Salazar, C., Linares, M., and Gilbert, L.E. (2007a). No genomic mosaicism in a putative hybrid butterfly species. Proc. Biol. Sci. *274*, 1255–1264.

Kronforst, M.R., Young, L.G., and Gilbert, L.E. (2007b). Reinforcement of mate preference among hybridizing *Heliconius* butterflies. J. Evol. Biol. *20*, 278–285.

Kulathinal, R.J., Stevison, L.S., and Noor, M.A. (2009). The genomics of speciation in *Drosophila*: diversity, divergence, and introgression estimated using low-coverage genome sequencing. PLoS Genet. *5*, e1000550.

Lawniczak, M.K., Emrich, S.J., Holloway, A.K., Regier, A.P., 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*, 512–514.

Lunter, G., and Goodson, M. (2011). Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res. *21*, 936–939.

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

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

Mallet, J., and Gilbert, L.E. (1995). Why are there so many mimicry rings: correlations between habitat, behavior and mimicry in *Heliconius* butterflies. Biol. J. Linn. Soc. Lond. *55*, 159–180.

Mallet, J., Barton, N., Lamas, G., Santisteban, J., Muedas, M., and 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.

Mallet, J., McMillan, W.O., and Jiggins, C.D. (1998). Mimicry and warning color at the boundary between races and species. In Endless Forms: Species and Speciation, D.J. Howard and S.H. Berlocher, eds. (Oxford, UK: Oxford University Press), pp. 390–403.

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

Marais, G., Nouvellet, P., Keightley, P.D., and Charlesworth, B. (2005). Intron size and exon evolution in *Drosophila*. Genetics *170*, 481–485.

Martin, A., Papa, R., Nadeau, N.J., Hill, R.I., Counterman, B.A., Halder, G., Jiggins, C.D., Kronforst, M.R., Long, A.D., McMillan, W.O., and Reed, R.D. (2012). Diversification of complex butterfly wing patterns by repeated regulatory evolution of a *Wnt* ligand. Proc. Natl. Acad. Sci. USA *109*, 12632–12637.

Martin, S., Dasmahapatra, K., Nadeau, N.J., Salazar, C., Walters, J.R., Simpson, F., Blaxter, M., Manica, A., Mallet, J., and Jiggins, C.D. (2013). Genome-wide evidence for speciation with gene flow in Heliconius butterflies. Genome Res. Published online October 3, 2013. http://dx.doi.org/10.1101/gr.159426.113.

Matute, D.R., Butler, I.A., Turissini, D.A., and Coyne, J.A. (2010). A test of the snowball theory for the rate of evolution of hybrid incompatibilities. Science *329*, 1518–1521.

Mavárez, J., Salazar, C.A., Bermingham, E., Salcedo, C., Jiggins, C.D., and Linares, M. (2006). Speciation by hybridization in *Heliconius* butterflies. Nature *441*, 868–871.

Mayr, E. (1963). Animal Species and Evolution (Cambridge, MA: Harvard University Press).

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

McKinnon, J.S., and Rundle, H.D. (2002). Speciation in nature: the threespine stickleback model systems. Trends Ecol. Evol. *17*, 480–488.

Merrill, R.M., Gompert, Z., Dembeck, L.M., Kronforst, M.R., McMillan, W.O., and Jiggins, C.D. (2011). Mate preference across the speciation continuum in a clade of mimetic butterflies. Evolution *65*, 1489–1500.

Merrill, R.M., Wallbank, R.W., Bull, V., Salazar, P.C., Mallet, J., Stevens, M., and Jiggins, C.D. (2012). Disruptive ecological selection on a mating cue. Proc. Biol. Sci. *279*, 4907–4913.

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

Moyle, L.C., and Nakazato, T. (2010). Hybrid incompatibility "snowballs" between *Solanum* species. Science *329*, 1521–1523.

Nadeau, N.J., Whibley, A., Jones, R.T., Davey, J.W., Dasmahapatra, K.K., Baxter, S.W., Quail, M.A., Joron, M., ffrench-Constant, R.H., Blaxter, M.L., 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.

Nadeau, N.J., Martin, S.H., Kozak, K.M., Salazar, C., Dasmahapatra, K.K., Davey, J.W., Baxter, S.W., Blaxter, M.L., Mallet, J., and Jiggins, C.D. (2013). Genome-wide patterns of divergence and gene flow across a butterfly radiation. Mol. Ecol. *22*, 814–826.

Naisbit, R.E., Jiggins, C.D., and Mallet, J. (2001). Disruptive sexual selection against hybrids contributes to speciation between *Heliconius cydno* and *Heliconius melpomene*. Proc. Biol. Sci. *268*, 1849–1854.

Naisbit, R.E., Jiggins, C.D., Linares, M., Salazar, C., and Mallet, J. (2002). Hybrid sterility, Haldane's rule and speciation in *Heliconius cydno* and *H. melpomene*. Genetics *161*, 1517–1526.

Neafsey, D.E., Lawniczak, M.K., Park, D.J., Redmond, S.N., Coulibaly, M.B., Traoré, S.F., Sagnon, N., Costantini, C., Johnson, C., Wiegand, R.C., et al. (2010). SNP genotyping defines complex gene-flow boundaries among African malaria vector mosquitoes. Science *330*, 514–517.

Nielsen, R., and Wakeley, J. (2001). Distinguishing migration from isolation: a Markov chain Monte Carlo approach. Genetics *158*, 885–896.

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

Nosil, P., Gompert, Z., Farkas, T.E., Comeault, A.A., Feder, J.L., Buerkle, C.A., and Parchman, T.L. (2012a). Genomic consequences of multiple speciation processes in a stick insect. Proc. Biol. Sci. *279*, 5058–5065.

Nosil, P., Parchman, T.L., Feder, J.L., and Gompert, Z. (2012b). Do highly divergent loci reside in genomic regions affecting reproductive isolation? A test using next-generation sequence data in Timema stick insects. BMC Evol. Biol. *12*, 164.

Papa, R., Martin, A., and Reed, R.D. (2008). Genomic hotspots of adaptation in butterfly wing pattern evolution. Curr. Opin. Genet. Dev. *18*, 559–564.

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

Parsch, J. (2003). Selective constraints on intron evolution in *Drosophila*. Genetics *165*, 1843–1851.

Rambaut, A., and Grassly, N.C. (1997). Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. *13*, 235–238.

Reed, R.D., Papa, R., Martin, A., Hines, H.M., Counterman, B.A., Pardo-Diaz, C., Jiggins, C.D., Chamberlain, N.L., Kronforst, M.R., Chen, R., et al. (2011). *optix* drives the repeated convergent evolution of butterfly wing pattern mimicry. Science *333*, 1137–1141.

Rogers, A.R., and Huff, C. (2009). Linkage disequilibrium between loci with unknown phase. Genetics *182*, 839–844.

Salazar, C., Baxter, S.W., Pardo-Diaz, C., Wu, G., Surridge, A., Linares, M., Bermingham, E., and Jiggins, C.D. (2010). Genetic evidence for hybrid trait speciation in *heliconius* butterflies. PLoS Genet. *6*, e1000930.

Seehausen, O. (2004). Hybridization and adaptive radiation. Trends Ecol. Evol. *19*, 198–207.

Sheppard, P.M., Turner, J.R.G., Brown, K.S., Benson, W.W., and Singer, M.C. (1985). Genetics and the evolution of Muellerian mimicry in *Heliconius* butterflies. Philos. Trans. R. Soc. Lond. B Biol. Sci. *308*, 433.

Smiley, J. (1978). Plant chemistry and the evolution of host specificity: new evidence from *heliconius* and *passiflora*. Science *201*, 745–747.

