Article

Transitions from Single- to Multi-locus Processes during Speciation

Martin P. Schilling [1,†,*], Sean P. Mullen [2], Marcus Kronforst [3], Rebecca J. Safran [1], Patrik Nosil [4], Jeffrey L. Feder [5], Zachariah Gompert [4], and Samuel M. Flaxman [1,†]

[1] Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, CO
[2] Department of Biology, Boston University, Boston, MA 02215
[3] Department of Ecology & Evolution, University of Chicago, Chicago, IL
[4] Department of Biology & Ecology Center, Utah State University, Logan, UT 84322
[5] Department of Biological Sciences, University of Notre Dame, South Bend, IN 46556
* Correspondence: schimar@gmail.com
† Current address: Department of Ecology and Evolutionary Biology, N211 Ramaley Hall, Campus Box 334, University of Colorado, Boulder, CO 80309

Abstract: During speciation-with-gene-flow, a transition from single-locus to multi-locus processes can occur, as strong coupling of multiple loci creates a barrier to gene flow. Testing predictions about such transitions with empirical data requires building upon past theoretical work and the continued development of quantitative approaches. We simulated genomes under different evolutionary scenarios of gene flow and divergent selection, extending previous work with the additions of neutral sites and coupling statistics, allowing us to investigate if and how selected and neutral sites differ in the conditions they require for transitions during speciation. As the per-locus strength of selection grew and/or migration decreased, it became easier for selected sites to show divergence – and thus to rise in linkage disequilibrium (LD) with each other as a statistical consequence – farther in advance of the conditions under which neutral sites could diverge. Indeed, even very low rates of gene flow were sufficient to prevent differentiation at neutral sites. However, once strong enough, coupling among selected sites eventually reduced gene flow at neutral sites as well. To explore whether similar transitions might be detectable in empirical data, we used published genome resequencing data from three taxa of *Heliconius* butterflies. We found that allele-frequency outliers and F_{ST} outliers exhibited stronger patterns of LD than the genomic background, as expected. The statistical characteristics of LD – likely indicative of the strength of coupling of barrier loci – varied between chromosomes and taxonomic comparisons. Broad qualitative agreement between the patterns we observed in the empirical data and our simulations suggests that selection drives rapid genome-wide transitions to multi-locus coupling, illustrating how divergence and gene flow interact along the speciation continuum.

Keywords: gene flow; sympatry; parapatry; simulation model; population genomics; *Heliconius*; coupling; nonlinear transitions

1. Introduction

Understanding the genetic basis of speciation – long a central goal of evolutionary biology – has been greatly advanced by high-throughput sequencing (HTS) methods. High quality data sets from a wealth of empirical studies of speciation-in-action now abound, and we can now point to many excellent examples demonstrating that combinations of factors – selection, drift, ecology, geography, hybridization, recombination, and more – shape divergence and reproductive isolation [e.g. 1,2]. But the current abundance of data also underscores the vast gulf that still often exists between "having the data" and "having the answers." Signs of varied evolutionary processes can sometimes be unambiguously detected, but predicting and testing for the patterns of aggregate, genome-wide processes along the speciation continuum (i.e, at varying points of divergence and differentiation) remains challenging. Therefore, additional work is needed to expand the scale of prediction and analysis from sets of "speciation genes" to genome-wide statistical patterns.

As Butlin and Smadja [3] recently highlighted, theory based upon coupling is a promising foundation for this work. In this context, "coupling" refers to any process that combines barrier effects from multiple loci, leading to strengthening of the overall barrier to gene flow. Coupling involves alleles at different loci becoming statistically non-independent with respect to their evolutionary dynamics. For example, two loci that are subject to divergent selection may together, when coupled, reach allele frequency differences between demes that are larger than either allele would reach at the corresponding uncoupled, single-locus migration-selection balance. In general, the potential for loci to be coupled is strongly dependent upon the relative strengths of selection and recombination [4, more on this below]. It is also important to note that the notion of coupling as defined here can include but is not limited to allele frequency clines becoming coincident in space.

Multi-locus processes involved in speciation have been studied by theoreticians for decades [1,5], providing an excellent foundation to understand the buildup and maintenance of differentiation under selection, gene flow, and genetic drift [3–10]. A number of theoretical studies have also cast light upon the roles of linkage and genomic architecture in speciation [11–17]. Much of this work has emphasized the parameter space – combinations of selection strength, migration rates, recombination rates, numbers of loci, etc. – in which loci would become coupled with one another [e.g. 4,6,8,9,18].

Recent work has additionally emphasized the temporal properties of the emergence of coupling, and indicates that the buildup of divergence can be strongly nonlinear in time [12,17,19–22]. Specifically, these studies suggest that many alleles with individually small effects may rapidly (in evolutionary terms) transition from an uncoupled to a coupled state, a process in which highly divergently adapted multi-locus genotypes "congeal" out of what was previously a well-mixed gene pool [17]. However, the latter work did not incorporate neutral sites and thus was silent about their dynamics. Southcott and Kronforst [23] used forward-time simulations which included neutral sites in a large (100 kb) region of the genome containing a single site under selection. Their work suggested that the genomic patterns produced by neutral and non-neutral processes may not be easily distinguishable. This raises important questions: Are there ways to reliably distinguish barrier loci under selection from neutral genomic background? Can aggregate, genome-wide statistical patterns offer insights about neutral and non-neutral processes? What, if any, patterns produced by neutral versus non-neutral processes are robust and detectable?

Toward this end, we investigated the temporal dynamics of coupling in two-deme simulations under divergent selection and gene flow. We built on a previous model ["*bu2s*": 17] that considered the de novo buildup of large numbers (100s) of sites under selection and extended that model to incorporate neutral sites. Following the results about the difficulty of using traditional summary statistics to distinguish non-neutral and neutral evolutionary processes [23], we used results from the model to compare the potential power of multiple population genetic statistics to differentiate between neutral and selected sites. We computed coupling statistics derived from multi-locus cline theory [4,8] to quantitatively describe transitions from single- to multi-locus processes during speciation with gene flow. Consistent with previous work, selected sites reached high coupling between loci, and we now provide quantitative predictions about the time and parameter lag between the transition seen for selected sites and the analogous transition for neutral sites.

Finally, to connect empirical data and theory, we compared the general (qualitative) patterns from our simulation results to previously published population genomic data sampled from *Heliconius* species spanning various stages of evolutionary divergence along the speciation continuum [24]. In this empirical system, there is strong evidence of divergence with gene flow [24–29], despite multiple forms of reproductive isolation arising as both direct [30–33], and indirect [34–38] consequences of selection on mimetic Müllerian wing color patterns. Therefore, we predicted that evidence for coupling would be most pronounced for putatively selected loci, which we identified as allele frequency outliers in the empirical data. In addition, we predicted that evidence of coupling between recently derived *Heliconius* species, that differ only in aspects of color pattern (due to selection on only a single or small number of loci; [39]), would be restricted to chromosomes housing color patterning genes, and then a rapid transition to genome-wide coupling as phylogenetic distance increased (reflecting reduced effective migration resulting from ongoing selection on color pattern and other forms of ecological divergence). In a purely allopatric scenario, neutral sites would readily diverge (between demes) along with the selected sites, due to the effects of hitchhiking and drift. However, with even small amounts of gene flow and recombination, neutral divergence would be delayed until after barrier loci had already become strongly coupled and diverged. Hence, we predicted that the genomic background would show little divergence in most of our comparisons of taxa and regions of the genome. However, for taxa showing the greatest divergence (i.e., those farthest along the speciation continuum), effects of coupling should be detectable even in the genomic background (i.e., the non-outlier sites).

2. Materials and Methods

2.1. Simulations

We performed forward-time, individual-based simulations with *bu2s* version 3.6.1. The details of this software have been previously described [14,17]. Source code is freely available at *GitHub* [40]. Data sets used in this publication are archived at [insert link upon publication]. The main extension of the model not present in previous versions [17] is the addition of neutral sites. Since the rest of the workings of the *bu2s* model have been described in multiple previous publications [14,17,22], we give only a brief overview of *bu2s* here, focused on key elements and new features, and we refer the interested reader to those previous works for more details.

Table 1 gives a summary of key parameters and the value(s) of each for the results shown here. These were chosen to includes cases with $s > m$ (per locus), in which divergence of individual loci is expected to be relatively smooth and linear in time, as well as cases with $s < m$, in which abrupt, nonlinear transitions in divergence and coupling are expected and individual mutations are often swamped by migration [17,22,41]. Altogether, we show results from a total of 300 simulation runs, with 50 runs each of 6 different parameter combinations (Table 1).

Table 1. Parameter values used in simulations.

parameter	Notation (if applicable)	Value(s) used (and units if applicable)
Mean selection coefficient for divergently selected mutations (mean of exponential distribution from which new mutations' coefficients were drawn)	s	0.005, 0.01, 0.02
Migration rate	m	0.01, 0.1 (probability per individual per generation)
Total population size	N	5000 individuals
Mutations per generation (population mutation rate)		10 per generation
Number of chromosomes in a genome (haploid number)	c	4
Recombination length of each individual chromosome	l	50 cM
Ratio of neutral:selected mutations		10:1

2.1.1. Model overview and life cycle

bu2s is a forward-time, mutation-based, stochastic model of divergence with gene flow, starting from a point with zero differentiation and zero segregating variation. Space is discrete, there are two demes, and individuals are diploid, hermaphroditic, and obligate outcrossers. Evolution results from the combination of mutation, selection, migration, recombination, and drift. In this setup, new mutations can either be neutral or subject to divergent selection. A life cycle diagram of the *bu2s* model is shown in Fig. 1 A.

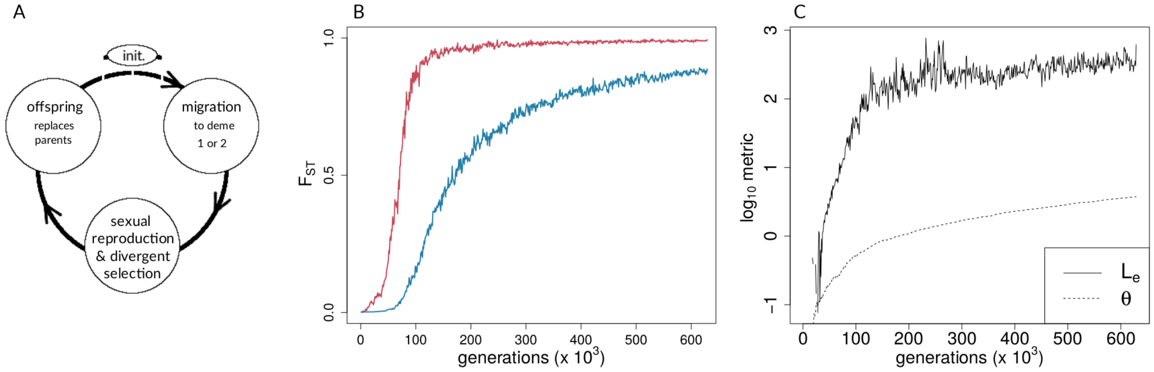

Figure 1. Forward-time simulations with (A) life-cycle diagram for the simulation model used, (B) population differentiation (F_{ST}) over time for neutral (blue) and selected (red) sites, and (C) coupling measures for single simulation run. In (B) and (C), $s = 0.005$ and $m = 0.01$. See Methods for definitions of the coupling coefficient, θ, and the effective number of loci, L_e.

