1 **Speciation in *Heliconius* Butterflies: Minimal Contact Followed** 2 **by Millions of Generations of Hybridisation**

3 Simon H. Martin^{*1}, Anders Eriksson^{*1,2}, Krzysztof M. Kozak¹, Andrea Manica¹ and
4 Chris D. Jiggins¹

5 ¹ Department of Zoology, University of Cambridge, Downing Street, Cambridge, CB2
6 3EJ, United Kingdom

7 ² Integrative Systems Biology Laboratory, King Abdullah University of Science
8 and Technology, Thuwal 23955-6900, Kingdom of Saudi Arabia

9 * These authors contributed equally

10 Corresponding Author: S.H. Martin, shm45@cam.ac.uk

11 Running Head: Change in the rate of gene flow during butterfly speciation

12 Abstract

13 Documenting the full extent of gene flow during speciation poses a challenge, as
 14 species ranges change over time and current rates of hybridisation might not reflect
 15 historical trends. Theoretical work has emphasized the potential for speciation in the
 16 face of ongoing hybridisation, and the genetic mechanisms that might facilitate this
 17 process. However, elucidating how the rate of gene flow between species may have
 18 changed over time has proved difficult. Here we use Approximate Bayesian
 19 Computation (ABC) to fit a model of speciation between the Neotropical butterflies
 20 *Heliconius melpomene* and *Heliconius cydno*. These species are ecologically divergent,
 21 rarely hybridize and display female hybrid sterility. Nevertheless, previous genomic
 22 studies suggests pervasive gene flow between them, extending deep into their past, and
 23 potentially throughout the speciation process. By modelling the rates of gene flow
 24 during early and later stages of speciation, we find that these species have been
 25 hybridising for hundreds of thousands of years, but have not done so continuously since
 26 their initial divergence. Instead, it appears that gene flow was rare or absent for as long
 27 as a million years in the early stages of speciation. Therefore, by dissecting the timing
 28 of gene flow between these species, we are able to reject a scenario of purely sympatric
 29 speciation in the face of continuous gene flow. We suggest that the period of minimal
 30 contact early in speciation may have allowed for the accumulation of genomic changes
 31 that later enabled these species to remain distinct despite a dramatic increase in the rate
 32 of hybridisation.

33 Introduction

34 Speciation is widely viewed as the development of reproductive isolation between
 35 lineages. However, there is now considerable evidence that reproductive isolation is not
 36 necessarily a genome-wide phenomenon, but rather that species integrity can be
 37 maintained despite gene flow affecting a considerable proportion of the genome [1–6].
 38 What remains less clear is the importance of gene flow (or lack thereof) for the
 39 establishment of new species. Theory has shown that it may be possible, under certain
 40 genetic and selective conditions, for species to become established in the face of
 41 ongoing gene flow [7–14]. To test this theory, it is necessary either to observe speciation
 42 in real time, or to reconstruct the historical extent and timing of gene flow between
 43 existing species.

44 In geographic terms, speciation can be described as sympatric, parapatric or
 45 allopatric. We follow Mallet et al. [15] in defining these terms: sympatric populations
 46 share the same geographic area (but not necessarily the same niche), such that
 47 individuals from the two populations are liable to encounter one-another frequently over
 48 much of their range. Parapatric populations “occupy separate but adjoining geographic
 49 regions,” such that only a small fraction of individuals at the edge of each range are
 50 liable to encounter the other. Allopatric populations are geographically separated, such
 51 that encounters between them are very rare or impossible. Despite the abundance of
 52 closely related sympatric species, there are very few cases in which it can be stated with
 53 any certainty that speciation occurred in sympatry [e.g. crater-lake cichlids [16]]. In
 54 terms of gene flow, we can predict that sympatric speciation might involve a gradual
 55 decline over time, with higher rates of historic than contemporary gene flow [17]. In
 56 allopatric speciation, gene flow would be absent until the populations came into
 57 secondary contact. Parapatric speciation might fall somewhere in between these
 58 extremes, with a low level of gene flow throughout, potentially decreasing over time,
 59 but possibly also increasing if the populations later return to a sympatric distribution.

60 Genomic data now offers the exciting possibility of reconstructing the history of gene
61 flow between existing species, illuminating the roles of gene flow and geography in the
62 origin of new species.

63 A number of methods exist to fit a model of “isolation with migration” (IM) using
64 patterns of DNA sequence variation, thereby testing for post-speciation gene flow. This
65 can be achieved by maximising the likelihood of observed genetic data in a coalescent
66 framework, either directly [18–22] or using Markov Chain Monte Carlo (MCMC)
67 approximation [23–26]. However, these approaches lack power and accuracy to
68 examine change in the rate of gene flow over time [27,28], owing to characteristics of
69 the standard IM model itself [29]. This limitation could be overcome through the
70 implementation of more complex models [30], but this is currently not feasible in a
71 likelihood framework. Approximate Bayesian Computation (ABC) offers a tractable
72 means to fit such complex genetic models by avoiding the need to derive likelihoods
73 [31]. ABC is therefore suited to the problem of reconstructing changes in the rate of
74 gene flow during speciation [32,33], offering the potential to resolve long-standing
75 debates in the speciation literature.

76 Here we investigate the history of gene flow during speciation in *Heliconius*
77 butterflies. This Neotropical genus is well known for its broad diversity of aposematic
78 wing patterns, and multiple instances of Müllerian mimicry – where unrelated species
79 converge in wing pattern, providing a unified signal of toxicity to predators. Closely
80 related species usually differ in wing pattern, and it is thought that pattern divergence
81 between populations adapting to mimic different locally-abundant patterns could lead to
82 parapatric speciation [34,35]. We examine *Heliconius melpomene* and *Heliconius*
83 *cydno*, closely related species that have diverged in wing pattern and other ecological
84 traits during the past million years, but continue to hybridise at low frequency
85 [36] where their ranges overlap in the western parts of South America and Central
86 America. The ability to compare sympatric and allopatric populations of *H. cydno* and

H. melpomene provides an ideal opportunity to detect the genetic signatures of recent gene flow. Indeed, whole-genome studies have found evidence of abundant gene flow between these species, affecting a large proportion of the genome and extending deep into the past [3,4]. However, it has remained uncertain whether this pair diverged in the face of ongoing gene flow, or experienced a period of isolation early during speciation.

