Selection and gene flow define polygenic barriers between incipient butterfly species

Steven M. Van Belleghem[1*], Jared M. Cole[2,3], Gabriela Montejo-Kovacevich[4], Caroline N. Bacquet[5], W. Owen McMillan[6], Riccardo Papa[1,7] and Brian A. Counterman[2+]

[1]Department of Biology, University of Puerto Rico, Rio Piedras, Puerto Rico.
[2]Department of Biological Sciences, Mississippi State University, Mississippi State, USA.
[3] Department of Integrative Biology, University of Texas at Austin, Austin, Texas, USA.
[4] Department of Zoology, University of Cambridge, Cambridge, UK.
[5] Universidad Regional Amazonica Ikiam, Tena, Ecuador.
[6] Smithsonian Tropical Research Institute, Panamá, Panama.
[7] Molecular Sciences and Research Center, University of Puerto Rico, San Juan, PR.

[*]e-mail: Steven.Van@upr.edu
[+]e-mail: BCounterman@biology.msstate.edu

Abstract

Characterizing the genetic architecture of species boundaries remains a difficult task. Hybridizing species provide a powerful system to identify the factors that shape genomic variation and, ultimately, identify the regions of the genome that maintain species boundaries. Unfortunately, complex histories of isolation, admixture and selection can generate heterogenous genomic landscapes of divergence which make inferences about the regions that are responsible for species boundaries problematic. However, as the signal of admixture and selection on genomic loci varies with recombination rate, their relationship can be used to infer their relative importance during speciation. Here, we explore patterns of genomic divergence, admixture and recombination rate among hybridizing lineages across the *Heliconius erato* radiation. We focus on the incipient species, *H. erato* and *H. himera*, and distinguish the processes that drive genomic divergence across three contact zones where they frequently hybridize. Using demographic modeling and simulations, we infer that periods of isolation and selection have been major causes of genome-wide correlation patterns between recombination rate and divergence between these incipient species. Upon secondary contact, we found surprisingly highly asymmetrical introgression between the species pair, with a paucity of *H. erato* alleles introgressing into the *H. himera* genomes. We suggest that this signal may result from a current polygenic species boundary between the hybridizing lineages. These results contribute to a growing appreciation for the importance of polygenic architectures of species boundaries and pervasive genome-wide selection during the early stages of speciation with gene flow.

Introduction

Disentangling the factors that drive genomic divergence is necessary for advancing our understanding of speciation. Targets of selection, for example, may be responsible for adaptive differences between species; and their number, distribution and effect on gene flow along the genome define the architecture of species boundaries. In population genomic studies, targets of selection are expected to show elevated divergence between species and reduced genetic diversity within species [1]. These highly divergent loci often reflect local adaptation and/or incompatibilities between species, and can be considered the loci that define the species [2,3]. This is because natural selection acts as a local genomic "barrier" to gene flow between hybridizing species [4–6]. In contrast, the rest of the genome, which is not under such selective pressures, may be expected to exchange more freely between the species (i.e. admixture). However, genetic variation at these latter genomic regions can be greatly impacted by neutral demographic processes (i.e. population size and migration) and the indirect effects of nearby targets of selection (i.e. linked selection). Local recombination rates can further influence how these processes impact genetic variation, which collectively result in highly heterogenous patterns of genome-wide divergence [1,4,7–10]. Thus, the challenge is to distinguish those targets of selection and demographic processes that generate the genomic landscape of divergence.

Here, we reconstruct the history of demographic isolation and characterize the extent to which selection has shaped genomic divergence between two closely related, hybridizing *Heliconius* species. More precisely, we first use a demographic modeling approach to reconstruct the history of population sizes and isolation during divergence of these species [11–18]. Next, with knowledge of the most likely demographic history, we use coalescent simulations to test for the importance of selection as the underlying mechanism driving heterogeneous patterns of genomic divergence. The coalescent simulation approach allows us to explore other genomic factors, such as recombination rate, on genomic divergence. Specifically, we can test for genome-wide impacts of linked selection by using the expectation that linked selection is higher in regions of the genome where recombination rate is lower [19]. Hence, we would expect a negative association between recombination rate and divergence across the genome [8,9,20,21]. Similarly, we expect a positive association between recombination rate and admixture, if the species continue to hybridize [4,22], but not necessarily if they diverged in isolation [8,23]. Thus, the relationships of recombination rate with divergence and admixture can be used to infer the relative importance of different evolutionary processes during speciation.

To provide a relative perspective of the divergence between our focal hybridizing *Heliconius* species, we first investigate the relationship between reproductive isolation and genomic divergence across 15 pairs of increasingly divergent populations and species in the *H. erato* clade. Next, we use the incipient species *Heliconius erato* and *Heliconius himera* that hybridize across three geographically distinct contact zones to test for the relative contribution of demographic and selective factors in the evolution of the divergence landscape. *Heliconius himera* is found in dry forest areas of southern Ecuador and northern Peru [24]. It comes into contact with *Heliconius erato cyrbia* on the western slopes of the Ecuadorian Andes and with *Heliconius erato favorinus* and *Heliconius erato emma* on the eastern slopes of the Andes, both areas with wet forest (Figure 1A). Hybridization is ongoing and hybrids are easily identifiable by their wing color patterns. In Ecuador, hybrids compose approximately 5% of the population in the contact zone [25], and, although poorly characterized, hybridization is known to occur in the other contact zones. The two species show strong premating isolation but little or no postmating reproductive barriers [25]. The eastern and western *H. erato* populations that hybridize with *H. himera* do not come into contact with each other and show deep genetic divergence in the *H. erato* clade [26].

Our findings demonstrate the importance of both isolation and selection in establishing the heterogeneous genomic landscape of divergence that characterizes our hybridizing species. The reconstructed demographic history supports a complex dynamics of population fluctuations and varying migration rates, that with selection, resulted in the observed heterogeneous patterns of genomic divergence. Further, our coalescent simulations well fit the observed relationship of genomic divergence and recombination when we consider genome-wide impacts of recent and strong selective events in the absence of gene flow. Finally, we show that the species boundary between *H. himera* and *H. erato* is highly porous and that gene flow is highly asymmetrical and distinct across each of the three contact zones. We suggest that this asymmetrical signal of gene flow may be the result of polygenic species boundaries that restrict introgression in *H. himera*. Overall, our results highlight that the study of heterogeneous landscapes of divergence can help us understand how demographic and selective processes drive speciation.

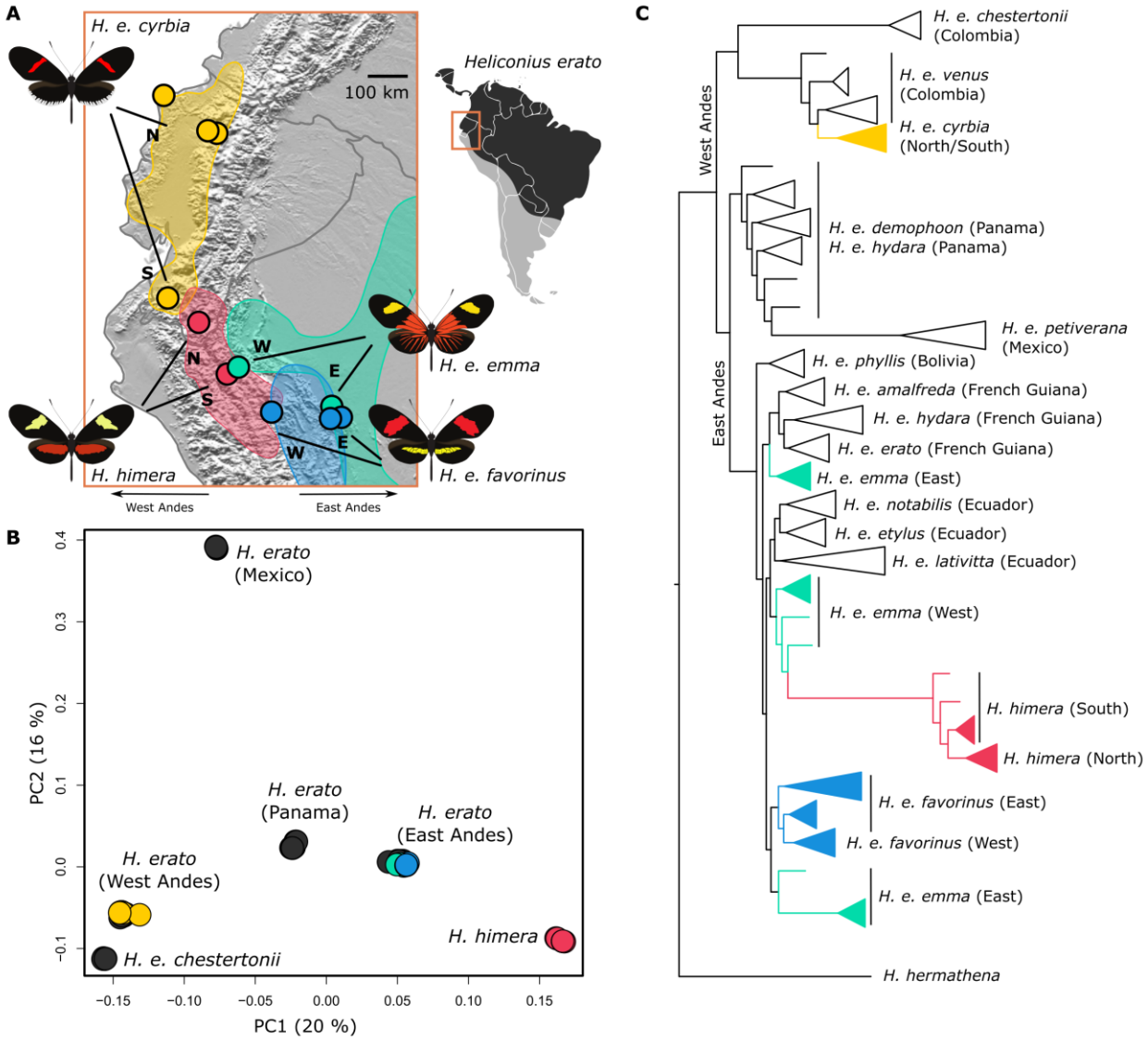


Figure 1. Geographical distribution, population structure and phylogeny of the focal populations in relation to the *Heliconius erato* radiation. (A) We sampled two populations of *H. himera*, *H. e. cyrbia*, *H. e. emma* and *H. e. favorinus*. The distribution of *H. himera* (red) covers dry valleys in the Andes of South Ecuador and North Peru. In the North, *H. himera* (N) comes into contact with a *H. e. cyrbia* (S) population. In the South, *H. himera* comes into contact with a *H. e. emma* (W) and *H. e. favorinus* (W) population. **(B)** Principal Component Analysis (PCA) of the focal samples (colored points) among all the available whole genome data for the *H. erato* radiation (black points). **(C)** Maximum likelihood tree built using FastTree and using only autosomal sites from 121 whole genome resequenced individuals (see Figure S1 for the uncollapsed tree). Nodes in the tree that represent the major clades within *H. erato* (east and west of Andes) obtained high support (= 1) from the Shimodaira-Hasegawa test.