2.1.2. Migration

In each generation, each individual can migrate to the other deme with probability m.

2.1.3. Selection, fitness, and reproduction

Following migration, reproduction occurs with the N_k parents in deme k (k = 1 or 2) giving rise to N_k offspring in that deme (i.e., soft selection). Selection occurs during reproduction: an individual's relative fitness is linearly proportional to the probability that it contributes a gamete to the formation of each offspring. Fitness is calculated multiplicatively across selected loci, with each locus' contribution to fitness given by Table 2. Though fitness is multiplicative, there is no epistasis in the sense of incompatibilities. As the fitness scheme in Table 2 implies, this is a "two-allele" model of barrier loci [sensu 42].

Table 2. Fitness scheme for loci under selection.

Genotype at locus i	Fitness contribution of locus in Deme 1	Fitness contribution of locus in Deme 2
A_iA_i	$1 + s_i$	1
A_iB_i	$1 + 0.5s_i$	$1 + 0.5s_i$
B_iB_i	1	$1 + s_i$

In Table 2, A_i is the ancestral allele at locus i, B_i is the derived allele at locus i (originated by mutation during the simulation), and s_i is the selection coefficient associated with locus i. Note that a "locus" is synonymous with a "site" in this infinite sites model; any "locus" is a spot in the genome analogous to a single nucleotide polymorphism (SNP) in modern genomic sequencing data. The fitness of the jth individual, W_j, is calculated as:

$$W_j = \prod_{i=1}^{L} w_{ij}(g_{ij}) \,, \tag{1}$$

where $w_{ij}(g_{ij})$ is the contribution of locus i to individual j's fitness, as a function of its genotype at that locus, g_{ij} (Table 2), L is the total number of selected loci with segregating variants, and the product is over all such loci (i.e., ignoring neutral sites).

2.1.4. Recombination

Recombination occurs during gamete formation (meiosis). Recombination locations are individually identically distributed along the length of the genome. In results shown here, there were four chromosomes of equal length, with each chromosome having an expected number of 0.5 recombination events per meiosis (i.e., each was 50 cM long). The total number of recombination events thus follows a Poisson distribution with mean equal to the length of the entire genome expressed in Morgans (= 2 for results shown here).

2.1.5. Mutation

A fixed number of mutations are introduced to the population in each generation (10/generation in results shown here). Each mutation is introduced in a randomly chosen offspring at a uniformly randomly chosen location in the genome. In results shown here, neutral and divergently selected mutations were introduced in a 10:1 ratio (i.e., a new mutation could be neutral with probability $\sim$0.909 or subject to divergent selection with probability $\sim$0.0909). Given that neutral mutations are more likely to be lost by drift, this ratio was chosen to allow a large number of both types of variants to build up standing variation. Additionally, it is realistic to expect more mutations to be neutral than beneficial. Global positive selection and epistatic incompatibilities were not considered, in order to focus on the effects of genome-wide, divergent adaptation. Thus, this is a model in which reproductive isolation evolves by divergent adaptation only. Selection coefficients, s_i, are drawn from an exponential distribution with mean s [see 14,17, for discussion]. All mutations arise de novo; there is no standing variation at the beginning of a simulation. In the course of a full-length simulation run with parameters used here (Table 1), $\sim 10^7$ neutral mutations and $\sim 10^6$ selected mutations would be introduced, but, because of drift and migration, only a tiny fraction of these would establish.

2.1.6. Data derived from simulations and metrics computed from simulation data

Each simulation run outputs standard population genetic metrics as time series, including global and deme-specific allele frequencies, F_{ST}, allele frequency spectra, samples of individual fitness values, and more (see [17] for additional metrics). F_{ST} is calculated as $F_{ST} = (H_T - H_S)/H_T$, where H_T is the total observed heterozygosity at a given locus at a given time step and H_S is the expected heterozygosity based upon each deme's observed heterozygosity [43]. Reproductive isolation in this model is quantified by the expected effective backward migration rate [44], m_e, defined as the expected proportion of reproduction in a deme attributable to immigrants. This quantity shrinks as populations become increasingly isolated due to differential adaptation (i.e., as immigrants have lower and lower fitness relative to residents). For results shown here, simulations were run until 15,000,000 mutations had been introduced (a maximum feasible run time of about 168 hours of CPU time in some parameter combinations) or until effective migration dropped below an a priori threshold (whichever came first). We arbitrarily defined the latter threshold as $Nm_e < 0.0001$, i.e., a less than 1 in 10,000 chance of having a single immigrant successfully reproduce. We ran 300 simulations under different combinations of s and m, (Table 1), yielding 50 independent replicates of each specific parameter combination.

To connect our simulations to theoretical expectations about the fate of selected and neutral alleles derived from seminal work on hybrid zones, we calculated Barton's coupling coefficient [4], defined as $\theta = s/r$. We note here that, unlike previous analytic work, in the *bu2s* model r is not a fixed parameter but rather a dynamic variable that is a function of time, as is typical of forward-time simulations. For this reason, θ changes over time in our simulations. Additionally, while s is a fixed parameter in our simulations (Table 1), we note that it is the mean of a distribution from which mutation effect sizes, s_i, are drawn. The mutations that actually establish, especially early in a simulation, will be a non-random set of these because those with large values of s_i will have a greater probability of establishment. Thus, the value of "s" used in our calculations of θ is an average from the segregating, selected alleles present at a given time rather than the fixed value of the parameter s per se. Henceforth, we denote the parameter (fixed value) as s, and we denote the mean of the s_i that are actually present at a given time as $\bar{s}$; Additional explanations of these metrics and their calculation for the *bu2s* model can be found in the online supplementary material associated with Nosil et al. 2017 [45].

The coefficient θ is a quantitative description of the *potential* for coupling in the system. The actual amount of coupling at a given time can be measured by the effective number of loci, L_e [4], which, in our discrete-space model, is the number of barrier loci that would have to be maximally coupled (i.e., in maximum linkage disequilibrium) to produce the observed between-deme difference in allele frequencies of barrier loci. If all barrier loci are evolving independently (i.e., completely uncoupled; zero linkage disequilibrium), we expect alleles at each locus to exhibit single-locus migration-selection balance, which would result in $L_e = 1$. As loci become coupled, their alleles aid each other in reaching higher frequencies in their favored deme, causing departures from single-locus migration-selection balance and increases in L_e. Following [4], we compute the effective number of loci as

$$L_e = \frac{s^*}{\bar{s}}\,, \tag{2}$$

where s^* is the value of the selection coefficient for a single locus that would cause that locus to have the observed frequency at migration-selection balance. Note that when loci are not coupled at all, we should find $s^* = \bar{s}$ and $L_e = 1$. However, as loci become coupled, we expect $s^* > \bar{s}$ and thus $L_e > 1$. For the fitness scheme used in the *bu2s* model, and considering a single locus independently of all other loci, the equilibrium frequency of an allele under migration-selection balance in the favored deme is:

$$p = \frac{m(0.5 + 0.75s_i) - 0.125s_i - 0.125\sqrt{s_i^2 - 4ms_i^2 + 4m^2(2 + s_i)^2}}{(-0.25 + m)s_i}\,, \tag{3}$$

[22, a *Mathematica* notebook with this solution and its derviation can be downloaded from [46]]. Let $\bar{p}$ be the mean observed frequency of barrier alleles in their favored deme. Substituting $\bar{p}$ for p, and s^* for s_i in equation 4, we can then rearrange that equation to solve for s^*, yielding:

$$s^* = \frac{m(\bar{p} - 0.5)}{(0.25 - 0.25 * \bar{p})\bar{p} + m(0.5 + \bar{p}(\bar{p} - 1.5))}\,, \tag{4}$$

which provides the numerator in equation 2 to calculate L_e.

For comparisons between simulation results and empirical data, we focused on statistics that are indicative of coupling and its effects, namely, linkage disequilibrium (LD). For the simulation results, LD was calculated as the correlation of allelic states (a value between 0 and 1; empirical LD methods described below). This was done for two different sets of loci: (i) the average pairwise value of LD between selected sites across the genome, and (ii) the average pairwise value of LD between neutral sites across the genome. In both cases, LD is calculated over the whole population, i.e, across demes. We used *R* [47] for the calculations of θ with the additional package *rhdf5* [48] for efficient parsing of simulation runs [insert github repository before publication].

To quantify times and conditions at which transitions occurred in simulations, we fitted generalized logistic models to results on allele frequency differences (AFD) and LD. This choice of functional form (a logistic form) works well because AFD and LD are each bounded on the interval $[0, 1]$. Below we plot allele-frequency differences as a function of θ, and the median of LD across runs as a function of time. For the former, we fit the data using

$$y(x) = \frac{1}{(1 + e^{-a(x-b)})}\,, \tag{5}$$

where y is AFD, and x is $\log_{10}(\theta)$. For LD as a function of time, we added one parameter to the logistic function, namely the asymptote (z), yielding

$$y(x) = \frac{z}{(1 + e^{-a(x-b)})}\,, \tag{6}$$

where y is LD and x is time. For both θ versus AFD and LD versus time, the coefficient b provides an estimate of the value of θ or time at which the change in AFD or LD is most rapid (the inflection point of the logistic curve). The coefficient a is a shape parameter, with larger magnitudes of a indicating a steeper slope at the inflection point. Comparing values of a between selected and neutral sites gives a measure of how quickly each type of site undergoes a transition (once it has already begun). The value of b gives a measure of how long it takes transitions to begin, and comparing values of b gives an estimate of the difference or lag between transition points for the types of sites. Model fitting was performed with the *nls* function in *R* version 3.4.3 [47].