We used ABC to reconstruct the genealogical history of three populations: *H. cydno* from Panama, *H. melpomene rosina* from Panama (sympatric with *cydno*) and *H. melpomene melpomene* from French Guiana (allopatric) (Fig. 1A). Our model allowed for hybridisation between *H. cydno* and *H. melpomene* throughout speciation, and accounted for the possibility of a change in the rate of hybridisation during this time by considering two separate periods with distinct migration rates, the duration or which could vary. This enabled us to test various hypotheses under a single model, including a clean split without gene flow, continuous gene flow throughout speciation or gene flow restricted to ancient or recent time periods.

Figure 1. Species distributions, sample locations, summary statistic and model design. **A.** Distributions (shaded) of *H. melpomene* (light red) and *H. cydno* (grey), based on Rosser et al. [37]. Sampling locations in Panama and French Guiana are indicated. **B.** The composite summary statistic used consisted of the proportions of the 13 possible biallelic genotype combinations, where one individual is sampled from each population (given to the right). '0' and '2' indicate alternative homozygous states and '1' indicates the heterozygous state. Given that four individuals were sampled from each population, the proportion of biallelic SNPs carrying each pattern was averaged over all 64 possible sets of three samples. Although 25 SNP states are theoretically possible, twelve of these can be folded if we ignore major and minor alleles (e.g. 2-0-1 is equivalent to 0-2-1) and so these were counted together, to give 13 unique states. **C.** The model had ten free parameters: four population sizes (N_a , N_c , N_{mr} and N_{mm}) (the ancestral *H. melpomene* population size was assumed to be the average of N_{mr} and N_{mm}); migration rates between *H. cydno* and *H. melpomene* in Periods 1 and 2 (M_{cm1} and M_{cm2}) the time dividing Periods 1 and 2 (t_d); migration rate between the two *H. melpomene* populations (M_m), and the split times for the two species (t_1) and the two *H. melpomene* populations (t_2).

Results

Genotype data and summary statistics

Whole genome resequence data from twelve wild-caught butterflies, with four representatives from each of the three sampled populations, *H. cydno*, *H. m. rosina* and *H. m. melpomene*, were used for model fitting (S1 Table, data from Martin et al. 2013). Only intergenic regions, as defined by the *Heliconius melpomene* reference genome annotation v1.1 [38], with high-quality genotype calls (see Materials and Methods) for all twelve samples, were considered. We also excluded all scaffolds on the Z chromosome, which is known to experience strongly reduced gene flow [4], as well as putative CpG clusters, which can have unusual mutation rates. These criteria gave ~60 million sites (22% of the genome), of which approximately 10% were polymorphic (Single Nucleotide Polymorphisms, SNPs). The composite summary statistic used for model fitting consisted of the proportion of bi-allelic sites carrying each possible combination of genotypes among three diploid individuals. These proportions were averaged over all possible triplets, where each population is represented by one

individual, and folded such that major and minor alleles were not distinguished (Fig. 1B, see Materials and Methods for details). This composite summary statistic is similar to a three-dimensional site frequency spectrum. It provides a nearly exhaustive summary of the available SNP data among the ingroup taxa, is independent of linkage effects and scalable to any number of sites.

As expected, the most common SNP patterns were singletons, where one individual was heterozygous and the other two were homozygous for the same allele (0-0-1, 0-1-0, and 1-0-0; Fig. 1B). The most common pattern overall was 1-0-0, where *H. cydno* is heterozygous and both *H. melpomene* individuals are homozygous. This is unsurprising, given the longer branch leading to *H. cydno* (Fig. 1B). The pattern 0-0-1, where *H. m. melpomene* from French Guiana is heterozygous, was also considerably more common than 0-1-0, where *H. m. rosina* is heterozygous. This is consistent with increased shared variation between *H. cydno* and the sympatric *H. m. rosina*. Similarly, 1-1-0 was more common than 1-0-1, and 0-0-2 was more common than 0-2-0.

Estimating the timing and extent of gene flow

We consider a model with three populations (Fig. 1C), corresponding to *H. cydno* (which splits from the ancestral *melpomene* population at time t_1), and *H. m. rosina* and *H. m. melpomene* (which split at time t_2). Each lineage has a separate population size, except for the ancestral *melpomene* population, which is assumed to have a size equal to the mean of the two *melpomene* populations. Because the two *melpomene* populations represent extremes of a somewhat continuous range, migration between them is allowed at a continuous rate M_m . Migration is also allowed between *H. cydno* and *H. melpomene*, although after the split between the *melpomene* populations (t_2) only *H. m. rosina* is able to exchange migrants with *H. cydno*. Two distinct periods of between-species migration are modelled, with rates M_{cm1} and M_{cm2} . These periods are divided at a time t_d such that Period 1 begins at t_1 and ends at t_d , and Period 2 runs from t_d to the present. This model had ten free parameters: four population sizes, two split times, three

migration rates and one time dividing the migration periods. Model parameters were estimated using Approximate Bayesian Computation (ABC) based on the summary statistics described above (see Materials and Methods for details). Uniform priors were used for all parameters except for t_1 and t_2 , for which prior distributions were estimated by analysis of mitochondrial sequence data (see Materials and Methods for details).

