Demographically explicit scans for barriers to gene flow using gIMble

Dominik R Laetsch[a,†] , Gertjan Bisschop [a,†] , Simon H Martin [a], Simon Aeschbacher [b], Derek Setter [a] and Konrad Lohse [a,*]

a Institute of Evolutionary Biology, University of Edinburgh, EH9 3FL Edinburgh, Scotland
b Department of Evolutionary Biology and Environmental Studies, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland

∗ **Corresponding author: konrad.lohse@ed.ac.uk**
† **these authors contributed equally**

Abstract

Identifying regions of the genome that act as barriers to gene flow between recently diverged taxa has remained challenging given the many evolutionary forces that generate variation in genetic diversity and divergence along the genome and the stochastic nature of this variation. Here we implement a composite likelihood approach for quantifying barriers to gene flow. This analytic framework captures background selection and selection against maladaptive alleles, i.e. genomic barriers, in a model of isolation with migration (IM) as heterogeneity in effective population size (N_e) and effective migration rate (m_e) respectively. Variation in both effective demographic parameters is estimated in sliding windows via pre-computed likelihood grids. We have implemented genomewide IM blockwise likelihood estimation (`gIMble`) as a modular tool, which includes modules for pre-processing/filtering of genomic data and performing parametric bootstraps using coalescent simulations. To demonstrate the new approach, we reanalyse data from a well-studied sister species pair of tropical butterflies with a known history of post-divergence gene flow: *Heliconius melpomene* and *H. cydno*. Our example analysis uncovers both large effect barrier loci (including well known wing pattern genes) and a genome-wide signal of polygenic barrier architecture.

Introduction

How reproductive isolation between species builds up at the genomic level remains a major question in evolutionary biology [1]. Population genomic studies inferring explicit demographic models of speciation for an increasing number of taxa suggest that species divergence in the face of gene flow is common. This poses a conundrum: gene flow homogenizes populations and — in the presence of recombination — leads to the breakdown of co-adapted combinations of alleles, which directly opposes divergent selection [2]. However, divergent selection for locally beneficial alleles, if sufficiently strong, may overcome the influx of maladapted migrant alleles [3, 4]. It has been suggested that such barrier loci of large effect may act as nuclei around which further barrier loci can accumulate [5, 6]. The clustered genetic differentiation that may result from this feedback between migration, local selection and linkage has been referred to using metaphors of "speciation islands" [7] or "genomic islands of divergence" [6, 8]. Despite the fact that we now have access to vast volumes of whole genome sequence (WGS) data, it remains an open question whether the clustered genomic barriers envisaged by these verbal models are important in generating and maintaining new species in nature and, if so, how exactly they arise. More generally, it remains unclear how often species barriers are due to a few loci of large effect and how often barriers have a highly polygenic architecture. Progress on these questions has been stymied by a lack of methods that can extract the relevant information about the population processes involved in the build up of barriers from WGS data. Current speciation-genomic inference approaches can be classed into three broad categories [9] (summary statistics scans, demographic inference, and analyses of the relationship between recombination rate and divergence) which have proven challenging to connect.

1. Summary statistics scans

The search for speciation islands has led to an industry of genomic outlier scans that — initially — readily interpreted regions of increased differentiation between species as "speciation islands" [7, 8, 10, 11]. Outlier scans are generally based on summary statistics, in particular differentiation as measured by F_{ST} [1, 12]. However, despite their conceptual appeal, genome scans in general and F_{ST} scans in particular have been criticized on several grounds. Since F_{ST} is a relative measure of divergence, it is sensitive to any force that locally reduces diversity. Thus, even in the absence of gene flow, F_{ST} may be increased simply as a result of the direct and indirect effects of past positive selection [13, 14] and background selection (BGS) [13, 15] on linked neutral sites [16, 17]. In fact, outlier scans trace back to Lewontin and Krakauer [18] who proposed the variance in F_{ST} as a general test for selection. While a number of other summary statistics have been developed to overcome the problems of F_{ST} [19, 20], the fundamental problem of all scans based on summary statistics is that they are insufficient to distinguish between the key population genetic processes

we wish to learn about [14, 21, 22]. In other words, it is unclear how often one expects outliers in a summary statistic to arise simply by chance due to genetic drift or under scenarios that have nothing to do with the establishment of barriers to gene flow. As a result, distinguishing interesting outliers from noise has remained a major challenge [22].

2. Model based demographic inference

An arguably more successful research program has been to use genome data to reconstruct the demographic history of speciation. A large range of inference methods exist that fit more or less complex models of demographic history to whole-genome data (see [23–26] for reviews). An important starting point for quantifying gene flow during speciation has been the Isolation with Migration (IM) model [27] which assumes a constant rate of gene flow from the time of the divergence until the present. Estimates of gene flow under the IM and related models have been obtained for many recently diverged species and have led to a reappraisal of the role of gene flow as a ubiquitous ingredient of the speciation process, rather than the exception [28, 29].

3. Modelling the relationship between recombination and divergence

A third category of speciation genomic analysis, which has so far only been applied to a small number of taxa, exploits the genome-wide relationship between genetic divergence and recombination to extract information about the selection on and density of barrier loci [9]. In contrast to the positive correlation between genetic diversity and recombination expected under background selection [15] (and globally positive selection), divergent selection against locally maladapted alleles results in a negative correlation between genetic divergence and recombination [30]. This prediction is based on theory showing that selection against deleterious immigrant alleles reduces the effective rate of gene flow at linked neutral sites [2] below the rate m at neutral sites that are not affected by barriers. The extent of this reduction can be captured by the effective migration rate m_e, which decreases with the strength of selection and increases with the map distance from selective targets [2, 31]. Integrating over the unknown genomic positions of barrier loci and conditioning on a recombination map, Aeschbacher et al. [9] developed an approach for jointly estimating the aggregate strength of selection against immigrant alleles in the genome, the divergence time, and the baseline migration rate. However, although this method uses the expected correlation between recombination and genetic divergence under a demographically explicit model, it infers genome-wide parameters and does not allow scanning for barriers to gene flow.

Demographically explicit genome scans

Given the current state of inference approaches, speciation genomic analyses involve am awkward choice: one either infers a demographic history, which — however detailed — is uninformative about which genomic regions may act as barriers, or one visualizes the heterogeneity in differentiation along the genome via summary statistics, but sacrifices the ability to learn about the population processes that have given rise to this heterogeneity.

Ideally, we would like to reconstruct a fully specified model of the speciation process, including both the demographic history (parameterized in terms of genetic drift, divergence, gene flow), as well as all relevant selective forces, i.e. a plausible null model *sensu* [32]. Indeed, most selection is not likely related to the emergence of reproductive barriers [33], and even a very simple cartoon of the speciation process would have to capture both selection on variation that is globally beneficial or deleterious, and local selection that establishes and maintains barriers to gene flow. Unfortunately, it is currently impossible to infer such a detailed model from genome data.

An alternative and simpler approach is to model the selective processes that lead to the establishment of barriers to gene-flow via heterogeneity in effective migration rate m_e. This approach has been pioneered by Roux et al. [34] and has recently been implemented as an approximate Bayesian computation framework [35] that allows fitting models of heterogeneous gene flow to genome-wide data. However, this method estimates variation in m_e (and N_e) via genome-wide distributions and does not allow inference of m_e along the genome. The motivation of the present study is to fill this gap and develop an alternative framework based on likelihood calculation that allows interpreting the heterogeneity of divergence and diversity along the genome through the lens of an explicit, albeit necessarily simple, demographic history. We approximate the effects of background selection (BGS) and selection against migrants as variation in effective population size (N_e) and effective migration rate (m_e), respectively. Our method infers genome-wide histories of isolation with migration (IM) using a block-wise likelihood calculation. The composite likelihood (CL) calculation at the heart of our approach is based on previous analytic work [36–38] and is implemented as part of a modular, open source bioinformatic toolset: genome-wide IM blockwise likelihood estimation `gIMble` (https://github.com/LohseLab/gIMble). `gIMble` allows for an efficient inference of m_e and N_e in windows along the genome as a scan for barriers to gene flow, includes algorithms for the required partitioning, filtering and summarizing of genomic variation, and provides a simulation module for generating parametric bootstraps using the coalescent simulator `msprime` [39].

Heliconius melpomene and *H. cydno* as a test case

Throughout the paper, we describe our approach for demographically explicit scans for barriers to gene flow by way of a re-analysis of data from two sister species of *Heliconius* butterflies: *H.*

melpomene and *H. cydno*. This classic model of speciation research provides an excellent test case for our approach given that the genome-wide heterogeneity in divergence and diversity has been studied extensively in this species pair [40–45] and the genetic basis of key barrier traits (wing patterns and preference for them) is known [41–43, 46]. Despite strong behavioural, ecological, and postzygotic isolation (female hybrid sterility), there is overwhelming evidence for ongoing hybridisation and gene flow between these species that affects a large proportion of the genome [37, 42]. Martin et al. [42] found that F_{ST} between *H. melpomene* and *H. cydno* is highly heterogeneous across the genome, but concluded that this heterogeneity does not provide direct evidence of barriers to introgression. Indeed, F_{ST} between *H. melpomene* and *H. cydno* is similarly heterogeneous and strongly correlated in both sympatric and allopatic comparisons, despite the fact that allopatric populations are thought to experience no contemporary gene flow [42]. Many F_{ST} outliers between *H. melpomene* and *H. cydno* may therefore simply reflect increased rates of within-species coalescence [14]. Using this well studied *Heliconius* data set as a test case, we showcase `gIMble` and address the following questions:

1. What fraction of the genome likely acts as a complete barrier ($m_e = 0$) between these taxa?
2. What is the overlap/correlation between barriers defined via m_e and outliers in terms of F_{ST}?
3. Do m_e barriers include known true positives, i.e. major effect loci controlling phenotypes involved in reproductive isolation?
4. Is there a genome wide relationship between m_e and the rate of recombination, as predicted by models that assume polygenic barriers?

Results

The inference problem

The framework implemented in `gIMble` models the genomic landscape of speciation by constructing scans of two kinds of effective demographic parameters: background selection (BGS) is captured via heterogeneity in N_e, and barriers to gene flow are modeled via heterogeneity in m_e. Unlike previous approaches [9] which rely on the relationship between genetic variation and recombination, our inference is based on sequence variation alone and does not require knowledge of the recombination map.

We focus on an IM history between two species A and B. We assume divergence at a time T (measured in generations) and a constant probability of unidirectional migration, i.e. lineages sampled in species A may trace back to B (backward in time) via migration at rate m_e. Effective

population sizes are allowed to differ between species, i.e. the most complex model assumes three different N_e parameters: N_A, N_B and N_{anc}.

One could in principle co-estimate all five parameters of the IM model (N_A, N_B, N_{anc}, m_e, T) in sliding windows along the genome. However, in practice, this maximally complex model in which all parameters vary freely along the genome is both hard to infer — given that IM parameters are known to have correlated errors [27] — and difficult to interpret. Thus, we conduct a two-step inference (Fig 1): first, we identify the best-fitting *global* IM history which assumes no heterogeneity in demographic parameters. In a second step, *local* variation in the N_e and m_e parameters is estimated in sliding windows along the genome. This inference is conditioned on the T estimate under the global model, i.e. we assume that the onset of species divergence is a global event that is shared by the whole genome (see Discussion).

Implementation

The basic unit of data in the inference implemented in `gIMble` is a block of sequence of a fixed length in which we record sequence variation in a heterospecific pair of individuals (we will refer to this as a "pair-block"). An important property of the algorithm for defining blocks is that pair-blocks are created independently for each pair of individuals. Crucially, we assume that blocks are only indirectly affected by selection at nearby linked sites and otherwise evolve neutrally and under a constant mutation rate. While these assumptions allow us to treat blocks within the same window as statistically exchangeable, they necessitate a careful and consistent filtering strategy for variant and invariant sites. In particular, we have sought to avoid sample size dependent filters (e.g. restricting inference to data that meet filtering criteria in all individuals). To facilitate comparability across datasets, we have implemented a minimal set of standard filters (`preprocess` module, see Methods).

Inference is based on the block-wise site frequency spectrum (the bSFS *sensu* [37]), i.e. the tallies of SFS mutation types in blocks (formally the counts of joint, folded site frequency spectra). Lohse et al. [36] show how to analytically compute the probability of bSFS counts under the IM and related demographic models in the absence of recombination using the generating function (GF) of genealogies. `gIMble` implements this calculation using `agemo` [38] (see Methods) and supports both global and local inference through a number of modules. The `optimize` module uses the bSFS data to find maximum composite likelihood estimates (MCLE) under the IM model using standard numeric optimization (Fig 1, (3)). In contrast, `gridsearch` infers IM parameters in sliding windows over a grid of bSFS probabilities pre-computed using `makegrid` (Fig 1, (4) & (5)). Importantly, `gIMble` also provides a `simulate` module that generates replicate coalescent simulations for quantifying the uncertainty in both local and global parameter estimates through parametric bootstrapping. We provide a more detailed overview of `gIMble` modules in the methods.