2.2. Empirical data and analyses

2.2.1. Genotyping and descriptive population genetic statistics

We used whole genome resequencing data, previously published by Kronforst *et al.* [24]. Individuals from three species were sampled in Costa Rica. *H. cydno galanthus* & *H. pachinus* (which are more closely related to each other than to *H. melpomene rosina*) (Fig. 2 A) were sampled from the Caribbean and Pacific coastal drainages (Fig. 2 B), respectively (there is a contact zone between the two, which was not sampled there). *H. m. rosina* was sampled from overlapping sites with both *H. c. galanthus* and *H. pachinus*. Kronforst *et al.* [24] presented evidence of gene flow between the three species, and demonstrated signatures of (i) selection and adaptive introgression and (ii) elevated F_{ST} on Z chromosomes. Additionally, they showed that known wing pattern loci are involved in initial divergence in these *Heliconius* species. We note that the set of samples does not include hybrids between any of the three species [24], so we do not necessarily capture the genetic variation across (continuous) clines. Moreover, Kronforst *et al.* [24] found no evidence of admixture between these individuals, thus making it less likely that downstream analyses of within-species LD are strongly biased by LD created through admixture between taxa. Previous work indicates very little population structure in the three focal taxa within Costa Rica [28], so we are confident that population stratification is not a concern.

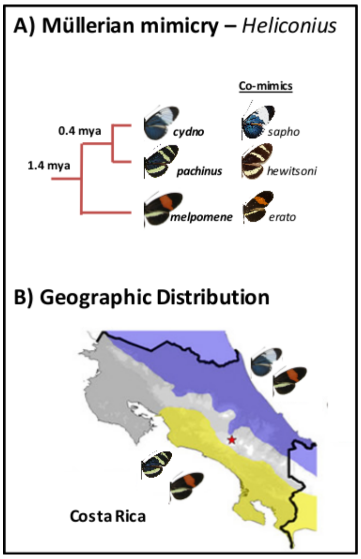

Figure 2. A) Phylogenetic tree for *Heliconius* taxa and B) geographic distribution in Costa Rica, with *H. melpomene rosina* and *H. cydno galanthus* occurring in sympatry on the Caribbean coastal drainage, and *H. pachinus* and *H. melpomene rosina* co-occurring on the Pacific coastal drainage.

Reads of *H. melpomene*, *H. c. galanthus* and *H. pachinus* (n = 10 individuals/species) were aligned to the *H. melpomene melpomene* reference assembly version 2.5 [49] with *bwa mem* version 1.15.0 [50,51]. We used *GATK* [52] version 3.8 to call variants, with heterozygosity for prior likelihood calculation per locus of 0.001, and we ignored sequences with mapping quality < 20. The minimum phred-scaled confidence threshold for variants to be called was set to 50. After genotype calling, we further filtered the data (see respective *python* script name in [53] in parentheses) to only contain variants with a distance of > 3 bp between neighbouring variants (*dropCloseVars.py*), and with depth of coverage $\leq$ mean + 3*sd (*dropHighCovVars.py*). Further, we kept variants with a minimum absolute value of -8 in base quality rank sum tests, minimum absolute value of the mapping quality rank sum test of -12.5, minimum absolute value of the read position rank sum test of -8, minimum ratio of variant confidence to non-reference read depth of 2, and finally, maximum phred-scaled p-value (using Fisher's Exact Test to detect strand bias) of 60 (*vcfFilter.py*). Even with these steps of filtering, the number of sites retained was too large for exhaustive analysis. Hence, given the computational limitations and practical limitations of manuscript length, we focused on comparisons involving 5 chromosomes, which included two autosomes (2, 7) not implicated in color patterning, two autosomes containing extensively studied color patterning loci (10, 18), and the Z chromosome (21).

For each taxon and chromosome, we calculated nucleotide diversity (π) and Tajima's D using *vcftools* [54] version 0.1.15. Additionally, for each of the three taxon pairs (*H. c. galanthus* & *H. m. rosina*; *H. pachinus* & *H. m. rosina*; and *H. c. galanthus* & *H. pachinus*), we obtained estimates of absolute divergence (d_{xy}) and population differentiation (F_{ST}) [55] with *vcftools* for non-overlapping windows of 10kbp size per chromosome. We extracted genotype likelihoods (*vcf2gl.py*) for all included variable sites and further obtained the Bayesian posterior probability distribution for genotypes using EM estimates of population allele frequencies following [56] to empirically define Hardy-Weinberg priors. We then took the mean of the posterior for each locus (and individual) as a point estimate of the genotype, which is not constrained to be an integer, ranging from zero to two. From here on, we will refer to these as genotype estimates.

2.2.2. LD between *Heliconius* loci

Based on genotype estimates, we calculated F_{ST} [sensu 57] and AFDs between taxa for each locus. In both cases, we calculated quantiles for each respective distribution, in order to determine sets of loci for the calculation of pairwise correlations between said loci. In the results section, we will focus on allele frequency differences, since the procedure for F_{ST} and AFDs is the same and the results for the F_{ST} calculations were very similar to those seen for AFDs (results based on F_{ST} can be found in the supplementary material). Loci were designated as outliers if the absolute AFD between two taxa lay above the 99th quantile for each chromosome. We expected outlier sites to be enriched for potential barrier loci, whereas non-outlier sites may represent putatively neutral sites but probably also contain weakly differentiated barrier alleles.

The resulting outliers are based on individual SNPs, and not on windows containing multiple SNPs. We did this for two reasons: (i) We wanted to present a general approach that can be used in most systems, regardless of the quality of the genome assembly used to obtain SNPs, and (ii) due to the practicality of using single SNPs for the calculation of LD, without having to account for whether or not SNPs come in blocks of elevated differentiation. It will be interesting, however, to see whether the patterns of LD for individual loci found here are similar for blocks of loci with increased differentiation.

After determining outliers for the absolute AFDs, we obtained estimates of LD among different pairwise sets of loci (outliers and non-outliers), to find potential signatures of coupling for putatively neutral and selected sites. Within each of the five chromosomes, as well as between chromosomes, we looked at simple measures of statistical non-independence. For outliers and non-outliers of both the single locus F_{ST} values as well as the absolute allele frequency differences, we calculated Pearson's correlation coefficients (r^2) of genotype estimates for six different groups of sites. Note that the outlier designation is made from the comparison of two taxa; the values, however, then come from correlations calculated within each taxon. Here, we will use allele frequency differences to illustrate the procedure (see also Fig. 3). Of the six groups of sites considered here (Fig. 3), three involved sites from the same chromosome, i.e. (1) correlations among outliers on the same chromosome (Fig. 3 B), (2) among non-outliers on the same chromosome (Fig. 3 C), and (3) between outliers and non-outliers on the same chromosome (Fig. 3 D). Additionally, we calculated pairwise r^2 values between (4) outliers on the given chromosome and outliers on all other chromosomes (Fig. 3 E), (5) between non-outliers on the given chromosome and non-outliers on other chromosomes (Fig. 3 F), and finally, (6) between outliers on the given chromosome and non-outliers on the other chromosomes (Fig. 3 G). For the first two groups, we thus obtained pairwise correlation matrices. For the remainder of groups, however, we calculated r^2 for each of the target loci with genotype estimates of a sample of loci from the respective other set. For instance, for group 3 (outlier vs non-outlier within the given chromosome - Fig. 3 D), we randomly sampled genotype estimates of non-outlier sites. For each species and outlier site, we calculated correlation coefficients with said non-outlier sites, and sampled every second correlation value from the resulting correlation matrix, to obtain a total number of r^2 values equal to the number of elements in the upper triangle of the correlation matrices for groups 1 & 2 (i.e. $(L_o^2 - L_o)/2$), with L_o = the number of outlier sites). For the correlations between SNPs on a given chromosome with loci on other chromosomes, we similarly sampled sites, although here, sites were sampled from all other chromosomes. Additionally, we limited the number of included SNPs per set to ≤ 5000. Subsampling was necessary, as the large number of variants would very quickly lead to vectors with sizes too large for downstream analyses. This sampling scheme still gave us more than 10 million correlation coefficients per respective set of loci. All of the coupling calculations as well as plotting were performed in *R* version 3.4.3 [47], with the additional package *ggplot2* [58].

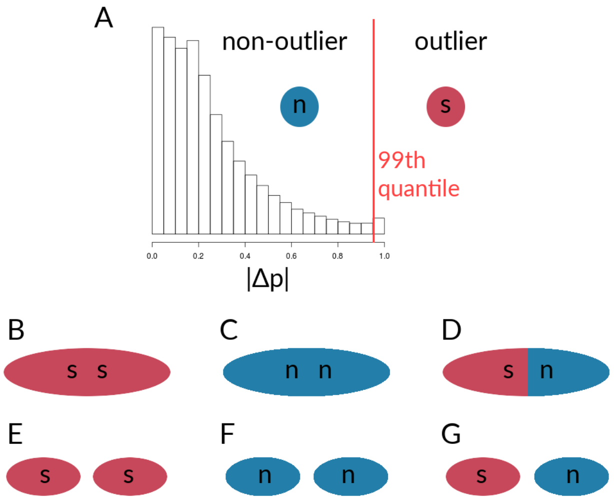

Figure 3. Schematic of comparisons for the calculation of correlation coefficients between sets of loci within species. A) designation of outlier loci based on absolute allele frequency differences, which were used for the calculation of r^2 for B) outlier loci within the given chromosome, C) non-outlier loci within the given chromosome, D) outlier loci with non-outlier loci within the given chromosome. Further, r^2 were calculated for sets of loci on the given chromosome with respective sets sampled from all other chromosomes, with E) outlier loci vs. outliers from other chromosomes, F) non-outlier loci vs. non-outliers from other chromosomes, and G) outlier loci vs. non-outliers from other chromosomes.

3. Results

3.1. Investigating coupling and its effects on selected versus neutral sites using simulations

Of the 300 simulation runs with 6 different parameter sets (with 50 runs each), only the 50 runs with $s = 0.005$ and $m = 0.1$ (i.e. runs with the highest ratio of m/s) did not reach the reproductive isolation threshold (which therefore ran for 1.5 million generations with 15 million mutations introduced); indeed, divergence never gained any traction in this parameter combination and there were no hints of coupling (Fig. 4 D, 5 D). Aggregate summary statistics for all 300 simulations are given in the supplementary material (Table S1).