Results & discussion

Genomic landscape of divergence among hybridizing races and incipient species

We first sought to describe the general patterns of genome-wide divergence and how they varied based on the varying degrees of reproductive isolation. To do this, we compared genome-wide patterns of divergence across 15 contact zones in the *H. erato* clade that have varying degrees of reproductive isolation (*RI*, Figure 2; Table S1). Correlations of *RI* with genome-wide estimates of relative divergence (F_{ST}) show that divergently selected color patterns between hybridizing races of *H. erato* with absence of other pre- or post-mating barriers are not sufficient to drive genome-wide increases in divergence (Figure 2A). In these hybridizing *H. erato* races, there are only narrow peaks of divergence largely centered over the loci known to be responsible for color pattern differences (Figure 2B). In contrast, between the incipient species *H. erato cyrbia* and *H. himera*, which have strong differences in mate preference, divergence is much higher across the entire genome. Divergence between these species has increased to the extent that F_{ST} peaks near the known color pattern loci *WntA* (chr 10), *cortex* (chr 15), and *optix* (chr 18) are not detectable (Figure 2B). The *H. himera* and *H. erato* contact zones on the eastern Andes (*H. e. emma* and *H. e. favorinus*) also show elevated genome-wide divergence, but much lower overall than in the western Andes contact zone. As expected, genomic divergence was highest between *H. e. venus* and *H. e. chestertonii* from Colombia, where both mate preference and hybrid sterility have been reported [27]. We also note a dramatic increase in divergence on the Z chromosome relative to the autosomes (Figure 2). This is in line with previous work on *H. e. chestertonii*, which suggested the important role played by the Z chromosome as a barrier to gene flow [28].

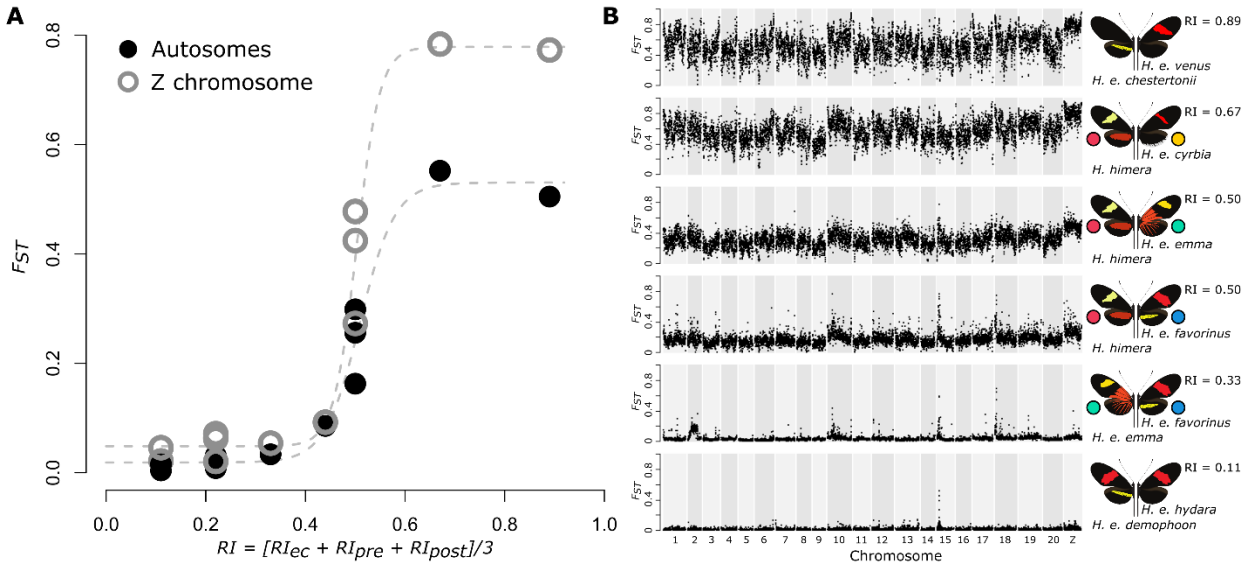


Figure 2. Reproductive isolation and divergence among *Heliconius erato* populations. (A) Genome-wide averages of relative divergence (F_{ST}) show a sharp increase with increasing measures of reproductive isolation between parapatric *H. erato* populations. The measure of reproductive isolation (*RI*) was obtained by equally weighting ecological (RI_{ec}), pre-isolation (RI_{pre}) and post-isolation (RI_{post}) components (Table S1). Higher relative divergence on the Z chromosome can be observed for the more divergent parapatric comparisons, however, incompatibilities that are potentially Z-linked have only been suggested for *H. e. chestertonii* crosses [27,28]. **(B)** Plots of relative divergence (average F_{ST} in 50 kb windows) between parapatric *H. erato* populations along the genome. Plots are ordered according to the measure of reproductive isolation (*RI*). Colored circles match color codes used for the focal populations in this study. Divergence peaks on chromosome 10, 15 and 18 correspond to the divergently selected color pattern genes *WntA* (affecting forewing band shape), *cortex* (affecting yellow hindwing bar), and *optix* (affecting red color pattern elements), respectively [26].

Demographic change and selection jointly drive genomic divergence among incipient species

Collectively, our results provide a view into the genomic landscape among lineages with increasing degrees of reproductive isolation (Figure 2)[29,30]. However, the increase in divergence does not seem to be a linear process as there is a marked increase in divergence between the incipient species that are known to still frequently and continuously hybridize over many generations. To understand what drives these elevated patterns of divergence, we have to recognize that each of these contact zones reflect hybridization between evolutionary distinct lineages. In this regard, the genomic landscapes do not reflect a continuum of genomic divergence throughout speciation, but rather they each are the evolutionary outcomes of various neutral and adaptive processes that shaped each of the populations coming into contact. For example, we find increased divergence between *H. himera* and *H. erato* from the western Andes slopes compared to *H. erato* from eastern Andes slopes. As seen in the PCA and phylogenetic inference, this results from a deeper split between *H. himera* and *H. e. cyrbia* from the western Andes slope compared to *H. himera* and the *H. erato* races that are found east of the Andes,

147 including *H. e. emma* and *H. e. favorinus* (Figure 1B, C). This is consistent with previous studies that
148 placed *H. himera* nested within the *H. erato* clade and not as a sister species to *H. erato* [26,31].

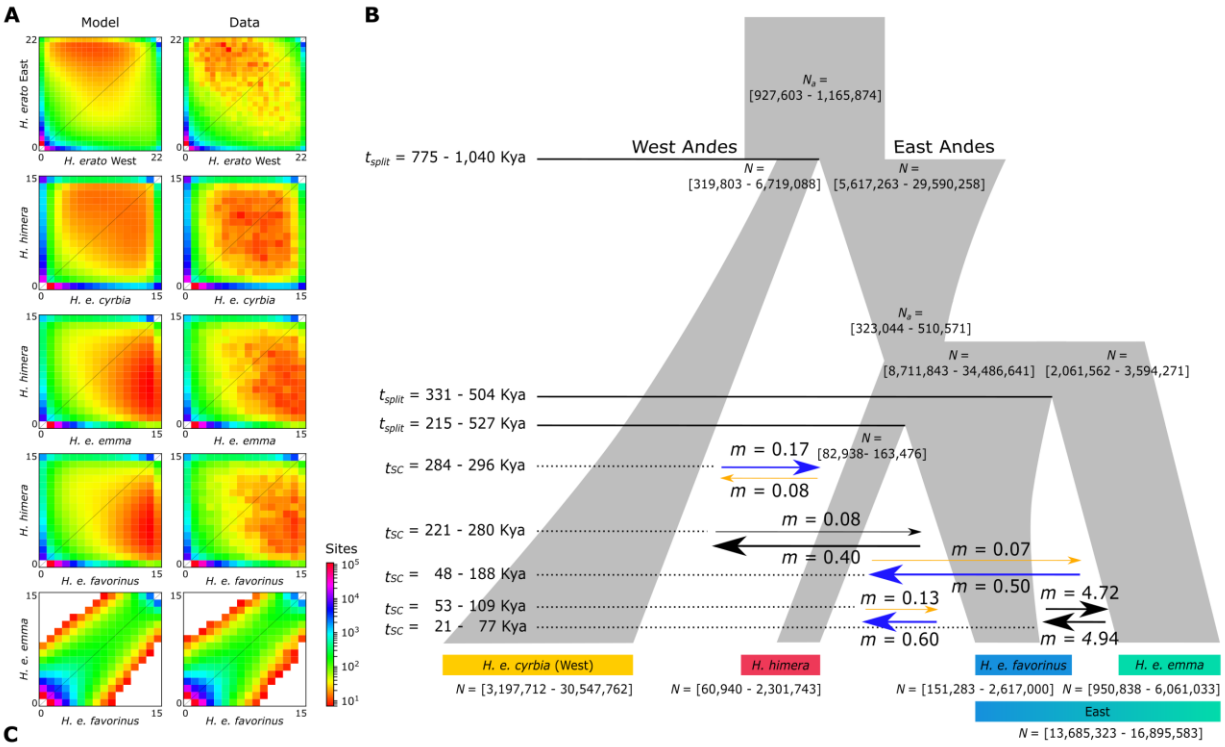
149 To understand what forces are likely driving differences in patterns of genome-wide divergence
150 between *H. erato and H. himera*, we fit the estimated joint site-frequency spectrum (JSFS) for three
151 geographically distinct contact zones to 26 alternative demographic scenarios that varied in split times,
152 migration rates and population sizes (Figure 3). All three *H. erato* and *H. himera* contact zones best fit
153 models that included secondary contact (SC) after a period of isolation without gene flow (Figure S2-3;
154 Table S2). For all three zones we found support for asymmetrical migration. In each case, migration
155 rates were predominantly in one direction, with on average 0.5 to 0.6 migrants per generation moving
156 from *H. erato* into *H. himera*, compared to 0.07 to 0.13 moving from *H. himera* into *H. erato*. This result
157 is consistent with the effective migration rates being driven by the marked population size differences
158 between the two species (Figure 3B).