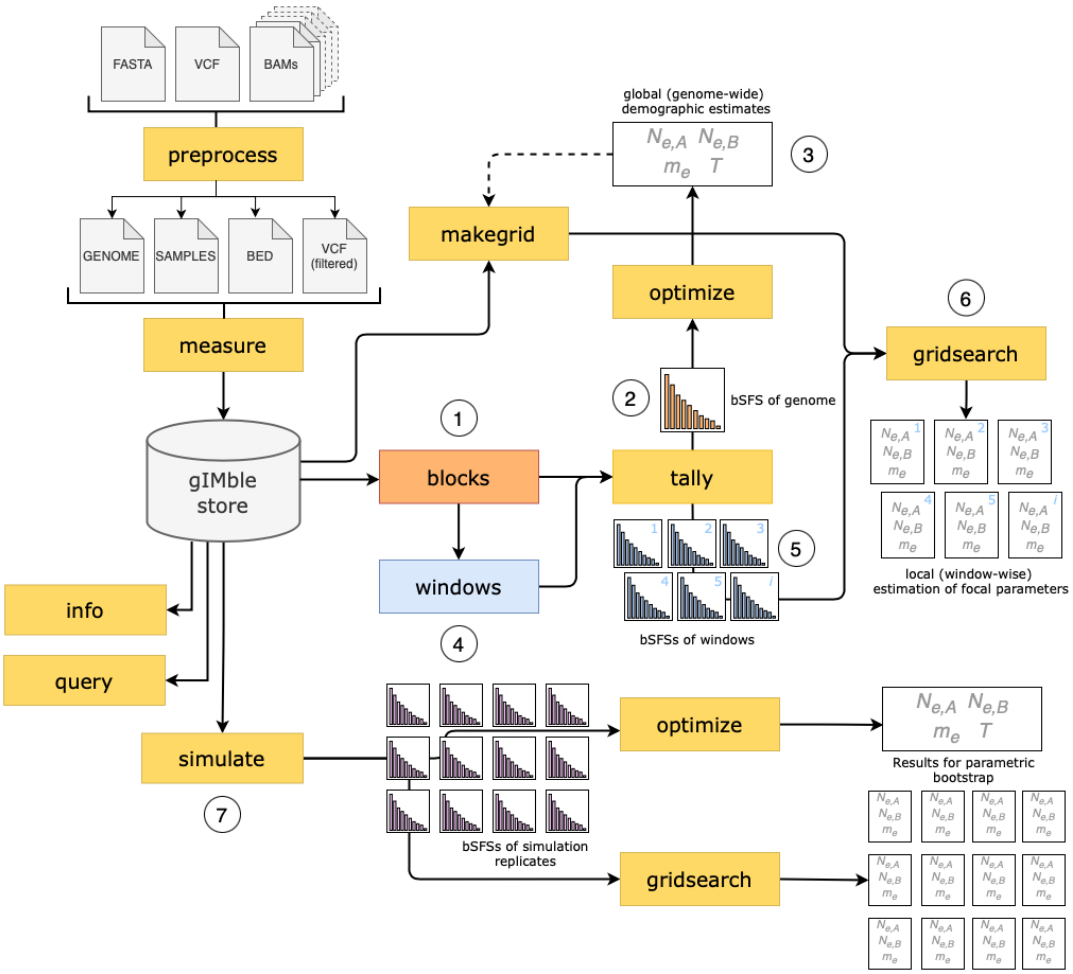


Fig 1. `gIMble` modelling workflow: Blocks for sample pairs (1) are tallied in terms of their mutation configuration (bSFS) (2); global (genome-wide) demographic estimates are obtained using optimization (3). Grouping blocks into windows (4 & 5) allows local (window-wise) estimation of focal parameters (6) by means of a pre-calculated grid. Estimates of uncertainty for both global and local estimates may then be obtained via parametric bootstrap (7).

A butterfly test case

To illustrate `gIMble`, we re-analysed WGS data of 10 wild-caught individuals each of *H. melpomene* and *H. cydno*, sampled from sympatric populations in Panama (of the subspecies *H. melpomene rosina* and *H. cydno chioneous* respectively, Table S1). This dataset that has been analysed extensively using summary statistics [42,47]. To maximise the density of variable sites, we focused all analyses on intergenic sequence, which we define as any sequence without gene or repeat annotation. For the sake of simplicity, we ignore the Z chromosome, which is known to have greatly reduced genetic diversity and gene flow in this and several other *Heliconius* taxa [42,47].

After filtering for coverage and quality, and partitioning the data into pair-blocks of 64 bases (see Methods), the total block-wise dataset comprised an average length of 72 Mb of sequence per heterospecific sample pair. This corresponds roughly to 24% of the *H. melpomene* genome and 69% of all intergenic sequence. Average heterozygosity within each species (H) in intergenic blocks is comparable to the mean pairwise genetic diversity [48] across all sample pairs. However, we find that H is lower for intergenic sites than four-fold degenerate sites (Table 1), which suggests that intergenic sequence blocks — as defined for the purpose of this analysis — are under some level of selective constraint.

We summarised blockwise genetic variation in windows of length 50,000 pair-blocks with a 20% offset between adjacent windows. This window size corresponds to a minimum window span (i.e. pre-filtering length) of 500×64 bp = 32kb of "blocked" sequence assuming complete coverage across all 100 hetero-specific sample pairs (see Methods). However, since we are only considering intergenic sequence that meet coverage and quality filters, the span of windows is considerably greater (median 103kb, Fig. S1). Exploring window-wise variation, we recover two well known properties of this dataset: first, outliers of F_{ST} are mostly due to reduced diversity rather than an increase in d_{xy} (Fig. S2). Second, differentiation along the genome is noisy, with many peaks of F_{ST}, but no large genomic regions of increased F_{ST} (Fig. S3).

Fitting a global background model of divergence and gene flow

Our initial aim was to identify a plausible, yet simple, global background demography for this *Heliconius* sister species pair. We compared support for IM models with gene flow in either direction, and also quantified support for a number of simpler nested histories. Specifically, we fitted IM histories in which the effective size of the ancestral population is equal to that of one of the daughter species (IM_2, $N_{\mathrm{anc}} = N_{\mathrm{mel}}$ or $N_{\mathrm{anc}} = N_{\mathrm{cyd}}$), a strict divergence history (DIV, $m_e = 0$) and models of migration only, i.e. long-term population structure (MIG, $T = \infty$) (see Table S2).

The scenario that best fit the global bSFS is an IM model with three different N_e parameters (N_{mel}, N_{cyd} and N_{anc}), involving gene flow at rate $m_e = 7.38 \times 10^{-7}$ ($M = 4N_{\mathrm{mel}}m_e = 1.61$

Table 1. Average heterozygosity (H) across all individuals within each species and divergence (d_{xy}) between *H. melpomene* and *H. cydno*. Summaries (generated using `gIMble` *info*) are based on 64 base intergenic blocks sampled in heterospecific pairs of individuals (top row). Estimates from a similar analysis for a single pair of individual genomes and longer blocks (150 bases) [37] (second row) and four-fold degenerate (4D) sites (last two rows) are shown for comparison.

data	H_{Hcyd}	H_{Hmel}	d_{xy}	F_{ST}
intergenic 64b	0.0178	0.0171	0.0242	0.167
intergenic [37]	0.0169	0.0150	0.0218	0.155
4D	0.023	0.022		
4D [45]		0.025	0.041	

Table 2. Divergence and gene flow estimates for *H. melpomene* and *H. cydno* inferred using `gIMble` *optimise*. By default `gIMble` parameter estimates are scaled in absolute units, i.e. number of individuals and generations given a user specified μ. We have converted T to millions of years (MYA) assuming four generations per year. The bottom row gives estimates from a previous analysis [37], which was based on a single pair of individual genomes for comparison.

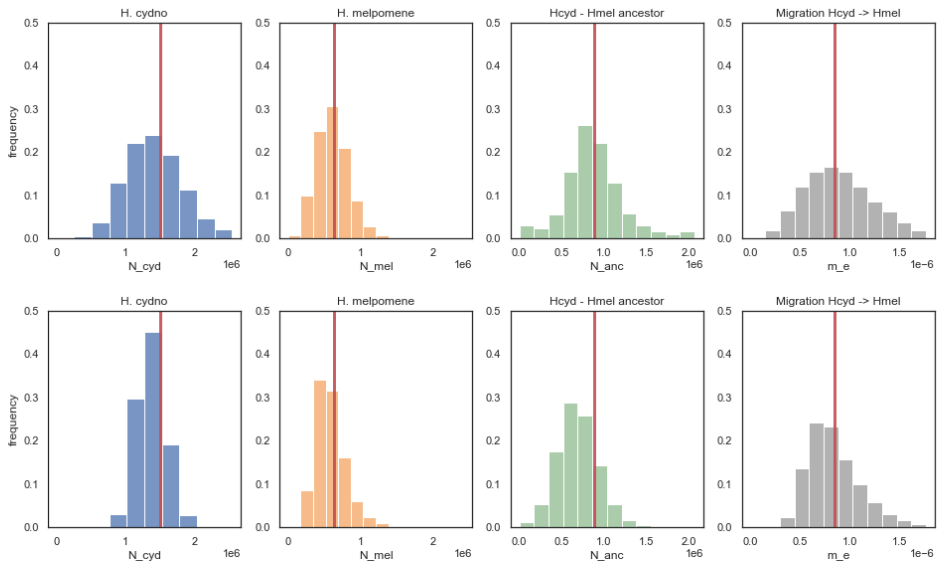
model/dataset	N_{mel}	N_{cyd}	N_{anc}	T	m ($M = 4N_e m$)
$IM_{\rightarrow\mathrm{mel}}$	5.45×10^5	1.41×10^6	9.23×10^5	1.06 MYA	7.38×10^{-7} (1.61)
$IM_{2,\rightarrow\mathrm{mel}}$ [37]	1.10×10^6	2.85×10^6	N_{cyd}	1.04 MYA	3.41×10^{-7} (1.50)

individuals) per generation from *H. cydno* into *H. melpomene* (forwards in time) ($IM_{\rightarrow mel}$ Table, 2). We will refer to this model as "the global model" from here on. Our estimates under the global model are both consistent both with previous inferences on this species pair [37, 41] (Table 2) and the observation that backcrosses of F1 hybrids are only viable in the *H. melpomene* direction in the lab. Assuming four generations per year and the spontaneous mutation rate estimated for *Heliconius* of 2.9×10^{-9} per bp and generation [49], the maximum composite likelihood estimates (MCLE) under the best fitting IM model correspond to effective population sizes (N_e) of roughly 0.5, 1.5 and 1 $\times 10^6$ for *H. melpomene*, *H. cydno* and their shared ancestor, respectively, and a divergence time of approximately 1 MYA (Table 2).

Strong support for heterogeneity in N_e and m_e

We explored variation in N_{mel}, N_{cyd}, N_{anc} and m_e along the *Heliconius* genome by exhaustively searching parameter combinations in a $12 \times 12 \times 12 \times 16$ pre-computed grid (see *makegrid*, Methods). The grid was centred on the MCLEs under the global model (Table 2) and T, the species divergence

Fig 2. The distribution of N_{mel}, N_{cyd}, N_{anc} and m_e from *H. cydno* into *H. melpomene* estimated in sliding windows of 32kb of blocked sequence, for the real data (top row) and bootstrap simulations without m_e and N_e heterogeneity (bottom row). Simulations were conditioned on the *Heliconius* recombination map (see Methods). The red lines indicate estimates under the global model. Note that histogram bins correspond to points in the $12 \times 12 \times 12 \times 16$ parameter grid.



time, was fixed to the global estimate (Table 2).

We found that window-wise (i.e. local) MCLEs of all three N_e parameters varied by a factor of three (Fig 2). Similarly, local estimates of m_e have a wide distribution. Importantly, we found that the heterogeneity in local estimates does not result solely from the stochasticity of the coalescent and mutational processes: applying the same window-wise inference to data simulated under an IM model without heterogeneity (under the estimates for the global model, Table 2) gave much narrower distributions of local estimates (Fig 2, bottom).

We can quantify the overall support for heterogeneity in N_e and m_e by summing the support ($\ln CL$) across windows: e.g. the support for a maximally heterogeneous model that allows for variation in N_{mel}, N_{cyd}, N_{anc} and m_e is the sum of $\ln CL$ associated with all local MCLEs. Allowing for heterogeneity in all three effective parameters greatly improved the overall fit of the $IM_{\rightarrow mel}$ model (Table 3).

Table 3. Model fit measured as $\Delta \ln CL$ relative to the best fitting global model, $IM_{\rightarrow mel,fix}$ (Table 2) which assumes that all effective parameters are fixed.

Model	DIV_{fix}	$MIG_{\rightarrow mel,fix}$	$IM_{\rightarrow mel,fix}$	$IM_{\rightarrow mel,hetN_e}$	$IM_{\rightarrow mel,hetN_e,m_e}$
$\Delta \ln CL$	-5,286,383	-970,922	0	19,672,897	25,526,074

Detecting barriers to gene flow

Estimating $m_{e,i}$ for each window i (shorthand m_i) along the genome immediately allows us to detect genomic barriers as regions with reduced gene flow. One possibility is to interrogate the point estimates of m_i along the genome. However, in quantifying the support for reduced m_e within each window i, we also want to account for the confounding effects of locally varying effective population sizes N_i. To this end, we define Δ_B as the relative support for a model in which m_i is reduced below some threshold m_e^* relative to a model in which m_i is fixed to $\hat{m}_e$, the value that best fits the data from all windows under the global model.

$$\Delta_B = \max_{\substack{m_i \leq m_e^* \in \mathbf{M} \\ N_i \in \mathbf{N}}} \left\{ \ln CL(m_i, \hat{T}, N_i) \right\} - \max_{N_i \in \mathbf{N}} \left\{ \ln CL(\hat{m}_e, \hat{T}, N_i) \right\} \quad (1)$$

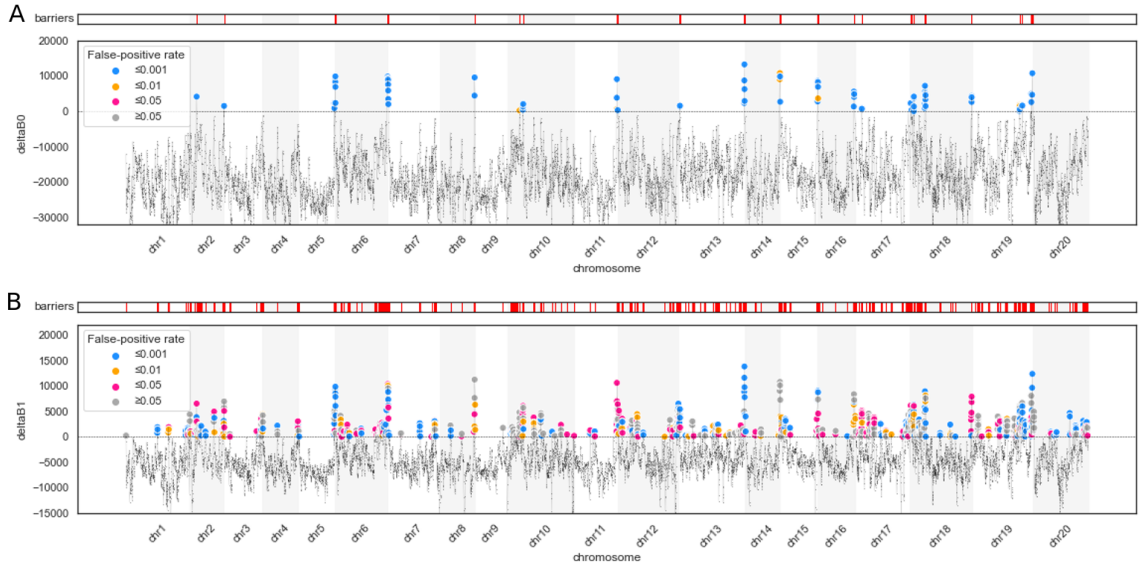
Here, $\hat{m}_e$ and $\hat{T}$ are the best supported parameter estimates globally and $\mathbf{M}$ and $\mathbf{N}$ are the grids of m_e and N_e values used for the window-wise optimization. The first term represents the highest support ($\ln CL$) obtained for window i across all possible population sizes N_i and migration rates m_i below the threshold value m_e^*. The second term gives the support of the best fitting model when the migration rate is fixed to the global estimate $m_i = \hat{m}_e$ and only N_i may vary. We can think of $\Delta_{B,0} = 0$ as defining a sea level for barriers to gene flow, i.e. windows with a positive value have support for reduced gene flow.