A representative run with $s = 0.005$ and $m = 0.01$ is shown in Fig. 1 B & C, which reflects lower migration than the case above. It can be seen that population differentiation (F_{ST}) between demes (Fig. 1 B) increased rapidly during the transition from single-locus to multi-locus divergence, where selected loci differentiated rapidly after ∼46,000 generations, and neutral sites showed strong differentiation after ∼75,000 generations (Fig. 1 B). By the time that reproductive isolation between demes reached our a priori threshold, in generation 629,000, selected sites had reached an average allele frequency difference between demes of 0.99, whereas neutral sites had reached an average allele frequency difference of 0.89. At this same point in time, F_{ST} for neutral sites had not reached a maximum (or equilibrium) value. At the time when selected sites transition to being strongly coupled, estimators of coupling show a change in slope, and L_e rises very quickly (Fig. 1 C). Note that there are two dimensions of differences between selected and neutral sites seen in Fig. 1B. First, the vertical distance between the points for selected and neutral sites at any point in time (i.e., at a given value on the x-axis) measures a gap in divergence at that time. We henceforth refer to these vertical distances as divergence "gaps." Second, the horizontal distance between the selected and neutral points at a given level of divergence (i.e., at a given value on the y-axis) measures a lag in time. We refer henceforth to these horizontal distances as "lags" in divergence.

Aggregating runs and looking across combinations of selection and migration (Figure 4), selected sites can show strong coupling at values of the coupling coefficient, θ, for which neutral sites show negligible differentiation. This can be seen in all panels of Figure 4 except for Fig. 4 D and 4 E by noting values of θ for which neutral sites have allele frequency differences very close to zero but selected sites have risen above zero. An exception is seen in Fig. 4 E, in which differentiation for selected and neutral sites takes off at about the same value of θ for each. This is the set of conditions (among those explored) in which differentiation was the most difficult but still possible ($s = 0.001 << m = 0.1$).

Neutral sites began to show signs of differentiation for $\theta > 10^{-0.5} (= \sim 0.32)$, regardless of the values of s and m (note where blue points begin to rise from zero in all panels of Fig. 4 except 4 D). They reached their maximum rate of differentiation for values of $\theta \geq 1$ (note positions of right-most vertical lines in Fig. 4; there is no line shown for neutral sites in panel C because it was off the scale; see Table S2 for exact values). Intuitively, selected sites showed signs of differentiation for lower values of θ than neutral sites (note red points above blue points in Fig. 4). Strong acceleration of differentiation indicates the occurrence of coupling among selected sites. The point of the maximum rate of change for selected sites as a function of theta depended strongly upon migration, m, and much less so upon s. When $m = 0.01$, this point occurred for $\theta \sim 10^{-0.65}$ (= 0.22; see Table S2) regardless of s (Fig. 4 A-C). When m was increased, this point increased as well, to $\theta \sim 10^{-0.23} (= 0.59)$, which again varied little with a two-fold change in s (Fig. 4 E,F). To summarize, there was a general association between how difficult it is for selected sites to differentiate and how similar the dynamics are for selected versus neutral sites, but there was no singular value of θ for which all scenarios showed transitions of either selected or neutral sites.

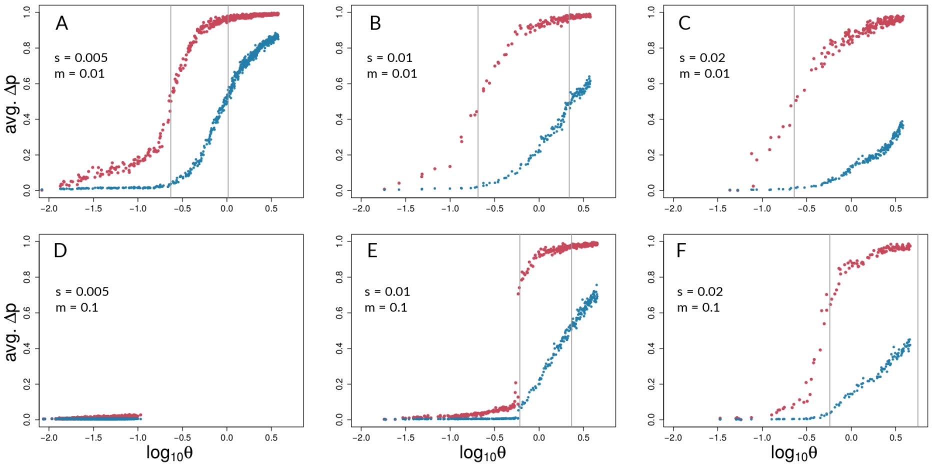

Figure 4. Barton's coupling coefficient (θ) and average allele frequency differences between demes, for different combinations of s and m (values given in each panel). Each plot shows data points for selected (red) and neutral (blue) sites from 50 independent simulation runs with equal parameters. Grey lines indicate the point of highest slope (i.e. inflection point) by nonlinear least squares model fitting across the 50 respective runs for selected and neutral sites (Table S2). Note that grey lines could fall outside the range of respective data points if the predicted inflection point was not reached.

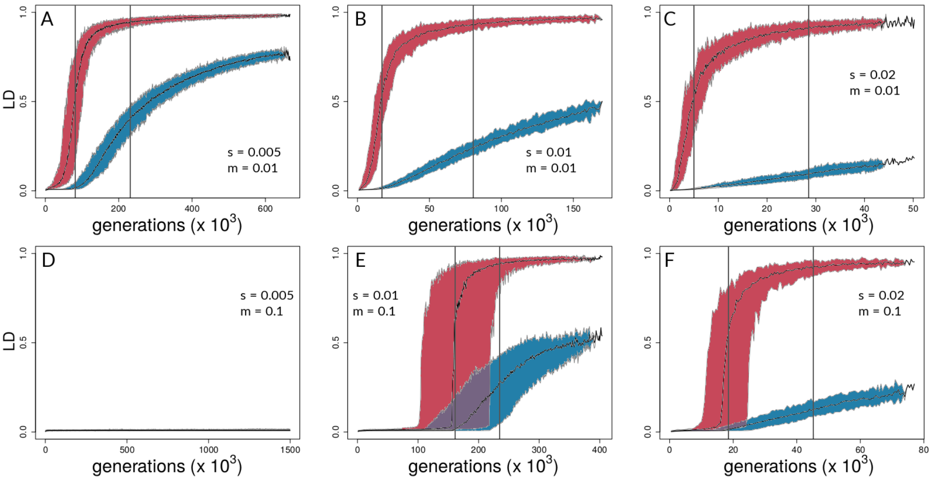

Figure 5. LD time series for the same combinations of *s* and *m* as in Fig. 4. Each plot shows the median result across 50 independent simulation runs and 95% quantiles for selected (red) and neutral sites (blue). Grey lines indicate the point of highest slope by nonlinear least squares model fitting for medians across the 50 respective runs (Table S3 for details).

These differences between selected and neutral sites are borne out in time as well (Figure 5). Once there is sufficient potential for coupling, LD between selected sites increases rapidly in all parameter combinations that lead to divergence. LD between neutral sites eventually shows increases in all these cases as well (Fig. 5 A, B, C, E, & F), but the divergence gap between the neutral and selected sites is pronounced for long periods of time. The size of the gap (vertical distance between data points for selected and neutral sites in Fig. 5) is related intuitively to the strength of selection: larger values of *s* lead to a greater gap in levels of divergence of selected versus neutral sites. This is seen by noting that the blue points reach lower levels as one moves from left to right across the panels within any one row of Figs. 4 and 5. However, it is important to note here that these gaps are not measured at the same point in time across all conditions. Rather, what we are highlighting here is relative to the time at which strong multi-locus coupling causes a transition to a highly differentiated state among selected sites (note different time scales on panels of Fig. 5). The lag between selected and neutral sites can be measured by the difference in inflection points of the fitted curves, i.e., where the rise in differentiation is predicted to be the steepest. This occurs for θ values that are consistently at least half an order of magnitude larger for neutral sites (compare x-axis distance between gray, vertical lines in Fig. 4, and see Tables S2). In time (Fig. 5), this translates to 10s or 100s of thousands of generations, depending upon parameter values (see Table S3).

Note that Figs. 4D and 5D underscore the existence of thresholds for transitions in divergence with gene flow: migration was so strong relative to selection ($m = 20s$) in this scenario that sufficient standing variation in selected sites could never build up to raise θ high enough to cause coupling. Instead, selected and neutral variation was consistently eliminated in this scenario by the combination of migration and drift, resulting in an undifferentiated pseudo-equilibrium state of mutation-drift balance (true over all of the 50 replicates).

3.2. *Empirical data and analyses*

3.2.1. Genotyping and descriptive population genetic statistics

For all 21 chromosomes, we found a total of 12,739,517 variants after alignment to *Hmel 2.5*, genotyping and further quality filtering. Across the 5 chosen chromosomes (2, 7, 10, 18, & 21), there are a total of 3,262,190 variants, with the number of variants and the number of scaffolds per chromosome given in Table S4.

Consistent with Kronforst *et al.* [24], we found that nucleotide diversity (π) was highest for *H. c. galanthus*, intermediate for *H. m. rosina*, and the lowest levels of polymorphism were found in *H. pachinus* (see Table S5 for values per chromosome and species). In all three taxa, π was consistently lowest on the Z chromosome (21). *H. c. galanthus* and *H. pachinus* showed lower levels of population differentiation from each other than from the more distant *H. m. rosina*, which can be seen from both levels of absolute divergence (d_{xy}) as well as F_{ST} (see Table S6). Also consistent with Kronforst *et al.* [24], for both the comparisons involving the more distant *H. m. rosina*, values of d_{xy} and F_{ST} are considerably higher on the Z chromosome, indicating higher species differentiation, when compared to the autosomes. For the remaining pair of *H. c. galanthus* with *H. pachinus*, however, we did not observe higher values of either d_{xy} or F_{ST} (see Table S6).

3.2.2. LD between *Heliconius* loci

For the three pairs of taxa (i.e. *H. c. galanthus* - *H. m. rosina*, *H. pachinus* - *H. m. rosina*, and *H. c. galanthus* - *H. pachinus*) across the five chromosomes, we designated a total of 32,621 outliers above the 99th quantile of AFDs and F_{ST}, respectively. Quantiles of AFDs across chromosomes showed average values of 0.97 (sd = 0.024) for the *c. galanthus* - *m. rosina* pair, 0.99 (sd = 0.013) for *pachinus* and *m. rosina*, and 0.433 (sd = 0.016) for *c. galanthus* and *pachinus*. Conversely, quantiles for F_{ST} across chromosomes showed average values of 0.961 (sd = 0.028) for the *c. galanthus* - *m. rosina* pair, 0.985 (sd = 0.016) for *pachinus* and *m. rosina*, and 0.419 (sd = 0.022) for *c. galanthus* - *pachinus* (see also Table S6 and Fig. S1). Both species pairs involving *H. m. rosina* share many outlier sites (18.1% on the Z chromosome, and 27.9, 33, 27.2 and 26.9% with AFDs for chromosomes 2, 7, 10, and 18, respectively). On the other hand, outlier sites between *H. c. galanthus* and *H. pachinus* are predominantly private, i.e. do not occur in either of the pairs involving *H. m. rosina* (see also Figure S1).