Our model pointed toward a dramatic change in the rate of inter-specific migration (i.e. hybridisation resulting in gene flow) from early to later stages of speciation (Fig. 2). Migration was minimal in Period 1 ($M_{cm1} \sim 0.08$ migrants per year [posterior mean]), and around tenfold greater in Period 2 ($M_{cm2} \sim 0.81$ migrants per year) (Fig. 2A, Table 1). The date of transition between these two periods (t_d) had a fairly wide posterior distribution, with a mean of 0.5 million years ago (Ma), but a 90% posterior density interval extending from 0.1 Ma to 1.2 Ma (Fig. 2A, Table 1). This was nevertheless considerably more recent than the inferred time of speciation (t_i), which had a fairly narrow posterior distribution centred around 1.5 Ma (Fig. 2A, Table 1). Therefore, our results support a case of stronger isolation during the early stages of speciation, with a large increase in the rate of hybridisation later. The posterior mean of 0.5 Ma for the date dividing the two periods would imply that hybridisation was rare or absent during the first two thirds of the time since initial divergence (~ 1 million years). However, we note that this transition cannot be dated with great accuracy, and indeed our model does not allow us to infer whether the increase in hybridisation was sharp or gradual. Nevertheless, it is notable that the posterior mean for the onset of more frequent hybridisation coincides roughly with the split between the two *H. melpomene* populations ($t_2 \sim 0.53$ Ma, Fig. 2A). Interestingly, the inferred rate of gene flow between *H. cydno* and *H. m. rosina* during this second period is several times greater than that between the two *H. melpomene* populations (Fig. 2A, Table 1).

Figure 2. Posteriors for time and migration rate parameters, and a schematic representation of the inferred model. A. Posterior distributions for the three time parameters (left) and three migration rates (right). Densities are scaled for ease of comparison. See S2 Figure for all posterior distributions. Posterior means are indicated by vertical dashed lines. **B.** A schematic phylogeny, where the posterior mean for each parameter is indicated. Colours correspond to those in A. Times are indicated by horizontal dashed lines. Posteriors for the times are given on the right-hand-side for reference. Migration rates are indicated by arrows, with the width of the arrow scaled according to the migration rate. Relative population sizes are indicated by branch widths. See Table 1 for all values.

Table 1. Properties of posterior distributions for the ten model parameters

Parameter	Mean	Mode	Median	HPD90 Lower ¹	HPD90 Upper ¹
t_1 (Ma)	1.5	1.4	1.5	1.2	1.8
t_2 (Ma)	0.53	0.53	0.53	0.42	0.65
t_d (Ma)	0.5	0.3	0.4	0.1	1.2
M_{cm1} (ind. /year)	0.08	0.01	0.06	0.01	0.86
M_{cm2} (ind. /year)	0.81	0.95	0.84	0.36	1.6
M_m (ind. /Year)	0.26	0.34	0.28	0.06	0.77
N_c (M ind.)	5.3	5.3	5.3	4.6	5.9
N_{mr} (M ind.)	1.8	1.8	1.8	1.2	2.4
N_{mm} (M ind.)	3.9	3.9	3.9	3.4	4.4
N_A (M ind.)	1.1	1.3	1.1	0.3	2.0

¹ Upper and lower bounds of the 90% Highest Posterior Density (HPD) interval. The HPD is defined as the set of values making up 90% of the density distribution and within which all values have higher density than outside.

Estimates of N_e were generally larger than previous estimates for these species [3], with relatively narrow posterior distributions (Table 1, S2 Figure). This difference may be driven by a lower mutation rate used in the present study. Consistent with this previous study, the estimated N_e for *H. cydno* (~5.3 M individuals) was larger than both *H. m. rosina* (~1.8 M) and *H. m. melpomene* (~3.9 M).

Overall, simulated summary statistics from the retained parameter combinations matched the observed summary statistic well (S3 Figure). To investigate the robustness of our conclusions, we repeated the ABC analysis using the split time priors from model F (S3 Table and see above). The general findings were largely the same (data not shown), indicating that our conclusions are not strongly influenced by the split time priors.

209 Discussion

210 The existence of distinct species that share genetic material through hybridisation, or
 211 have done so in the recent past, is no longer disputable. Genome-scale data have
 212 provided overwhelming evidence of pervasive gene flow between species, in taxa as
 213 diverse as fruit flies [39], flycatchers [2], and hominids [19,40]. However,
 214 understanding the timing of gene flow during speciation, and the importance of
 215 geographical isolation for species establishment remain difficult. Here, we combine a
 216 large-scale genomic dataset with Approximate Bayesian Computation (ABC) to
 217 reconstruct speciation in *Heliconius* butterflies. Our findings support recent studies
 218 showing abundant gene flow between *H. melpomene* and *H. cydno* [3,4], and suggest
 219 that gene flow has been ongoing for approximately half a million years (two million
 220 generations). However, we also find that hybridisation was rare or absent during the
 221 roughly the first million years of divergence between these species, a factor that may
 222 have played an important role in their establishment.

223 In order to reduce the parameter space explored, we used a fairly narrow joint prior
 224 distribution for the two split times, t_1 and t_2 . This joint distribution was inferred using
 225 mitochondrial sequence data, under the assumption that mitochondrial introgression
 226 between *H. cydno* and *H. melpomene* should be unlikely. This is reasonable given the
 227 fact that female hybrids are sterile [41]. Although one *H. melpomene* population from
 228 Colombia is known to carry *H. cydno*-like mitochondrial haplotypes [42], an analysis of
 229 all 449 unique haplotypes available on Genbank has not indicated any other
 230 mitochondrial introgression events between these species (see Materials and Methods).
 231 Our *cydno-melpomene* split time of 1.5 Ma is within the range of values inferred in
 232 previous analyses using IM-based methods [3,42,43]. In addition, posterior distributions
 233 for both split times were not strongly skewed toward the edge of the priors. It is
 234 nevertheless possible that the mitochondrial split times provide inadequate estimates of
 235 the nuclear split times. Even if this were the case, it is likely that our general conclusion

236 of reduced migration earlier in speciation will still hold. Indeed, when we repeated the
237 ABC analysis with a different, but overlapping, set of split time priors, the results were
238 largely unchanged.