Given that `gIMble` estimates m_e across sites that might be direct targets of selection as well as at linked neutral sites, m_i may be > 0 even for a window that contains a complete barrier to gene flow. We therefore explored the *Heliconius* dataset using two m_e thresholds to define barriers, $m_e^* = 0$ or $m_e^* = 0.2\hat{m}_e$. We refer to these analyses as $\Delta_{B,0}$ and $\Delta_{B,0.2}$ scans respectively. Only 0.78% of windows (88 out of 11,218) met the strictest barrier definition $\Delta_{B,0} > 0$. To define barrier regions, we merged all overlapping barrier windows above the "sea-level" of $\Delta_{B,0} = 0$: the $\Delta_{B,0}$ scan gave 25 barrier regions of a total length 4.96Mb (1.9% of autosomal sequence) which range in length from 90kb to 409kb (Fig 3, top). A less stringent barrier definition, $\Delta_{B,0.2} > 0$, identified 7.8% of windows, spanning a total 47.4Mb (18.2% of autosomal sequence), as barriers to gene flow (Fig 3, bottom).

Since Δ_B depends both on the locally best fitting set of $\mathbf{N}$ and the local recombination rate,

we use parametric bootstraps (generated using `gIMble` *simulate*, see Methods) to obtain the false positive rate (FPR) for $\Delta_{B,0}$ and $\Delta_{B,0.2}$ scans (Fig 3). We find that in both cases the average genome-wide FPR is far lower than the observed fraction of barrier windows: 0.065% and 1.7% for $\Delta_{B,0}$ and $\Delta_{B,0.2}$ respectively.

Fig 3. Barriers to gene flow between *H. melpomene* and *H. cydno* inferred using `gIMble`. The m_e threshold used to diagnose barriers is relaxed from top ($\Delta_{B,0}$, A) to bottom ($\Delta_{B,0.2}$, B). Regions above the "sea-level" of $\Delta_B = 0$ fit a history of reduced m_e better than a model assuming the global estimate $\hat{m}_e$. Windows $\Delta_B > 0$ have been coloured to reflect their expected FPR.



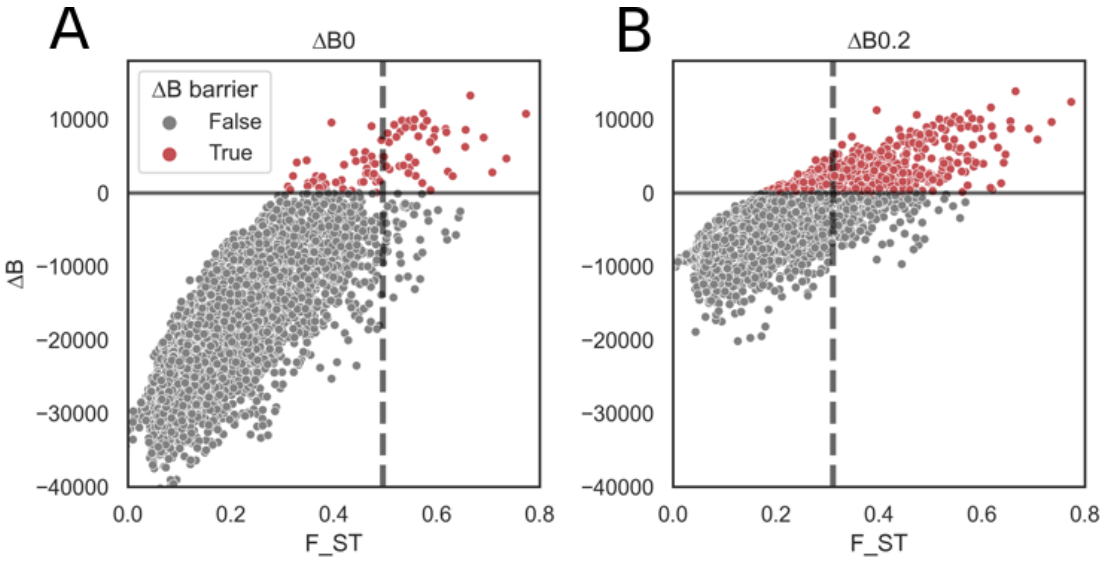
Barriers only partially overlap F_{ST} outliers but coincide with known wing patterning genes

Although F_{ST} and Δ_B are correlated ($\rho = 0.783$, $p < 0.000001$, Fig 4), we find that irrespective of which m_e^* threshold is used to define barriers windows, barriers as identified via Δ_B only partially overlap F_{ST} outliers. For example, only 42 of the 88 $\Delta_{B,0}$ barrier windows are contained in the analogous F_{ST} tail (Fig 4). In other words, the demographically explicit scan identified novel and different genomic regions which are not necessarily outliers in F_{ST}.

To investigate if the demographically explicit scan recovers previously identified barrier loci, we focussed on the three large-effect genes controlling wing-pattern differences between *H. melpomene* and *H. cydno* which have been studied in detail [43,50–52]: *wnt-A*, *cortex* and *optix* on chromosomes 10, 15 and 18 respectively. We find that all three of these known positives are barriers as defined by

$\Delta_{B,0.2}$ and two (*wnt-A* and *optix*) are barriers under the most stringent criterion of $\Delta_{B,0} > 0$ (Fig 5). While *optix* is also contained in the analogous tail of F_{ST} outliers (i.e. the 0.78% of windows with highest F_{ST}), both *wnt-A* and *cortex* are not. Intriguingly, the two $\Delta_{B,0}$ barrier regions downstream of *optix* on chromosome 18 perfectly coincide with a QTL peak of species-specific visual preference behaviours for the red wing pattern controlled by *optix* [40] which includes five candidate genes [53], specifically cis regulatory differences in *Grik2* and *regulcalcin2* (highlighted in orange in Fig 5).

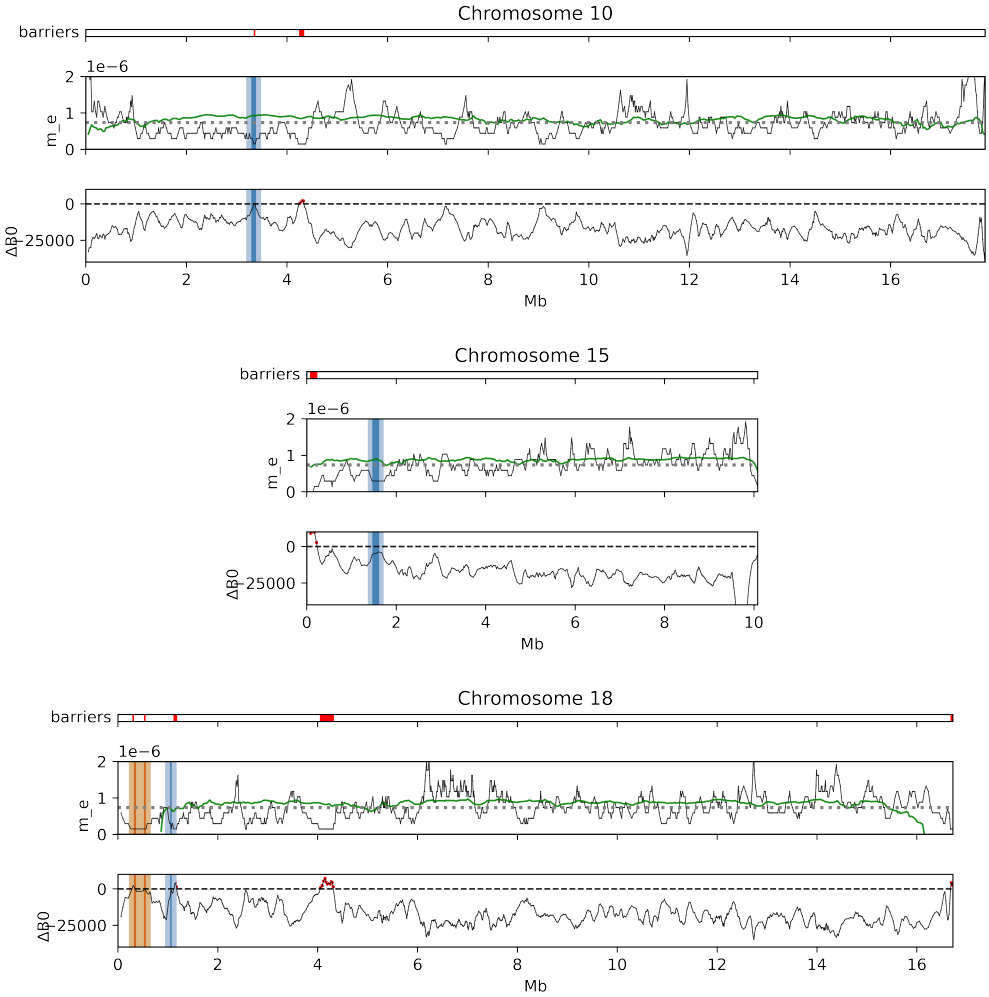
Fig 4. F_{ST} and support for barriers to gene flow between *H. melpomene* and *H. cydno* as measured by Δ_B (eqn. 1) are partially correlated. Each dot represents an autosomal window of effective length 32kb. Two m_e thresholds used to distinguish barriers are shown: $\Delta_{B,0}$ (A) and $\Delta_{B,0.2}$ (B) which identify 0.78% and 7.8% of windows as barriers. The dashed vertical lines delineate the corresponding percentiles in the F_{ST} distribution.



Barriers are concentrated in regions of low recombination and high coding density

Population-genetic models of species barriers predict a reduction of m_e in the genomic vicinity of barrier loci. This reduction scales with the strength of divergent selection at the barrier locus and the map distance to the barrier locus [9, 56, 57]. In regions of low recombination there is thus an increase in both the physical scale at which barrier loci reduce gene flow in flanking neutral regions and in the chance that new barrier loci arise [58, 59]. However, since the effects of all types of

Fig 5. Local point estimates of $m_{e,i}$ (top of each subplot) and $\Delta_{B,0}$ (bottom of each subplot) for the three chromosomes containing large effect wing patterning genes (shown in blue). $\Delta_{B,0}$ identifies *wn-A* [54] on chromosomes 10 (top) and *optix* on chromosome 18 (bottom) as barriers. Given that causal sites may be situated in regulatory regions which in the case of *optix* extend to ≈ 100kb away from the gene [55], we have highlighted the 100kb up and downstream of each gene (lighter blue). $\Delta_{B,0} > 0$ barriers for each chromosomes are shown in red on top. Another $\Delta_{B,0} > 0$ barrier on chromosome 18 coincides with *Grik2* and *regucalcin2* (orange), genes that are associated with male wing pattern preference. The wing pattern gene *cortex* on chromosome 15 (centre) is not a barrier according to $\Delta_{B,0}$. The predicted m_e under a polygenic model is shown in green.



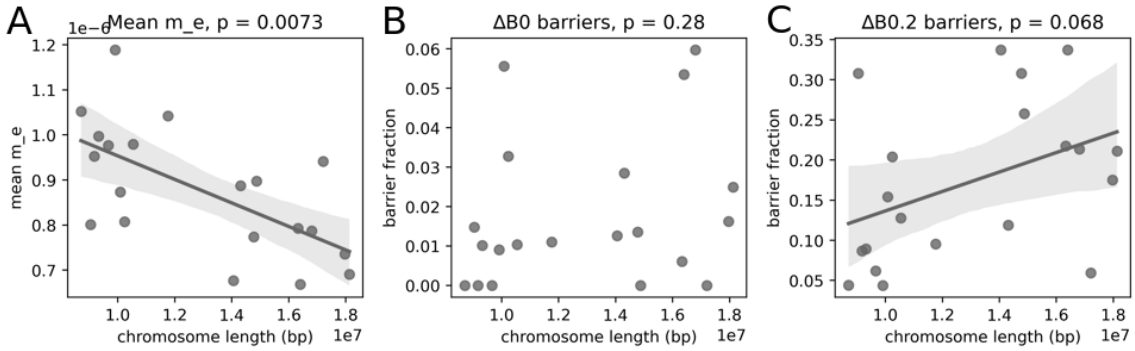
selection on nearby neutral sequence depend on recombination, quantifying the association between barriers and the recombination rate has proven hard using summary statistics that are sensitive to multiple types of selection (but see [60]). In particular, F_{ST} outliers are expected to be concentrated in regions of low recombination simply as a result of the ubiquitous positive correlation between genetic diversity and recombination rate caused by BGS and positive selection [13, 61]. Thus, the negative correlation between F_{ST} and recombination rate we observe in the *Heliconius* species pair examined here (Fig. S4) is not necessarily informative about the nature of species barriers.