159 Nearly all the models that included exponential population growth (G) best fit the JSFS, with the
160 exception of the *H. himera* and *H. e. emma* comparison. Estimates of ancestral and contemporary
161 population sizes suggest strong expansions in *H. himera* and *H. erato*. These inferred changes in
162 population sizes broadly fit previous results obtained from pairwise sequentially Markovian coalescent
163 (PSMC) analysis (Van Belleghem et al. 2018), which suggested an overall population growth in *H. erato*
164 east and west of the Andes in the past 1 My and size reduction for *H. himera* in the past 200 Ky (Figure
165 S3). However, we found that estimates of contemporary population sizes varied greatly depending on
166 the population comparison (Figure 3B, C), a result possibly explained by unaccounted population
167 structure and difficulties in estimating growth (G). For *H. e. favorinus,* estimates of contemporary
168 population size were generally much smaller than the ancestral population. This result fits the
169 observation that *H. e. favorinus* is a smaller Andean population of the "postman" color pattern (i.e. red
170 forewing band), which has a much larger distribution throughout the Neotropics. Collectively, the
171 models support a history that includes periods of allopatry, followed by lineage specific changes in
172 population size that coincide with more recent gene flow.

173 To investigate if the JSFS contained evidence of selection driving patterns of divergence across
174 the contact zones, we incorporated heterogeneity in population size (2N) and migration rate (2M) into
175 the models, similar to what was done by Rougeux *et al.* 2017 [32] and Tine *et al.* 2014 [16], respectively.
176 The 2N model allows heterogeneity in population size estimates across loci that result from the
177 differences in allelic variation caused by linked selection (*ls* = effective population size of locus relative to

178 neutral loci; Q = proportion of the genome affected by *ls*). We found that all contact zones between *H.*
179 *erato* and *H. himera* well supported 2N models, suggesting the effect of linked selection in shaping
180 patterns of genomic variation and divergence between the incipient species. The strongest *ls* was
181 observed for the population comparisons of *H. erato* East and West (*ls* = 0.10; *Q* = 0.35) and *H. himera*
182 and *H. erato* West (*ls* = 0.15; *Q* = 0.90) and the lowest observed between *H. e. emma* and *H. e. favorinus*
183 (*ls* = 0.07; *Q* = 0.04) (Figure 3C, Table S3).

184 The result of selection on locally adapted alleles is a heterogeneous landscape with regions
185 containing these variants showing much lower rates of admixture compared to the rest of the genome.
186 The 2M models allow for this type of heterogeneity in migration rates across the genome. Both contact
187 zones in the eastern Andes supported these models for *H. himera* and *H. erato*, suggesting that the
188 model fits the eastern Andean populations having genomic regions with much lower rates of
189 introgression than other parts of the genome.



C

Comparison	Best Model	t_{split} (years)	t_{SC} (years)	$m_{1<-2}$	$m_{2<-1}$	$me_{1<-2}$	$me_{2<-1}$	P	ls	Q
H. erato East vs. West Andes	SC2NG	[889531 - 1039826]	[221048 - 279532]	[0.081 - 0.084]	[0.396 - 0.407]	-	-	-	[0.090 - 0.100]	[0.340 - 0.350]
H. himera vs. *H. e. cyrbia*	SC2NG	[774568 - 824039]	[284457 - 296405]	[0.160 - 0.170]	[0.075 - 0.077]	-	-	-	[0.140 - 0.160]	[0.860 - 0.900]
H. himera vs. *H. e. emma*	SC2N2m	[215208 - 527389]	[48373 - 187901]	[0.400 - 0.580]	[0.060 - 0.080]	[0.69 - 1.76]	[0.270 - 4.220]	[0.360 - 0.670]	[0.260 - 2.360]	[0.190 - 0.900]
H. himera vs. *H. e. favorinus*	SC2N2mG	[224160 - 337996]	[53421 - 108780]	[0.500 - 0.690]	[0.100 - 0.160]	[0.58 - 0.98]	[3.960 - 8.320]	[0.510 - 0.620]	[0.180 - 1.580]	[0.070 - 0.870]
H. e emma vs *H. e. favorinus*	SC2N2mG	[331076 - 504316]	[29944 - 112173]	[1.840 - 7.600]	[0.970 - 8.900]	[0.00 - 2.92]	[0.010 - 14.91]	[0.000 - 0.870]	[0.050 - 0.090]	[0.030 - 0.040]

Comparison	Best Model	N_a	N1 [at split]	N2 [at split]	g1	g2	N1 [at sampling]	N2 [at sampling]
H. erato East vs. West Andes	SC2NG	[927603 - 941642]	[26341817 - 29590258]	[319803 - 406187]	[0.520 - 0.570]	[9.990 - 12.74]	[13685323 - 16895583]	[3197712 - 5176219]
H. himera vs. *H. e. cyrbia*	SC2NG	[1146766 - 1165874]	[5617263 - 7292527]	[5669976 - 6719088]	[0.280 - 0.320]	[4.360 - 4.540]	[1554981 - 2301743]	[24746873 - 30547762]
H. himera vs. *H. e. emma*	SC2N2m	[421200 - 482382]	[82938 - 197144]	[2484021 - 5543987]	-	-	-	-
H. himera vs. *H. e. favorinus*	SC2N2mG	[323044 - 348395]	[102651 - 163476]	[8711843 - 34486641]	[0.590 - 0.800]	[0.020 - 0.080]	[60940 - 130182]	[151283 - 2617000]
H. e emma vs *H. e. favorinus*	SC2N2mG	[498501 - 510571]	[2061562 - 3594271]	[17935128 - 25379445]	[0.460 - 1.690]	[0.010 - 0.050]	[950838 - 6061033]	[224427 - 1167647]

Figure 3. Secondary Contact (SC) best demographic model of the *H. himera* and *H. erato* population history. (A) Joint Site Frequency Spectra (JSFS) for data and best model (see Table S2 and Figure S3-4 for AIC values (Akaike Information Criterion)). **(B)** Reconstruction of historical demography of *H. himera* and *H. erato* populations using models with best AIC scores. All best models included a period of isolation and secondary contact. Arrows indicate effective migration rates ($2N_am$). Migration from *H. himera* into *H. erato* is indicated in orange, migration from *H. erato* into *H. himera* is indicated in blue. **(C)** Table with parameter ranges obtained from five best scoring models out of twenty runs. N_a = ancestral population size, N_1 = Size of population 1, N_2 = size of population 2, $g1$ = growth coefficient of population 1, $g2$ = growth coefficient of population 2, *ls* = linked selection, *Q* = proportion of the genome with a reduced effective size due to linked selection (*ls*), $m_{1<-2}$ = migration from population 2 into 1, $m_{2<-1}$ = migration from populations 1 into 2, t_{split} = split time, t_{SC} = time of secondary contact, *P* = proportion of the genome evolving neutrally.

Simulations of linked selection and secondary contact predict correlation of recombination rate with divergence, not admixture

The demographic models provide a comprehensive reconstruction of the evolutionary history of divergence between *H. erato* and *H. himera* and allow us to estimate time intervals of lineage splits, size changes and migration changes that span the past million year of divergence between *H. himera* and *H. erato* in the Andes (Figure 3). We next use these estimates to inform simulations and conduct further

tests for the impact of selection and demography during genomic divergence of the incipient species. To do this, we simulated the expected relationship of recombination rate to relative divergence (F_{ST}) and admixture (f_d) at a locus linked to a site under selection. We used two distinct scenarios of migration broadly applicable *to H. erato* and *H. himera* and compared the simulation results to real data (Figure 4A, B).

As demonstrated by other studies, in a scenario of "isolation with migration" (IM) with long periods of migration and divergent selection, our simulations predict a strong negative correlation between recombination rate and F_{ST} [20,21,33] and a strong positive correlation between recombination rate and f_d [4,8] (Figure 4C, left). This pattern reflects the degree of linkage between neutral sites and loci under divergent selection [23,34]. In contrast, our simulations show that the relationship between recombination rate and f_d is reduced when migration is more recent or low (Figure 4C, left). As expected, in a "secondary contact" (SC) scenario that is characterized by a more recent onset of migration, the simulations show that F_{ST} values are generally high and f_d values close to zero (Figure 4C, right). This results in the absence of a relationship between recombination rate and F_{ST} or f_d in most SC scenarios. However, when selection in the genome is recent (~selective sweeps), a negative relationship between recombination rate and divergence, but not admixture, emerges in the SC scenario (Figure 4C). This relationship arises due to linked selection that reduces diversity within populations and increases the relative divergence, F_{ST}, between populations [35]. The relationship holds for lower selection strengths at relatively short distances from the selected site (s = 0.02; Figure S5).