239 Allopatric speciation is often the null hypothesis in speciation studies [44]. One
240 scenario that would be consistent with our results is that speciation began in allopatry,
241 possibly with the emerging species separated by the Andes mountains. Subsequent
242 range expansion of *H. melpomene* into the western Andes and Central America would
243 have lead leading to secondary contact. The lower N_e of *H. m. rosina* compared to *H. m.*
244 *melpomene* is consistent with range expansion from east to west, which has been
245 proposed previously [45]. However, our results may also be consistent with parapatric
246 speciation, which is probably more common in *Heliconius* [35,46]. In fact, most pairs of
247 sister species in the genus are sympatric or parapatric (Rosser et al. In Review).
248 Allopatric populations of extant species, for example those on Caribbean islands, tend
249 not to display phenotypic and ecological divergence from their mainland progenitors. In
250 contrast, many species, including *H. melpomene* and *H. cydno*, are divided into
251 numerous parapatric wing-pattern races across their mainland ranges. *H. melpomene*
252 and *H. cydno* are also partially segregated by altitude, so it is plausible that parapatric
253 adaptation to altitude in the Andes played a role in their speciation. The evolution of
254 strong assortative mating associated with wing pattern might then have led to nearly
255 complete reproductive isolation between the parapatric populations. Indeed, loci
256 affecting both mate preference and hybrid sterility are known to be physically linked to
257 wing patterning loci in these species, which might have enhanced reproductive isolation
258 following divergence in wing pattern [47,48]. Ecological divergence, most notably in
259 host plant use but perhaps also microhabitat preference [49], would then have followed
260 later, permitting sympatric coexistence without competition [34]. The inferred increase
261 in gene flow later in speciation might therefore reflect increased contact associated with
262 the transition from parapatric to sympatric ranges. One final piece of evidence for
263 parapatric speciation is the existence of several other species pairs that may represent an

intermediate step in this process. The best studied are *Heliconius himera* and *H. erato*, which are largely parapatric with only narrow zones of overlap. They are strongly differentiated genetically [50] and display assortative mating based on colour pattern [51]. However, they have not diverged in host plant usage [52], which perhaps prevents their sympatric coexistence through competitive exclusion.

Regardless of its cause, we can speculate that an initial period of reduced gene flow contributed to the formation of these species. Reduced gene flow can facilitate the accumulation of Dobzhansky-Muller incompatibilities [53,54], which would help to maintain species integrity even after the rate of hybridisation increased. For example, gene flow is minimal across the entire Z chromosome [4], consistent with a high density of incompatibility loci in this part of the genome. Interestingly, in these species there are also genetic associations between wing pattern, host preference and mate preference loci which likely facilitate coexistence in sympatry [47,48]. However it is unclear whether such associations have arisen since hybridisation became widespread, or whether they fortuitously pre-dated the period of extensive contact. Finer-scale analysis of the patterns of introgression across the genome, combined with mapping incompatibility loci and structural differences in the genome will help to dissect the various factors contributing to species persistence.

Despite our finding that hybridisation was rare or absent for approximately two thirds of the time since speciation, this nevertheless implies that the hybridisation has now been ongoing for around two million generations. Our model assumes a single change in the rate of gene flow, while it is highly probable that this rate has changed more gradually through time. The wide posterior distribution for the time of this transition is consistent with a gradual increase over perhaps hundreds of thousands of generations. There is also reason to believe that the rate of gene flow has recently begun to decrease. The existence of character displacement (sympatric males display stronger mate discrimination than allopatric males [55], suggests that selection may have acted to

reinforce reproductive isolation in sympatry. Nevertheless, there remains strong evidence that gene flow continues today, both in the occurrence of natural F1 hybrids [36], and in geographic patterns of shared variation [4]. Specifically, *H. cydno* samples from Panama share an excess of variation with *H. melpomene* samples from the same location compared to those from about 100 km away [4]. A recent study of the hybridising mouse subspecies *Mus musculus musculus* and *M. m. domestica* supported a similar scenario to that described here, with gene flow occurring over the last 25% of time since initial divergence, although at a lower rate (<0.2 ind/gen) [32]. Our estimate of ~ 0.84 migrants per generation represents the effective number of hybrids, but it is certain that the number of actual F1 hybrid butterflies produced exceeds this value considerably, for two reasons. Firstly, in accordance with Haldane's Rule, female (ZW) F1 hybrids are sterile [41], and therefore do not contribute to observed gene flow. Secondly, F1 hybrids are subject to increased predation owing to their non-mimetic wing patterns [56].

Finally, it is worth considering the consequences of continued hybridisation between these species. Although a whole-genome phylogeny groups the *H. melpomene* populations as monophyletic, currently 40% of 100 kb windows group *H. m. rosina* with *H. cydno*, to the exclusion of the French Guianan *H. m. melpomene* [4]. Nevertheless, these sympatric populations retain the phenotypic, behavioural and ecological traits specific to their respective species, implying that species integrity is surprisingly resilient to gene exchange. It is certain that gene flow is inhibited by selection in some genomic regions, most obviously the wing pattern loci. However, natural selection has also favoured the occasional exchange of wing pattern alleles between certain populations of these clades, producing the paired mimetic races of *H. melpomene* and *H. timareta* found on the eastern slopes of the Andes [38,57]. It seems likely that much of the genome is neutral with respect to gene flow, and that most of the signal seen here is due to neutral exchange of alleles in sympatry, although we have not attempted to test for evidence of adaptive introgression. It is therefore possible that

ongoing hybridisation, even at a low rate, might eventually lead to a situation where the majority of the genome clusters populations by geography rather than by species, making one or both species paraphyletic. It seems inevitable that genomic studies will reveal such species pairs in the near future, posing a challenge to species definitions based on aggregate genetic ancestry.

Materials and Methods

Samples and genotyping

We used published whole-genome resequence data for twelve wild-caught butterflies (S1 Table, data from Martin et al. [4], www.datadryad.com doi:10.5061/dryad.dk712). Details of the sequencing, mapping and genotyping procedures are described by Martin et al. [4]. Briefly, 100bp paired-end Illumina reads were mapped to the *H. melpomene* reference genome [38], version 1.1, using Stampy [58]. Local realignment around indels and genotyping were both performed using The Genome Analysis Toolkit (GATK) [59]. For the purpose of this study, we considered only intergenic SNPs, identified based on the *H. melpomene* genome annotation, version 1.1. CpG islands were identified using the program CpGcluster [60], and these sites were excluded. Only high quality genotype calls were considered. High quality genotypes met the following conditions: quality (QUAL) ≥ 30 , $10 \leq$ read depth per individual ≤ 200 , and GQ ≥ 30 for SNPs. Processed genotype calls data are available from www.datadryad.com doi:XXX.