In contrast, our framework of estimating effective parameters and defining barriers of m_e via Δ_B allows us to disentangle selection acting on barriers from BGS (and positive selection not contributing to species differentiation). After controlling for heterogeneity in N_e, we found a clear association between local m_e and recombination rate over two physical scales. At an inter-chromosomal scale, we found a strong negative correlation between m_e and physical chromosome length (Fig 6, Pearson's $\rho = -0.66$, $p = 0.0073$). This pattern is expected under polygenic selection against interspecific gene flow because longer chromosomes tend to have lower r_{bp} in *Heliconius* [62]. A negative correlation between chromosome length and introgression – measured in terms of f_d, a summary statistic that draws information from taxon trios – has also been found in a previous study of the same *Heliconius* species pair [60]. Reassuringly, we found estimates of local N_e (both N_{cyd} and N_{mel}) to be even more strongly correlated with chromosome length (Fig. S5) than m_e, which is expected given the general effect of selection on linked neutral sites [63]. The fact that our estimates of both effective gene flow and effective population sizes correlated with recombination in the expected direction suggests that we have successfully decomposed the effects of selection at barrier loci vs. BGS and positive selection.

The fraction of chromosomes inferred to be covered by strong barrier regions (i.e. $\Delta_{B,0} > 0$) varies widely between chromosomes. For instance, chromosomes 1, 4, 9, 14 and 20 harbour no $\Delta_{B,0}$ barrier windows at all, while chromosome 18 appears to have 6% of its physical length covered by strong barriers. However, the fraction of chromosomes covered by barrier regions did not correlate with physical chromosome length (Fig 6, B & C). This is interesting for two reasons. First, under the simplifying assumptions that each barrier region is caused by a single barrier locus and that selection acting on individual barrier loci does not differ among chromosomes, the width of barriers is predicted to scale with $1/r_{\mathrm{bp}}$ [4]. If the rate at which mutations give rise to potential barrier alleles is uniform across chromosomes, we would thus expect longer chromosomes to show a larger physical proportion covered by barriers. Second, theory also predicts that existing barrier loci facilitate the establishment of further barrier loci in close linkage [58, 59]. This clustering effect would be stronger on long chromosomes (lower r_{bp}) compared to short chromosomes (higher r_{bp}), and should thus further contribute to a positive correlation between chromosome length and the fraction of chromosomes covered by barrier regions. The most likely explanation for why we did not observe

this predicted correlation for $\Delta_{B,0} > 0$ barrier regions is a lack of statistical power (i.e. there are only 25 $\Delta_{B,0}$ barrier regions in total). Indeed, when we consider the less-stringent barrier definition (i.e. the 172 barrier regions identified by $\Delta_{B,0.2} > 0$), the expected positive relation between chromosome length and the fraction of chromosomes covered by barriers starts to emerge (Fig 6, C).

Fig 6. A) Average effective migration rates $m_{e,i}$ are negatively correlated with chromosome length. B) and C) There is no significant relationship between the relative length of chromosomes defined as barrier regions and their total length irrespective of whether barriers are defined via $\Delta_{B,0}$ (B) or $\Delta_{B,0.2}$ (C).

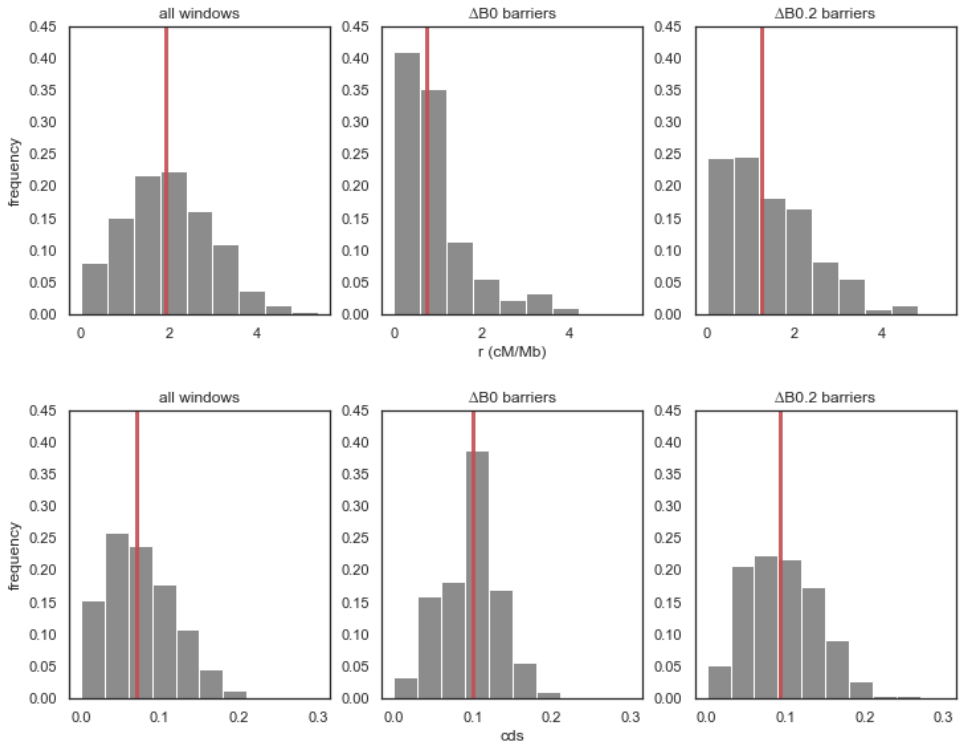


At the second, intra-chromosomal scale, we found that window-wise estimates of m_e are strongly and positively correlated with r_{bp} [62] (estimated directly from lab crosses, see Methods), again as predicted under a polygenic architecture of barriers [64]. Similarly, barriers to gene flow defined either by $\Delta_{B,0}$ or $\Delta_{B,0.2}$ are concentrated in regions of reduced recombination: r_{bp} is significantly lower in barrier windows compared to the average (0.7465 vs 1.9405 cM/Mb respectively) (Fig 7 and Table 4). Assuming that selective targets (both of BGS and selection on barrier loci) reside in or near coding DNA sequence (CDS), we also expected a negative relationship between the density of CDS and local N_e and m_e estimates. We found clear support for both relationships (Fig. S6, e.g. for m_e and CDS measured at a 250kb scale $\rho = -0.164$, $p < 0.000001$). Thus, overall, and in agreement with a model of polygenic selection against inter-specific gene flow, we found that barriers to gene flow are concentrated in regions of low recombination and high CDS density.

How many barriers are there?

Aeschbacher et al. [65] proposed a framework to quantify the strength of selection against introgression under a simple model in which barrier loci are uniformly distributed across the genome with density

Fig 7. Top: genomic windows diagnosed as barriers to gene flow by $\Delta_{B,0}$ (centre) and $\Delta_{B,0.2}$ (right) have lower than average (left) recombination rates per base pair (r_{bp}). Median r_{bp} for each type of window is shown in red. Bottom: $\Delta_{B,0}$ and $\Delta_{B,0.2}$ barriers are enriched for coding sequence (CDS).



ν per base pair. Under this model, the expected local reduction in m_e at a genomic window i relative to the background rate of gene flow (i.e. $m_{e,i}/m$) is a function solely of the per-base pair recombination rate (r_{bp}) and the "selection density" $\sigma = \nu\ s$, where s is the selection coefficient at each barrier locus [65, eqn. 1.9 in Appendix 1]. We used a version of this model in which the reduction in gene flow at the focal genomic window depends only on the closest flanking barrier locus on each side of that window. To account for variation in CDS as a proxy for the density of barrier loci, we modified this model by scaling ν by the inverse of the CDS density. Thus, we could infer σ by minimising the sum of squared differences between the expected local m_e (given r_{bp} and CDS) and the m_e estimates inferred by `gIMble` across all genomic windows. Using r_{bp} estimated at a scale of 1Mb around each window (see Methods) and CDS measured at a 250kb scale, we obtained estimates of $\hat{\sigma} \approx 2.77 \times 10^{-9}$ (95% CI: $\left[2.44 \times 10^{-9},\ 3.09 \times 10^{-9}\right]$) and of the background migration

Table 4. Average properties and N_e estimates for genomic barrier windows defined via Δ_B for *H. melpomene* and *H. cydno*.

Partition	# of windows	# of regions	Total (Mb)	%CDS	$r(cM/Mb)$	N_{cyd}	N_{mel}
$\Delta_{B,0}$	88	25	4.9	0.0958	0.904	1,030,000	430,000
$\Delta_{B,0.2}$	873	172	47.4	0.0956	1.433	1,285,000	515,000
autosomes	11,218	n/a	259	0.0765	2.011	1,506,000	630,000

$\hat{m} \approx 9.64 \times 10^{-7}$.

The compound nature of σ means that similar values can be produced by a small number of barrier loci under strong selection and a large number of barrier loci under weaker selection. One question we can ask is what strength of selection s per barrier locus would be required to produce our estimate of σ, assuming the signal was caused only by the 25 strong barrier regions inferred ($\Delta_{B,0}$). For the sake of simplicity, we assume that each barrier locus has experienced the same selection, and that each barrier region contains exactly one barrier locus. Assuming a total length of 259 Mb for *H. melpomene* autosomes, we find $\hat{s} = \hat{\sigma}/(25/[259 \times 10^6]) = 0.0286$, which would imply very strong selection at each individual barrier locus. However, the estimate of σ is clearly driven by the genome-wide relationship between m_e and r_{bp} (and CDS) – excluding $\Delta_{B,0}$ outliers has little effect on our estimate of σ, and so there must be many more barrier loci than regions of $\Delta_{B,0}$. We can obtain an upper bound for the number of barrier loci if we consider that $2N_e s > 1$ for each barrier locus for selection to be stronger than genetic drift. At the same time, selection at each barrier locus has to be stronger than the background migration rate to some approximation if the barrier is to be maintained (i.e. $s > m$) [3]. Given our estimates of $\hat{N}_{mel} \approx 430,000$ (Table 4) and $\hat{m} \approx 9.64 \times 10^{-7}$, the former threshold ($2N_e s > 1$) is the limiting one, and the maximum number of barrier loci implied by $\hat{\sigma} = 2.77 \times 10^{-9}$ is $\approx 600,000$. Thus, given this upper limit for the number of barrier loci, we cannot rule out that the overall species barrier between *H. melpomene* and *H. cydno* has an omnigenic architecture. In other words, while a moderate number of strong barriers – including those associated with the candidate genes *optix*, *regucalcin*, and *wnt-A* – appear to locally reduce m_e to a large extent, a much larger number of weaker barrier loci are likely contributing to reproductive isolation in this study system.

Discussion

Progress in speciation genomics has been impeded by a conceptual and methodological divide between approaches that model speciation history under neutral scenarios, and approaches aiming to identify

loci under divergent selection that act as barriers to gene flow. Our approach bridges this divide by using a single framework to model both the overall species history and the genomic landscape of speciation. Unlike previous approaches that describe species barriers using summary statistics like F_{ST} that tend to be affected by all forms of selection, the effective demographic parameters inferred by `gIMble` have straightforward interpretations in terms of evolutionary processes: while heterogeneity in N_e largely reflects BGS, $m_{e,i}$ measures the long term, combined effects of barriers to gene flow at each position in the genome [56]. The ability of our framework to identify barriers to gene flow based on estimates of m_e holds promise for comparative studies of species barriers [31] and for the joint inference of both the position and strength of barriers.

The genetic architecture of barriers

The case study of *Heliconius* data we performed here illustrates the use of `gIMble` and highlights the potential of approaches that quantify variation in m_e to enhance future speciation genomic analyses. Specifically, the output of `gIMble` can be used to answer two different but complementary sets of questions about the speciation process. On the one hand, one may assume that reproductive isolation is conferred at least in parts by a small to moderate number of large-effect barrier loci [1] that can be identified individually from their strong local effect on m_e. Because barriers inferred via `gIMble` are necessarily defined relative to a background m_e (and the global parameter T), a `gIMble` scan based on $\Delta_{B,0}$ yields a meaningful estimate of the fraction of genome in this category. We note that this estimate does not involve any arbitrary thresholds, unlike analyses of outliers defined in terms of summary statistics such as F_{ST}. However, the power to estimate barriers to gene flow is a function of the background demography (e.g. the higher the background m_e, the greater to power to identify $\Delta_{B,0}$ barriers), and of the recombination rate. Importantly, this power can be quantified readily using `gIMble` *simulate.*

On the other hand, reproductive isolation may be largely if not entirely polygenic, with a large number of weak barrier loci cooperating to reduce gene flow genome-wide (ref.). In this case, it may be impossible to identify individual barriers [66]. Instead, the main objective when analysing variation in m_e will be to understand the genomic predictors/correlates of m_e. One advantage of `gIMble` is that the inference of point estimates does not rely on recombination information (we assume blocks are non-recombining and exchangeable) Thus, the interaction between recombination rate and m_e can be modelled in a second step, and exploited to make aggregate inference on the nature of the polygenic species barrier. There are of course many other data types one may want to intersect with m_e variation, e.g. gene expression data or information about barriers to gene flow from crossing experiments or hybrid zone studies. We have provided an example of such a second-step inference using the framework of [9], extended here to account for variation in CDS density as a

proxy for the density of barrier loci.

Combining these two approaches, we found that *H. melpomene* and *H. cyndo* are separated by a mixture or both large effect and polygenic barriers. Three out of four known large effect loci controlling barrier phenotypes in this *Heliconius* species pair are contained in the 2% of the genome identified as a barrier by the most strict definition of no effective migration, $\Delta_{B,0}$. Yet at the same time, $\Delta_{B,0}$ barriers are not the main drivers of the genome-wide correlation we find between m_e and r_{bp}. The strength of the correlation is little affected by the exclusion of $\Delta_{B,0}$ barriers (or indeed restricting the analysis to the five autosomes devoid of $\Delta_{B,0}$ barriers). Simple scaling arguments lead us to conclude that, apart from the major barrier loci identified, up to hundreds of thousands of weak barrier loci scattered across the genome might contribute to the overal reproductive isolation between *H. melpomene* and *H. cyndo*. Comparisons of model-based genome scans across taxa are needed to determine whether this combination of a small number of strong barriers and a genome-wide, highly polygenic signal of barrier architecture is ubiquitous.

Limitations and future extensions

It is important to reiterate that modelling the effect of barriers via demographic parameters is a simplification. While the approximation implemented in `gIMble` does capture the effect of selection *maintaining* barriers once they have been established, it does not capture the processes by which alleles at barrier loci (and locally beneficial alleles more generally) become established. While Bateson-Dobzhansky-Muller incompatibilities can arise through drift alone only in the absence of migration [67], their establishment in the face of gene flow involves selection favouring locally adaptive alleles. Thus, we believe that developing methods that capture the joint effect of selection establishing barriers and the reduction in m_e that results remains an important goal of future work.