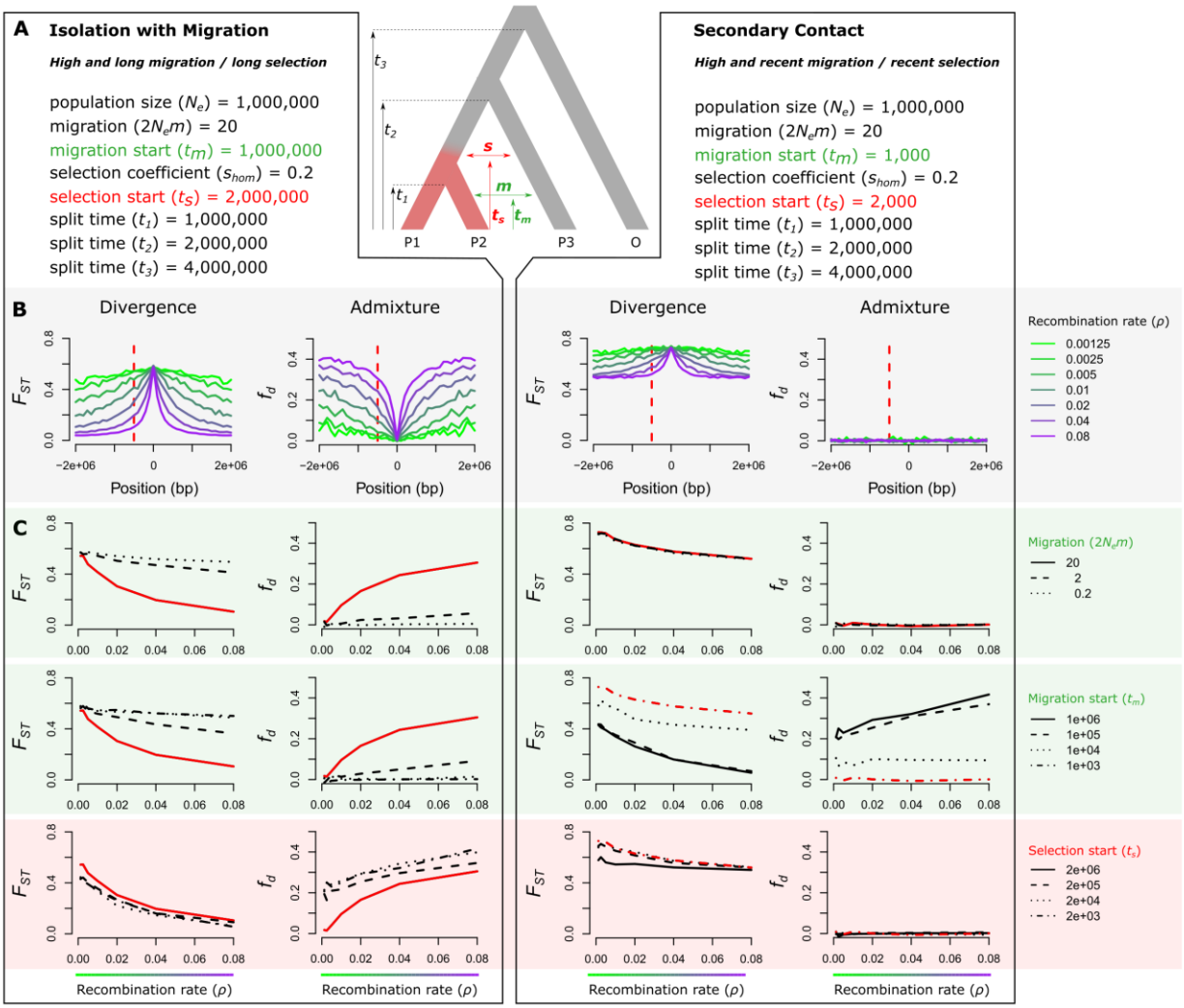


Figure 4. Expected relationship of recombination rate with divergence (F_{ST}) and admixture (f_d) near a divergently selected locus. (A) The population tree shows the simulated scenario with the onset of divergent selection on a derived allele indicated in red and in which migration rate (m) and migration time (t_m) between P2 and P3 and selection start time (t_s) are varied. Left and right of the simulated scenario are parameter combinations for two extreme scenarios that both include linked selection; on the left a scenario with Isolation with Migration (IM) and on the right a scenario reflecting Secondary Contact (SC). **(B)** Effect of population recombination rate (ρ) on relative divergence (F_{ST}) and admixture (f_d) near a divergently selected locus for the two simulated scenarios with parameter combinations as in panel A. The selected allele occurs at position 0. The dashed red line indicates a locus at 500 kb from the selected locus at which the relationship between ρ, divergence and admixture is assessed in panel C. **(C)** The effect of migration start time (t_m) and selection start time (t_s) on the relation between ρ, divergence and admixture. Apart from the respective parameters being evaluated, other parameters were fixed as in panel A, with the red lines indicating the exact parameter combinations as in panel A and B. For simulations with a lower selection coefficient (s = 0.02), see Figure S5.

Patterns of divergence are shaped by linked selection

The differences in the expected relationships between recombination rate and F_{ST} and between recombination rate and f_d under the IM and SC scenarios provide specific predictions that we can use to

244 determine how linked selection may have shaped genomic divergence between *H. erato* and *H. himera*.
245 We first explored this relationship of recombination rate, relative divergence and admixture across the
246 *optix* locus using the expected patterns obtained from our simulations. The *optix* locus controls red wing
247 pattern differences among *H. erato* races, as well as between *H. erato* and *H. himera* and is a target of
248 strong selection (Figure 5A) [26,31,36]. We found recombination rate (ρ) estimates were markedly lower
249 upstream, compared to downstream of the *optix* gene (Figure 5C). Such sharp decreases in
250 recombination rate near chromosome ends have been observed in other *Heliconius* species [4] and likely
251 explain the decrease in recombination rate upstream of the *optix* gene. To test if selection at the *optix*
252 locus has produced the expected relationship of recombination rate with divergence and admixture, as
253 predicted above in our simulations (Figure 4C), we sampled sites 500 kb from the center of the peak of
254 the divergence at *optix* and plotted the recombination rate and divergence estimates for those sites
255 (Figure 5D). As expected from the scenario of SC with linked selection, we found negative relationships
256 between recombination rate and divergence (F_{ST}) and no correlations between recombination rate and
257 admixture (f_d). These findings demonstrate that our simulated patterns of recombination, divergence
258 and admixture, reflect real observed patterns in a region known to be under strong selection.

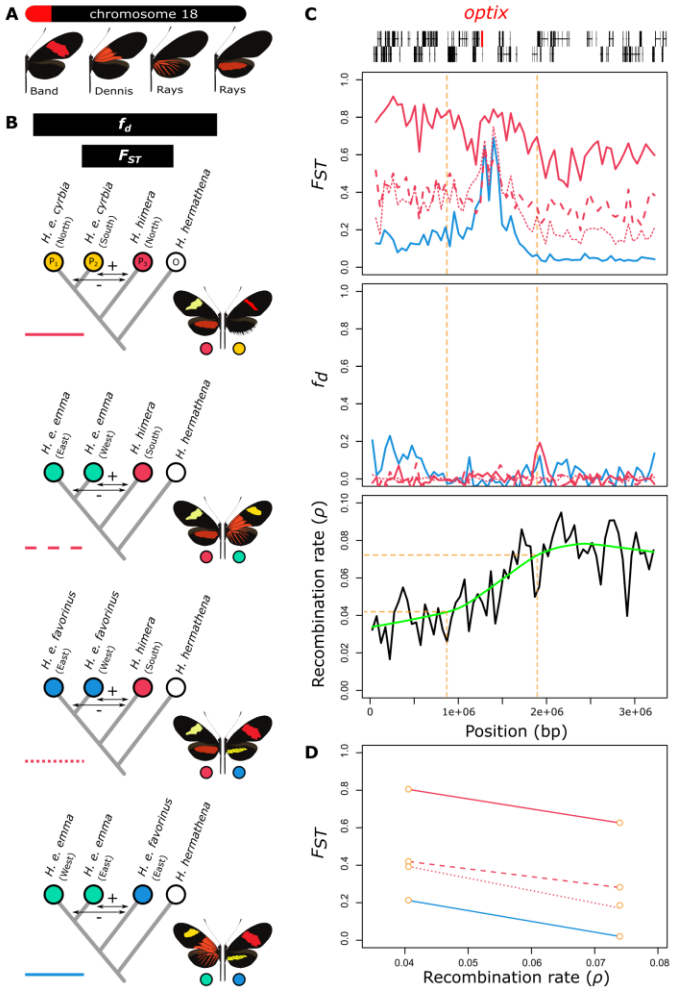


Figure 5. Divergence (F_{ST}), admixture (f_d) and recombination rate (ρ) near the red color pattern gene *optix*. (A) The *optix* gene is located near the start of chromosome 18 and has been demonstrated to control the expression of red color pattern elements in *Heliconius* wings [37]. **(B)** Relative divergence (F_{ST}) and admixture (f_d) comparisons performed between *H. himera*, *H. e. cyrbia*, *H. e. emma* and *H. e. favorinus*. Colored circles match color codes in Figure 2. **(C)** Lines show F_{ST}, f_d and recombination rate (ρ) calculated in 50 kb non-overlapping windows. The green line in the bottom plot shows a loess fit of the recombination rate. Gene models including the location of the *optix* gene are presented at the top. **(D)** Relationship between F_{ST} and recombination rate (ρ) 500 kb left and right from the center of the *optix* regulatory sequence divergence peak.

To test for evidence that linked selection has driven genome-wide patterns of divergence, we compared recombination rates with divergence and admixture across the whole genome. We found a significant negative association between recombination rate (ρ) and relative divergence (F_{ST}) between *H. himera* and *H. erato* populations (Figure 6A) but no association with admixture (f_d) (Figure 6B). These genome-wide patterns are identical to those observed at the *optix* locus (Figure 5D) and the simulations of SC with linked selection (Figure 4C). Further, we observed a positive association between proportion of coding sequence and relative divergence, which again suggests the importance of linked selection for

genome divergence (Figure S6). Additionally, we found a significant relationship between F_{ST} and f_d in the *H. himera – H. erato* hybrid zones, which suggests that, although admixture can partly explain reduced F_{ST} (Figure 6C), the rates of migration have been too low or recent to build up a significant association with recombination rate. Finally, we note that the observed patterns of linked selection may include the effects of both genetic hitchhiking and background selection. While our genomic dataset does not allow us to differentiate between them, our simulations suggest that the observed patterns can be explained by genetic hitchhiking alone and other studies suggest background selection may be too subtle to cause these patterns [9].