Summary statistic

The summary statistic used for model fitting was a composite of the proportion of sites representing each of the possible combinations of bi-allelic genotypes among three diploid individuals, with one individual representing each population (Fig. 1B). For example, a SNP would be assigned the pattern 0-1-2, if the *H. cydno* individual was homozygous, carrying zero copies of the minor allele, the *H. m. rosina* individual was heterozygous, carrying one copy, and the *H. m. melpomene* individual was homozygous

with two copies of the minor allele. The counts of all patterns were then folded, such that major and minor alleles were not taken into account. For example the pattern 0-1-2 was taken as equivalent to 2-1-0. This gave 13 unique SNP patterns (Fig. 1B). Because four individuals were sampled from each population, the counts of each pattern were averaged over all 64 possible triplets with one individual from each population. Custom scripts used to calculate and plot summary statistics are available from www.datadryad.com doi:XXX. Since having too many summary statistics is a known problem with ABC, we used Partial Least Squares [implemented in the findPLS.r script in the ABC Toolbox [61]] to find the eight most informative linear combinations of the original summary statistics.

Model

A three-population model of isolation with migration was used (Fig. 1C). An ancestral population divides at time t_1 into two populations (corresponding to *H. cydno* and *H. melpomene*, respectively). At time t_2 , the *melpomene* population further divides into the *H. m. rosina* and *H. m. melpomene* races, which remain connected by limited gene flow at a constant rate M_m . The two *H. melpomene* populations, the *H. cydno* population, and the ancestral population, all have unique population sizes, but the size of the ancestral *melpomene* population is assumed to be the average of the two *melpomene* populations. Migration is allowed between *H. cydno* and the ancestral *H. melpomene* populations. Migration is allowed between *H. cydno* and *H. m. rosina* after the two *H. melpomene* populations diverge. Two distinct periods of hybridisation are modelled, with rates M_{cm1} and M_{cm2} . These two periods occupy the entire speciation time from t_1 to the present, and are divided at time t_d . Hence, Period 1 runs from t_1 to t_d and Period 2 from t_d to the present. The division between the periods, t_d , may fall anywhere between t_1 and the present. A constant mutation rate of 1.9×10^{-9} per site per generation was used. This corresponds to the estimated per-generation mutation rate for *H. melpomene* [62], corrected for weak purifying selection on intergenic regions by multiplying by the relative level of interspecific divergence at intergenic and putatively neutral four-fold

374 degenerate sites (data not shown). A generation time of 0.25 years was assumed [63].

375 Priors for split times

376 To reduce the dimensionality of the model, we used fairly narrow priors for the two
377 split times t_1 and t_2 (S2 Table). These were inferred using analysis of mitochondrial
378 sequence data, which should be resistant to gene flow between these taxa. This
379 assumption was first tested with a Maximum Likelihood analysis of all 847 publicly
380 available sequences (449 unique haplotypes) in RAxML v.8 [64]. This identified a
381 single, previously known [42] case of mitochondrial introgression between these
382 species, which does not involve the populations considered here. Sequence data for
383 1606 bp of *Col/II* for 125 samples from several populations of *H. melpomene*, *H. cydno*
384 and the outgroup silvaniform clade were obtained from Genbank and the data of from
385 Martin et al. [4], and aligned with MUSCLE [65]. Strict and relaxed molecular clock
386 models and codon-partitioning schemes were fitted to the data in BEAST v. 1.8.
387 [66] and compared with a posterior analog of AIC (AICM) and Bayes Factors calculated
388 by the Stepping Stone Analysis [67] (S2 and S3 Table). For root-calibrated analyses, the
389 split time between the *H. melpomene* and Silvaniform caldes inferred by Kozak et al.
390 [68] was used. For fixed rate analyses a mutation rate of 0.0024 per million years was
391 used, as inferred under a relaxed-clock model applied to the complete *Heliconiini*
392 alignment of Kozak et al. [68]. While the exact split dates varied, all approaches
393 converged on the same topology and a similar ratio of split times t_1/t_2 (S3 Table). Bayes
394 Factors and AICM [67] favoured a strict clock model (model E, S2 Table) with a
395 separate partition for third codon positions. We present results obtained using the split
396 times from model E, although we also tested the split times from model F (S3 Table).
397 The resulting joint posterior distributions for the two split times formed the priors for
398 the ABC simulations, with pairs of times for t_1 and t_2 being drawn together from this
399 joint distribution.

400 **Model fitting using ABC**

401 Approximate Bayesian Computation (ABC) was used to estimate parameters of the
402 model. Briefly, ABC fits a model by evaluating the distance between observed and
403 simulated summary statistics, allowing the estimation of posterior probability
404 distributions without calculating likelihood functions.

405 Uniform priors were used for all parameters except for the split times t_1 and t_2 (see
406 above). Two million parameter combinations were generated over the parameter space
407 by sampling from the prior distributions randomly and independently (except for
408 parameters t_1 and t_2 , which had a joint prior distribution). A custom program (written in
409 C/C++) was then used to simulate 100,000 unlinked SNPs from our model under the
410 standard coalescent framework [69] for each sampled parameter set, and then to
411 calculate the summary statistic. SNPs were simulated independently (i.e. unlinked) as
412 the composite summary statistic used here, which is based on the genome-wide joint
413 frequency spectrum, should not be strongly influenced by linkage disequilibrium,
414 especially given the fairly rapid decline in LD in *Heliconius melpomene* [4]. Using the
415 standard ABC method, we used the Euclidian distance between simulated and observed
416 values to identify parameter combinations that fit the data well. We used a cut-off of
417 0.01 for accepting parameter combinations, yielding 27377 good parameter
418 combinations for ABC. To account for variation in goodness of fit obtained among
419 retained parameters, the distribution of retained parameters was adjusted using the
420 General Linear Model method of [70], as implemented by ABCestimator of the ABC
421 Toolbox [61].