While `gIMble` is focused on minimal subsamples of pairs of diploid genomes, its modular design facilitates the implementation of generalisations to larger samples and more complex demographic models. There are several obvious and useful avenues for future extensions. First, `gIMble` assumes a constant rate of unidirectional gene flow until the present. Since many taxon pairs that are of interest in speciation research are likely to have completed speciation (i.e. there is no contemporary gene flow), it would be useful to generalise the method to allow for a time at which gene flow ceases — the Isolation with initial migration (IIM) model [68] — as well as bi-directional gene flow. While a feasible approximate solution to the problem of bi-directional gene flow has been described [69], we currently do not know of a closed-form analytic solution to accounts for the multiple discrete demographic events in the IIM model as required by `agemo` [38] (see Methods). This limitation applies equally to other useful extensions of model space involving multiple discrete events, such as inference under models of instantaneous admixture or the inclusion of explicit selective events [70].

Second, `gIMble` currently does not make use of phased sequence data. While this makes it useful for studies where phasing is not achievable (e.g. when short-read sequencing is used and sample sizes are too small to allow statistical phasing), extension of `gIMble` to phased data would increase its statistical power. In practice, this would involve extending both the modules that summarise the data into block-wise mutational configurations and the module computing likelihoods to the case of fully labelled samples.

A further and more profound extension would be to incorporate haplotype information in the "blocking" algorithm. Currently, blocks of fixed length are defined by blind discretisation of the genome. Given that blocks are assumed not to have experienced recombination, they should ideally be defined via topologically informative sites (e.g. using the four-gamete-test and other topology information) to ensure that each block is compatible with the assumption of a single genealogy [71]. Recently, progress has been made on approaches that infer ancestral recombination graphs (ARGs) from phased data. In particular, several minimal assumption heuristics exist to infer topology-only ARGs [72–74]. These methods use information from neighbouring haplotypes to reconstruct marginal tree topologies and would allow blocks to be defined by the spans of marginal trees. Clearly, sampling blocks at random (and assuming no recombination within blocks), as we have done here, sacrifices information. Thus, developing sampling strategies and likelihood calculations that incorporate topology and haplotype information will undoubtedly be a priority of future work.

Conclusion

In conclusion `gIMble` represents a first step towards the goal of model-based characterisation of species barriers. There is extensive room for further development in this area, both in the complexity of the models and the optimal exploitation of genomic data. Given that much of the past progress in speciation research has come from comparative analyses [28,75], it is exciting that even the relatively simple model-based scans we have implemented here open the door to truly comparative studies of species barriers.

Materials and Methods

Implementation

The `gIMble` analysis workflow is designed for WGS data in which all samples have been aligned to the a single reference and requires minimal deviation from established bioinformatic pipelines and file formats. Figure 1 shows the basic workflow we have used to analyse the *Heliconius* data. Currently, `gIMble` only supports analyses of data from samples belonging to two populations/species.

The different steps of the workflow are implemented as distinct modules within `gIMble` allowing users to string together analysis steps in a way and order they deem fit for a particular dataset and research question. The modules are:

- `preprocess` applies `gIMble`-specific quality filters to the input files, thereby assuring consistency with the assumptions underpinning the analytical `gIMble` framework
- `measure` reads the pre-processed input files into a compressed (ZARR) datastore
- `blocks` partitions genomic variation data into pair-blocks based on population membership of samples
- `windows` creates (sliding) windows of pair-blocks along the genome
- `info` displays information for the output of `blocks` and `windows` data
- `tally` prepares `blocks`/`windows` data for modules `optimize` and `gridsearch`
- `optimize` performs parameter optimization given a tally and a bounded parameter space for a specified model
- `makegrid` pre-computes and stores probabilities for a defined parameter space under a specified model
- `gridsearch` computes and stores likelihoods of `tally` data given a pre-computed `makegrid`
- `query` writes output files describing `tally`, `optimize`, and `gridsearch` data
- `simulate` creates tallies through `msprime` [39] simulations

Below, we describe the pre-processing of data, the organisation of `gIMble` analyses into a `zarr` datastore, the modules that allow the fitting of global and local models, and the parametric bootstrapping MCLEs.

Data processing

`gIMble preprocess`

The `gIMble preprocess` module allows for a simple and consistent filtering of input files for subsequent analyses. The purpose of this module is to standardise the quality of data analysed by gIMble using a minimal set of filters that are applicable to both variants and invariant sites while maximizing the proportion of the data retained for analysis.

The required input for `gIMble preprocess` consists of the reference genome (FASTA format), read sets of each sample aligned to the reference (BAM format), and the called variants (VCF format).

The output is composed of the following four files:

- **genome file** recording the name and length of each sequence in the FASTA file.
- **sample file** specifying the name of each sample in the VCF file.
- **VCF file** containing those variants that passed the filter criteria: only SNPs with a minimum distance from non-SNP variants, adequate read depth (based on min/max thresholds inferred from BAM files), minimum genotype quality, and minimum read mapping (SAF, SAR, RPR, RPL) parameters are retained.
- **BED file** describing the "callable" genomic regions for which each sample BAM has adequate coverage (based on min/max thresholds). All variants in the VCF which failed filtering are subtracted from these regions.

The genome and sample files give the user control over what data are processed in subsequent analysis steps of `gIMble`: the genome file delimits the parts of the reference to analyse and the sample file specifies the samples (and their population membership) from which block-wise data will be tabulated.

The VCF and BED files generated by `gIMble preprocess` describe the genomic variation across samples satisfying the defined quality thresholds and can therefore serve as a starting point for any complementary or alternative analyses the user may wish to run on the same data.

`gIMble measure`

This module parses the four input files generated by `gIMble preprocess` (see above). Based on the parsed data, a compressed `zarr` datastore is created which is central to all subsequent analyses.

`gIMble blocks`

The output of this step determines which parts of the sampled sequences are available for analysis. The module `blocks` uses the "callable" regions specified in the BED file and the variants contained in the VCF file to define blocks of contiguous sequence containing a fixed number of callable sites. Defined this way, blocks must also span un-callable regions and therefore have variable physical span in the genome. To maximise the genomic regions sampled, blocks are constructed independently for each sample pair (hence our term pair-blocks). This ameliorates the asymmetry in coverage profiles among the samples due to stochastic variation in sequencing depth between samples and/or reference

bias. If, instead, blocks were conditioned on overall coverage (and other filtering criteria) across all samples, blocking success would be dominated by poorly covered samples and the "block-able" fraction of genome would be increasingly biased/reduced with sample size.

Blocks are constructed for all pairwise sample combinations between the two populations (X) and within each population (A and B). However, only pair-blocks between the two populations (X) are used in the inference steps. The blocking of genomic data is controlled by the parameters `--block_length` — the number of callable sites in each block — and `--block_span` — the maximum distance between the first and last site in a block. While `gIMble` only considers biallelic genotypes in subsequent steps, the block-cutting algorithm allows for multiallelic and missing genotypes, which are controlled with the parameters `--max_multiallelic`, and `--max_missing`. It should be noted that only one set of blocks can be stored in a `gIMble` datastore at a given time.

`gIMble` windows

Windows are constructed by traversing each sequence of the reference from start to end, incorporating the heterospecific pair-blocks (X) as they appear based on their start positions. The parameter `--blocks` controls how many blocks are incorporated into each window and the parameter `--steps` by how many blocks the next window is shifted. Since coverage differences or mapping biases between samples can cause uneven distributions of blocks along the genome, there is no guarantee that blocks from all heterospecific pairs are equally represented within a window. However, by choosing a large enough value for `--blocks` this problem can be mitigated. Analogous to blocks, only one set of windows can be stored in a `gIMble` datastore at a given time.

`gIMble info`

Information about the data contained in a `gIMble` datastore can be summarised by running the `gIMble info` module. Once blocks have been constructed, `gIMble info` may be used to summarise variation via standard population genetic summary statistics (π, d_{xy} and H mean heteozygosity) based on blocks of samples between populations (X) and within populations (A and B).

`gIMble` query

Data contained in the `gIMble` datastore can be queried using this module and written into different BED and TSV files. Currently, this functionality is supported by the data generated from the modules `windows`, `tally`, `optimize`, and `gridsearch`.

`gIMble` tally

Having created blocks (with `gIMble blocks`) and windows (with `gIMble windows`), the module `gIMble tally` generates the bSFSs for heterospecific sample pairs, i.e. a single sample from each population/species A and B. The bSFS [69] is a tally of the mutation configurations of blocks which are themselves described by vectors of the form $\underline{k}_i$, which count the mutation types found within a block.

- m_1 sites with homozygous genotype in A, heterozygous genotype in B
- m_2 sites with heterozygous genotype in A, homozygous genotype in B
- m_3 sites with shared heterozygous genotype in both populations
- m_4 sites with distinct homozygous genotype in both populations

The argument `--maxk` limits the counts of mutations of each type in the bSFS tally. Counts of mutations inside blocks (m_1, m_2, m_3, m_4) exceeding the values in `--maxk` (default: $2, 2, 2, 2$) are combined. Thus the choice of `--maxk` involves a trade-off between capturing the probabilities of rare mutation configurations fully and computational complexity. The bSFS tallies of heterospecific data in blocks can be generated for all blocks (`-d blocks`), per window (`-d windows`), or for all blocks that are included in windows (`-d windowsum`).

Calculating likelihoods

The likelihood under the coalescent with arbitrary demographic history can be calculated for block-wise data using the method described in [36, 37]. In this framework, the generating function (GF) for the distribution of genealogical branch lengths is described recursively as the convolution of independent Laplace-transformed exponential random variables. The GF for a particular model and sampling configuration can be represented most simply as a directed graph involving all possible ancestral states of the sample. [37] have previously shown that the probability of any bSFS configuration $\underline{k}_i$ can be calculated as a partial derivative of the GF.

Previous automation of this likelihood calculation relied on rather inefficient use of *Mathematica*'s [76] symbolic algebra functionality. `gIMble` now uses `agemo`, an open source and user-friendly re-implementation of the GF approach which extracts the probabilities of bSFS configurations efficiently from the GF by means of a graph traversal algorithm [38].

Since both the number of possible bSFS configurations and the coalescent state space graph grow exponentially with sample size n, computing full probabilities becomes prohibitive for $n = n_A + n_B > 6$. `gIMble` accommodates large samples by considering the composite likelihood (CL)

across all $n_A \times n_B$ heterospecific sample pairs (each pair j consists of single diploid individual each from A and B). Using $n_{i,j}$ as shorthand notation for the number blocks in sample pair j with bSFS configuration $\underline{k}_i$, the joint composite likelihood is given by:

$$\ln CL(\Theta) = \sum_i \sum_{j=1}^{n_A \times n_B} n_{i,j} \ln p(\underline{k}_i|\Theta) \tag{2}$$

where Θ is the vector of model parameters. The composite likelihood can be calculated dynamically during an optimization run for a given data set (see `optimize`). However, in some cases it might be more efficient to cast a grid over a specific part of parameter space and pre-compute $\ln p(\underline{k}_i|\Theta_h)$ for all $\underline{k}_i$ and each possible possible parameter combination h in the grid (see `makegrid`). Composite likelihoods can then be computed efficiently for `window` tallies using `gridsearch`.

`gIMble optimize`

This module optimizes model parameters for a given data tally using the Nelder-Mead simplex algorithm [77] with bounded constraints, as implemented in `nlopt` [78]. Given a set of bounds as specified in the config file, optimization can be initiated either at the midpoint of the bounded parameter space or at a random start point.

`gIMble makegrid` and `gridsearch`

Pre-computing the probabilities of bSFS configurations in a grid across parameter space is computationally efficient (relative to using an optimization algorithm) and therefore useful when we wish to interrogate data in replicate (across windows or simulation replicates). Grids can be pre-computed using `makegrid`, are saved in the `zarr` store, and can be used to interrogate any dataset contained within the same datastore using `gridsearch`. Grid searches can be used either for an initial exploration of the likelihood surface (i.e. prior to defining the parameter space across which to run *optimize*), or to fit models to the window-wise data.

Parametric bootstrap

Maximizing $\ln CL$ across blocks and subsamples ignores both linkage between blocks and the pseudo-replication inherent in subsampling heterospecific pairs. To facilitate estimating uncertainty in demographic parameter estimates, critical values for $\Delta \ln CL$ (when comparing models) and measuring power of Δ_B scans, we implemented a wrapper for `msprime` [39].

`gIMble simulate`

The module simulates tallies given an IM type model and a set of parameter values. Each independent simulation replicate results in a bSFS that is saved to the `zarr` store with the necessary metadata. Simulated data can be analysed in the same way as real data. This allows users to perform parametric bootstraps on both global (genome-wide) and local (window-wise) estimates and to quantify the power and potential biases (due to model misspecification) of `gIMble` analyses, in particular the assumption of no recombination within blocks. Parametric bootstraps on local estimates are set up by providing the key of the corresponding gridded analysis. Our example analyses of the *Heliconius* data include simulation based bootstraps and power analyses (see below).

The module *simulate* relies on `demes` [79] for unequivocal model specification. `gIMble` can also output the results of `optimize` and `gridsearch` as a `demes`-readable yaml-file, allowing users to connect `gIMble` to the growing list of other population genetics tools [39, 80–84] supporting the `demes` format.

Power analysis

While the choice of block length involves a trade-off between power and bias (discussed above), the definition of windows trades off power and and resolution. While our analysis scanned the *Heliconius* genomes for barriers using windows on the scale of 100 kb (i.e. $50,000 \times 64$ base blocks), we find that given the *Heliconius* background demography, there is high power to detect genomic barriers for a wide range of window sizes (Fig. S7).