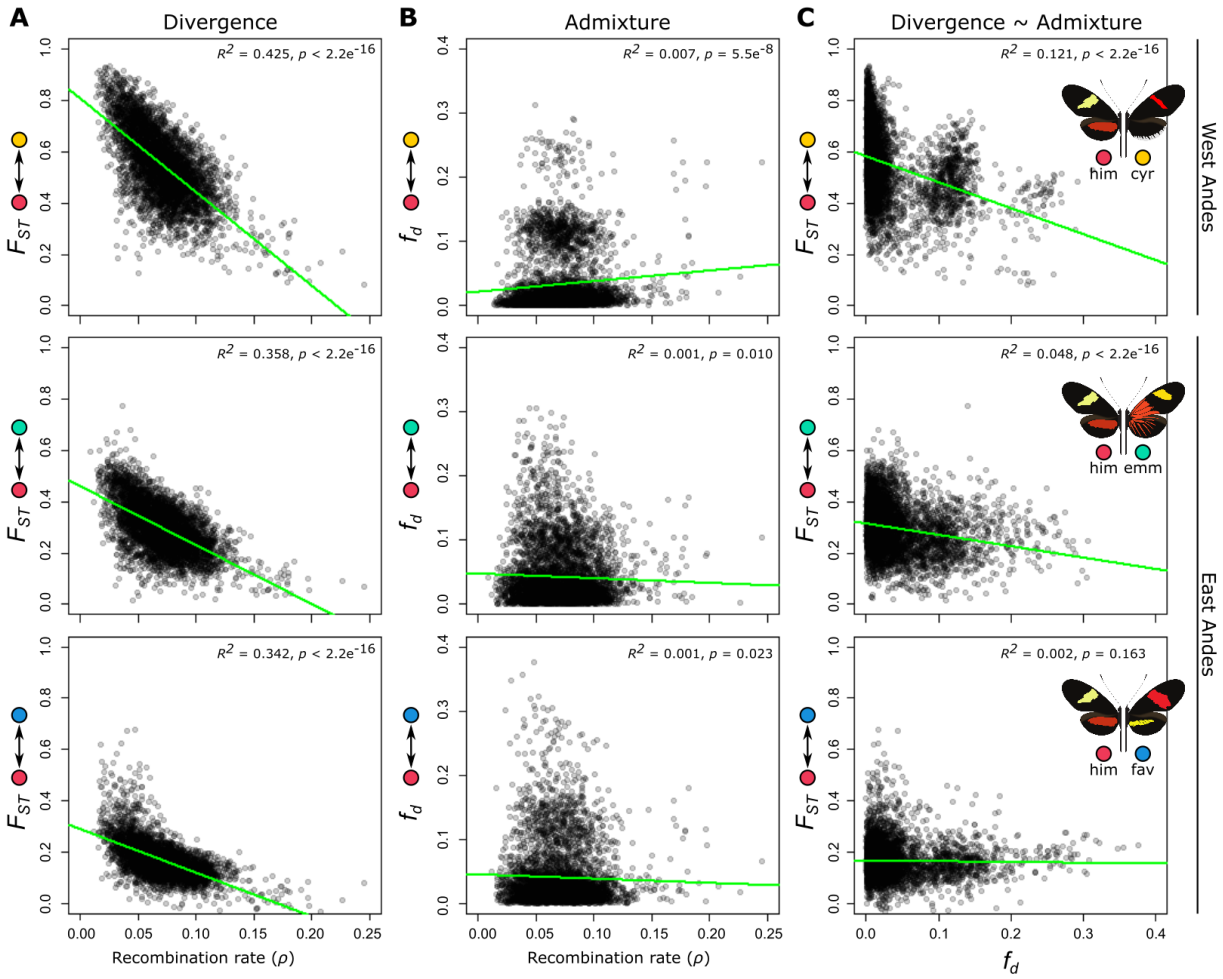


Figure 6. Correlations of divergence and admixture proportions with recombination rate in the three *H. himera – H. erato* contact zones. (A) Relative divergence (F_{ST}) versus recombination rate (ρ). **(B)** Admixture proportion (f_d) versus recombination rate (cM/Mb). **(C)** Relative divergence (F_{ST}) versus admixture proportion (f_d). Statistics were calculated in 50 kb non-overlapping windows. Recombination rates were calculated from each *H. erato* population separately and averaged over populations (see methods) and showed a genome-wide average of ρ equal to 0.071 (SD = 0.026; $\rho = 4N_e r$). Colored circles match geographic distributions and contact zones in Figure 2.

Asymmetrical admixture suggests a polygenic species boundary

The combination of empirical and simulation data suggests that periods of isolation and linked selection within populations have played a major role in shaping the divergence landscape between *H. erato* and *H. himera*. To explore how these factors influence the finer details of genomic divergence we investigated patterns of admixture along chromosomes. We were both interested in the patterns of admixture across our three replicate hybrid zones and inferring the direction of admixture. To determine if patterns of admixture were similar across the three contact zones, we used a modified four-taxon *D*-statistic called f_d [38]. Across several chromosomes we observed large blocks of increased f_d, particularly in the hybrid zone between *H. himera* and *H. e. cyrbia* west of the Andes (see chromosomes 3, 4, 6, 7, 8, 12 and 18 in Figure 7A). The large size of these admixture tracks indicates that these signals are recent and there has not been sufficient time for recombination to break down the haplotypes. In contrast, high admixture values (f_d) between *H. himera* and *H. e. emma* and *H. himera* and *H. e. favorinus* east of the Andes are distributed more evenly along the chromosomes (Figure 7A), which would agree with older admixture events between these populations. In general, there was a lack of overlap between genomic regions with high f_d in the different hybrid zone comparisons further reinforcing the idea that we are examining independent admixture events.

To determine the directionality of introgression across the genome we used a five-taxon D_{FOIL}-statistic [39]. This test considers all available taxa (i.e. two populations of *H. himera* and two populations of *H. erato*) and calculates all possible four-taxon *D*-statistic comparisons to infer admixture as well as directionality of admixture. While the D_{FOIL}-statistic is calculated using a single genome for each taxon, we performed this test on all possible combinations of available samples and represented the D_{FOIL} signal as the proportion of significant comparisons (Figure 7A). Comparing f_d and D_{FOIL} results revealed that among loci that show strong evidence of admixture, there is a relative paucity of loci in *H. himera* individuals carrying *H. erato* alleles (Figure 7B). This asymmetry suggests *H. erato* is more porous to introgressed alleles from *H. himera* than vice-versa. Consistent with a well characterized "large X(Z) effect" in speciation, on the sex (Z) chromosome there are only a few loci that show signals of admixture (high f_d) and again the D_{FOIL} tests only show evidence of introgression of *H. himera* alleles into *H. erato* (Figure 7A).

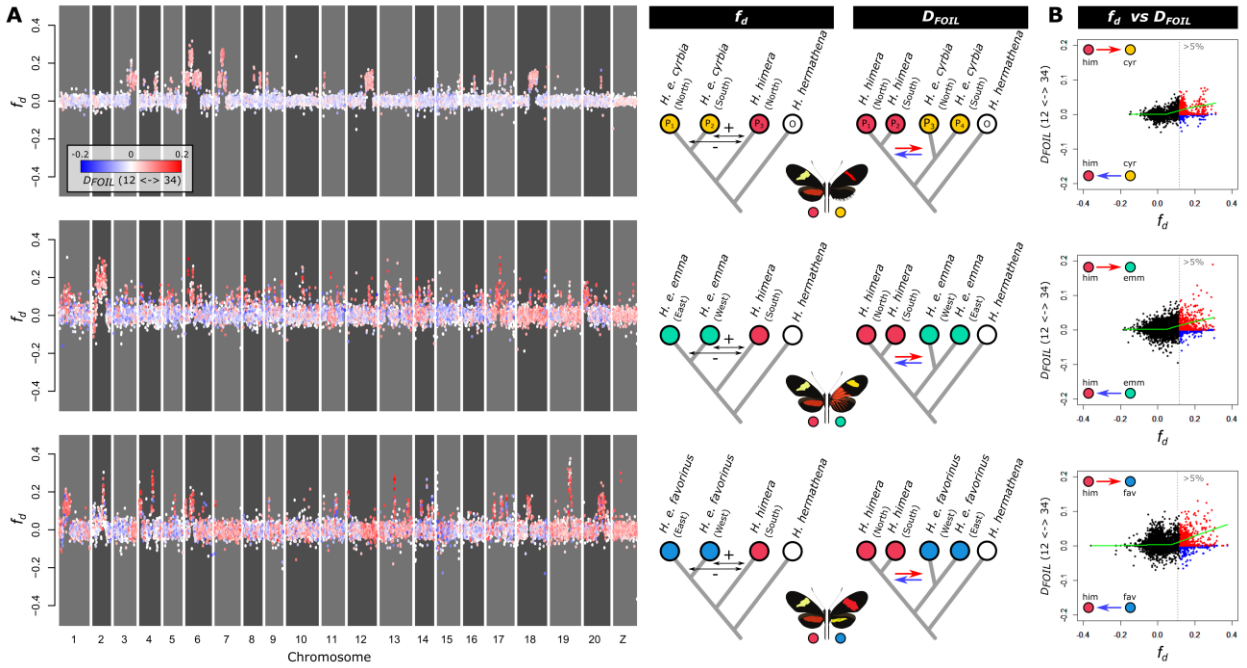


Figure 7. Admixture (f_d) and admixture directionality (D_{FOIL}) between *H. himera* and *H. erato* in three contact zones. (A) Points show admixture (f_d) values, whereas coloring shows directionality (D_{FOIL}) in 50 kb non-overlapping windows for the contact zones *H. himera – H. e. cyrbia* (top), *H. himera – H. e. emma* (middle) and *H. himera – H. e. favorinus* (bottom). Blue indicates predominant admixture from *H. erato* into *H. himera* (12 <- 34), whereas red indicates predominant admixture from *H. himera* into *H. erato* (12 -> 34) based on the D_{FOIL} tests. **(B)** Summary of admixture versus directionality with points above the 95% quantile indicated in blue (12 <- 34) and red (12 -> 34) demonstrates that the majority of windows with high f_d indicate admixture from *H. himera* into *H. erato*. The green line indicates a loess fit of the data. Colored circles match color codes in Figure 2.

We propose that the multitude of loci with asymmetrical gene flow may represent the genetic signal of a polygenic species barrier. These barriers could result from co-adapted loci in *H. himera* that cannot be replaced by *H. erato* alleles without fitness consequences. The inference that this pattern results from a polygenic species barrier is further strengthened by two observations. First, *H. himera* males have been shown to mate more frequently with F1 hybrids compared to *H. erato* males [25], which should result in an opposite asymmetric admixture pattern than we found [40], with more alleles introgressing from *H. erato* into *H. himera*. Second, the smaller effective population size estimates of *H. himera* should also result in an opposite pattern of asymmetric admixture [13], with more alleles introgressing from *H. erato* into *H. himera.* This latter expectation is observed in the demographic modeling results, which show greater migration of alleles from *H. erato* into *H. himera* and corresponds to their estimated differences in population size between the species (Figure 3B), but is not seen in the more recent signals of admixture as measured by f_d.

Different histories can generate similar heterogeneous divergence patterns

Although the genomic landscape of divergence between hybridizing taxa reflects the history of selection and demographic changes they have experienced, different histories can generate strikingly similar heterogeneous patterns. This can greatly mimit our ability to make inferences about the evolutionary processes driving genomic change [2]. For example, in *Heliconius melpomene* there are strikingly similar heterogeneous landscapes of divergence to those we report here for *H. erato* and *H. himera*, despite known differences in their evolutionary histories [41,42]. Here, we show that through a comprehensive set of analytical approaches we can disentangle these evolutionary histories and reveal that different evolutionary processes have resulted in similar divergence landscapes among the *H. erato* and *H. melpomene* clades.