422 **Acknowledgements**

423 We thank Nick Barton, Aylwyn Scally and Konrad Lohse, for helpful comments on
424 earlier drafts of the manuscript. Computational support was provided by Jenny Barna,
425 School of Life Sciences, University of Cambridge.

426 References

- 427 1. Kulathinal RJ, Stevison LS, Noor M a F (2009) The genomics of speciation in *Drosophila*:
428 diversity, divergence, and introgression estimated using low-coverage genome
429 sequencing. *PLoS Genet* 5: e1000550. doi:10.1371/journal.pgen.1000550.
- 430 2. Ellegren H, Smeds L, Burri R, Olason PPI, Backström N, et al. (2012) The genomic
431 landscape of species divergence in *Ficedula* flycatchers. *Nature* 491: 756–760.
432 doi:10.1038/nature11584.
- 433 3. Kronforst MRR, Hansen MEB, Crawford NGG, Gallant JRR, Zhang W, et al. (2013)
434 Hybridization reveals the evolving genomic architecture of speciation. *Cell Rep* 5: 666–
435 677. doi:10.1016/j.celrep.2013.09.042.
- 436 4. Martin SH, Dasmahapatra KK, Nadeau NJ, Salazar C, Walters JR, et al. (2013) Genome-wide
437 evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Res* 23: 1817–
438 1828. doi:10.1101/gr.159426.113.
- 439 5. Fontaine MC, Pease JB, Steele A, Waterhouse RM, Neafsey DE, et al. (2014) Extensive
440 introgression in a malaria vector species complex revealed by phylogenomics. *Science*
441 347: 1258524 – . doi:10.1126/science.1258524.
- 442 6. Lamichhaney S, Berglund J, Almén MS, Maqbool K, Grabherr M, et al. (2015) Evolution of
443 Darwin's finches and their beaks revealed by genome sequencing. *Nature advance on*.
444 doi:10.1038/nature14181.
- 445 7. Maynard Smith J (1966) Sympatric speciation. *Am Nat* 100: 637–650.
- 446 8. Lande R (1982) Rapid Origin of Sexual Isolation and Character Divergence in a Cline.
447 *Evolution* (N Y) 36: 213–223.
- 448 9. Slatkin M (1982) Pleiotropy and Parapatric Speciation. 36: 263–270.
- 449 10. Kirkpatrick M, Ravigné V (2002) Speciation by natural and sexual selection: models and
450 experiments. *Am Nat* 159 Suppl : S22–S35. doi:10.1086/338370.
- 451 11. Navarro A, Barton NH (2003) Accumulating postzygotic isolation genes in parapatry: a new
452 twist on chromosomal speciation. *Evolution* 57: 447–459.
- 453 12. Gavrillets S (2004) Fitness landscapes and the origin of species. Princeton University Press.
- 454 13. Kirkpatrick M, Barton N (2006) Chromosome inversions, local adaptation and speciation.
455 *Genetics* 173: 419–434. doi:10.1534/genetics.105.047985.
- 456 14. Van Doorn GS, Dieckmann U, Weissing FJ (2004) Sympatric speciation by sexual selection:
457 a critical reevaluation. *Am Nat* 163: 709–725. doi:10.1086/383619.
- 458 15. Mallet J, Meyer a., Nosil P, Feder JL (2009) Space, sympatry and speciation. *J Evol Biol* 22:
459 2332–2341. doi:10.1111/j.1420-9101.2009.01816.x.
- 460 16. Barluenga M, Stölting KN, Salzburger W, Muschick M, Meyer A (2006) Sympatric
461 speciation in Nicaraguan crater lake cichlid fish. *Nature* 439: 719–723.
462 doi:10.1038/nature04325.
- 463 17. Wu C (2001) The genic view of the process of speciation. *J Evol Biol* 14: 851–865.
- 464 18. Lohse K, Harrison RJ, Barton NH (2011) A general method for calculating likelihoods under
465 the coalescent process. *Genetics* 189: 977–987. doi:10.1534/genetics.111.129569.
- 466 19. Lohse K, Frantz L a F (2014) Neandertal admixture in eurasia confirmed by maximum-
467 likelihood analysis of three genomes. *Genetics* 196: 1241–1251.
468 doi:10.1534/genetics.114.162396.
- 469 20. Wilkinson-Herbots HM (2008) The distribution of the coalescence time and the number of