We quantified the impact of window size, recombination rate and sample size on the power to detect barriers using *simulate* and two extreme barrier definitions: the complete absence of gene flow ($m_e^* = 0$, $\Delta_{B,0}$), or a very relaxed definition of a reduction in m_e by half: $m_e^* < \hat{m}_e/2$. Note that the latter is even more permissive than the weaker barrier definition we used to scan the *Heliconius* genomes ($\Delta_{B,0.2}$). When fitting window-wise models, the split time T was fixed to the background value while N_e parameters were allowed to vary freely. We assumed a per base and generation recombination rate

Sequence data

Raw reads (Illumina 150 base paired-end) (Table S1) were downloaded from ENA and quality trimmed using `fastp 0.20.1` [85] (`--length_required 50`). Trimmed reads of each read set were aligned against the chromosome-level assembly of *H. melpomene* Hmel2.5 [62] using `bwa mem 0.7.17-r1188` [86]. Alignment files were sorted with `samtools 1.12` [87] and duplicated read pairs were marked with `sambamba markdup 0.8.1` [88]. Variants were called using `freebayes v1.3.3` [89] (`-k -w -j -T 0.02 -E -1 --report-genotype-likelihood-max`).

Data processing

For the *Heliconius* dataset the following `gIMble preprocess` parameters were specified: `--min_qual 1`, `--snpgap 2`, `--min_depth 8` (i.e. 8 reads), and `--max_depth 3` (i.e. 3 * mean coverage of each BAM file). Blocks were cut using the following parameters: `--block_length 64`, `--block_span 128`, `--max_multiallelic 3`, and `--max_missing 3`. These blocks were grouped into windows using `gIMble windows` with these parameters: `--blocks 50000` (on average, each of the 100 heterospecific sample pairs should occur in $50,000/100 = 500$ blocks within a window) and `--steps 10000`. Tallies were made using `gIMble tally` and `--maxk 2,2,2,2` for 'blocks', 'windows', and 'windowsum' (all blocks in windows).

Recombination map and parametric bootstrap

For each data window, we obtained a recombination rate r_{bp} for each window from the total number of crossover events observed in *H. melpomene* crosses by Davey et al [62] in a 1Mb region (centred on the window's mid-point) using a custom script `getrecombinationrate.py`.

To quantify variation in local parameter estimates expected in the absence of heterogeneity in m_e and N_e, we used *simulate* to simulate window-wise data under the global model and the local r_{bp}. We applied *gridsearch* to these simulated data analogous to the real analysis.

Acknowledgments

We thank Christelle Fraisse and Stuart Baird for discussions and for critical comments on the manuscript. We thank Alex Mackintosh, Noora Poikela and Sam Ebdon for many helpful discussions and for testing various pilot versions of `gIMble` and uncovering countless bugs. We also thank Mark Ravinet and the organisers of the SMBE satellite meeting on speciation genomics (Tjärnö) for inviting us to present a pilot version of this method. This work (including DRL, GB, DS, and KL) was supported by an ERC starting grant (ModelGenomLand, 757648). SM was supported by the Royal Society URF/R1/180682. KL was also supported by a Natural Environmental Research Council (NERC) UK Independent Research fellowship (NE/L011522/1).

References

1. Ravinet M, Faria R, Butlin RK, Galindo J, Bierne N, Rafajlović M, et al. Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. Journal of Evolutionary Biology. 2017;30(8):1450–1477.

2. Barton NH, Bengtsson BO. The barrier to genetic exchange between hybridising populations. Heredity. 1986;57(3):357–376.

3. Haldane JBS. A mathematical theory of natural and artificial selection. (Part VI, Isolation.). Mathematical Proceedings of the Cambridge Philosophical Society. 1930;26(2):220–230.

4. Barton NH. Multilocus Clines. Evolution. 1983;37(3):454–471. Available from: http://www.jstor.org/stable/2408260.

5. Wu CI. The genic view of the process of speciation. Journal of Evolutionary Biology. 2001 dec;14(6):851–865. Available from: http://doi.wiley.com/10.1046/j.1420-9101.2001.00335.x.

6. Feder JL, Nosil P. The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. Evolution. 2010 jun;64(6):1729–1747. Available from: http://doi.wiley.com/10.1111/j.1558-5646.2009.00943.x.

7. Turner TL, Hahn MW, Nuzhdin SV. Genomic islands of speciation in Anopheles gambiae. PLoS Biology. 2005 aug;3(9):1572–1578. Available from: https://dx.plos.org/10.1371/journal.pbio.0030285.

8. Nadeau NJ, Whibley A, Jones RT, Davey JW, Dasmahapatra KK, Baxter SW, et al. Genomic islands of divergence in hybridizing Heliconius butterflies identified by large-scale targeted sequencing. Philosophical Transactions of the Royal Society B: Biological Sciences. 2012 feb;367(1587):343–353. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22201164http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3233711http://rstb.royalsocietypublishing.org/lookup/doi/10.1098/rstb.2011.0198.

9. Aeschbacher S, Selby JP, Willis JH, Coop G. Population-genomic inference of the strength and timing of selection against gene flow. Proceedings of the National Academy of Sciences. 2017;114(27):7061–7066. Available from: http://www.pnas.org/lookup/doi/10.1073/pnas.1616755114.

10. Malinsky M, Challis RJ, Tyers AM, Schiffels S, Terai Y, Ngatunga BP, et al. Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake. Science. 2015 dec;350(6267):1493–1498. Available from: http://science.sciencemag.org/content/350/6267/1493.

11. Ellegren H. Genome sequencing and population genomics in non-model organisms. Trends in Ecology & Evolution. 2014;29(1):51 – 63.

12. Wolf JBW, Ellegren H. Making sense of genomic islands of differentiation in light of speciation. Nature Reviews Genetics. 2017;18(2):87–100. Available from: http://dx.doi.org/10.1038/nrg.2016.133.

13. Charlesworth B. Measures of divergence between populations and the effect of forces that reduce variability. Molecular Biology and Evolution. 1998 may;15(5):538–543. Available from: https://academic.oup.com/mbe/article-lookup/doi/10.1093/oxfordjournals.molbev.a025953.

14. Cruickshank TE, Hahn MW. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. Molecular Ecology. 2014;23(13):3133–3157.

15. CHARLESWORTH B, NORDBORG M, CHARLESWORTH D. The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. Genetical Research. 1997;70(2):155–174.

16. Charlesworth B, Morgan MT, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. Genetics. 1993;134(4):1289–303. Available from: http://www.ncbi.nlm.nih.gov/pubmed/8375663{%}0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1205596.

17. Hudson RR, Kaplan NL. Deleterious background selection with recombination. Genetics. 1995;141(4):1605–1617. Available from: http://www.genetics.org/content/141/4/1605.

18. Lewontin RC, Krakauer J. DISTRIBUTION OF GENE FREQUENCY AS A TEST OF THE THEORY OF THE SELECTIVE NEUTRALITY OF POLYMORPHISMS. Genetics. 1973;74(1):175–195. Available from: http://www.genetics.org/content/74/1/175.

19. Rosenzweig BK, Pease JB, Besansky NJ, Hahn MW. Powerful methods for detecting introgressed regions from population genomic data. Molecular Ecology. 2016;25(11):2387–2397. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.13610.

20. Martin SH, Jiggins CD. Interpreting the genomic landscape of introgression. Current Opinion in Genetics and Development. 2017;47:69–74. Available from: http://dx.doi.org/10.1016/j.gde.2017.08.007.

21. Noor MAF, Bennett SM. Islands of speciation or mirages in the desert Examining the role of restricted recombination in maintaining species. Heredity. 2009;103(6):439–444. Available from: http://dx.doi.org/10.1038/hdy.2009.151.

22. Lohse K. Come on feel the noise – from metaphors to null models. Journal of Evolutionary Biology. 2017;30(8):1506–1508.

23. Sousa VC, Carneiro M, Ferrand N, Hey J. Identifying Loci Under Selection Against Gene Flow in Isolation-with-Migration Models. Genetics. 2013;194(1):211–233. Available from: http://www.genetics.org/content/194/1/211.

24. Beichman AC, Huerta-Sanchez E, Lohmueller KE. Using Genomic Data to Infer Historic Population Dynamics of Nonmodel Organisms. Annual Review of Ecology, Evolution, and Systematics. 2018;49(1):433–456. Available from: https://doi.org/10.1146/annurev-ecolsys-110617-062431.

25. Marchi N, Schlichta F, Excoffier L. Demographic inference. Current Biology. 2021;31(6):R276–R279. Available from: https://www.sciencedirect.com/science/article/pii/S0960982221001184.

26. Jiao X, Flouri T, Yang Z. Multispecies coalescent and its applications to infer species phylogenies and cross-species gene flow. National Science Review. 2021 07;8(12). Nwab127. Available from: https://doi.org/10.1093/nsr/nwab127.

27. Hey J, Machado CA. The study of structured populations - new hope for a difficult and divided science. Nature Reviews Genetics. 2003;4(7):535–543.

28. Roux C, Fraïsse C, Romiguier J, Anciaux Y, Galtier N, Bierne N. Shedding Light on the Grey Zone of Speciation along a Continuum of Genomic Divergence. PLOS Biology. 2016 12;14(12):1–22. Available from: https://doi.org/10.1371/journal.pbio.2000234.

29. Pennisi E. Disputed islands. Science. 2014;345(6197):611–613. Available from: https://www.science.org/doi/abs/10.1126/science.345.6197.611.

30. Nachman MW, Payseur BA. Recombination rate variation and speciation: Theoretical predictions and empirical results from rabbits and mice. Philosophical Transactions of the Royal Society B: Biological Sciences. 2012 feb;367(1587):409–421. Available from: http://rstb.royalsocietypublishing.org/cgi/doi/10.1098/rstb.2011.0249.

31. Westram AM, Stankowski S, Surendranadh P, Barton N. What is reproductive isolation? Journal of Evolutionary Biology. 2022;35(9):1143–1164.

32. Johri P, Aquadro CF, Beaumont M, Charlesworth B, Excoffier L, Eyre-Walker A, et al. Recommendations for improving statistical inference in population genomics. PLoS Biology. 2022;20(5):e3001669.

33. Bisschop G, Setter D, Rafajlović M, Baird SJ, Lohse K. The impact of global selection on local adaptation and reproductive isolation. Philosophical Transactions of the Royal Society B. 2020;375(1806):20190531.

34. Roux C, Tsagkogeorga G, Bierne N, Galtier N. Crossing the Species Barrier: Genomic Hotspots of Introgression between Two Highly Divergent Ciona intestinalis Species. Molecular Biology and Evolution. 2013;30(7):1574–1587. Available from: http://dx.doi.org/10.1093/molbev/mst066.

35. Fraïsse C, Popovic I, Mazoyer C, Spataro B, Delmotte S, Romiguier J, et al. DILS: Demographic inferences with linked selection by using ABC. Molecular Ecology Resources. 2021;21(8):2629–2644.

36. Lohse K, Harrison RJ, Barton NH. A general method for calculating likelihoods under the coalescent process. Genetics. 2011;189(3):977–987.

37. Lohse K, Chmelik M, Martin SH, Barton NH. Efficient strategies for calculating blockwise likelihoods under the coalescent. Genetics. 2016;202(2):775–786.

38. Bisschop G. Graph-based algorithms for Laplace transformed coalescence time distributions. PLOS Computational Biology. 2022 09;18(9):1–13. Available from: https://doi.org/10.1371/journal.pcbi.1010532.

39. Baumdicker F, Blows MW, Goldstein D, Gower G, Ragsdale AP, Tsambos G, et al. Efficient ancestry and mutation simulation with msprime 1.0. Genetics. 2022 mar;220(3). Available from: https://www.biorxiv.org/content/early/2021/09/21/2021.08.31.457499https://academic.oup.com/genetics/article/doi/10.1093/genetics/iyab229/6460344.

40. Merrill RM, Rastas P, Martin SH, Melo MC, Barker S, Davey J, et al. Genetic dissection of assortative mating behavior. PLOS Biology. 2019 02;17(2):1–21. Available from: https://doi.org/10.1371/journal.pbio.2005902.

41. Kronforst M, Hansen MB, Crawford N, Gallant J, Zhang W, Kulathinal R, et al. Hybridization Reveals the Evolving Genomic Architecture of Speciation. Cell Reports. 2013;5(3):666 – 677. Available from: http://www.sciencedirect.com/science/article/pii/S2211124713005652.

42. 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 Research. 2013;23(11):1817–1828. Available from: http://genome.cshlp.org/content/23/11/1817.abstract.

43. Naisbit RE, Jiggins CD, Linares M, Salazar C, Mallet J. Hybrid Sterility, Haldane's Rule and Speciation in Heliconius cydno and H. melpomene. Genetics. 2002;161(4):1517–1526. Available from: https://www.genetics.org/content/161/4/1517.

44. Davey JW, Barker SL, Rastas PM, Pinharanda A, Martin SH, Durbin R, et al. No evidence for maintenance of a sympatric Heliconius species barrier by chromosomal inversions. Evolution Letters. 2017 aug;1(3):138–154. Available from: http://doi.wiley.com/10.1002/evl3.12.

45. 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(1):525–541. Available from: http://www.genetics.org/content/203/1/525.

46. Merrill RM, Van Schooten B, Scott JA, Jiggins CD. Pervasive genetic associations between traits causing reproductive isolation in ¡i¿Heliconius¡/i¿ butterflies. Proceedings of the Royal Society B: Biological Sciences. 2011;278(1705):511–518. Available from: https://royalsocietypublishing.org/doi/abs/10.1098/rspb.2010.1493.

47. Van Belleghem SM, Baquero M, Papa R, Salazar C, McMillan WO, Counterman BA, et al. Patterns of Z chromosome divergence among Heliconius species highlight the importance of historical demography. Molecular ecology. 2018;27(19):3852–3872.

48. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989;123(3):585–595.

49. Keightley PD, Pinharanda A, Ness RW, Simpson F, Dasmahapatra KK, Mallet J, et al. Estimation of the Spontaneous Mutation Rate in Heliconius melpomene. Molecular Biology and Evolution. 2015;32(1):239–243.

50. Reed RD, Papa R, Martin A, Hines HM, Counterman BA, Pardo-Diaz C, et al. ¡i¿optix¡/i¿ Drives the Repeated Convergent Evolution of Butterfly Wing Pattern Mimicry. Science. 2011;333(6046):1137–1141.

51. Martin A, Papa R, Nadeau NJ, Hill RI, Counterman BA, Halder G, et al. Diversification of complex butterfly wing patterns by repeated regulatory evolution of a Wnt ligand. Proceedings of the National Academy of Sciences. 2012;109(31):12632–12637. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.1204800109.