Heliconius melpomene and *H. erato* are co-mimics and experience similar strong selective pressures on wing color patterns throughout their distributions. In the *H. melpomene* clade*, H. melpomene* comes into contact and hybridizes with *H. cydno* in Panama and occasionally with *H. timareta* in Ecuador and northern Peru. Although at first sight the genomic landscapes appear similar, correlation analyses reveal differences in the relative importance of admixture. The *H. melpomene* comparisons showed strong correlations of recombination rate with divergence as well as admixture, which supports a longer history of divergent selection with gene flow (Martin et al. 2019). In contrast, *H. himera* and *H. erato* comparisons showed recombination rate correlated with divergence, but not admixture. We suggest the lack of correlation may be because the secondary contact is recent and the rates of gene flow are low between *H. erato* and *H. himera*. Thus, there has not been enough time for a correlation between recombination rate and admixture to arise. Instead, we argue that differences accumulated in periods of isolation have had more profound effects on the divergence landscape of *H. erato* and *H. himera* whereas gene flow has likely been more continuous throughout the divergence history of the young species pair in the *H. melpomene* clade. These differences in the history of divergence among co-mimetic species highlights the power of approaches like those applied here to resolve the roles of different evolutionary processes in generating seemingly similar heterogeneous patterns of genomic divergence.

370 *Conclusions*

371 We observed genome-wide increase of divergence as reproductive barriers increase between
372 hybridizing populations or species. Nonetheless, it remains difficult to determine if peaks of divergence
373 result from barrier loci that are resistant to ongoing gene flow (heterogeneous gene flow), from recent
374 selective sweeps in isolated populations (heterogeneous selection), or both. Fortunately, using a
375 combination of demographic modeling, simulations, and correlation analyses we can characterize the
376 evolutionary processes responsible for the heterogeneous landscapes of divergence. For the incipient
377 species *H. erato* and *H. himera*, our results suggest a disproportionately large effect of the Z
378 chromosome on the evolution of species barriers in *H. erato*. The data also favor a scenario of periods of
379 isolation accompanied by both selection and gene flow. The overall patterns of asymmetric admixture
380 suggest that during periods of isolation, selection at multiple loci may have resulted in a polygenic
381 species boundary. This finding adds to a number of studies that illustrate how fluctuating gene flow and
382 pervasive selection along the genome lead to the evolution of polygenic architectures of species
383 boundaries [4,8,21,43,44].

Methods

Sampling

We obtained whole genome resequence data for a total of 122 *Heliconius* butterflies (Tables S4). These include northern (North Ecuador, n = 10) and southern (South Ecuador *H. himera* contact zone, n = 4) *H. e. cyrbia*, western (Peru *H. himera* contact zone, n = 4) and eastern (Peru *H. e. favorinus* contact zone, n = 7) *H. e. emma* and western (Peru *H. himera* contact zone, n = 4) and eastern (Peru *H. e. emma* contact zone, n = 8) *H. e. favorinus* used to study admixture patterns with northern (Ecuador *H. e. emma/favorinus* contact zone, n = 5) and southern *H. himera* (Peru *H. e. cyrbia* contact zone, n = 4). Additionally, samples from *H. e.* petiverana (Mexico, n = 5), *H. e. demophoon* (Panama, n = 10), *H. e. hydara* (Panama, n = 6 and French Guiana, n = 5), *H. e. erato* (French Guiana, n = 6), *H. e. amalfreda* (Suriname, n = 5), *H. e. notabilis* (Ecuador *Heliconius erato lativitta* contact zone, n = 5 and Ecuador *H. e. etylus* contact zone, n = 5), *H. e. etylus* (Ecuador, n = 5), *H. e. lativitta* (Ecuador, n = 5), *H. e. phyllis* (Bolivia, n = 4), *H. e. venus* (Colombia, n = 5) and *H. e. chestertonii* (Colombia, n = 7) were used for contact zone divergence analysis and population structure visualization as well as samples of *H. hermathena* (Brazil, n = 3) as an outgroup to root the phylogenetic inference and polarize site frequency spectra. All data have been previously published [26,28], apart from the ten *H. e. cyrbia* from North Ecuador.

Sequencing and genotyping

Genotypes were obtained as in [28]. In short, whole-genome 100-bp paired-end Illumina resequencing data from *H. erato* samples were aligned to the v1 [26] reference genome, using BWA v0.7 [45]. PCR duplicated reads were removed using PICARD v1.138 (http://picard.sourceforge.net) and sorted using SAMTOOLS [46]. Genotypes were called using the genome analysis tool kit (GATK) Haplotypecaller [47]. Individual genomic VCF records (gVCF) were jointly genotyped using GATK's genotype GVCFs. Genotype calls were only considered in downstream analysis if they had a minimum depth (DP) ≥ 10, maximum depth (DP) ≤ 100 (to avoid false SNPs due to mapping in repetitive regions), and for variant calls, a minimum genotype quality (GQ) ≥ 30. Specific data filtering steps for running population structure, phylogenetic and demographic analysis are explained in the respective sections.

412 *Population structure and phylogenetic relationships*

413 We estimated levels of relative divergence (F_{ST}) [48] between populations in nonoverlapping 50
414 kb windows using python scripts and egglib [49]. For this analysis, we only considered windows for
415 which at least 10% of the positions were genotyped for at least 75% of the individuals within each
416 population. On average 96% of windows met this criterium. To discern population structure among the
417 *H. erato* and individuals, we performed principal component analysis (PCA) using EIGENSTRAT SmartPCA
418 [50]. For this analysis, we only considered autosomal biallelic sites that had coverage in all individuals
419 and excluding the Z chromosome (5,058,785 SNPs). Using the same filtering but including the outgroup
420 *H. hermathena* (4,927,152 SNPs), we used FastTree v2.1 [51] to infer an approximate maximum
421 likelihood phylogeny using the default parameters.

422

423 *Demographic modeling*

424 We performed demographic analyses on the joint site-frequency spectra (JSFS) of five
425 population comparisons using a modified version of *∂a∂i* v1.7 [11]. Genotype calls were filtered for
426 biallelic autosomal SNPs with a threshold of at least 50 % minimum genotype calls for each population
427 of interest. In order to ensure demographic analyses were performed on unlinked loci, we subsampled
428 our data so that a single SNP was selected at least every ~1000 bp (based on linkage disequilibrium maps
429 from [52] in Figure S1). Unfolded joint site-frequency spectra (JSFS) were created from the filtered calls
430 data using *H. hermathena* as an outgroup, including on average 316090.3 SNPS (Table S3). The
431 proportion of accurate SNP orientation (O) was consistent across all pairwise comparisons (~97 %),
432 which suggests that ancestral state identification was correct for the majority of SNPs using *H.*
433 *hermathena* as an outgroup.

434 Models tested include four basic scenarios in addition to 22 extensions of the basic models that
435 allowed for independent assessment of additional selective and demographic parameters [32] (Table
436 S5). Basic model scenarios included strict isolation (SI), ancient migration (AM), isolation-with-migration
437 (IM), and secondary contact (SC). The standard four models involved the divergence of an ancestral
438 population (N_{ref}) at *t* generations into two resulting daughter populations with an effective size of N_1 and
439 N_2, respectively. Migration occurred in IM, AM, and SC at rate m_{12} from population 2 into population 1
440 and m_{21} in the opposite direction. Model extensions included growth rate parameters (*g*) which account
441 for changes in the effective population size over time (bottlenecks and expansions) by taking a ratio of
442 the contemporary and ancient population sizes. Note that *∂a∂i* cannot infer both heterogeneous

443 migration and genetic drift when gene flow is not temporally localized (i.e. IM model). To capture the
444 effects of linked selection that are due to sweeps and background selection (Hill and Robertson, 1966;
445 Maynard Smith and Haigh, 1974; Charlesworth et al, 1993), further model extensions incorporated two
446 categories of loci (2*N*) that occur in proportions of *Q* and 1-*Q* in order to account for heterogeneity in N_e
447 across the genome. A Hill-Robertson scaling factor (*hrf*) was included with these models that relates the
448 effective size of loci experiencing selection to that of neutral loci. Models that infer migration (IM, AM,
449 SC) were extended to include parameters that capture heterogeneous migration rates (2*m*) across the
450 genome resulting from selection on barrier loci during adaptive divergence [16]. These extensions
451 allowed for the estimation of a proportion of loci (*P*) experiencing standard migration rates (m_{12} and m_{21})
452 and a second category of loci (1-*P*) undergoing rates me_{12} and me_{21}. Given that the JSFS was polarized
453 using *H. hermathena* as the outgroup for all comparisons, a SNP orientation parameter (*O*) was included
454 in all models to account for ancestral state misidentification. Additionally, the effective mutation rate (ϑ)
455 of N_{ref} was estimated as a free parameter in all comparisons.

456 To check for model convergence, a total of 20 independent optimizations were performed for
457 each model on each population comparison. When running these models, consistency in the likelihood
458 scores generally increased as the best optimized parameters from previous steps were incorporated into
459 subsequent steps. To score the models and account for overparameterization, the Akaike Information
460 Criterion (AIC: Burnham and Anderson 2004) was used and parameters from the top five scoring runs for
461 each model were averaged. The models with the best average AIC score for each comparison were
462 retained for each comparison. The highest and lowest optimized parameter values in the top five
463 replicate runs for each model were used to construct intervals to estimate uncertainty.

464 Model parameter estimates for effective population sizes, migration rates and divergence times
465 were transformed into absolute units using a *Heliconius* mutation rate of 2×10^{-9} per generation (i.e.
466 spontaneous *Heliconius* mutation rate corrected for selective constraint; Keightley et al. 2014; Martin,
467 Eriksson, et al. 2015) and assuming a generation time of 0.25 years. Ancestral effective population size
468 was calculated using the optimized theta value for each model comparison ($N_{ref} = \vartheta/4\mu L$), where L
469 represents the effective sequence length and μ the estimate of the mutation rate. Effective sequence
470 length for each pairwise comparison is estimated as $L = (x/y)z$, where x is the number of SNPs used in the
471 ðaði demographic analysis and y is the number of segregating sites detected in the original sample out
472 of z total sites. Migration rates are presented in units of $2N_{ref}$ to represent the number of individuals per
473 generation that migrate into each population.