Smith, J., and Kronforst, M.R. (2013). Do *Heliconius* butterfly species exchange mimicry alleles? Biol. Lett. *9*, 20130503.

Staubach, F., Lorenc, A., Messer, P.W., Tang, K., Petrov, D.A., and Tautz, D. (2012). Genome patterns of selection and introgression of haplotypes in natural populations of the house mouse (*Mus musculus*). PLoS Genet. *8*, e1002891.

Turner, T.L., Hahn, M.W., and Nuzhdin, S.V. (2005). Genomic islands of speciation in *Anopheles gambiae*. PLoS Biol. *3*, e285.

Via, S. (2009). Natural selection in action during speciation. Proc. Natl. Acad. Sci. USA *106*(*Suppl 1*), 9939–9946.

Woerner, A.E., Cox, M.P., and Hammer, M.F. (2007). Recombination-filtered genomic datasets by information maximization. Bioinformatics *23*, 1851–1853.

Supplemental Information for:

Hybridization reveals the evolving genomic architecture of speciation

Marcus R. Kronforst, Matthew E. B. Hansen, Nicholas G. Crawford, Jason R. Gallant, Wei Zhang, Rob J. Kulathinal, Durrell D. Kapan & Sean P. Mullen

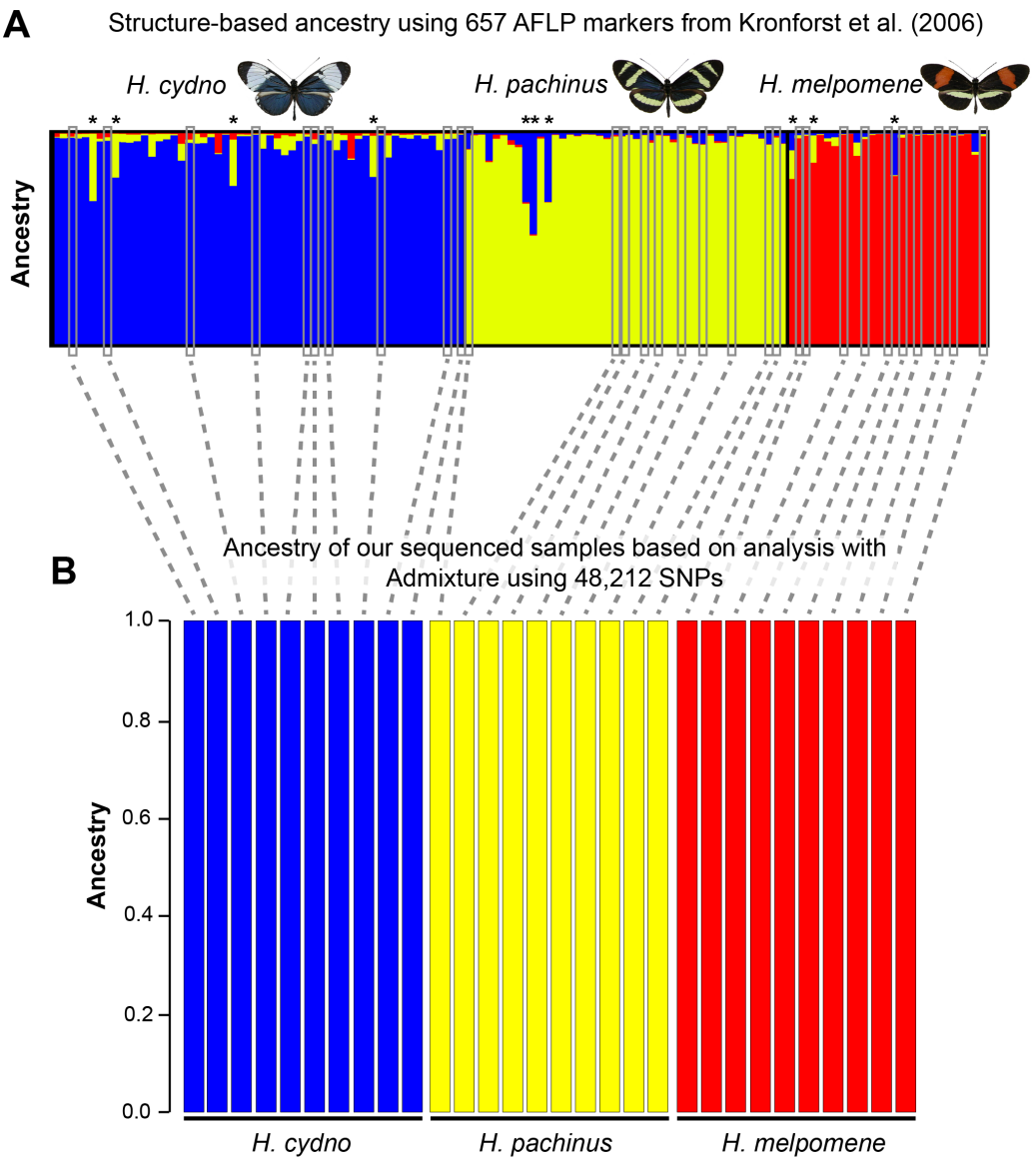

Figure S1. A) Previous work detected widespread admixture among *H. cydno*, *H. pachinus* and *H. melpomene* in Costa Rica (Kronforst et al., 2006b). We specifically selected samples from this set that did not show evidence of recent admixture so as to not bias our estimates of divergence and gene flow among species. (B) Subsequent analysis based on 48K of our genotyped SNPs verified that sequenced samples did not have recent hybrid ancestry. This analysis, run with the program ADMIXTURE 1.2, was based on a small subset of our total SNP dataset because we sampled polymorphisms with a minimum spacing of 5 kbp, a value set by the extent of LD in the genome (Figure S6).

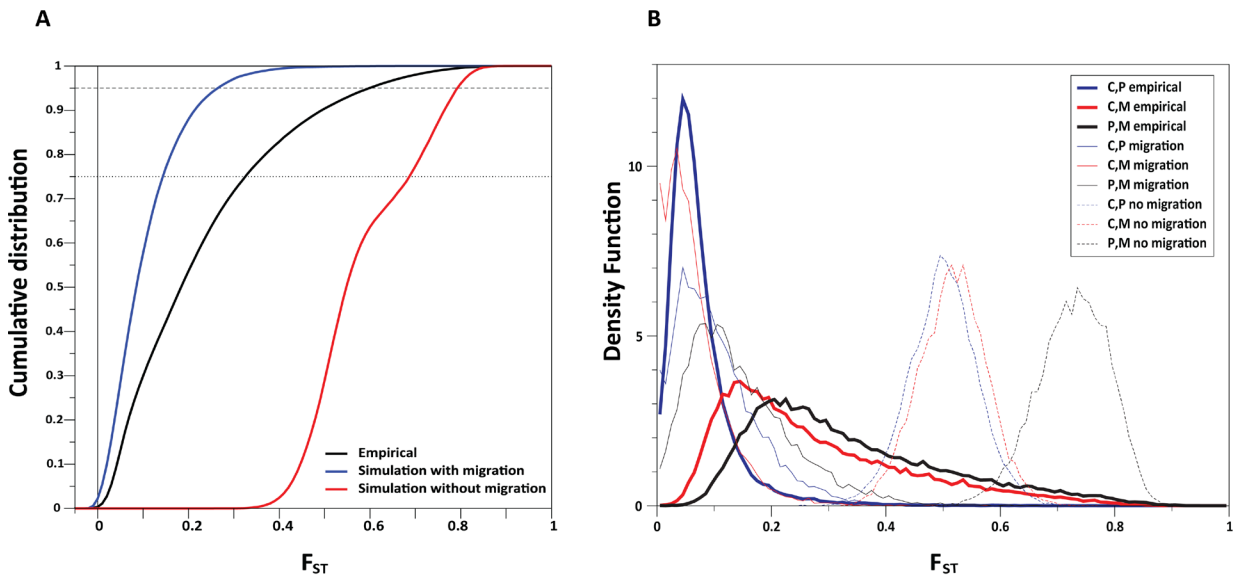

Figure S2. A) Cumulative F_{ST} distribution for 1) empirical data (blue), 2) simulated data under a Fisher-Wright neutral model with migration (black), and 3) simulated data under a neutral model without migration (red). Dashed lines represent the 95% (-) and 75% (˙) distribution thresholds, respectively. B) Density function of 5 kbp window F_{ST} values for each pairwise species comparison: empirical values are shown in solid (bold) lines, simulated values under a Fisher-wright neutral model with migration are also shown as solid (unbolded) lines, simulated values under a Fisher-Wright neutral model without migration are displayed as dashed lines. Iterative experimental manipulations of simulation parameters indicate that the left-shifted F_{ST} values observed for our *H. cydno-H. melpomene* data simulated under a neutral no migration model result from the large N_e of *H. cydno* relative to *H. melpomene*.