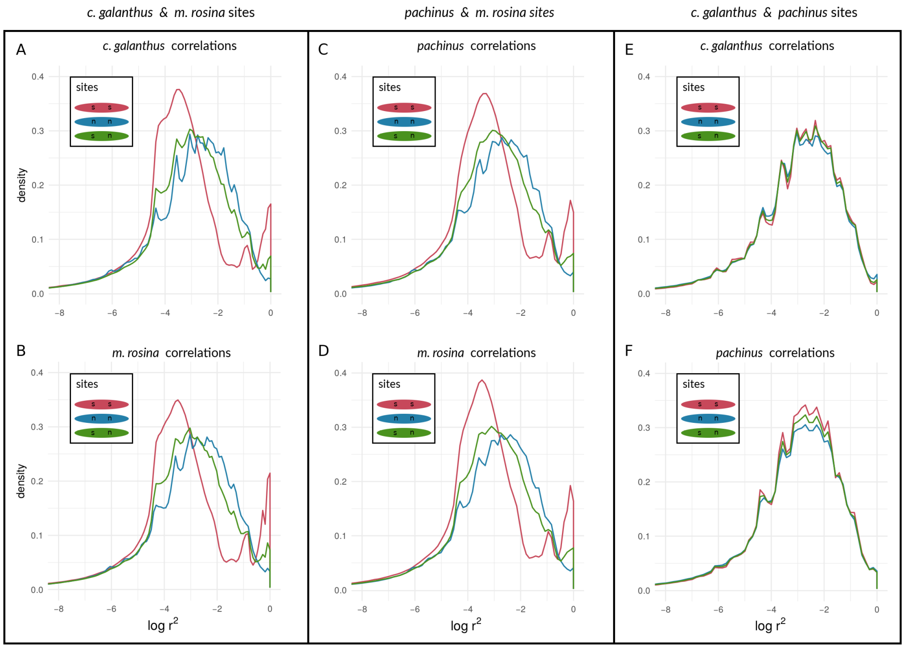

Figure 6. Density curves of linkage disequilibrium for loci at different types of sites on chromosome 2 (representative of autosomes), determined by outliers of allele frequency differences. $\log(r^2)$ values are shown between selected loci, between neutral loci, and between selected and neutral loci for species pairs of *H. cydno galanthus* and *H. melpomene rosina* (A & B), *H. pachinus* and *H. melpomene rosina* (C & D), and *H. cydno galanthus* and *H. pachinus* (E & F).

We found a unimodal distribution of $\log(r^2)$ between loci for most non-outlier sites. For the two taxon-pairs of (1) *H. c. galanthus* and *H. m. rosina*, as well as for (2) *H. pachinus* and *H. m. rosina*, we found bimodal distributions of $\log(r^2)$ between non-outlier loci on autosomes (Fig. 6 A - D for chromosome 2, and Figures S2 - S4) and the Z chromosome (21) (Fig. 7 A - D), which deviate from the distributions of $\log(r^2)$ between either non-outlier sites as well as the correlations of outliers between non-outliers. Furthermore, $\log(r^2)$ between loci on the Z chromosome appear to have shifted to higher values, especially for correlations between outlier sites, but we could also see that the peak of neutral sites seems to have decreased in the *H. m. rosina* outliers for both pairs involving *H. m. rosina* (Fig. 6 B & D), with what appears to be a slowly differentiating peak at higher $\log(r^2)$ values. In the third pair, *H. c. galanthus* with *H. pachinus*, we did not see a deviation of $\log(r^2)$ between sites, neither for outliers nor for non-outliers in any of the autosomes (Fig. 6 E & F, and Fig. S2 - S4) or the Z chromosome (Fig. 7 E & F).

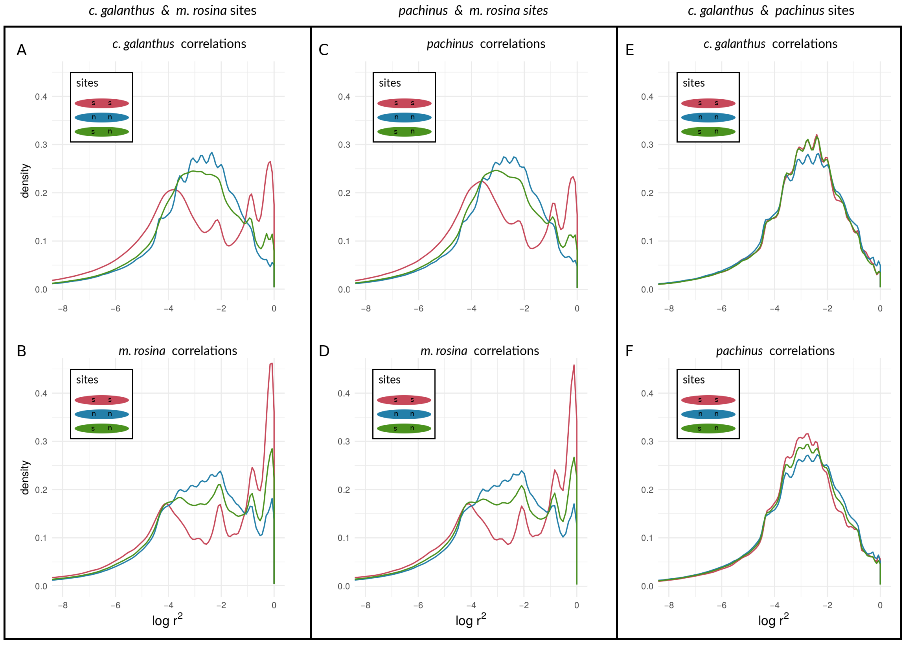

Figure 7. Density curves of linkage disequilibrium for loci at different types of sites on chromosome 21, determined by outliers of allele frequency differences. $\log(r^2)$ values are shown between selected loci, between neutral loci, and between selected and neutral loci for species pairs of *H. cydno galanthus* and *H. melpomene rosina* (A & B), *H. pachinus* and *H. melpomene rosina* (C & D), and *H. cydno galanthus* and *H. pachinus* (E & F).

Turning towards coupling of respective sites with sites on other chromosomes, we found that both species pairs involving *H. m. rosina* are indicative of coupling of outliers with outliers from other chromosomes, both for autosomes (Fig. S5 A - D for chromosomes 2; and see Fig. S6 - S8), as well as for the Z chromosome (Fig. S9 A - D). Similarly to coupling within chromosomes, the *c. galanthus* - *pachinus* pair did not exhibit signs of coupling (Fig. S5 - S9, E & F). In plain terms, when a second, upper mode appears in a distribution of correlation coefficients, as seen in cells A and B of Figures 6 and 7 (and see Figures S2 - S9), this is indicative of the occurrence of non-random associations between large numbers of sites. That is, this bimodality indicates that a transition has occurred in which at least some portion of sites in the genome are statistically non-independent, and thus coupled with each other. Note that with the sample sizes of numbers of outliers, there are thousands of sites near a given mode. Different types of sites (outlier vs. non-outlier), different chromosomes within a taxonomic comparison, and different taxonomic comparisons all show variation in the presence/absence and (if present) prominence of the upper mode of the distribution. This variation indicates how different taxa and different portions of the genome are at different points along the speciation continuum. We can hypothesize that these different points are proxies for changes in time, as Figure 8 attempts to illustrate. Moving from the curves in sequence from "1" to "4" in Figure 8 A, we see increasing levels of coupling among selected loci. Looking at the same comparisons for neutral sites in Figure 8 B, we see the lag of neutral sites behind selected sites, but that the neutral sites do eventually begin to show signs of being coupled as well. Results based on F_{ST} quantiles were very similar to AFDs (see Figures S10 - S19 for all species pairs and respective chromosomes).

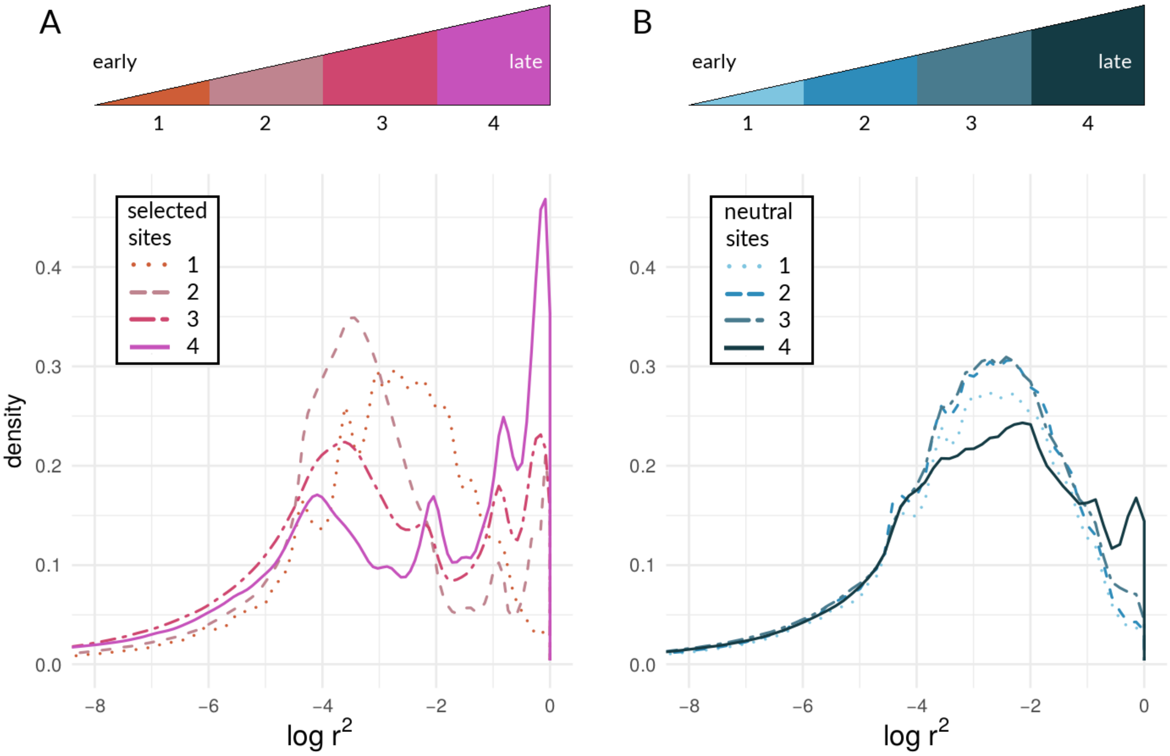

Figure 8. Density curves of linkage disequilibrium for (A) outliers and (B) non-outlier loci. Individual lines correspond to $\log(r^2)$ values of 1) *pachinus* locus pairs for the comparison of *c. galanthus* and *pachinus* on chromosome 2, 2) *m. rosina* locus pairs for the comparison of *pachinus* and *m. rosina* on chromosome 2, 3) *pachinus* locus pairs for the comparison of *pachinus* and *m. rosina* on chromosome 21, 4) *m. rosina* locus pairs for the comparison of *c. galanthus* and *m. rosina* on chromosome 21.