52. Nadeau NJ, Pardo-Diaz C, Whibley A, Supple MA, Saenko SV, Wallbank RWR, et al. The gene cortex controls mimicry and crypsis in butterflies and moths. Nature. 2016;534(7605):106–110. Available from: http://www.nature.com/doifinder/10.1038/nature17961.

53. Rossi M, Hausmann AE, Thurman TJ, Montgomery SH, Papa R, Jiggins CD, et al. Visual mate preference evolution during butterfly speciation is linked to neural processing genes. Nature Communications. 2020;11(1):2041–1723.

54. Hanly JJ, Robertson ECN, Corning OBWH, Martin A. Porcupine/Wntless-dependent trafficking of the conserved WntA ligand in butterflies. Journal of Experimental Zoology Part B: Molecular and Developmental Evolution. 2021;336(6):470–481.

55. Wallbank RWR, Baxter SW, Pardo-Diaz C, Hanly JJ, Martin SH, Mallet J, et al. Evolutionary Novelty in a Butterfly Wing Pattern through Enhancer Shuffling. PLoS Biology. 2016;14(1):1–16.

56. Barton N, Bengtsson BO. The barrier to genetic exchange between hybridising populations. Heredity. 1986;57(3):357–376.

57. Flaxman SM, Wacholder AC, Feder JL, Nosil P. Theoretical models of the influence of genomic architecture on the dynamics of speciation. Molecular Ecology. 2014;23(16):4074–4088.

58. Aeschbacher S, Bürger R. The effect of linkage on establishment and survival of locally beneficial mutations. Genetics. 2014;197(1):317–336.

59. Yeaman S, Aeschbacher S, Bürger R. The evolution of genomic islands by increased establishment probability of linked alleles. Molecular Ecology. 2016;25(11):2542–2558. Available from: `http://dx.doi.org/10.1111/mec.13611`.

60. Martin SH, Davey JW, Salazar C, Jiggins CD. Recombination rate variation shapes barriers to introgression across butterfly genomes. PLOS Biology. 2019 02;17(2):1–28. Available from: `https://doi.org/10.1371/journal.pbio.2006288`.

61. Cruickshank TE, Hahn MW. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. Molecular Ecology. 2014;23:3133–3157.

62. Davey JW, Barker SL, Rastas PM, Pinharanda A, Martin SH, Durbin R, et al. No evidence for maintenance of a sympatric *Heliconius* species barrier by chromosomal inversions. Evolution letters. 2017;1(3):138–154.

63. Smith JM, Haigh J. The hitch-hiking effect of a favourable gene. Genetical Research. 1974 feb;23(1):23–35. Available from: `https://www.cambridge.org/core/product/identifier/S0016672300014634/type/journal{_}article`.

64. Bengtsson B. The flow of genes through a genetic barrier. Evolution: essays in honour of John Maynard Smith. 1985;1:31–42.

65. Aeschbacher S, Selby JP, Willis JH, Coop G. Population-genomic inference of the strength and timing of selection against gene flow. Proceedings of the National Academy of Sciences. 2017;114(27):7061–7066. Available from: https://www.pnas.org/content/114/27/7061.

66. Jiggins CD, Martin SH. Glittering gold and the quest for Isla de Muerta. Journal of Evolutionary Biology. 2017;30(8):1509–1511. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/jeb.13110.

67. Bank C, Bürger R, Hermisson J. The Limits to Parapatric Speciation: Dobzhansky–Muller Incompatibilities in a Continent–Island Model. Genetics. 2012 jul;191(3):845–863. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10747061http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1461023http://www.genetics.org/lookup/doi/10.1534/genetics.111.137513.

68. Wilkinson-Herbots HM. The distribution of the coalescence time and the number of pairwise nucleotide differences in the "isolation with migration" model. Theoretical Population Biology. 2008;73(2):277–288.

69. Lohse K, Chmelik M, Martin SH, Barton NH. Efficient strategies for calculating blockwise likelihoods under the coalescent. Genetics. 2016;202(2):775–786. Available from: http://www.genetics.org/content/202/2/775.

70. Bisschop G, Lohse K, Setter D. Sweeps in time: Leveraging the joint distribution of branch lengths. Genetics. 2021;219(2).

71. Shipilina D, Stankowski S, Pal A, Chan YF, Barton N. On the origin and structure of haplotype blocks. Authorea. 2022;.

72. Kelleher J, Wong Y, Wohns AW, Fadil C, Albers PK, McVean G. Inferring whole-genome histories in large population datasets. Nature Genetics. 2019 sep;51(9):1330–1338. Available from: https://doi.org/10.1038/s41588-019-0483-yhttp://www.nature.com/articles/s41588-019-0483-y.

73. Ignatieva A, Lyngsø RB, Jenkins PA, Hein J. KwARG: parsimonious reconstruction of ancestral recombination graphs with recurrent mutation. Bioinformatics. 2021;37(19):3277–3284.

74. Rasmussen DA, Guo F. Espalier: Efficient tree reconciliation and ARG reconstruction using maximum agreement forests. bioRxiv. 2022;.

75. Coyne JA, Allen Orr H. The evolutionary genetics of speciation. Philosophical Transactions of the Royal Society B: Biological Sciences. 1998;353(1366):287–305. Available from: http://rstb.royalsocietypublishing.org/cgi/doi/10.1098/rstb.1998.0210.

76. Wolfram Research I. Mathematica, Version 11.1.1.0; 2017.

77. Nelder JA, Mead R. A Simplex Method for Function Minimization. The Computer Journal. 1965;7(4):308–313.

78. Johnson SG. The NLopt nonlinear-optimization package;. Available from: http://github.com/stevengj/nlopt.

79. Gower G, Ragsdale AP, Gutenkunst RN, Hartfield M, Noskova E, Struck TJ, et al. Demes: a standard format for demographic models. bioRxiv. 2022;Available from: https://www.biorxiv.org/content/early/2022/06/01/2022.05.31.494112.

80. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. PLoS Genetics. 2009 oct;5(10):e1000695. Available from: https://dx.plos.org/10.1371/journal.pgen.1000695.

81. Haller BC, Messer PW. SLiM 3: Forward Genetic Simulations Beyond the Wright-Fisher Model. Molecular Biology and Evolution. 2019;36(3):632–637.

82. Thornton KR. A c++ template library for efficient forward-time population genetic simulation of large populations. Genetics. 2014;198(1):157–166.

83. Noskova E, Ulyantsev V, Koepfli KP, O'Brien SJ, Dobrynin P. GADMA: Genetic algorithm for inferring demographic history of multiple populations from allele frequency spectrum data. GigaScience. 2020;9(3):giaa005.

84. Ragsdale AP, Gravel S. Models of archaic admixture and recent history from two-locus statistics. PLoS genetics. 2019;15(6):e1008204.

85. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884–i890.

86. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:13033997. 2013;.

87. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021 02;10(2). Giab008. Available from: https://doi.org/10.1093/gigascience/giab008.

88. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. Bioinformatics. 2015;31(12):2032–2034.

Table S1. *Heliconius* Illumina read sets downloaded from ENA. COV-mean and COV-sd represent mean and standard deviation of read coverage when mapping quality-trimmed reads against the Hmel2.5 assembly.

Read group ID	ENA run accession	sex	taxon	COV-mean	COV-sd
chi.CAM25091.f	ERR2298220	F	*H. cydno chioneus*	27.71	78.67
chi.CAM25137.f	ERR2298222	F	*H. cydno chioneus*	29.6	77.41
chi.CAM580.m	ERR2298196	M	*H. cydno chioneus*	23.41	44.5
chi.CAM582.m	ERR2298197	M	*H. cydno chioneus*	25.7	48.83
chi.CAM585.m	ERR2298199	M	*H. cydno chioneus*	24.96	52.37
chi.CAM586.m	ERR2298200	M	*H. cydno chioneus*	24.3	49.73
chi.CJ553.m	ERR260295	M	*H. cydno chioneus*	37.62	68.71
chi.CJ560.m	ERR260296	M	*H. cydno chioneus*	37.11	67.13
chi.CJ564.m	ERR260297	M	*H. cydno chioneus*	41.27	78.71
chi.CJ565.m	ERR260298	M	*H. cydno chioneus*	48.36	76.16
ros.CAM1841.m	ERR2298205	M	*H. melpomene rosina*	28.7	47.35
ros.CAM1880.m	ERR2298206	M	*H. melpomene rosina*	31.94	52.82
ros.CAM2045.m	ERR2298208	M	*H. melpomene rosina*	23.19	30.51
ros.CAM2059.m	ERR2298209	M	*H. melpomene rosina*	26.35	34.71
ros.CAM2519.m	ERR2298210	M	*H. melpomene rosina*	28.48	31.07
ros.CAM2552.m	ERR2298211	M	*H. melpomene rosina*	25.16	56.86
ros.CJ2071.m	ERR260280	M	*H. melpomene rosina*	38.2	57.43
ros.CJ531.m	ERR260277	M	*H. melpomene rosina*	28.26	44.68
ros.CJ533.m	ERR260278	M	*H. melpomene rosina*	28.27	43.12
ros.CJ546.m	ERR260279	M	*H. melpomene rosina*	27.97	45.36

89. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:12073907. 2012;.

Supplementary Information

Table S2. Maximum composite likelihood estimates (MCLEs) of parameters and overall support for various global demographic histories of divergence and/or geneflow for *H. melpomene* and *H. cydno* assuming no heterogeneity in N_e and m_e. The best fitting model is shown in bold; $\Delta \ln CL$ (last row) shows support relative to this model.

Model	N_{cyd}	N_{mel}	N_{anc}	T	$m_{\rightarrow cyd}$	$m_{\rightarrow mel}$	$4N_e m$	$\ln CL$	$\Delta \ln CL$
DIV	1,486,722	1,117,681	1,416,636	970,873				-394,862,755	1,078,758
$MIG_{\rightarrow mel}$	1,292,936	528,814				7.8010^{-7}	1.65	-393,973,616	189,619
$MIG_{\rightarrow cyd}$	695,480	1,254,882			7.8610^{-7}		2.19	-395,772,748	1,988,751
$IM_{\rightarrow cyd}$	1,393,771	1,142,726	1,401,906	1,047,537	8.0210^{-8}		0.45	-394,858,825	1,074,828
$\mathbf{IM}_{\rightarrow mel}$	1,406,917	545,018	923,270	4,256,400		7.3810^{-8}	1.61	-393,783,997	0

Fig. S1 The *Heliconius* genome was analysed in sliding windows comprised of $L = 50,000$ (64 base) pair-blocks of intergenic sequence. Since each block is sampled in a single heterospecific pair, this corresponds to a minimum window span of $(L * 64)/(n_A * n_B) = 32kb$ (black dashed line). Windows show a wide distribution of spans with a median (103 kb, red line) substantially greater than this minimum.

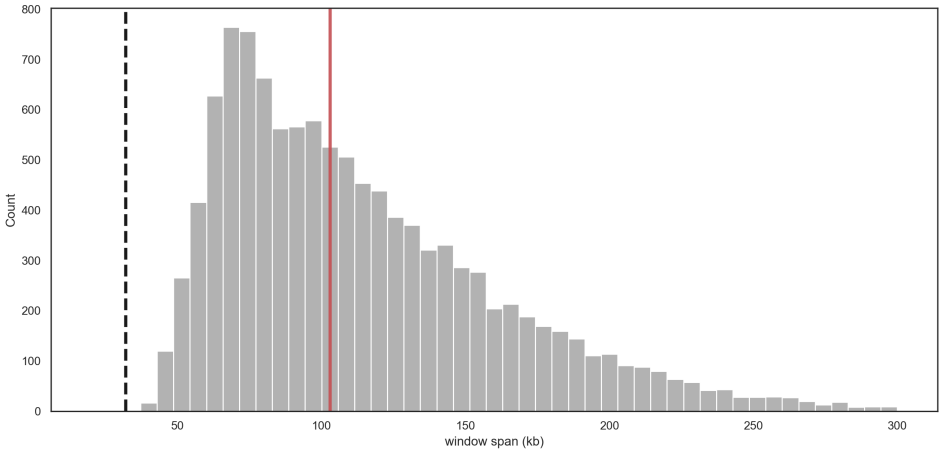


Fig. S2 Mean heterozygosity H in *H. cydno* (left) and *H. melpomene* (right) against pairwise divergence (d_{xy}) in intergenic windows (see Fig. S1). The upper 5% tail of window-wise F_{ST} is shown in blue.

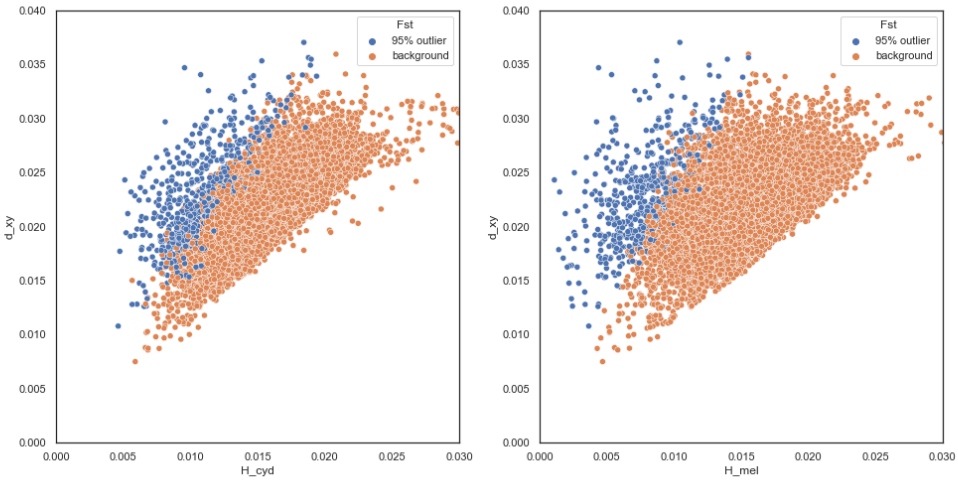


Fig. S3 **Top**: Divergence (d_{xy}, black), heterozygosity in *H. cydno* (H_{cyd}, green) and *H. melpomene* (H_{mel}, blue) along the genome in windows of intergenic sequence (50,000 × 64b blocks). Mean values are shown as horizontal lines. **Bottom**: F_{ST} computed from the summaries above. The upper 5% tail of window-wise F_{ST} is delineated by the black dashed line, the median is shown in gray.