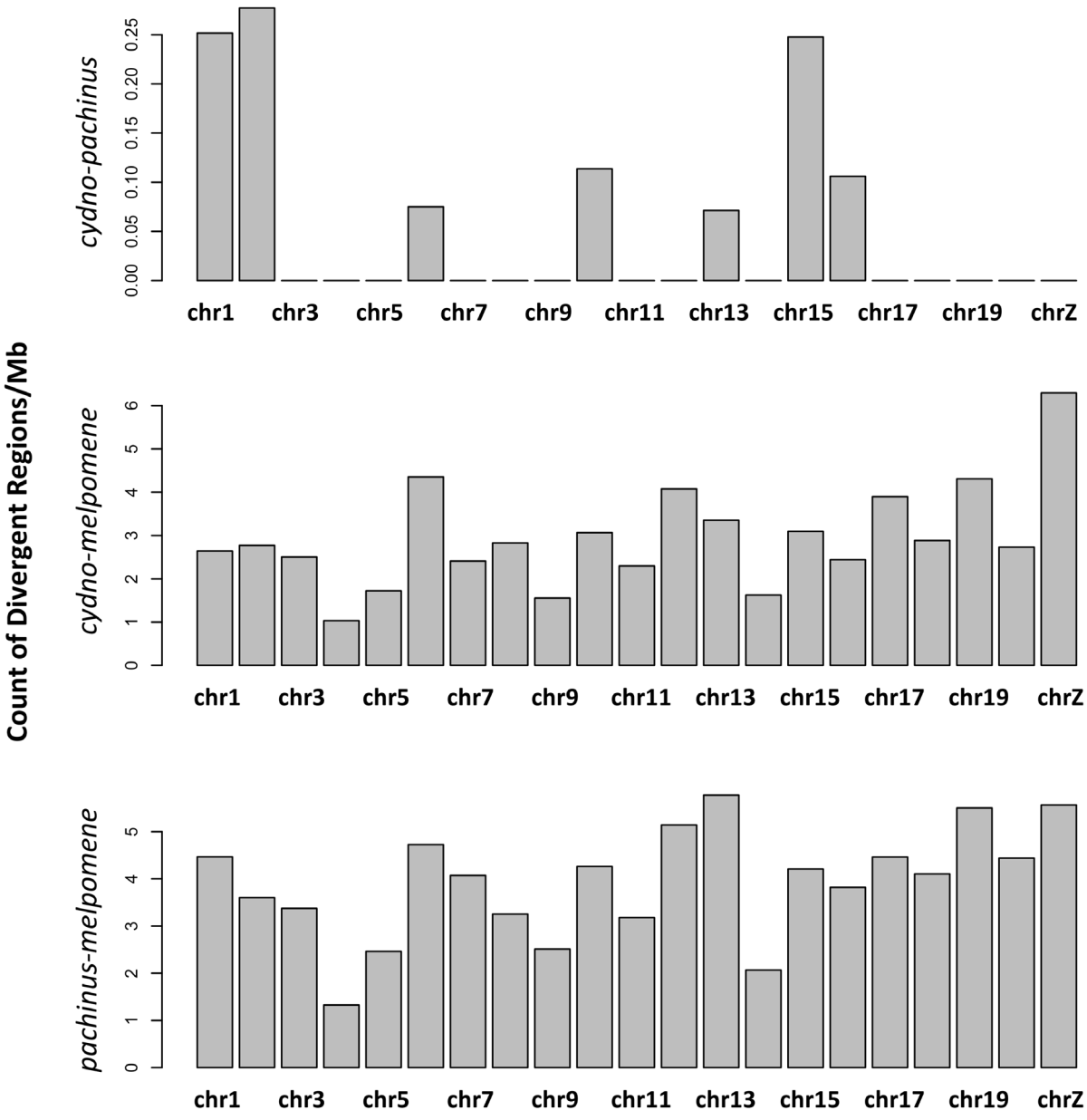

Figure S3. Clustering of highly divergent regions by chromosome for 1) *cydno-pachinus* (top), 2) *cydno-melpomene* (middle), and 3) *pachinus-melpomene* (bottom).