- pairwise nucleotide differences in the “isolation with migration” model. *Theor Popul Biol* 73: 277–288. doi:10.1016/j.tpb.2007.11.001.
21. Wang Y, Hey J (2010) Estimating divergence parameters with small samples from a large number of loci. *Genetics* 184: 363–379. doi:10.1534/genetics.109.110528.
22. Zhu T, Yang Z (2012) Maximum likelihood implementation of an isolation-with-migration model with three species for testing speciation with gene flow. *Mol Biol Evol* 29: 3131–3142. doi:10.1093/molbev/mss118.
23. Nielsen R, Wakeley J (2001) Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics* 158: 885–896.
24. Hey J, Nielsen R (2004) Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* 167: 747–760.
25. Hey J, Nielsen R (2007) Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proc Natl Acad Sci U S A* 104: 2785–2790. doi:10.1073/pnas.0611164104.
26. Hey J (2010) Isolation with Migration Models for More Than Two Populations. *Mol Biol Evol* 27: 905–920.
27. Becquet C, Przeworski M (2009) Learning about modes of speciation by computational approaches. *Evolution (N Y)* 63: 2547–2562.
28. Strasburg JL, Rieseberg LH (2011) Interpreting the estimated timing of migration events between hybridizing species. *Mol Ecol* 20: 2353–2366. doi:10.1111/j.1365-294X.2011.05048.x.
29. Sousa VC, Grelaud A, Hey J (2011) On the nonidentifiability of migration time estimates in isolation with migration models. *Mol Ecol* 20: 3956–3962.
30. Strasburg JL, Rieseberg LH (2013) Methodological challenges to realizing the potential of hybridization research. *J Evol Biol* 26: 259–260. doi:10.1111/jeb.12006.
31. Beaumont M a. (2010) Approximate Bayesian Computation in Evolution and Ecology. *Annu Rev Ecol Evol Syst* 41: 379–406. doi:10.1146/annurev-ecolsys-102209-144621.
32. Duvaux L, Belkhir K, Boulesteix M, Boursot P (2011) Isolation and gene flow: inferring the speciation history of European house mice. *Mol Ecol* 20: 5248–5264. doi:10.1111/j.1365-294X.2011.05343.x.
33. Li J-W, Yeung CKL, Tsai P-W, Lin R-C, Yeh C-F, et al. (2010) Rejecting strictly allopatric speciation on a continental island: prolonged postdivergence gene flow between Taiwan (*Leucodiotron taewanus*, Passeriformes Timaliidae) and Chinese (*L. canorum canorum*) *hwameis*. *Mol Ecol* 19: 494–507. doi:10.1111/j.1365-294X.2009.04494.x.
34. Mallet J, McMillan WO, Jiggins CD (1998) Mimicry and warning color at the boundary between races and species. *Endless forms species Speciat*: 390–403.
35. Jiggins C (2008) Ecological speciation in mimetic butterflies. *Bioscience* 58: 541–548.
36. Mallet J, Beltrán M, Neukirchen W, Linares M (2007) Natural hybridization in heliconiine butterflies: the species boundary as a continuum. *BMC Evol Biol* 7: 28. doi:10.1186/1471-2148-7-28.
37. Rosser N, Phillimore AB, Huertas B, Willmott KR, Mallet J (2012) Testing historical explanations for gradients in species richness in heliconiine butterflies of tropical America. *Biol J Linn Soc* 105: 479–497. doi:10.1111/j.1095-8312.2011.01814.x.
38. The Heliconius Genome Consortium 1 (2012) Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* 487: 94–98. doi:10.1038/nature11041.

39. Garrigan D, Kingan SSB, Geneva AJ, Andolfatto P, Clark AG, et al. (2012) Genome sequencing reveals complex speciation in the *Drosophila simulans* clade. *Genome* ... 22: 1499–1511. doi:10.1101/gr.130922.111.
40. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, et al. (2010) A draft sequence of the Neandertal genome. *Science* 328: 710–722. doi:10.1126/science.1188021.
41. Naisbit RE, Jiggins CD, Linares M, Salazar C, Mallet J (2002) Hybrid Sterility, Haldane's Rule and Speciation in. *Race* 1526: 1517–1526.
42. Salazar C, Jiggins CD, Taylor JE, Kronforst MR, Linares M (2008) Gene flow and the genealogical history of *Heliconius heurippa*. *BMC Evol Biol* 8: 132. doi:10.1186/1471-2148-8-132.
43. Bull V, Beltrán M, Jiggins CD, McMillan WO, Bermingham E, et al. (2006) Polyphyly and gene flow between non-sibling *Heliconius* species. *BMC Biol* 4: 11. doi:10.1186/1741-7007-4-11.
44. Coyne JA, Orr HA, others (2004) *Speciation*. Sinauer Associates Sunderland, MA.
45. Quek S-P, Counterman B a, Albuquerque de Moura P, Cardoso MZ, Marshall CR, et al. (2010) Dissecting comimetic radiations in *Heliconius* reveals divergent histories of convergent butterflies. *Proc Natl Acad Sci U S A* 107: 7365–7370. doi:10.1073/pnas.0911572107.
46. Jiggins CD, Estrada C, Rodrigues a (2004) Mimicry and the evolution of premating isolation in *Heliconius melpomene* Linnaeus. *J Evol Biol* 17: 680–691. doi:10.1111/j.1420-9101.2004.00675.x.
47. Merrill RM, Van Schooten B, Scott J a, Jiggins CD (2011) Pervasive genetic associations between traits causing reproductive isolation in *Heliconius* butterflies. *Proc Biol Sci* 278: 511–518. doi:10.1098/rspb.2010.1493.
48. Merrill RM, Naisbit RE, Mallet J, Jiggins CD (2013) Ecological and genetic factors influencing the transition between host-use strategies in sympatric *Heliconius* butterflies. *J Evol Biol* 26: 1959–1967. doi:10.1111/jeb.12194.
49. Estrada C, Jiggins CD (2002) Patterns of pollen feeding and habitat preference among *Heliconius* species. *Ecol Entomol* 27: 448–456. doi:10.1046/j.1365-2311.2002.00434.x.
50. Jiggins CD, McMillan WO, King P, Mallet J (1997) The maintenance of species differences across a *Heliconius* hybrid zone. *Heredity (Edinb)* 79: 495–505. doi:10.1038/sj.hdy.6882230.
51. Merrill RM, Chia A, Nadeau NJ (2014) Divergent warning patterns contribute to assortative mating between incipient *Heliconius* species. *Ecol Evol* 4: 911–917. doi:10.1002/ece3.996.
52. Jiggins CD, McMillan WO, Mallet J (1997) Host plant adaptation has not played a role in the recent speciation of *Heliconius himera* and *Heliconius erato*. *Ecol Entomol* 22: 361–365. doi:10.1046/j.1365-2311.1997.00067.x.
53. Kondrashov AS (2003) Accumulation of Dobzhansky-Muller incompatibilities within a spatially structured population. *Evolution* 57: 151–153. doi:10.1111/j.0014-3820.2003.tb00223.x.
54. Bank C, Bürger R, Hermisson J (2012) The limits to parapatric speciation: Dobzhansky-Muller incompatibilities in a continent-island model. *Genetics* 191: 845–863. doi:10.1534/genetics.111.137513.
55. Jiggins CD, Naisbit RE, Coe RL, Mallet J (2001) Reproductive isolation caused by colour pattern mimicry. *Nature* 411: 302–305. doi:10.1038/35077075.
56. Merrill RM, Wallbank RWR, Bull V, Salazar PC a, Mallet J, et al. (2012) Disruptive