4. Discussion

Despite sustained and deep interest in the processes that lead to the evolution of diversity at multiple levels of biological organization, the study of speciation remains challenging because of the interacting mechanisms involved, the breadth and complexity of the potential demographic scenarios that encompass divergence, and the difficulty of integrating theoretical approaches with emergent empirical patterns observed in natural populations. The emphases on "speciation genes" or "islands of divergence", in particular, have prevented more holistic and integrative considerations of genome-wide statistical patterns of divergence and gene flow throughout the speciation process in empirical studies of speciation [see also 22,59–61]. To address this limitation, we built upon previous theoretical results suggesting that linkage disequilibrium (LD) among numerous weakly-selected loci leads to rapid, genome-wide congealing during divergence by incorporating neutral sites into our model. In addition, we focused on classical signatures of multi-locus coupling, derived from hybrid zone theory [e.g. 4,8], to investigate how divergent selection directly and indirectly affects the transition from independently evolving loci, which theory predicts should predominate early during speciation, to more extensive statistical linkage among loci at genome-wide scales. Finally, to investigate the qualitative agreement between the results of our theoretical simulations with patterns of divergence and LD in empirical data, we calculated measures of LD for loci both within and between chromosomes for three species of hybridizing *Heliconius* butterflies representing both early and later stages of the speciation process.

4.1. Simulations

We first investigated how patterns of divergence change over generations and over parameter conditions in our forward-time simulations of two demes under different levels of divergent selection and migration between demes (Figs. 1, 4, 5). In Felsenstein's seminal paper [42], he argued that in two-allele models, as presented here, the build up of reproductive isolation should be constrained by migration because recombination is expected to break down associations among different selected loci. As seen here, however, if the repeated establishment of differentially adapted alleles is possible, a threshold can be reached at which strong divergence evolves and effective migration between demes is reduced, even at neutral sites [4,6,9,12,17,41] (Figs 4 & 5). This transition is characterized by a shift from single loci acting independently, to a joint effect of loci on the increasingly diverging genome. Such transitions in speciation occur when divergent selection and linkage disequilibrium reach a critical threshold, triggering genome-wide differentiation [17,22], but previous work has not thoroughly explored how these dynamics impact neutral sites temporally.

The results of our simulations suggest that while both selected and neutral sites experience a similar build up of genome-wide differentiation, selected loci experience this transition at lower values of θ (Fig. 4) and earlier in time (Fig. 5) than neutral sites. Even when transitions for selected and neutral sites begin at nearly coincident points in time (e.g., Fig. 5 E) , the rates of increase in and magnitude of differentiation are quite different, as shown by the steepness and (for Fig. 5) asymptotes of the fitted generalized logistic models (see Tables S2 & S3). These findings suggest that genome-wide statistical patterns may indeed offer substantial power to distinguish barrier from non-barrier loci, under a wide range of selection and migration conditions (e.g., across the panels of Figs. 4 & 5), at least once there is a transition to a state of multi-locus divergence (i.e., not in conditions like Fig. 4 D, 5 D).

Although historically defined as the ratio of total selection and total recombination [4], the definition of coupling has recently been extended to include any process that leads to a coincidence of barrier effects and, hence, stronger barriers to gene flow due to the buildup of LD among loci [3]. The use of a coupling statistic inspired by previous theory [4], θ, allowed us to encapsulate the potential for coupling in simulated systems. However, there was not a single approximate value of θ at which coupling of selected sites occurred. The sharp uptick in allele-frequency differences – a dynamic change indicative of coupling – and the point of the maximum rate of change spanned a range of θ values (Fig. 4). The uptick for differentiation at neutral sites was consistently at larger θ values, intuitively, but varied as well. Comparing inflection points for neutral and selected sites as a function of θ, we observed the greatest difference in simulations with higher selection coefficients (s) and lower migration rates (m) (Figs. 4 C, 5 C; Table S2).That is, even with tiny rates of gene flow, differentiation at neutral sites is very difficult to achieve and coupling among selected sites must be very strong to cut down gene flow at neutral sites. However, as s grows and/or m shrinks, it becomes easier and easier for selected sites to show divergence – and thus to rise in LD with each other as a statistical consequence – far in advance of the conditions at which neutral sites can diverge. Note that these differences in threshold θ values do not translate into absolute differences in time; divergence for both types of sites is faster in absolute terms of time as s increases (x-axis scales in Fig. 5).

Interestingly, the value of θ at which selected sites showed strong coupling and the most rapid differentiation depended strongly upon m but less so on s. This is seen by comparing the panels of Fig. 4: in the upper row of panels, m is constant at 0.01, and although s is varied four-fold, the inflection point (coefficient b in model fits) stays at about the same value of θ (see Table S2). When m is increased but s is held constant (compare Fig. 4B to E, or C to F), the value of theta at the inflection point increases. One way to interpret this is that increases in m increase the effective amount of recombination between genotypes from different demes. As such, the potential for coupling, as encapsulated by θ, has to reach greater levels before coupling can actually occur.

We note that these predictions focus on divergence with gene flow. Though we did not study strict allopatric scenarios here, prior work [14,17] indicates that, for selected sites, when $s >> m$ (as is true in allopatry when $m = 0$), divergence from de novo mutations is relatively linear in time: differentiation between demes at selected sites will accumulate at a more-or-less constant rate dictated by the mutation rate, drift, and the strength of selection. For neutral sites, however, zero migration is likely to be very different than even small amounts of migration because – as results above indicate – even very low rates of effective gene flow and recombination are sufficient to prevent differentiation at neutral sites (i.e., neutral divergence did not show strong upticks until multi-locus barriers were very strong and the coupling coefficient reached high values). In a strict allopatry scenario, there would be no such impediment to neutral differentiation, and thus neutral sites would accumulate differentiation at a rate determined by mutation and drift, starting as soon as allopatry began. Furthermore, in a strict allopatry scenario, all de novo mutations would be private alleles, and thus between-deme LD would be at its maximum among all mutations that established, selected and neutral. This would produce very different distributions and time series of LD compared to our scenarios with gene flow.

Understanding additional quantitative dynamics of the differences between selected and neutral sites, including scenarios of strict allopatry, will require additional future work, but our results generally suggest that differences may exist in the emergence of significant LD among sites for selected and neutral loci at different stages of the speciation process and that heterogeneity in patterns of LD, as measured by coupling, might reveal important details of where along the speciation continuum two hybridizing lineages currently reside.

4.2. Examining the speciation continuum in Heliconius by using LD as a proxy for coupling

To empirically investigate speciation transitions related to coupling, we also assessed LD among three *Heliconius* species pairs, spanning various stages of evolutionary divergence along the speciation continuum [24,28], which are known to hybridize. While LD is not required for coupling [3], it is expected when coupling occurs among loci that have "two-allele" effects [sensu 42]. Specifically, we investigated patterns of LD between different categories of loci identified as outliers or non-outliers based on population allele frequencies. Our implicit assumption is that outliers should be enriched for loci that act as barriers, such as those under divergent selection. However, we do not explicitly argue that non-outlier sites are neutral. Instead, we argue that the comparison of outlier sites with the genomic background (non-outliers which might have experienced varying direct or indirect selective pressures) represents a reasonable approach to investigate empirical patterns of divergence and coupling. We purposely chose to examine LD among individual sites – rather than attempting to examine regions of differentiation (e.g. known color pattern loci in *Heliconius*) – to maximize the generalizability to other systems where knowledge about the specific targets of divergent natural selection leading to reproductive barriers may be unavailable.

Our approach was not meant to identify "speciation genes" per se. Rather, we examined the aggregate statistical properties – the statistical distributions of LD values seen in Figs. 6-7 – of chromosomes and the genome with respect to divergence. We argue that the categorical difference between a unimodal distribution (e.g., lines in Fig. 6 E) and a bimodal distribution (e.g., red line in Fig. 6 A) are powerful indicators of populations and/or portions of the genome that are at different stages along the speciation continuum. Specifically, the appearance of an upper mode, close to the maximum possible value of LD, is only plausibly explained by coupling of barrier loci and its effects (though we emphasize that we are not attempting to infer the mechanism that produced this coupling in the empirical data). We found that signatures of LD for outlier sites do coincide with prior findings of differentiation of both *H. c. galanthus* and *H. pachinus* with *H. m. rosina*. Moreover, within both of the two-species comparisons involving *H. m. rosina*, there is variation in LD among different chromosomes, with the Z chromosome exhibiting the strongest signature of coupling (which is the most differentiated in both *c. galanthus* and *m. rosina*, and has the lowest nucleotide diversity (see Tables S5 & S6 and [24]). The third pair of taxa (*c. galanthus* and *pachinus*) does not exhibit signs of elevated LD for the outliers. In other words, we found that the taxa which are more closely related to each other do not show LD, whereas taxa that have diverged further from each other, showed signatures of LD, both within chromosomes as well as between chromosomes. Further, we found variation in LD for different chromosomes, where the Z chromosome exhibits the strongest signal (coinciding with higher differentiation at the Z chromosome).

The different patterns of LD between different loci for the three pairwise species-comparisons could be seen as different stages along the speciation continuum and in the process of transitioning from uncoupled to coupled (Fig. 8 A), and this could be true for the genomic background as well (Fig. 8 B). However, it is also possible that sites in the genomic background – non-outliers that are less likely to be barriers – may also be under selection, and coupled, but simply haven't reached high enough allele frequencies in the respective taxa, to be identified as outliers.