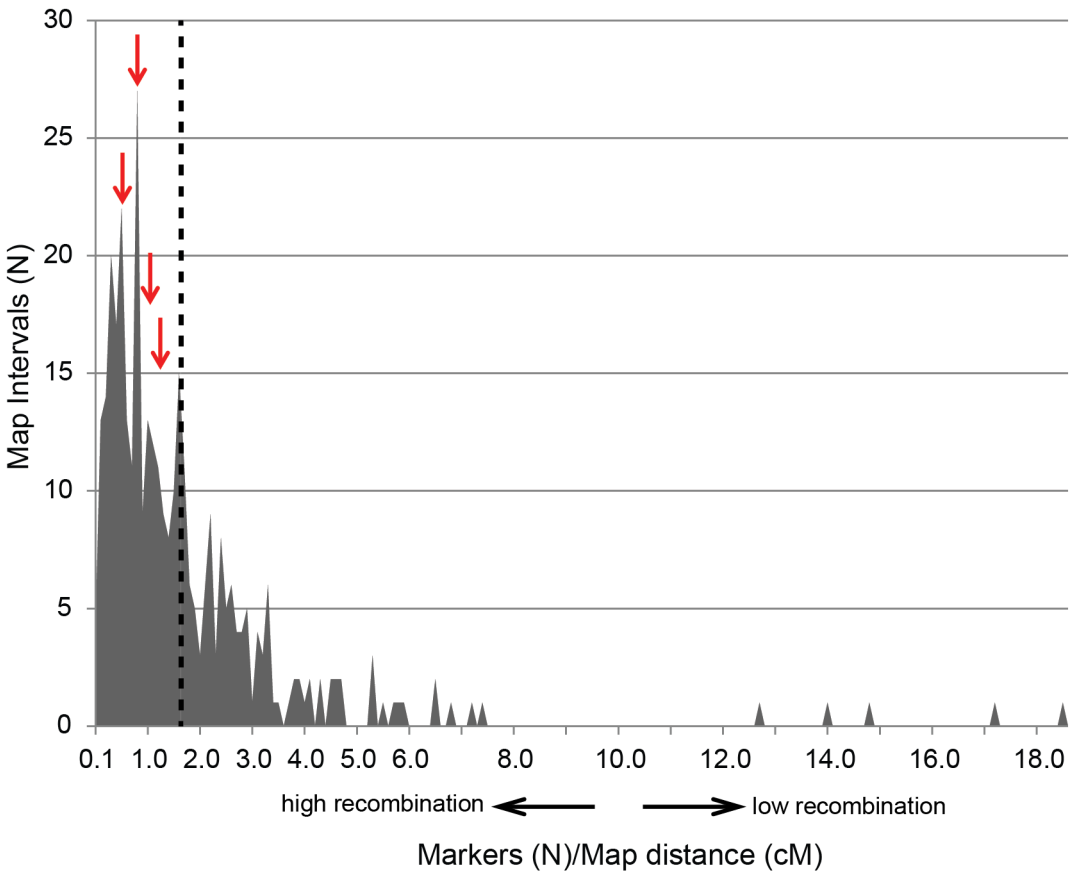

Figure S4. Linkage mapping reveals that *Heliconius* mimicry loci are not in regions of reduced recombination. We quantified recombination across the genome based on our full genome linkage map (Kronforst et al., 2006a; Kronforst et al., 2006c). Areas of reduced recombination will emerge as clusters of tightly-linked markers, quantified here as a high marker/cM ratio (x-axis). The mean marker/cM ratio across the genome is 1.7 (dashed line) and all four of the major wing pattern mimicry loci (red arrows) are located in regions with lower values, indicating high levels of recombination. Since the mimicry loci stand out as the most strongly divergent regions in our analyses, this suggests that observed signatures of divergence and selection are not a by-product of reduced recombination.

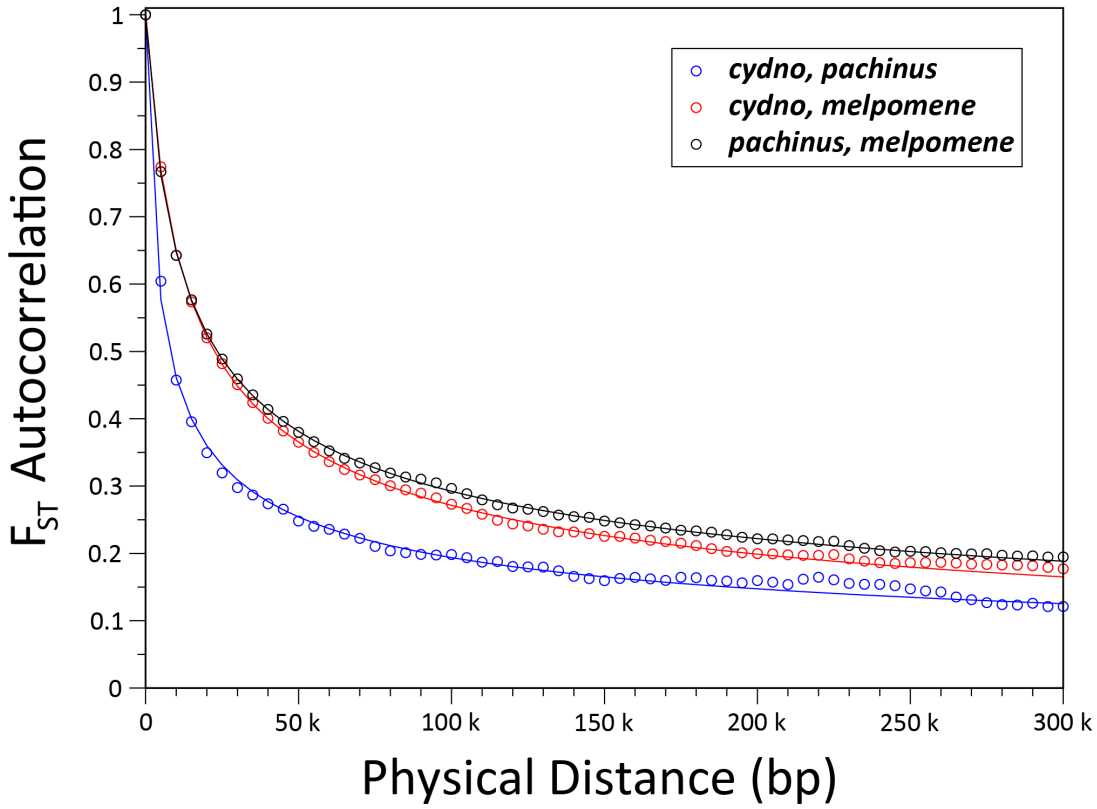

Figure S5. F_{ST} autocorrelation among sites fit to an exponentially declining power law (exponent -0.4 to -0.47 across the three comparisons) enforcing the ACF constraint y(x=0) = 1. For all three pairwise ACFs the correlation coefficient (r) exceeds 0.998.
Figure S6. Pairwise Linkage Disequilibrium (LD) measured as the squared correlation coefficient (r^2) between two SNPS for each species within (red) and outside (black) divergent genomic regions.

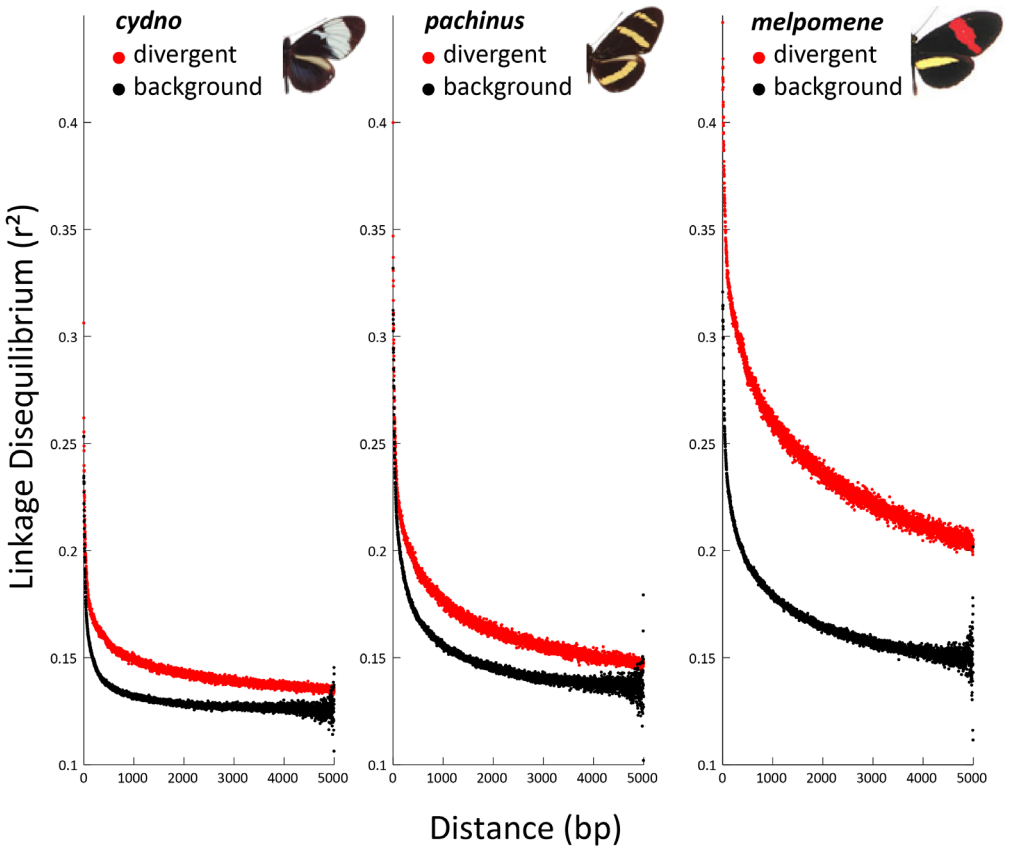

Figure S6. Pairwise Linkage Disequilibrium (LD) measured as the squared correlation coefficient (r^2) between two SNPS for each species within (red) and outside (black) divergent genomic regions.