- ecological selection on a mating cue. *Proc Biol Sci.* doi:10.1098/rspb.2012.1968.
57. Pardo-Diaz C, Salazar C, Baxter SW, Merot C, Figueiredo-Ready W, et al. (2012) Adaptive introgression across species boundaries in *Heliconius* butterflies. *PLoS Genet* 8: e1002752. doi:10.1371/journal.pgen.1002752.
58. Lunter G, Goodson M (2011) Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res* 21: 936–939. doi:10.1101/gr.111120.110.
59. DePristo M a, Banks E, Poplin R, Garimella K V, Maguire JR, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43: 491–498. doi:10.1038/ng.806.
60. Hackenberg M, Previti C, Luque-Escamilla PL, Carpena P, Martínez-Aroza J, et al. (2006) CpGcluster: a distance-based algorithm for CpG-island detection. *BMC Bioinformatics* 7: 446. doi:10.1186/1471-2105-7-446.
61. Wegmann D, Leuenberger C, Neuenschwander S, Excoffier L (2010) ABCtoolbox: a versatile toolkit for approximate Bayesian computations. *BMC Bioinformatics* 11: 116.
62. Keightley PD, Pinharanda A, Ness RW, Simpson F, Dasmahapatra KK, et al. (2014) Estimation of the Spontaneous Mutation Rate in *Heliconius melpomene*. *Mol Biol Evol*: 1–5. doi:10.1093/molbev/msu302.
63. Mallet J (1986) Hybrid zones of *Heliconius* butterflies in Panama and the stability and movement of warning colour dines. *Heredity (Edinb)* 56: 191–202.
64. Stamatakis A (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30: 1312–1313. doi:10.1093/bioinformatics/btu033.
65. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32: 1792–1797. doi:10.1093/nar/gkh340.
66. Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol* 29: 1969–1973.
67. Baele G, Lemey P, Bedford T, Rambaut A, Suchard MA, et al. (2012) Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Mol Biol Evol* 29: 2157–2167. doi:10.1093/molbev/mss084.
68. Kozak KM, Wahlberg N, Neild A, Dasmahapatra KK, Mallet J, et al. (2015) Multilocus Species Trees Show the Recent Adaptive Radiation of the Mimetic. *Syst Biol*.
69. Kingman J (1982) The coalescent. *Stoch Proc Appl* 13: 235–248.
70. Leuenberger C, Wegmann D (2010) Bayesian computation and model selection without likelihoods. *Genetics* 184: 243–252.

Financial Disclosure

This work was funded by the Biotechnology and Biological Sciences Research Council (www.bbsrc.ac.uk/) grants BB/H01439X/1 to CDJ and H005854/1 to AM; and the European Research Council (erc.europa.eu/) SpeciationGenomics grant to CDJ. SHM is funded by St John's College, Cambridge (www.joh.cam.ac.uk/). KMK was supported by the Herchel Smith and Balfour Studentships to the Department of

603 Zoology, University of Cambridge (www.zoo.cam.ac.uk/). The funders had no role in
604 study design, data collection and analysis, decision to publish, or preparation of the
605 manuscript.

606 List of supplemental files

607 **S1 Table. Sample information and sequencing read depth**

608 **S2 Table. Model comparison of various strategies for estimating the divergence**
609 **times in BEAST.** Models E and F are equally good based on Bayes Factors. Log Bayes
610 Factors (BF) are calculated based on the log Marginal Likelihood estimates (MLE) from
611 Path Sampling (PS) and Stepping Stone Analysis (SSA) (Baele et al. 2012). The
612 molecular clock rates were calibrated by either modelling the age of the root or setting a
613 constant rate of substitution. UCLD= Uncorrelated Lognormal clock.

614 **S3 Table. Parameter values for the three best Bayesian models of divergence**
615 **between the *CoI/II* sequences.**

616 **S1 Figure. Bayesian phylogeny of the *Cytochrome Oxidase I/II (CoI/II)* haplotypes**
617 **from the *Heliconius melpomene* / *H. cydno* clade.** This tree was estimated under the
618 relaxed molecular clock model (F) with calibrated age of the root. The sampling is based
619 on a balanced design with 25 samples per group, including *H. timareta* as the sister
620 species of *H. cydno*. Blue: *H. melpomene* from Central America; green: *H. melpomene*
621 from French Guiana (clade including allopatric samples used in the analysis of the
622 nuclear genomes); pink: *H. cydno*; grey: Silvaniform outgroups. Time scale in millions of
623 years, bars represent the 95% HPD intervals around split ages. Samples with whole
624 mitogenome data from Martin et al. (2013) indicated in capital letters. NCBI GI numbers
625 are provided for all sequences obtained from GenBank.

626 **S2 Figure. Posterior density plots inferred by ABC for the 10 model parameters.**
627 Parameters labels match Fig. 1 in the main text, with the three time parameters given in
628 the left-hand column (A-C), migration rates in the middle column (D-F), and population
629 sizes in the right-hand column (G-J). Grey shading indicates the 90% highest posterior
630 density (HPD) interval, and posterior means are indicated by vertical dashed lines.

631 **S3 Figure. Summary statistics from retained parameter combinations compared to**
632 **the observed values.** Patterns 1-13 correspond respectively to 0-0-1, 0-0-2, 0-1-0, 0-1-1,

633 0-1-2, 0-2-0, 0-2-1, 2-0-0, 1-0-0, 1-0-1, 1-0-2, 1-1-0 and 1-1-1. Box plots indicate the
634 distribution of simulated values from the 27377 retained parameter combinations. Red
635 points indicate outliers (outside of the 10th and 90th percentiles). Yellow diamonds
636 indicate observed frequencies, as in Fig. 1B. Here the y-axis indicates the the scaled
637 frequency of each pattern, in units of million generations. This can be obtained by
638 dividing the observed absolute frequency of each pattern (as in Fig. 1B) by the mutation
639 rate per site per million years.