One potential caveat to these results is that we estimated allele frequencies for each species using individual samples of each taxon rather than discrete populations. Thus, unsampled population structure could potentially confound our allele frequency estimates. However, Kronforst and Gilbert [28] previously investigated these *Heliconius* species for evidence of population stratification, and found extremely low pairwise F_{ST} between populations (*H. cydno*: 0.002-0.018; *H. melpomene*: 0.037-0.136; *H. pachinus*: 0-0.0016), corresponding with prior evidence of little population genetic structure based on allozyme data [62–65]. Another potential confounding factor to consider is the possibility that large structural variation (e.g., inversions; [19]) or other recombination modifiers [16,66–69] segregating between species might lead to higher LD estimates. Feder et al. (2014) predicted that such modifiers would have to be large to be fixed in a population (i.e. on the order of megabases) [19]. However, recent work by Davey et al. (2017) using fine scale recombination maps, found no evidence of large-scale inversions between *H. melpomene* and *H. cydno*. Additionally, the occurrence of LD between sites on different chromosomes (e.g., Figs. S5-S9) would not be directly attributable to effects of structural variation (but is a good indicator of coupling across the entire genome).

5. Conclusions

Our simulations focused on the de novo build-up of divergent adaptation, neutral differentiation, and coupling in the face of gene flow. We note, however, that the question and our results are also germane to cases of understanding the dynamics of speciation following secondary contact. Our findings pertain to the potential breakdown or retention and build-up of coupled complexes of selected and neutral genes following contact of diverged populations, as well. Future theory will more explicitly explore the effects of varying periods of allopatry. Even so, making explicit quantitative comparisons between simulation results and empirical data is difficult, as we do not have temporal data from *Heliconius* species spanning large numbers of generations. Future work will attempt to surmount this with a combination of forward-time simulations, coalescent simulations, and model-based parameter inference. Additionally, it would be particularly useful to broaden the scope to a wider range of *Heliconius* taxa, as well as to other cases of adaptive radiations for comparative purposes. With data from multiple independent radiations, we could test the generality and scale(s) of repeatability of patterns of divergence.

Coupling between loci represents an informative component of the study of the genetics of speciation [3]. We need to examine a breadth of additional systems to gain insight into the potential generality of coupling measures to reveal speciation in action. Additionally – recognizing that our simulations focused on only divergent adaptation and neutral processes here – it is crucially important that theory continue to broaden to include the interactions of a variety of types of selection (background, balancing, positive, epistatic, etc.) and their interactions with realistic demographic processes. Doing so will enable refinement of the predictions we can make about the signatures of the multitude of processes shaping genome-wide variation in natural populations.

Supplementary Materials: The following are available online at www.mdpi.com/link,

Table S1: Summary statistics for the 300 included simulation runs,

Table S2: Summary for nonlinear least squares fit of Barton's coupling coefficient (θ) and average allele frequency differences between demes for selected and neutral loci,

Table S3: Summary for nonlinear least squares fit of average LD for selected and neutral loci,

Table S4: Numbers of scffolds and variants for each invluded *Heliconius* chromosome,

Table S5: Population genetic statistics for *Heliconius* species per chromosome,

Table S6: Population genetic statistics for *Heliconius* species pairs per chromosome,

Figure S1: Venn diagrams to illustrate the overlap for outliers of AFDs and F_{ST},

Figure S2: Density curves of coupling coefficients for loci at different types of sites on chromosome 7, determined by outliers of AFDs,

Figure S3: Density curves of coupling coefficients for loci at different types of sites on chromosome 10, determined by outliers of AFDs,

Figure S4: Density curves of coupling coefficients for loci at different types of sites on chromosome 18, determined by outliers of AFDs,

Figure S5: Density curves of coupling coefficients for loci at different types of sites on chromosome 2 and respective sites on all other chromosomes,

Figure S6: Density curves of coupling coefficients for loci at different types of sites on chromosome 7 and respective sites on all other chromosomes,

Figure S7: Density curves of coupling coefficients for loci at different types of sites on chromosome 10 and respective sites on all other chromosomes,

Figure S8: Density curves of coupling coefficients for loci at different types of sites on chromosome 18 and respective sites on all other chromosomes,

Figure S9: Density curves of coupling coefficients for loci at different types of sites on chromosome 21 and respective sites on all other chromosomes,

Figure S10: Density curves of coupling coefficients for loci at different types of sites on chromosome 2, determined by F_{ST} outliers,

Figure S11: Density curves of coupling coefficients for loci at different types of sites on chromosome 7, determined by F_{ST} outliers,

Figure S12: Density curves of coupling coefficients for loci at different types of sites on chromosome 10, determined by F_{ST} outliers,

Figure S13: Density curves of coupling coefficients for loci at different types of sites on chromosome 18, determined by F_{ST} outliers,

Figure S14: Density curves of coupling coefficients for loci at different types of sites on chromosome 21, determined by F_{ST} outliers,

Figure S15: Density curves of coupling coefficients for loci at different types of sites on chromosome 2 and respective sites on all other chromosomes, determined by F_{ST} outliers,

Figure S16: Density curves of coupling coefficients for loci at different types of sites on chromosome 7 and respective sites on all other chromosomes, determined by F_{ST} outliers,

Figure S17: Density curves of coupling coefficients for loci at different types of sites on chromosome 10 and respective sites on all other chromosomes, determined by F_{ST} outliers,

Figure S18: Density curves of coupling coefficients for loci at different types of sites on chromosome 18 and respective sites on all other chromosomes, determined by F_{ST} outliers,

Figure S19: Density curves of coupling coefficients for loci at different types of sites on chromosome 21 and respective sites on all other chromosomes, determined by F_{ST} outliers. .

Acknowledgments: MPS and SMF were supported by NSF DEB award 1627483. JLF was supported by funding from NSF and USDA NIFA grants. This work utilized the Janus supercomputer, which was supported by the National Science Foundation (award number CNS-0821794) and the University of Colorado Boulder. The Janus supercomputer was a joint effort of the University of Colorado Boulder, the University of Colorado Denver and the National Center for Atmospheric Research. Further computation was done on the University of Colorado Biofrontiers Fiji cluster.

Author Contributions: MPS analyzed all the data, made most of the figures, contributed ideas for analysis and framing, and co-wrote the paper. SPM and MK helped with data from Heliconius, contributed ideas for analysis and framing, and co-wrote the paper. RJS, JLF, PN, and ZG contributed ideas for analysis and framing, and co-wrote the paper. SMF wrote the simulation source code, produced the data from it, contributed ideas for analysis and framing, and co-wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

1. Seehausen, O.; Butlin, R.K.; Keller, I.; Wagner, C.E.; Boughman, J.W.; Hohenlohe, P.a.; Peichel, C.L.; Saetre, G.P.; Bank, C.; Brännström, A.; Brelsford, A.; Clarkson, C.S.; Eroukhmanoff, F.; Feder, J.L.; Fischer, M.C.; Foote, A.D.; Franchini, P.; Jiggins, C.D.; Jones, F.C.; Lindholm, A.K.; Lucek, K.; Maan, M.E.; Marques, D.a.; Martin, S.H.; Matthews, B.; Meier, J.I.; Möst, M.; Nachman, M.W.; Nonaka, E.; Rennison, D.J.; Schwarzer, J.; Watson, E.T.; Westram, A.M.; Widmer, A. Genomics and the origin of species. *Nat Rev Genet* **2014**, *15*, 176–92.
2. Harrison, R.G.; Larson, E.L. Hybridization, introgression, and the nature of species boundaries. *J Hered* **2014**, *105*, 795–809.
3. Butlin, R.K.; Smadja, C.M. Coupling, reinforcement, and speciation. *Am Nat* **2018**, *191*, 000–000.
4. Barton, N.H. Multilocus clines. *Evolution* **1983**, *37*, 454–471.
5. Abbott, R.; Albach, D.; Ansell, S.; Arntzen, J.W.; Baird, S.J.E.; Bierne, N.; Boughman, J.; Brelsford, A.; Buerkle, C.A.; Buggs, R.; Butlin, R.K.; Dieckmann, U.; Eroukhmanoff, F.; Grill, A.; Cahan, S.H.; Hermansen, J.S.; Hewitt, G.; Hudson, a.G.; Jiggins, C.; Jones, J.; Keller, B.; Marczewski, T.; Mallet, J.; Martinez-Rodriguez, P.; Möst, M.; Mullen, S.; Nichols, R.; Nolte, a.W.; Parisod, C.; Pfennig, K.; Rice, a.M.; Ritchie, M.G.; Seifert, B.; Smadja, C.M.; Stelkens, R.; Szymura, J.M.; Väinölä, R.; Wolf, J.B.W.; Zinner, D. Hybridization and speciation. *J Evol Biol* **2013**, *26*, 229–46.
6. Barton, N.H.; Bengtsson, B.O. The barrier to genetic exchange between hybridising populations. *Heredity* **1986**, *56*, 357–376.
7. Baird, S.J.E. A simulation study of multilocus clines. *Evolution* **1995**, *49*, 1038–1045.
8. Kruuk, L.E.B.; Baird, S.J.E.; Gale, K.S.; Barton, N.H. A comparison of multilocus clines maintained by environmental adaptation or by selection against hybrids. *Genetics* **1999**, *153*.
9. Barton, N.H.; De Cara, M.A.R. The evolution of strong reproductive isolation. *Evolution* **2009**, *63*, 1171–1190.
10. Yeaman, S.; Otto, S.P. Establishment and maintenance of adaptive genetic divergence under migration, selection, and drift. *Evolution* **2011**, *65*, 2123–2129.
11. Feder, J.L.; Nosil, P. The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. *Evolution* **2010**, *64*, 1729–1747.
12. Feder, J.L.; Gejji, R.; Yeaman, S.; Nosil, P. Establishment of new mutations under divergence and genome hitchhiking. *Phil Trans R Soc Lond B Biol Sci* **2012**, *367*, 461–474.
13. Via, S. Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-gene-flow. *Phil Trans R Soc Lond B Biol Sci* **2012**, *367*, 451–460.
14. Flaxman, S.M.; Feder, J.L.; Nosil, P. Genetic hitchhiking and the dynamic buildup of genomic divergence during speciation with gene flow. *Evolution* **2013**, *67*, 2577–2591.
15. Yeaman, S.; Whitlock, M.C. The genetic architecture of adaptation under migration–selection balance. *Evolution* **2011**, *65*, 1897–1911.
16. Yeaman, S. Genomic rearrangements and the evolution of clusters of locally adaptive loci. *Proc Natl Ac Sci* **2013**.
17. Flaxman, S.M.; Wacholder, A.C.; Feder, J.L.; Nosil, P. Theoretical models of the influence of genomic architecture on the dynamics of speciation. *Mol Ecol* **2014**, *23*, 4074–4088.
18. Bierne, N.; Welch, J.; Loire, E.; Bonhomme, F.; David, P. The coupling hypothesis: why genome scans may fail to map local adaptation genes. *Mol Ecol* **2011**, *20*, 2044–2072.
19. Feder, J.L.; Nosil, P.; Wacholder, A.C.; Egan, S.P.; Berlocher, S.H.; Flaxman, S.M. Genome-wide congealing and rapid transitions across the speciation continuum during speciation with gene flow. *J Hered* **2014**, *105*, 810–820.
20. Nosil, P.; Gompert, Z.; Farkas, T.E.; Comeault, a.a.; Feder, J.L.; Buerkle, C.a.; Parchman, T.L. Genomic consequences of multiple speciation processes in a stick insect. *Proc R Soc Lond Biol* **2012**, *June*, 5058–5065.
21. Nosil, P.; Feder, J.L. Genomic divergence during speciation: causes and consequences. *Phil Trans R Soc Lond B Biol Sci* **2012**, *367*, 332–342.
22. Nosil, P.; Feder, J.L.; Flaxman, S.M.; Gompert, Z. Tipping points in the dynamics of speciation. *Nat Ecol Evol* **2017**, *1*, 0001.
23. Southcott, L.; Kronforst, M.R. A neutral view of the evolving genomic architecture of speciation. *Ecol Evol* **2017**.