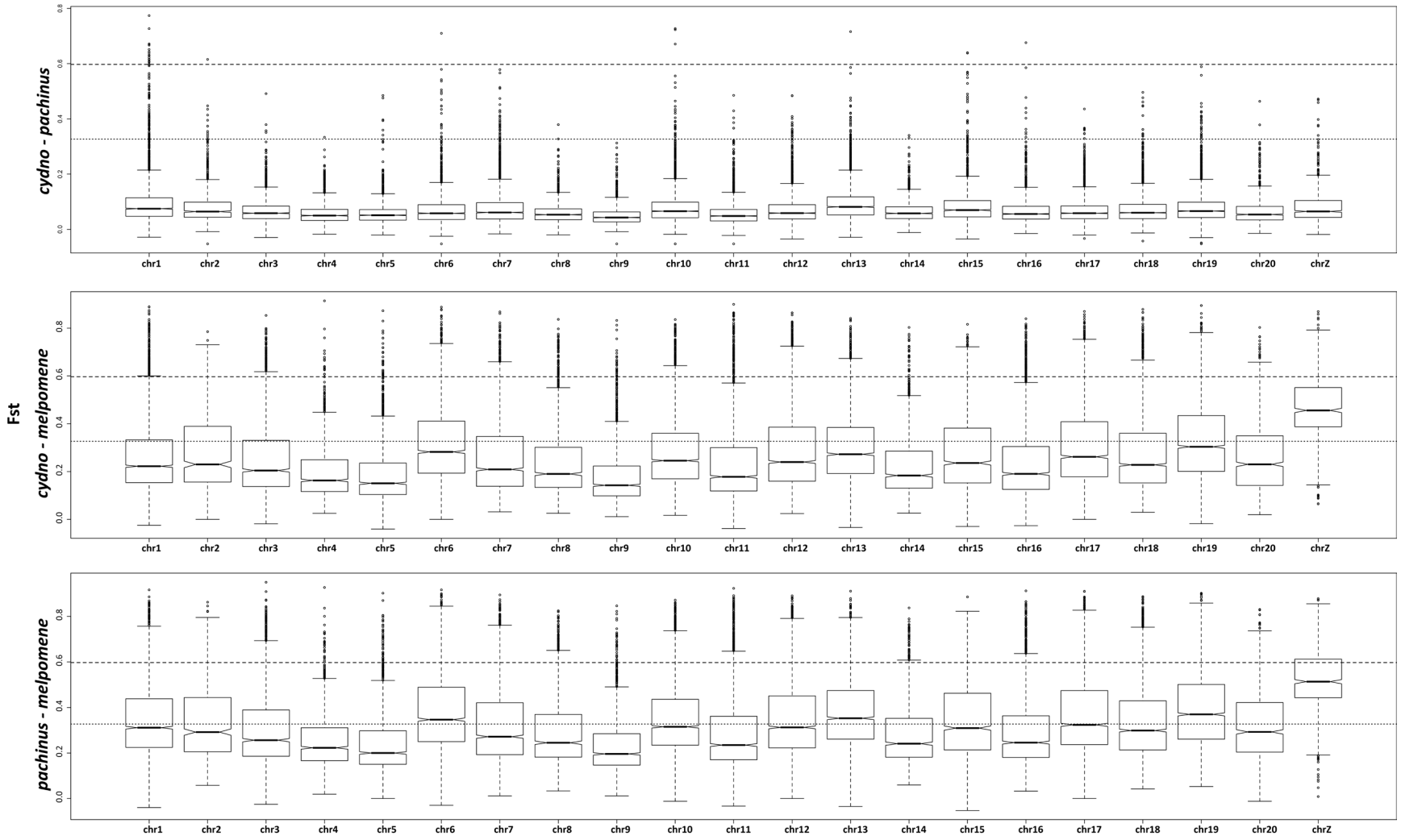

Figure S7. Pairwise F_{ST} represented as boxplots with whiskers between 1) *cydno-pachnius* (top), 2) *cydno-melpomene* (middle), and 3) *pachinus-melpomene* (bottom) for all 21 chromosomal regions highlighting elevated divergence on the Z-chromosome in comparisons with *H. melpomene*.

Table S1. Sample collection information.

Sample	Species	sex	location	Coast	Lat.	Long.
c511	*cydno galanthus*	female	Cariblanco	Caribbean	10° 16' N	84° 11' W
c512	*cydno galanthus*	male	PN Hitoy Cerere	Caribbean	9° 40' N	83° 2' W
c513	*cydno galanthus*	male	La Selva Biological Station	Caribbean	10° 26' N	83° 59' W
c514	*cydno galanthus*	female	Vesta	Caribbean	9° 43' N	83° 3' W
c515	*cydno galanthus*	female	Guacimo	Caribbean	10° 13' N	83° 41' W
c563	*cydno galanthus*	male	Guapiles	Caribbean	10° 13' N	83° 47' W
c614	*cydno galanthus*	male	Vesta	Caribbean	9° 43' N	83° 3' W
c630	*cydno galanthus*	female	La Selva Biological Station	Caribbean	10° 26' N	83° 59' W
c639	*cydno galanthus*	male	Guacimo	Caribbean	10° 13' N	83° 41' W
c640	*cydno galanthus*	female	Vesta	Caribbean	9° 43' N	83° 3' W
h665	*hecale*	male	La Selva Biological Station	Caribbean	10° 26' N	83° 59' W
i02_210	*ismenius*	male	PN Manuel Antonio	Pacific	9° 24' N	84° 10' W
m523	*melpomene rosina*	male	Vesta	Caribbean	9° 43' N	83° 3' W
m589	*melpomene rosina*	male	Selva Bananito	Caribbean	9° 52' N	83° 0' W

m676	*melpomene rosina*	male	Vesta	Caribbean	9° 43' N	83° 3' W
m682	*melpomene rosina*	male	La Selva Biological Station	Caribbean	10° 26' N	83° 59' W
m687	*melpomene rosina*	female	La Selva Biological Station	Caribbean	10° 26' N	83° 59' W
m524	*melpomene rosina*	female	Puriscal	Pacific	9° 51' 0N	84° 19' W
m525	*melpomene rosina*	male	Sirena Biological Station	Pacific	8° 28' N	83° 35' W
m675	*melpomene rosina*	male	PN Manuel Antonio	Pacific	9° 24' N	84° 10' W
m683	*melpomene rosina*	male	PN Manuel Antonio	Pacific	9° 24' N	84° 10' W
m689	*melpomene rosina*	female	Puriscal	Pacific	9° 51' 0N	84° 19' W
p516	*pachinus*	female	Puriscal	Pacific	9° 51' 0N	84° 19' W
p517	*pachinus*	male	Puriscal	Pacific	9° 51' 0N	84° 19' W
p518	*pachinus*	male	PN Manuel Antonio	Pacific	9° 24' N	84° 10' W
p519	*pachinus*	female	Sirena Biological Station	Pacific	8° 28' N	83° 35' W
p520	*pachinus*	male	Sirena Biological Station	Pacific	8° 28' N	83° 35' W
p591	*pachinus*	male	Colon	Pacific	9° 55' N	84° 15' W
p596	*pachinus*	male	PN Carara	Pacific	9° 47' N	84° 36' W
p690	*pachinus*	female	Puriscal	Pacific	9° 51' 0N	84° 19' W

p694	*pachinus*	female	PN Manuel Antonio	Pacific	9° 24' N	84° 10' W
p696	*pachinus*	female	Sirena Biological Station	Pacific	8° 28' N	83° 35' W

Table S2. Sequence coverage and SNP scoring statistics.