Divergence and admixture simulations

To compare patterns in our data to expectations, we simulated genealogies near a selected locus. Genealogies were simulated using the coalescent simulator *msms* [56] and from these genealogies 10 kb sequences were simulated with a mutation rate of $2e^{-9}$ (i.e. spontaneous *Heliconius* mutation rate corrected for selective constraint; Keightley et al. 2014; Martin, Eriksson, et al. 2015) and a Hasegawa–Kishino–Yano (HKY) substitution model using *seq-gen* v1.3.4 [57]. With *msms*, we simulated four populations with a split history given by (((P1, P2), P3), O) where t_1, t_2 and t_3 denote the split times between P1 and P2, (P1, P2) and P3 and ((P1, P2), P3) and O, respectively (Figure 4A). Population size (N_e) was fixed to 1,000,000 individuals and t_2 and t_3 were fixed to 0.5 (2,000,000 generations) and 1 (4,000,000 generations) coalescent units ($4N_e$), respectively. Within the range of selection on color pattern in *Heliconius* [58], we simulated divergent selection at a single locus with selection coefficients (s) of 0.2 and 0.02 for homozygous genotypes and 0.1 and 0.01 for heterozygous genotypes and selection strength specified in units of $2N_es$. Selection was set to work in opposite direction for P1 and P2 versus P3 and O and was set to start at time t_s. After population splits, migration (m) was restricted between P2 and P3 only, with symmetrical migration rates and a start time equal to t_m. In relevance to our demographic modeling results, we ran simulations by varying the parameters t_s (selection start time; $2e^3$ - $2e^6$ generations), t_m (migration start time; $2.5e^{-5}$ - $5e^{-1}$ generations), m (migration rate; $1e^{-7}$ - $1e^{-5}$ generations) and ρ (population recombination rate $4N_er$; r = probability of recombination per generation per bp; $6.25e^{-4}$ - $8e^{-2}$). A maximum ρ of $8e^{-2}$ was used for computational feasibility. The genealogies were sampled at 100 kb increments from the selected locus. This was achieved by using an infinite recombination sites model and changing the position of the selected locus in increments of 10 neutral locus units (i.e. 10 x 10 kb) away from the sampled locus. Divergence (F_{ST}) was calculated as in Hudson *et al.* 1992 and admixture (f_d) was calculated as in Martin *et al.* 2016 using python scripts and *egglib* v3 (De Mita & Siol, 2012). We investigated the correlation between F_{ST}, f_d and recombination rate at a distance of 500 kb from a selected locus but similar expectations are obtained from wide range of distances to the selected locus. Simulations were run with 100 replicates for each parameter combination. Pseudocode to run the *msms* command lines are provided in Tables S6.

Recombination rate estimates

We estimated fine-scale variation in population recombination rate ($\rho = 4N_er$; r = probability of recombination per generation per bp) along the *H. erato* chromosomes from linkage-disequilibrium in

population genetic data using *LDhelmet* v1.7 [59]. We phased quality filtered genotypes from thirteen *H. erato* populations (i.e. *H. e. cyrbia*, *H. e. venus*, *H. e. demophoon*, *H. e. hydara* (Panama), *H. e. emma*, *H. e. etylus*, *H. e. lativitta*, *H. e. notabilis*, *H. e. favorinus*, *H. e. phyllis*, *H. e. erato*, *H. e. hydara* (French Guiana) and *H. e. amalfreda*) using Beagle v4.1 [60] with default parameters. From the phased genotypes, fasta sequences were generated for 50 kb windows. These 50 kb windows were transformed to haplotype configuration files with the recommended window size of 50 SNPs used by *LDhelmet* to estimate composite likelihoods of the recombination rate. From the haplotype configuration files, lookup tables for two-locus pairwise recombination likelihoods and Padé coefficients were generated within the recommended value range. Transition matrices were calculated for each chromosome separately by comparing genotypes obtained from *H. erato demophoon* to the outgroup species *H. hermathena*. The likelihood lookup tables, Padé coefficients and transition matrices were used in the rjMCMC procedure of *LDhelmet* to estimate the recombination map. In this latter step, 1000,000 Markov chain iterations were run with a burn-in of 100,000 iterations, a window size of 50 SNPs and block penalty of 50. To reduce the potential effect of locus-specific changes in effective population size (N_e) on population recombination rate (ρ) estimates (e.g. due to population specific selective sweeps or background selection), we estimated ρ for each *H. erato* population separately and obtained averages for each 50 kb interval.

Admixture statistics

We estimated admixture proportions for 50 kb non-overlapping windows using the f_d statistic [38]. This statistic is based on the ABBA-BABA test or Patterson's *D* statistic which measures an excess of derived allele sharing between sympatric non-sister taxa [61]. This excess is tested by comparing the relative abundance of SNP patterns termed ABBAs and BABAs in a tree of three populations and an outgroup with the relationship (((P1, P2), P3), O). ABBAs are sites where a derived allele is shared between P2 and P3, whereas BABAs are sites where a derived allele is shared between P1 and P3. Under a neutral coalescent model, such sites are only expected to be found due to incomplete lineage sorting or recurrent mutation and a *D* statistic of 0 is expected. In the presence of admixture, however, an access of either ABBAs or BABAs can be observed and a D statistic that significantly deviates from 0 may be obtained. The f_d statistic is derived from the *D* statistic by calculating the difference between ABBA and BABA sites and normalizing this difference by a scenario of complete admixture. The estimator is dynamic in that for the complete admixture scenario used to normalize, a donor population for the

admixture is chosen with the highest frequency of the derived site. The resulting normalized measure is approximately proportional to the effective migration rate and has been evaluated not to be confounded by locus-specific changes in effective population size due to background selection or reductions in diversity due to selective sweeps [4,38]. The populations included as P1, P2, P3 and O are indicated in the figures. *Heliconius hermathena* samples were consistently used as the outgroup taxa.

Admixture directionality

By expanding the four-taxon *D* statistic to a five-taxon scenario, it is possible to obtain information on the directionality of admixture (i.e. donor versus recipient population). A set of statistical measures that use a five-taxon symmetric phylogeny to infer both the taxa involved in and the direction of admixture are called the D_{FOIL} statistics [39]. The D_{FOIL} statistics identify taxa involved in admixture by performing four possible *D* tests with different combinations of three ingroup taxa within a five-taxon phylogeny defined as (((P1, P2), (P3, P4)), O). These four *D* tests considered collectively can provide information on the directionality of admixture. This is because admixture does not only change the position of the donor sample in the topology but will also change the relationship of the donor's sister taxon to the other taxa in the phylogeny. For instance, if admixture occurs from P2 into P3, the sampled topology becomes (((P2, P3), P1), P4), O) and P1 will group more closely to (P2, P3) because of more recent sharing of variation with P2, whereas if admixture occurs from P3 into P2, the topology becomes (((P2, P3), P4), P1), O). For the latter instance, this will be reflected by a similar sign of the *D* test statistics that include (((P1, P2), P3), O) or (((P1, P2), P4), O) but a different sign for the D test that includes (((P3, P4), P2), O) and no significant D test for (((P3, P4), P1), O). Hence, by comparing the combinations of different signs (+, -, or 0) of the four *D* tests within the five-taxon topology, the directionality of admixture can be assessed [39].

We assessed directionality of admixture in 50 kb non-overlapping windows in the three *H. himera* contact zones with *H. erato* populations using the D_{FOIL} tests explained above using the available *dfoil* software (www.github.com/jbpease/dfoil). Samples from the *H. himera* North and South populations were specified as the P1 and P2 group, whereas samples from the considered *H. erato* populations were specified as P3 and P4. *Heliconius hermathena* was used as the outgroup taxon (ID hermathena_13 in Table S4). The D_{FOIL} statistics were calculated between each possible combination of available ingroup taxa (i.e. one sample for each taxon group); 800 combinations for *H. himera – H. e. cyrbia*, 500 for *H. himera – H. e. emma* and 480 for *H. himera – H. e. favorinus*. Among these sample

567 combinations, significant D_{FOIL} signatures (χ^2 goodness-of-fit test) were counted and used to obtain
568 heterogeneous patterns of admixture directionality along the genome.

569

570 **Data accessibility**

571 For GenBank accession numbers of whole genome resequence data see Table S4.

572

573 **Author contributions**

574 The study was conceived and designed by SVB and BAC in collaboration with RP and WOM.
575 Genomic analyses were performed by SVB and JC. Demographic modeling with *∂a∂I* was conducted by
576 JC. Samples of *H. e. cyrbia* from northern Ecuador were contributed by GMK, and CB assisted with
577 permits. RP, GMK and WOM provided input on results and manuscript preparation. The manuscript was
578 written and figures were made by SVB, JC, and BAC.

579

580 **Acknowledgments**

581 This work was funded by NSF grant (DEB 1257689) to BAC and RP and NSF EPSCoR RII Track-2
582 FEC (OIA 1736026) to RP and BAC. For sequencing and computational resources, we thank the University
583 of Puerto Rico, the Puerto Rico INBRE Grant P20 GM103475 from the National Institute for General
584 Medical Sciences (NIGMS), a component of the National Institutes of Health (NIH); and awards 1010094
585 and 1002410 from the EPSCoR program of the NSF. We thank the Ecuadorian Ministerio del Ambiente
586 (No. 005-13 ICFAU- DNB/MA), Peruvian Ministerio de Agricultura and Instituto Nacional de Recursos
587 Naturales (201-2013-MINAGRI-DGFFS/DGEFFSS) and Autoridad Nacional De Licencias Ambientales-ANLA
588 in Colombia (Permiso Marco 0530) for permission to collect butterflies. We thank Simon Martin for
589 sharing his code and useful discussions that shaped this paper. We thank Nicola Nadeau for access to
590 sequence data for northern *H. erato cyrbia* samples, which was funded by her UK Natural Environment
591 Research Council (NERC) fellowship (NE/K008498/1).