24. Kronforst, M.R.; Hansen, M.E.B.; Crawford, N.G.; Gallant, J.R.; Zhang, W.; Kulathinal, R.J.; Kapan, D.D.; Mullen, S.P. Hybridization reveals the evolving genomic architecture of speciation. *Cell Reports* **2013**, *5*, 666–677.
25. Beltrán, M.; Jiggins, C.D.; Bull, V.; Linares, M.; Mallet, J.; McMillan, W.O.; Bermingham, E. Phylogenetic discordance at the species boundary: comparative gene genealogies among rapidly radiating *Heliconius* butterflies. *Mol Biol Evol* **2000**, *19*, 2176–2190.
26. Bull, V.; Beltrán, M.; Jiggins, C.D.; McMillan, W.O.; Bermingham, E.; Mallet, J. Polyphyly and gene flow between non-sibling *Heliconius* species. *BMC biol* **2006**, *4*, 11.
27. Kronforst, M.R.; Young, L.G.; Blume, L.M.; Gilbert, L.E. Multilocus analyses of admixture and introgression among hybridizing *Heliconius* butterflies. *Evolution* **2006**, *60*, 1254–1268.
28. Kronforst, M.R.; Gilbert, L.E. The population genetics of mimetic diversity in *Heliconius* butterflies. *Proc R Soc Lond Biol* **2008**, *275*, 493–500.
29. Martin, S.H.; Dasmahapatra, K.K.; Nadeau, N.J.; Salazar, C.; Walters, J.R.; Simpson, F.; Blaxter, M.; Manica, A.; Mallet, J.; Jiggins, C.D. Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Res* **2013**, *23*, 1817–1828.
30. Jiggins, C.D.; Naisbit, R.E.; Coe, R.L.; Mallet, J. Reproductive isolation caused by colour pattern mimicry. *Nature* **2001**, *411*, 302.
31. Jiggins, C.D.; Salazar, C.; Linares, M.; Mavarez, J. Hybrid trait speciation and *Heliconius* butterflies. *Phil Trans R Soc Lond B Biol Sci* **2008**, *363*, 3047–3054.
32. Kronforst, M.R.; Kapan, D.D.; Gilbert, L.E. Parallel genetic architecture of parallel adaptive radiations in mimetic *Heliconius* butterflies. *Genetics* **2006**, *174*, 535–539.
33. Chamberlain, N.L.; Hill, R.I.; Kapan, D.D.; Gilbert, L.E.; Kronforst, M.R. Polymorphic butterfly reveals the missing link in ecological speciation. *Science* **2009**, *326*, 847–850.
34. Benson, W.W. Resource partitioning in passion vine butterflies. *Evolution* **1978**, *32*, 493–518.
35. Estrada, C.; Jiggins, C.D. Patterns of pollen feeding and habitat preference among *Heliconius* species. *Ecol Entomol* **2002**, *27*, 448–456.
36. Mallet, J.; Gilbert Jr, L.E. Why are there so many mimicry rings? Correlations between habitat, behaviour and mimicry in *Heliconius* butterflies. *Biol J Linnean Soc* **1995**, *55*, 159–180.
37. Smiley, J. Plant chemistry and the evolution of host specificity: new evidence from *Heliconius* and *Passiflora*. *Science* **1978**, *201*, 745–747.
38. Merrill, R.M.; Wallbank, R.W.; Bull, V.; Salazar, P.C.; Mallet, J.; Stevens, M.; Jiggins, C.D. Disruptive ecological selection on a mating cue. *Proc R Soc Lond Biol* **2012**, *279*, 4907–4913.
39. Kronforst, M.R.; Papa, R. The functional basis of wing patterning in *Heliconius* butterflies: the molecules behind mimicry. *Genetics* **2015**, *200*, 1–19.
40. Flaxman, S.M. https://github.com/flaxmans/bu2s, 2014. [Online; accessed 14-February-2018].
41. Barton, N. Does hybridization influence speciation? *J Evol Biol* **2013**, *26*, 267–269.
42. Felsenstein, J. Skepticism towards Santa Rosalia, or why are there so few kinds of animals? *Evolution* **1981**, *35*, 124–138.
43. Nei, M. Analysis of gene diversity in subdivided populations. *PNAS* **1973**, *70*, 3321–3323.
44. Vuilleumier, S.; Goudet, J.; Perrin, N. Evolution in heterogeneous populations: From migration models to fixation probabilities. *Theor Pop Biol* **2010**, *78*, 250–258.
45. Nosil, P.; Feder, J.L.; Flaxman, S.M.; Gompert, Z. https://media.nature.com/original/nature-assets/natecolevol/2017/s41559-016-0001/extref/s41559-016-0001-s1.pdf, 2017. [Online; accessed 14-February-2018].
46. Nosil, P.; Feder, J.L.; Flaxman, S.M.; Gompert, Z. https://raw.githubusercontent.com/flaxmans/NatureEE2017/master/figures-and-scripts/MigrationSelectionBalance.nb, 2017. [Online; accessed 14-February-2018].
47. Team, R.C. R: A language and environment for statistical computing **2017**.
48. Fischer, B.; Pau, G.; Smith, M. *rhdf5: HDF5 interface to R*, 2017. R package version 2.22.0.
49. http://download.lepbase.org/v4/sequence/Heliconius_melpomene_melpomene_Hmel2.5.scaffolds.fa.gz, 2017. [Online; accessed 14-February-2018].
50. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **2009**, *25*, 1754–60.

51. Li, Y.; Vinckenbosch, N.; Tian, G.; Huerta-Sanchez, E.; Jiang, T.; Jiang, H.; Albrechtsen, A.; Andersen, G.; Cao, H.; Korneliussen, T.; Grarup, N.; Guo, Y.; Hellman, I.; Jin, X.; Li, Q.; Liu, J.; Liu, X.; Sparsø, T.; Tang, M.; Wu, H.; Wu, R.; Yu, C.; Zheng, H.; Astrup, A.; Bolund, L.; Holmkvist, J.; Jørgensen, T.; Kristiansen, K.; Schmitz, O.; Schwartz, T.W.; Zhang, X.; Li, R.; Yang, H.; Wang, J.; Hansen, T.; Pedersen, O.; Nielsen, R.; Wang, J. Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nature Genet* **2010**, *42*, 969–972.
52. McKenna, A.; Hanna, M.; Banks, E.; Sivachenko, A.; Cibulskis, K.; Kernytsky, A.; Garimella, K.; Altshuler, D.; Gabriel, S.; Daly, M.; DePristo, M.a. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **2010**, *20*, 1297–303.
53. Schilling, M.P. https://github.com/schimar/hts_tools, 2018. [Online; accessed 14-February-2018].
54. 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.; others. The variant call format and VCFtools. *Bioinformatics* **2011**, *27*, 2156–2158.
55. Weir, B.S.; Cockerham, C.C. Estimating F-Statistics for the analysis of population structure. *Evolution* **1984**, *38*, 1358–1370.
56. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **2011**, *27*, 2987–93.
57. Wright, S. Variability within and among natural populations. In *Evolution and the Genetics of Populations*; Univ. of Chicago Press: Chicago, 1978.
58. Wickham, H. *ggplot2: Elegant graphics for data analysis*; Springer-Verlag New York, 2009.
59. Feder, J.; Nosil, P.; Gompert, Z.; Flaxman, S.; Schilling, M. Barnacles, barrier loci and the systematic building of species. *J Evol Biol* **2017**, *30*, 1494–1497.
60. Jiggins, C.; Martin, S. Glittering gold and the quest for Isla de Muerta. *J Evol Biol* **2017**, *30*, 1509–1511.
61. Lindtke, D.; Yeaman, S. Identifying the loci of speciation: the challenge beyond genome scans. *J Evol Biol* **2017**, *30*, 1478–1481.
62. Turner, J.; Johnson, M.S.; Eanes, W.F. Contrasted modes of evolution in the same genome: allozymes and adaptive change in *Heliconius*. *PNAS* **1979**, *76*, 1924–1928.
63. Jiggins, C.; McMillan, W.; King, P.; Mallet, J. The maintenance of species differences across a *Heliconius* hybrid zone. *Heredity* **1997**, *79*, 495.
64. Jiggins, C.D.; Davies, N. Genetic evidence for a sibling species of *Heliconius charithonia* (Lepidoptera; Nymphalidae). *Biol J Linnean Soc* **1998**, *64*, 57–67.
65. Mallet, J.; McMillan, W.O.; Jiggins, C.D. Mimicry and warning color at the boundary between races and species. *Endless forms: species and speciation* **1998**, pp. 390–403.
66. Butlin, R.K. Recombination and speciation. *Mol Ecol* **2005**, *14*, 2621–2635.
67. Kirkpatrick, M.; Barton, N. Chromosome inversions, local adaptation and speciation. *Genetics* **2006**, *173*, 419–434.
68. Feder, J.L.; Nosil, P. Chromosomal inversions and species differences: when are genes affecting adaptive divergence and reproductive isolation expected to reside within inversions? *Evolution* **2009**, *63*, 3061–3075.
69. Ortiz-Barrientos, D.; Engelstädter, J.; Rieseberg, L.H. Recombination rate evolution and the origin of species. *Trends Ecol Evol* **2016**, *31*, 226–236.