Sample	Sequencing coverage (X)	Effective Coverage (X)	Polymorphic Sites w/ Genotypes	Total SNPs/sample
c511	15.98	14.57	32,330,521	7,834,670
c512	16.13	14.67	32,339,964	7,880,908
c513	16.19	14.83	32,323,062	7,865,793
c514	16.12	14.73	32,302,561	7,805,667
c515	16.06	14.70	32,271,197	7,781,108
c563	15.14	13.83	32,242,026	7,831,760
c614	16.07	14.81	32,346,747	7,885,840
c630	15.24	13.83	32,258,915	7,787,160
c639	16.13	14.84	32,274,434	7,841,661
c640	16.15	14.82	32,276,539	7,799,394
h665	16.04	14.13	31,263,335	9,026,075
i02-210	15.90	13.61	30,292,045	8,522,654
m523	16.26	14.94	32,351,936	5,732,343
m524	16.06	15.03	32,386,948	5,838,627
m525	16.20	14.58	32,430,751	5,848,993
m589	16.11	14.92	32,299,253	5,704,349
m675	16.06	15.07	32,399,303	5,835,224
m676	14.95	13.63	32,218,837	5,631,994
m682	16.03	14.96	32,302,390	5,646,811
m683	15.01	14.19	32,346,131	5,808,734

m687	16.11	15.06	32,298,593	5,637,142
m689	15.97	14.71	32,398,193	5,835,221
p516	16.64	14.94	32,200,990	7,489,584
p517	16.63	15.26	32,220,174	7,516,119
p518	16.63	15.23	32,221,498	7,517,439
p519	16.62	15.15	32,180,303	7,454,787
p520	15.02	13.78	32,145,249	7,512,737
p591	15.12	13.84	32,148,217	7,518,538
p596	15.21	13.72	32,121,475	7,497,475
p690	16.63	14.85	32,143,341	7,439,867
p694	16.64	15.06	32,210,561	7,486,019
p696	16.66	15.19	32,214,529	7,478,731
mean	**15.99**	**14.61**	**32,180,001**	**7,134,170**

Table S3. Summary of demographic parameters inferred using IMa2. These are mean and 95% CI based on 10 independent runs of IMa2, using different loci in each run. Parameter estimates and 95% CI for each IMa2 run are shown in Table S7.

Population sizes	**Mean**	**CI**
cydno (C)	1.81E+06	3.53E+05
pachinus (P)	2.55E+05	4.85E+04
melpomene (M)	5.13E+05	4.56E+05
C, P MRCA	4.04E+04	2.43E+04
C, P, M MRCA	1.13E+06	6.03E+05
Divergence times		
C vs. P	4.29E+05	7.36E+04
M vs. C, P MRCA	1.36E+06	2.08E+05
Migration Rates		
C into P	0.546	0.218
P into C	9.905	2.374
M into C	0.018	0.033
C into M	0.054	0.056
M into P	0.000	NA
P into M	0.072	0.055
M into C, P MRCA	0.019	0.023
C, P MRCA into M	1.989	1.192

Table S4. Genome sequence data suggest very few potential inversions in the locations of the 12 genomic regions that are divergent between *H. cydno* and *H. pachinus*. 'Yes' and 'No' below refer to whether an inversion is indicated for each sample in the genomic intervals displayed in Figure 3.

Sample	**Chr1**	**Chr2**	**Chr6**	**Chr10a**	**Chr10b**	**Chr13**	**Chr15**	**Chr16**	**Chr18**
c511	No	Yes	No	No	No	No	No	No	No
c512	No	No	No	No	No	No	No	No	No
c513	No	No	No	No	No	No	No	No	No
c514	No	No	No	No	No	No	No	No	No
c515	Yes	Yes	No	No	No	No	No	No	No
c563	No	No	No	No	No	No	No	No	No
c614	No	Yes	No	No	No	No	No	No	No
c630	No	No	No	Yes	No	No	No	No	No
c639	No	No	No	No	No	No	No	No	No
c640	No	No	No	No	No	No	No	No	No
h665	No	No	No	No	No	No	No	No	No
i02_210	No	No	No	No	No	No	No	No	No
m523	No	No	No	No	No	No	No	No	No
m589	No	No	No	No	No	No	No	No	No
m676	No	No	No	No	Yes	No	No	No	No
m682	No	No	No	No	No	No	No	No	No
m687	No	No	No	No	No	No	No	No	No
m524	No	No	No	No	No	No	No	No	No

m525	No	No	No	No	No	No	No	No	No
m675	No	No	No	No	No	No	No	No	No
m683	No	No	No	No	No	No	No	No	No
m689	No	No	No	No	No	No	No	No	No
p516	No	No	No	No	No	No	No	No	No
p517	No	No	No	No	No	No	No	No	No
p518	No	No	No	No	No	No	No	No	No
p519	No	No	No	No	No	No	No	No	No
p520	No	Yes	No	No	No	No	No	No	No
p591	Yes	No	No	No	No	No	No	No	No
p596	No	No	No	No	No	No	No	No	No
p690	No	No	No	No	No	No	No	No	No
p694	No	No	No	No	No	No	No	No	No
p696	No	No	No	No	No	No	No	No	No

Table S5. Estimates of fixed differences between populations/species. For each pairing, five individuals were sampled at random without replacement from *cydno* and/or *pachinus*, and all five individuals were used for *melpomene* East and/or *melpomene* West. The number of sites with fixed differences were then computed. This was done 20 times, with different groupings of five individuals taken each time for *cydno* and *pachinus*. Only sites with complete data (genotypes for all 10 individuals in the pairing) were used.

Group 1	**Group 2**	**Mean Fraction of Fixed Differences**	**SD**
melpomene East	*melpomene* West	4.24E-07	NA[1]
cydno	*pachinus*	1.09E-04	1.41E-05
cydno	*melpomene* East	9.20E-03	1.43E-04
cydno	*melpomene* West	9.55E-03	1.02E-04
pachinus	*melpomene* East	1.34E-02	2.35E-04
pachinus	*melpomene* West	1.39E-02	2.65E-04

[1] The *melpomene* East vs. *melpomene* West pairing involved no resampling because there are a total of five samples in each group.

Table S6. GO term enrichment analysis, comparing divergent genome regions to the entire genome.

Ontology ID	Ontology term	Ontology type	Adjusted P
GO:0000339	RNA cap binding	molecular function	5.20E-11
GO:0046914	transition metal ion binding	molecular function	6.57E-11
GO:0044427	chromosomal part	cellular component	1.36E-07
GO:0032312	regulation of ARF GTPase activity	biological process	7.19E-07
GO:0051119	sugar transmembrane transporter activity	molecular function	1.95E-06
GO:0008513	secondary active organic cation transmembrane transporter activity	molecular function	2.85E-06
GO:0003824	catalytic activity	molecular function	1.96E-05
GO:0060250	germ-line stem-cell niche homeostasis	biological process	1.13E-04
GO:0008306	associative learning	biological process	8.87E-04
GO:0045466	R7 cell differentiation	biological process	4.31E-03
GO:0043035	chromatin insulator sequence binding	molecular function	6.74E-03
GO:0050803	regulation of synapse structure and activity	biological process	7.03E-03
GO:0051783	regulation of nuclear division	biological process	8.16E-03
GO:0005700	polytene chromosome	cellular component	8.92E-03
GO:0001076	RNA polymerase II transcription factor binding transcription factor activity	molecular function	1.10E-02
GO:0008060	ARF GTPase activator activity	molecular function	1.10E-02
GO:0040020	regulation of meiosis	biological process	1.52E-02
GO:0030703	eggshell formation	biological process	2.38E-02

GO:0045595	regulation of cell differentiation	biological process	2.90E-02
GO:0045742	positive regulation of epidermal growth factor receptor signaling pathway	biological process	3.04E-02
GO:0005049	nuclear export signal receptor activity	molecular function	3.39E-02
GO:0048598	embryonic morphogenesis	biological process	3.81E-02
GO:0030163	protein catabolic process	biological process	4.10E-02
GO:0044260	cellular macromolecule metabolic process	biological process	4.15E-02
GO:0045010	actin nucleation	biological process	4.42E-02
GO:0030659	cytoplasmic vesicle membrane	cellular component	4.59E-02

Table S7. Estimates and 95% CI for each parameter from the 10 IMa2 runs.