592

593 **References**

594 1. Wolf JBW, Ellegren H. Making sense of genomic islands of differentiation in light of speciation. Nat Rev Genet. Nature
595 Publishing Group; 2017;18: 87–100.
596 2. Ravinet M, Faria R, Butlin RK, Galindo J, Bierne N, Rafajlovic M, et al. Interpreting the genomic landscape of speciation:
597 a road map for finding barriers to gene flow. J Evol Biol. 2017;30: 1450–1477.
598 3. Wu C. The genic view of the process of speciation. J Evol Biol. 2001;14: 851–865.
599 4. Martin SH, Davey JW, Salazar C, Jiggins CD. Recombination rate variation shapes barriers to introgression across
600 butterfly genomes. PLOS Biol. 2019;17: e2006288.
601 5. Edelman NB, Frandsen PB, Miyagi M, Clavijo B, Davey J, Dikow R, et al. Genomic architecture and introgression shape a
602 butterfly radiation. Science. 2019;366: 24174–24183.
603 6. Tavares H, Whibley A, Field DL, Bradley D, Couchman M, Copsey L, et al. Selection and gene flow shape genomic islands

that control floral guides. Proc Natl Acad Sci U S A. 2018;115: 11006–11011.
7. Charlesworth B. Effective population size and patterns of molecular evolution and variation. Nat Rev Genet. 2009;10: 195–205.
8. Aeschbacher S, Selby JP, Willis JH, Coop G. Population-genomic inference of the strength and timing of selection against gene flow. Proc Natl Acad Sci. 2017;114: 7061–7066.
9. Stankowski S, Chase MA, Fuiten AM, Rodrigues MF, Ralph PL, Streisfeld MA. Widespread selection and gene flow shape the genomic landscape during a radiation of monkeyflowers. PLOS Biol. 2019;17: e3000391.
10. Burri R. Interpreting differentiation landscapes in the light of long-term linked selection. Evol Lett. 2017;1: 118–131.
11. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 2009;5: e1000695.
12. Lohse K, Chmelik M, Martin SH, Barton NH. Efficient strategies for calculating blockwise likelihoods under the coalescent. Genetics. 2016;202: 775–786.
13. Frantz LAF, Madsen O, Megens HJ, Groenen MAM, Lohse K. Testing models of speciation from genome sequences: Divergence and asymmetric admixture in Island South-East Asian Sus species during the Plio-Pleistocene climatic fluctuations. Mol Ecol. 2014;23: 5566–5574.
14. Roux C, Tsagkogeorga G, Bierne N, Galtier N. Crossing the species barrier: Genomic hotspots of introgression between two highly divergent ciona intestinalis species. Mol Biol Evol. 2013;30: 1574–1587.
15. Sousa VC, Carneiro M, Ferrand N, Hey J. Identifying loci under selection against gene flow in isolation-with-migration models. Genetics. 2013;194: 211–233.
16. Tine M, Kuhl H, Gagnaire PA, Louro B, Desmarais E, Martins RST, et al. European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. Nat Commun. 2014;5: 5770.
17. Kousathanas A, Leuenberger C, Helfer J, Quinodoz M, Foll M. Likelihood-free inference in high-dimensional models. Genetics. 2016;203: 893–904.
18. Rougeux C, Gagnaire PA, Bernatchez L. Model-based demographic inference of introgression history in European whitefish species pairs. J Evol Biol. 2019;32: 806=817.
19. Kaplan NL, Hudson RR, Langley CH. The “hitchhiking effect” revisited. Genetics. 1989;123: 887–899.
20. Keinan A, Reich D. Human population differentiation is strongly correlated with local recombination rate. PLoS Genet. 2010;6: e1000886.
21. Brandvain Y, Kenney AM, Flagel L, Coop G, Sweigart AL. Speciation and introgression between Mimulus nasutus and Mimulus guttatus. PLoS Genet. 2014;10: e1004410.
22. Barton N, Bengtsson BO. The barrier to genetic exchange between hybridising populations. Heredity. 1986;57: 357–376.
23. Nachman MW, Payseur BA. Recombination rate variation and speciation: Theoretical predictions and empirical results from rabbits and mice. Philos Trans R Soc B Biol Sci. 2012;367: 409–421.
24. Jiggins CD, Mcmillan O, Neukirchen W, Mallet J, Nw L. What can hybrid zones tell us about speciation? The case of *Heliconius erato* and *H. himera* (Lepidoptera: Nymphalidae). Biol J Linn Soc. 1996;59: 221–242.
25. McMillan WO, Jiggins CD, Mallet J. What initiates speciation in passion-vine butterflies? Proc Natl Acad Sci U S A. 1997;94: 8628–8633.
26. Van Belleghem SM, Rastas P, Papanicolaou A, Martin SH, Arias CF, Supple MA, et al. Complex modular architecture around a simple toolkit of wing pattern genes. Nat Ecol Evol. 2017;1: 52.
27. Muñoz AG, Salazar C, Castaño J, Jiggins CD, Linares M. Multiple sources of reproductive isolation in a bimodal butterfly hybrid zone. J Evol Biol. 2010;23: 1312–1320.
28. Van Belleghem SM, Salazar C, Jiggins CD, Baquero M, Papa R, Counterman BA, et al. Patterns of Z chromosome divergence among *Heliconius* species highlight the importance of historical demography. Mol Ecol. 2018;27: 3852–3872.
29. Kronforst MR, Hansen MEB, Crawford NG, Gallant JR, Zhang W, Kulathinal RJ, et al. Hybridization reveals the evolving genomic architecture of speciation. Cell Rep. The Authors; 2013;5: 666–677.
30. Roux C, Fraisse C, Romiguier J, Anciaux Y, Galtier N, Bierne N. Shedding light on the grey zone of speciation along a continuum of genomic divergence. PLoS Biol. 2016;14: e2000234.
31. Supple M, Papa R, Hines HM, McMillan WO, Counterman BA. Divergence with gene flow across a speciation continuum of *Heliconius* butterflies. BMC Evol Biol. 2015;15: 204.
32. Rougeux C, Bernatchez L, Gagnaire PA. Modeling the multiple facets of speciation-with-gene-flow toward inferring the divergence history of lake whitefish species pairs (*Coregonus clupeaformis*). Genome Biol Evol. 2017;9: 2057–2074.
33. Crawford JE, Riehle MM, Guelbeogo WM, Gneme A, Sagnon N, Vernick KD, et al. Reticulate speciation and barriers to introgression in the Anopheles gambiae species complex. Genome Biol Evol. 2015;7: 3116–3131.
34. Charlesworth B. Measures of divergence between populations and the effect of forces that reduce variability. Mol Biol Evol. 1998;15: 538–543.
35. Cruickshank TE, Hahn MW. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not

reduced gene flow. Mol Ecol. 2014;23: 3133–3157.
36. Counterman BA, Araujo-Perez F, Hines HM, Baxter SW, Morrison CM, Lindstrom DP, et al. Genomic hotspots for adaptation: the population genetics of Müllerian mimicry in *Heliconius erato*. PLoS Genet. 2010;6: e1000796.
37. Reed RD, Papa R, Martin A, Hinas HM, Counterman BA, Pard-Diaz C, et al. *optix* drives the repeated convergent evolution of butterfly wing pattern mimicry. Science. 2011;333: 1137–1141.
38. Martin SH, Davey JW, Jiggins CD. Evaluating the use of ABBA-BABA statistics to locate introgressed loci. Mol Biol Evol. 2015;32: 244–257.
39. Pease JB, Hahn MW. Detection and polarization of introgression in a five-taxon phylogeny. Syst Biol. 2015;64: 651–662.
40. Latour Y, Perriat-Sanguinet M, Caminade P, Boursot P, Smadja CM, Ganem G. Sexual selection against natural hybrids may contribute to reinforcement in a house mouse hybrid zone. Proc R Soc B Biol Sci. 2013;281: 20132733.
41. Martin SH, Dasmahapatra KK, Nadeau NJ, Salazar C, Walters JR, Simpson F, et al. Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. Genome Res. 2013;23: 1817–1828.
42. Quek SP, Counterman BA, De Moura PA, Cardoso MZ, Marshall CR, McMillan WO, et al. Dissecting comimetic radiations in *Heliconius* reveals divergent histories of convergent butterflies. Proc Natl Acad Sci U S A. 2010;107: 7365–7370.
43. Berner D, Roesti M. Genomics of adaptive divergence with chromosome-scale heterogeneity in crossover rate. Mol Ecol. 2017;26: 6351–6369.
44. Janousek V, Munclinger P, Wang L, Teeter KC, Tucker PK. Functional Organization of the Genome May Shape the Species Boundary in the House Mouse. Mol Biol Evol. 2015;
45. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. 2013; 1303.3997v1.
46. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078–2079.
47. Van der Auwera G a., Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. Curr Protoc Bioinforma. 2013;UNIT 11.10: 1–33.
48. Hudson RR, Slatkin M, Maddison WP. Estimation of levels of gene flow from DNA sequence data. Genetics. 1992;132: 583–589.
49. De Mita S, Siol M. EggLib: processing, analysis and simulation tools for population genetics and genomics. BMC Genet. BioMed Central Ltd; 2012;13: 27.
50. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick N a, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38: 904–909.
51. Price MN, Dehal PS, Arkin AP. FastTree 2 – Approximately maximum-likelihood trees for large alignments. PLoS One. 2010;5: e9490.
52. Martin SH, Möst M, Palmer WJ, Salazar C, McMillan WO, Jiggins FM, et al. Natural selection and genetic diversity in the butterfly Heliconius melpomene. Genetics. 2016;203: 525–541.
53. Burnham KP, Anderson DR. Multimodel inference: Understanding AIC and BIC in model selection. Sociol Methods Res. 2004;33: 261–304.
54. Keightley PD, Pinharanda A, Ness RW, Simpson F, Dasmahapatra KK, Mallet J, et al. Estimation of the spontaneous mutation rate in *Heliconius melpomene*. Mol Biol Evol. 2014;32: 239–243.
55. Martin SH, Eriksson A, Kozak KM, Manica A, Jiggins CD. Speciation in Heliconius Butterflies : Minimal Contact Followed by Millions of Generations of Hybridisation. BioRxiv. 2015; 1–24.
56. Ewing G, Hermisson J. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics. 2010;26: 2064–2065.
57. Rambaut A, Grass NC. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Bioinformatics. 1997;13: 235–238.
58. Moest M, Van Belleghem SM, James J, Salazar C, Martin S, Barker S, et al. Selective sweeps on novel and introgressed variation shape mimicry loci in a butterfly adaptive radiation. PLoS Biol. 2020;18: e3000597.
59. Chan AH, Jenkins PA, Song YS. Genome-wide fine-scale recombination rate variation in *Drosophila melanogaster*. PLoS Genet. 2012;8.
60. Browning BL, Browning SR. Genotype imputation with millions of reference samples. Am J Hum Genet. The American Society of Human Genetics; 2016;98: 116–126.
61. Durand EY, Patterson N, Reich D, Slatkin M. Testing for ancient admixture between closely related populations. Mol Biol Evol. 2011;28: 2239–2252.