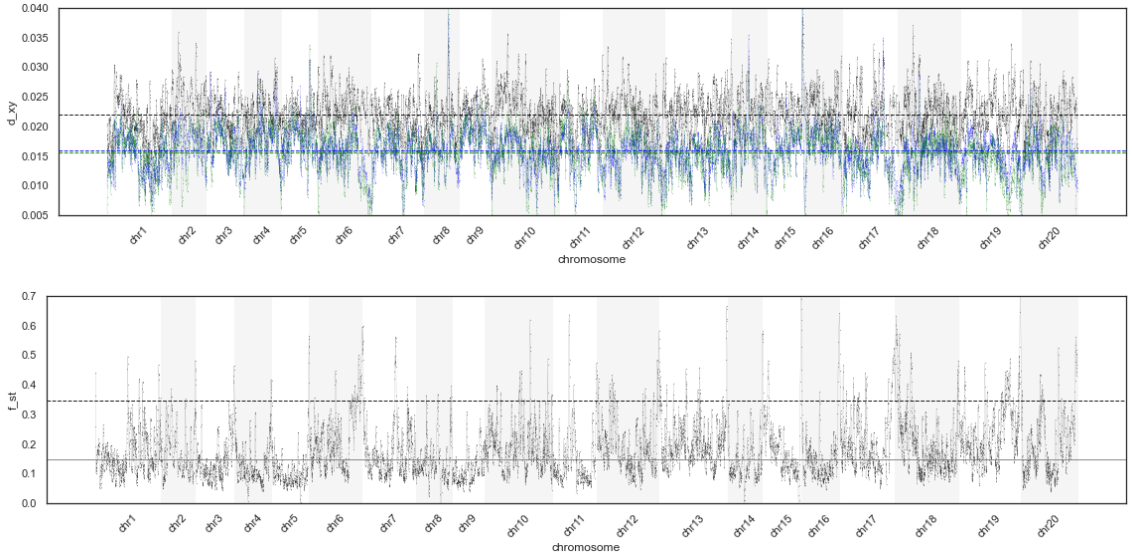


Fig. S4 F_{ST} between *H. melpomene* and *H. cydno* is negatively correlated with recombination rate.

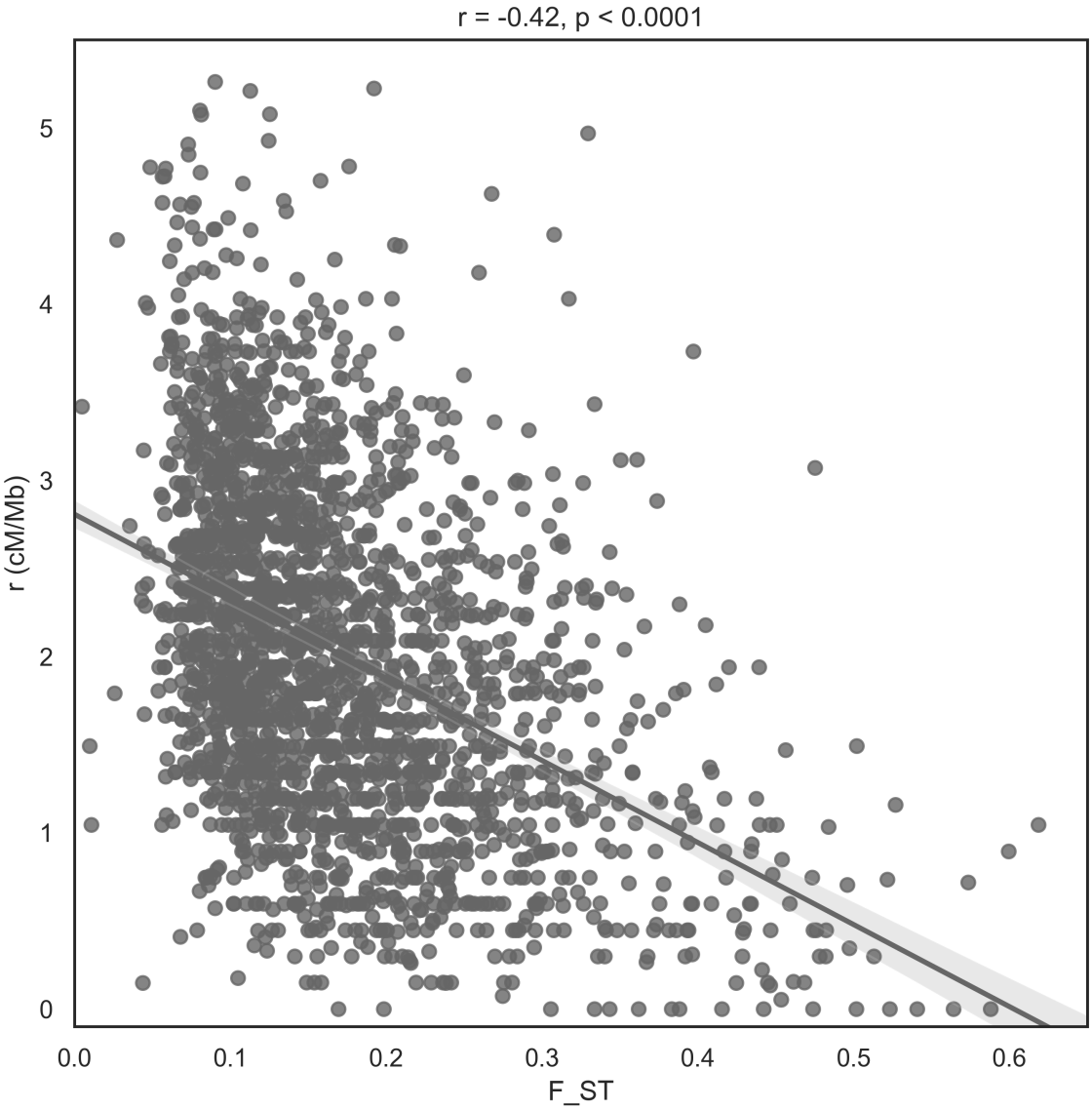


Fig. S5 Local estimates of N_e in *H. melpomene* (A) and *H. cydno* (B) are negatively correlated with chromosome length.

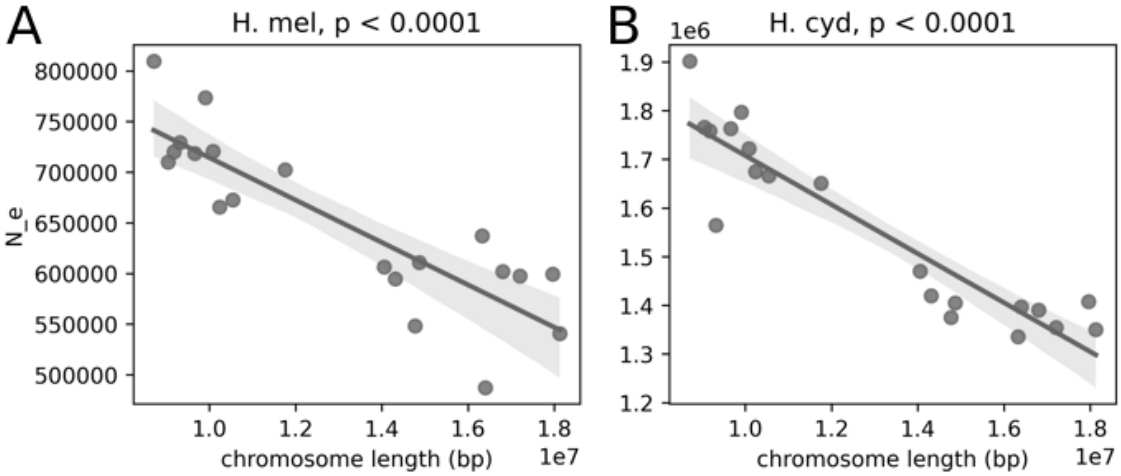


Fig. S6 Estimates of local N_e in *H. melpomene* and *H. cydno* and m_e are correlated positively with the recombination rate (top) and negatively with the density of CDS (bottom).

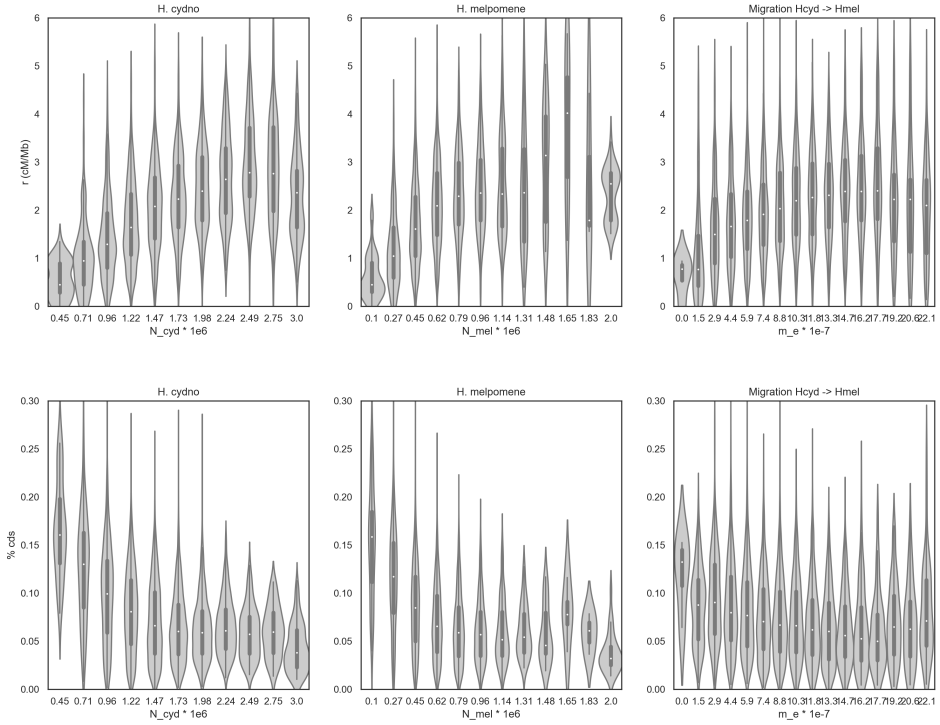


Fig. S7 The impact of window size on the power to detect a local reduction in m_e relative to the estimated background migration rate. The black line corresponds to the window size used in the *Heliconius* analysis.

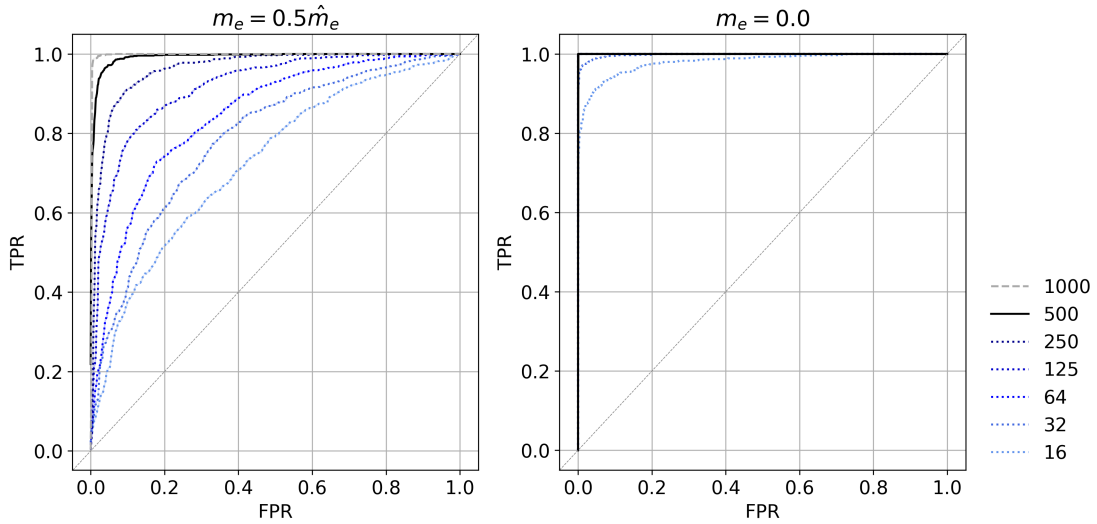
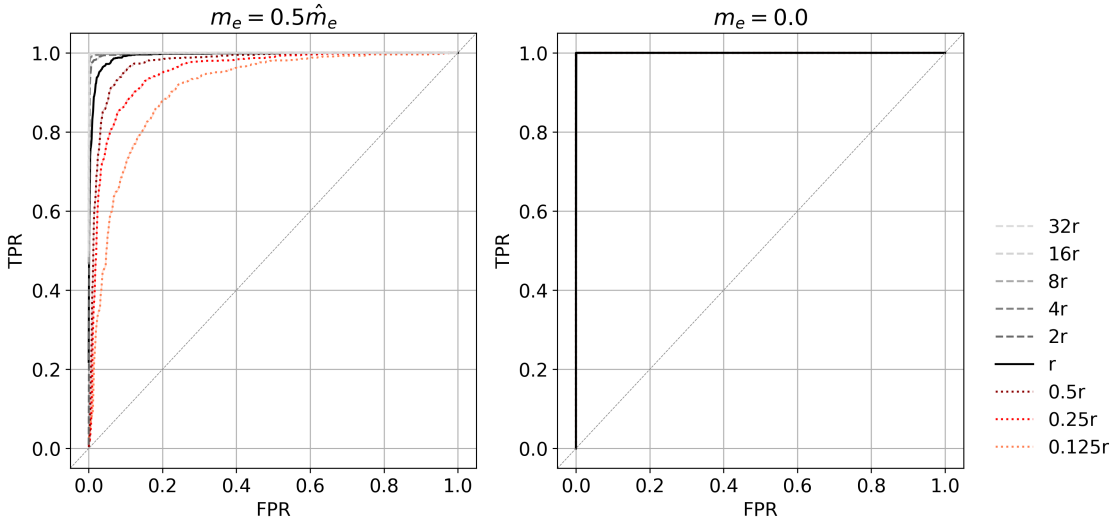



Fig. S8 The impact of r_{bp} on the power to detect a local reduction in m_e relative to the estimated background migration rate. The black line indicates the average recombination rate in *Heliconius*.