See Kronforst et al Table S7.xlsx.

Supplemental Experimental Procedures

Samples

We collected samples from 13 locations across Costa Rica (Table S1). Samples were collected in the field as adults, euthanized, and then wings were separated and placed in glassine envelopes while the bodies were stored in 95% ethanol. For each specimen, genomic DNA was extracted from a portion of thoracic tissue using a DNeasy Blood & Tissue Kit (Qiagen).

Sequencing

A custom Illumina sequencing library with a 500 bp insert was prepared for each sample and sequenced to an average depth of 16X coverage using an Illumina Hi-Seq 2000 (2 × 100 paired-end sequencing). Library preparation and sequencing were performed at BGI. In total, we generated 182.7 Gbp of data across the 32 libraries. Raw reads were preprocessed to trim adaptor sequences and remove low quality reads. These data were subsequently aligned to the Hmel 1.1 reference genome (Heliconius Genome Consortium, 2012) using Stampy (Lunter and Goodson, 2011). SNPs were called simultaneously for all samples using the multi-allelic calling function in GATK version 1.5 (DePristo et al., 2011; McKenna et al., 2010). Positions with a total SNP quality less than 40 were filtered from subsequent analyses. Females have only one Z chromosome but due to sequencing errors, our SNP calling pipeline occasionally, but rarely, scored females as heterozygotes for Z-linked SNPs. However, GATK gives a probability score to each allele and we found that these were not the same in the case of female Z-linked heterozygous sites, likely as a result one allele being due to sequencing error. We edited these sites to assign them single alleles by retaining the single, highest probability nucleotide at each site. The final dataset consisted of 33,061,085 SNPs, with 97% of these sites covered in each sample (Table S2).

Phylogeny

To calculate a genome wide species tree, 4,051 non-overlapping 50 kbp windows were drawn from the multi-allelic VCF file. Scaffolds less than 50 kbp were excluded. Windows were centered within scaffolds (e.g., for a scaffold of 60 kbp the window started at 5 kbp and ended at 55 kbp). Trees were calculated for each window with PhyML v3.0_360-500M using an HKY DNA substitution model. The summary tree is shown in Figure 1A.

Genome-wide Demographic Inference

Coalescent simulations, implemented in IMa2 (Hey, 2010; Nielsen and Wakeley, 2001), were used to generate neutral estimates of migration ($2Nm$), effective population size (θ), and divergence times ($t\mu$;TMRCA). To generate phased haplotype input files for

IMa2, bi-allelic SNPs were called independently for each species using GATK with the same filtering parameters described above for the multi-allelic analyses. The resulting VCF files were fed through the GATK walker "ProduceBeagleInputWalker" to generate likelihood files in BEAGLE format. Ten 10 kbp windows were then drawn randomly from each chromosome, for a total 210 windows (2.1 Mbp of sequence), and each window was phased using the software program BEAGLE version 3.3.2 (Browning and Browning, 2007). The phased SNPs were converted to FASTA formatted haplotypes and the longest non-recombining block within each window was identified with IMgc (Woerner et al., 2007). Each of the resulting ten, 21 locus (representing each chromosome) datasets was analyzed in IMa2 under an HKY model of mutation, using 100 geometrically heated chains (0.99, 0.75), and a pre-defined species tree. After discarding the first 150K steps as burn-in, each simulation was allowed to proceed until the parameter estimates from the first and second half of the run converged. We then used 10K sampled coalescent genealogies to estimate population sizes, bi-directional migration rates, splitting times, and mutation scalars. Results are summarized across the ten data sets in Figure 1C and Table S3.

Simulations

To assess the interplay between selection, drift, and demographic history, 10,000 coalescent gene trees were simulated under a neutral model, within the species level phylogeny for our three focal taxa (*H. cydno*, *H. pachinus* & *H. melpomene*; Figure 1A), using Hudson's program ms (Hudson, 2002). Mean estimates of individual demographic parameters were obtained via 10 independent, replicate IMa2 analyses (see above), and used to parameterize neutral models with and without migration. The full migration model, with population size changes, was modeled as: ms 60 10000 -t 34.6 -I 3 20 20 20 -ma x 11.53 11.53 0 x 12.56 0 4.89 x -n 1 0.35 –n 2 1.59 –n 3 0.22 -ej 0.761 3 2 –en 0.761 2 0.035 -ej 2.48 2 1 –en 2.48 1 1. Coalescent trees without migration were simulated using the following command line: ms 60 10000 -t 34.6 -I 3 20 20 20 -n 1 0.35 –n 2 1.59 –n 3 0.22 -ej 0.761 3 2 –en 0.761 2 0.035 -ej 2.48 2 1 –en 2.48 1 1.

60 5-kbp DNA segments, corresponding to 10 sampled diploid genomes for each of the three focal species, were then generated for each of the 10,000 coalescent gene trees using Seq-Gen (Rambaut and Grassly, 1997), assuming an HKY model of molecular evolution. The resulting simulated DNA sequences were used to determine the neutral distribution of F_{ST} for each of the three pairwise comparisons by calculating divergence across 10,000 5-kbp windows using a custom Perl bioinformatics pipeline and the program Arlequin 3.5.1.3 (Excoffier and Lischer, 2010). F_{ST} distributions under models with and without migration were then compared to our empirical distributions (Figure S1).

Identifying Divergent Genomic Regions

Every scaffold was divided into 5 kbp tiling windows, with the last window on each scaffold taking up the remainder (window sizes less than 5 kbp). F_{ST} values were calculated for each window in the following pairwise comparisons *H. cydno-H. pachinus*, *H. cydno-H. melpomene*, and *H. pachinus-H. melpomene*. The F_{ST} autocorrelation function (ACF) was computed across the entire genome for each population pair. Each ACF, over the distance range of 0 – 3 Mbp, is well fit by a modified power law of the form $y(x) = a(x+b)^p$ with the parameter constraint $y(0)=1$, yielding two free fitting parameters, as shown in Figure S5. The correlation coefficient (r) of each fit exceeds 0.988. The best fit parameters (a,b,p) for the three population pairs are *H. cydno-H. pachinus*: (19.6, 1691.5, -0.401), *H. cydno-H. melpomene*: (62.6, 6650.8, -0.470), and *H. pachinus-H. melpomene*: (34.6, 5349.1, -0.413).

To identify a common scale across which to compare genomic divergence and to reduce the statistical non-independence of F_{ST} comparisons for 5 kbp windows, we estimated empirical significance thresholds and linked adjacent windows that exhibited elevated differentiation. To do this, all F_{ST} values from the three pairwise comparisons were combined and used to identify global 95th and 75th percentiles of 0.598 and 0.325, respectively (Figure S2). We also compared the observed F_{ST} distributions to distributions obtained from simulations with and without interspecific gene flow (Figure S2). For our analyses, it was essential to apply the same F_{ST} threshold across pairwise comparisons, although each comparison exhibits a different amount of divergence, because we were focused on identifying and then comparing the location and physical size of divergent segments. These comparisons are only possible if we identify a common cut-off across the combined F_{ST} distributions and this cut-off is held constant. Windows with F_{ST} values greater than the 95th percentile ($F_{ST} \geq 0.598$) were treated as highly divergent windows. By comparing observed and simulated F_{ST} distributions, we found that values above the 95th percentile represented outliers relative to simulations with interspecific gene flow but no selection, making this a robust and conservative cut-off. For each pair of consecutive, though not necessarily adjacent, highly divergent windows, all the enclosed windows were classified as divergent if none of their F_{ST} values fell below the 75th percentile ($F_{ST} \geq 0.325$). After one iteration, this resulted in a set of divergent regions, each composed of windows with F_{ST} values at or above the 75th percentile and bounded by windows at or above the 95th percentile. Windows separated by more than 5 kbp in the chromosomal coordinates were not treated as consecutive and divergent regions were not allowed to cross such gaps between windows.

Population Genomics

Population Statistics - For most of our analyses, we grouped samples by species, *H. cydno*, *H. pachinus*, and *H. melpomene*. However, in two analyses we examined how genome-wide divergence evolves over time, focusing on both the cumulative portion of the genome contained in divergent regions and mean d_{XY} (Figure 5A). For these analyses only, we separated *H. melpomene* samples into east and west collecting locations, to estimate within species divergence at t=0 (IMa2 analyses consistently yielded a *melpomene* East vs. *melpomene* West divergence time equal to the smallest sampled parameter value which is effectively zero).

For all our analyses, the following population genetic statistics were calculated over each window using the command line version of Arlequin 3.5.1.3: F_{ST}, segregating sites, π within species, d_{XY}, and Tajima's D. The data were treated as unphased genotypic data for every individual, except for the Z chromosome where the window sequence data was input as two haplotypes for each male and a single haplotype for each female.

Comparing Divergent Regions to the Genome - We took the union of all divergent regions between the species pairs *H. cydno-H. pachinus*, *H. cydno-H. melpomene*, and *H. pachinus-H. melpomene* as a combined set, which was then compared to the remaining portion of the genome for a variety of population genetic statistics. This set consisted of 941 genomic regions, containing 6,637 windows, spanning 32,983,224 bp of the genome (14.6% of the mapped chromosomes). Comparisons between divergent regions and the rest of the genome focused on within species statistics calculated across the union of all divergent regions. The 97.5 and 2.5 percentile confidence intervals around the mean values were computed by bootstrap resampling from the entire set of windows 10,000 times. *P*-values, comparing the difference between the means for within and outside divergent regions, were estimated by bootstrap resampling and were adjusted to control for multiple tests (Benjamini and Hochberg, 1995).

Linkage Disequilibrium - Pairwise Linkage Disequilibrium (LD), within and outside divergent genomic regions, was calculated as the squared correlation coefficient (r^2) between allele counts observed at two SNPs using the VCFtools software package (Danecek et al., 2011). This approach is computationally feasible for large data sets since it does not require haplotype reconstruction, but it provides only an approximation of the true LD (Rogers and Huff, 2009). A total of 6,000 5 kbp windows, representing highly divergent regions across all comparisons of our three focal species, were analyzed, and LD within these regions was compared to LD within 6,000 randomly sampled 5 kbp windows representing the genomic background. Pairwise linkage disequilibrium for divergent regions and background windows were separately averaged in 1bp bins for each species (Figure S6). Clear differences in pairwise LD can be observed between divergent and background comparisons within species, and also

between species. r^2 and log (distance) was linearly regressed for each comparison, and 95% confidence intervals were calculated by bootstrapping with 1000 replicates. Results of these comparisons are shown in Figure 6D.

Derived Allele Frequency and Patterson's D Statistics - The mean derived allele frequency and Patterson's D statistic were computed over each window. Both statistics require polarizing the allelic type into ancestral and derived. For this purpose, the *H. ismenius* sample and the *H. hecale* sample were, together, used as the outgroup. Only sites where the outgroup was fixed and had at least two acceptable allele calls were retained within each window. The mean derived allele frequency per species per window was computed as the mean frequency of non-ancestral alleles at sites that segregate within the population. For Patterson's D statistic, we further reduced the sites in each window to those that were strictly bi-allelic. Patterson's D was calculated in true phylogenetic order (*H. cydno*, *H. pachinus*, *H. melpomene*, outgroup; Figure 1A), thus testing the impact of gene-flow between *H. pachinus* and *H. melpomene* (ABBA) and *H. cydno* and *H. melpomene* (BABA). Patterson's D generates a single statistic across the three species, so in this case we compared its absolute value as a measure of increased bi-directional gene-flow amongst the three sets of divergent regions, *H. cydno*-*H. pachinus* regions, *H. pachinus*-*H. melpomene* regions and *H. cydno*-*H. melpomene* regions (Figure 6H).

Clustering Analysis

H. cydno – H. pachinus - To test if the counts of divergent regions were overrepresented or underrepresented on any chromosome (versus a null expectation based on chromosome length) we used a Monte-Carlo simulated non-parametric paired Wilcoxon test ($Z = -1.949$, $p=0.05$). The probability of observing regions of high divergence between *H. cydno* and *H. pachinus* on a chromosome containing a known color-pattern locus is given by a contingency table of color pattern chromosomes (chr1, chr10, chr15 & chr18) versus not. Since there were so few observations in this comparison, we only counted the numbers by chromosome and did not take into account chromosome size (Fisher's exact test for number of diverged regions in these two chromosome classes, simulated $p < 0.01$).

Comparisons with *H. melpomene* - Equivalent tests for *H. cydno – H. melpomene* and *H. pachinus – H. melpomene* were performed using the non-parametric simulated paired Wilcoxon test, as above. In all cases, the distribution of diverged regions did not differ from the random expectation based on placing the same number of regions on chromosomes in proportion to their size (all $Z <= -5.06$, all $p > 0.61$). To test for enrichment of divergent regions on color-pattern chromosomes versus non-color pattern chromosomes for both comparisons (*H. cydno – H. melpomene* and *H. pachinus – H.*

melpomene) we tested a contingency table of regions on color pattern chromosomes (chr1, chr10, chr15 & chr18) versus not on these chromosomes normalized by chromosome length. The observed number did not differ from the null hypothesis of no clustering (Fisher's exact tests, $p > 0.538$ in both cases).

Structural variant detection

We used pair-end reads for gap alignment by BWA (-e 30) to detect indels (small Insertion/Deletion). Then we used SAMtools pileup to detect the indels from mapping reads with gap. Gaps that were supported by at least 3 paired-end reads were extracted. Inversions and other additional structural variants were identified with BreakDancer vsn 1.1 (Chen et al., 2009) and filtered based on the BWA quality score. We did not find any fixed inversion among species in the *H. cydno – H. pachinus* divergent regions (Table S4).

GO term enrichment analysis

To test for enrichment of specific gene ontology categories in divergent regions of the genome, gene sequences falling within C/P, C/M, and P/M regions were extracted from the Hmel1.1 gene annotations. Because the current gene ontology terms associated with Hmel1.1 are sparsely assigned, we blasted Hmel1.1 gene against known genes in the *Drosophila melanogaster* (Dmel Release 5.48, FlyBase). Hmel genes that blasted to Dmel genes with e-values < 0.05 were considered matched. We then used appropriate FlyBase accessions (e.g., accessions within vs. accessions outside of divergent regions) as the input for GO Elite (http://www.genmapp.org/go_elite/). We combined permuted probabilities from the merged GO Elite analysis for the three comparisons using Fisher's method and then adjusted the tests for multiple comparisons[12] based on the total number of genes in the comparison set, multiplied by 3 to further correct for the three non-independent comparisons (C/P, C/M & P/M) this method identifies GO terms enriched across all three comparisons (Table S6).

Supplemental References

Chen, K., Wallis, J.W., McLellan, M.D., Larson, D.E., Kalicki, J.M., Pohl, C.S., McGrath, S.D., Wendl, M.C., Zhang, Q.Y., Locke, D.P.*, et al.* (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat. Methods *6*, 677-681.