Full-likelihood analysis of genomic data clarifies a complex history of species divergence and introgression in the *erato-sara* group of *Heliconius* butterflies

Yuttapong Thawornwattana[1], Fernando A. Seixas[1], Ziheng Yang[2*], James Mallet[1*]

[1] Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA

[2] Department of Genetics, Evolution and Environment, University College London, London WC1E 6BT, UK

Corresponding authors: z.yang@ucl.ac.uk, jmallet@oeb.harvard.edu

Running head: INTROGRESSION HISTORY OF ERATO-SARA HELICONIUS

Abstract

Introgression plays a key role in adaptive evolution and species diversification in many groups of species including *Heliconius* butterflies. However, frequent hybridization and gene flow between species makes estimation of the species phylogeny challenging, especially for rapidly speciating species within adaptive radiations. Here, we analyze the genomic sequences from six members of the *erato-sara* clade of *Heliconius* butterflies to infer the species phylogeny and cross-species introgression events. To avoid heterozygote phasing errors in haploid sequences commonly produced by genome assembly methods, we processed and compiled unphased diploid sequence alignments, with the analytical methods averaging over uncertainties in heterozygote phase resolution. We use likelihood-based methods under the multispecies coalescent (MSC) model with and without gene flow to accommodate random fluctuations in genealogical history across the genome due to deep coalescence. There is robust evidence of introgression across the genome, both among distantly related species deep in the phylogeny and between sister species in shallow parts of the tree. We obtain estimates of population parameters such as introgression times and probabilities, species divergence times, and population sizes for modern and ancestral species. We confirm ancestral gene flow between the *sara* clade and an ancestral population of *H. telesiphe*, a likely hybrid origin of *H. hecalesia*, and gene flow between the sister species *H. erato* and *H. himera*. Our approach also demonstrates how introgression among ancestral species can explain the history of two chromosomal inversions deep in the phylogeny of the group. For the first time, we not only test for the presence of cross-species gene flow, but also estimate its direction, timing and magnitude, extracting rich historical information of species divergence and gene flow from genomic data.

Keywords: 3S, BPP, *erato*, gene flow, *Heliconius*, introgression, multispecies coalescent, inversion

INTRODUCTION

Thanks to increasing availability of genomic data and advances in analytical methods (Sousa and Hey 2013; Payseur and Rieseberg 2016), hybridization or introgression has been detected in a variety of species including *Anopheles* mosquitoes (Fontaine et al. 2015), *Panthera* cats (Figueiró et al. 2017) and cichlid fishes (Malinsky et al. 2018), as well as *Heliconius* butterflies (Dasmahapatra et al. 2012; Jay et al. 2018; Edelman et al. 2019; Kozak et al. 2021). Introgression is increasingly recognized as an important process that introduces genetic diversity into the receiving population and potentially contributes to adaptive evolution (Mallet et al. 2016; Taylor and Larson 2019; Edelman and Mallet 2021). Whether gene flow occurs and what role it plays during the process of species divergence is increasingly seen as important to a fuller understanding of speciation (Pinho and Hey 2010; Feder et al. 2012).

Heliconius is a rapidly radiating group of butterflies in tropical America. They are unpalatable to predators and are perhaps best known for mimicry in which a single species can have many different wing coloration patterns across its geographical range (classified as subspecies) while multiple unrelated sympatric species have converged to share the same wing pattern in local regions as a common warning sign to deter predators (Bates 1862; Müller 1879). The genus *Heliconius* comprises two major clades, the *erato-sara* clade and the *melpomene*-silvaniform clade, which diverged around 10–12 million years ago (Ma) in the Miocene (Kozak et al. 2015). Natural hybridization among species is well-documented within each clade (Mallet et al. 2007). The prevalence of introgression between species coupled with rapid radiation and diversification of species and geographic races makes estimation of the species phylogeny challenging. As a result, our understanding of the history of species divergence and introgression in *Heliconius* remains limited. This work aims to resolve the history of key speciation events and cross-species introgression in the *erato-sara* clade.

Most previous studies of the genus *Heliconius*, including the *erato* group, have focused on evolutionary relationships and gene flow at specific regions of the genome, especially the colour pattern loci responsible for phenotypic variation in the mimetic wing patterns, typically in a few species, and mainly in the *melpomene*-silvaniform clade where gene flow appears to be more prevalent (Dasmahapatra et al. 2012; Nadeau et al. 2013; Martin and Van Belleghem 2017; Jay et al. 2018; Balaban et al. 2019). Other studies have focused on wing-colour pattern loci between two species with comimic races, *H. erato* (*erato-sara* clade) and *H. melpomene* (*melpomene*-silvaniform clade) (Hines et al. 2011; Reed et al. 2011).

Earlier molecular phylogenetic studies of *Heliconius* were based on a small number of loci (Brower 1994; Brower and Egan 1997; Beltrán et al. 2002; Beltrán et al. 2007; Kozak et al. 2015), revealing variation in gene genealogies among loci. In particular, Kozak et al. (2015) employed BUCKy (Larget et al. 2010) and *BEAST (Heled and Drummond 2010) to account for the heterogeneity of gene genealogies across loci. Hybridization and introgression were acknowledged but not directly accounted for in their analytical methods. Kozak et al. (2021) analyzed genome-wide coding loci from >100 individuals from 40 *Heliconius* species to estimate the species tree using approximate multispecies coalescent (MSC) methods such as ASTRAL (Mirarab et al. 2014) and MP-EST (Liu et al. 2010), and to test for introgression using a range of summary methods including the *D* statistic (or ABBA-BABA test) (Patterson et al. 2012), f_4 (Reich et al. 2009), and an approximate multispecies coalescent with introgression (MSci) method PHYLONET/MPL (Yu and Nakhleh 2015). The *D* and f_4 statistics use site-pattern counts averaged over the genome, ignoring information in genealogical fluctuations across the genome, while PHYLONET/MPL takes gene trees as input without accounting for phylogenetic reconstruction errors and uncertainties. Van Belleghem et al. (2017) estimated the species phylogeny of the *erato* clade from concatenated autosomal SNP data, focusing on wing colour pattern loci among different geographic races of *H. erato*. Recently, Edelman et al. (2019) conducted phylogenetic analyses with sixteen new genome assemblies of *Heliconius* species using sliding windows and ASTRAL (Mirarab et al. 2014). Introgression was tested using *D* statistics

(Patterson et al. 2012) and a new test called QuIBL based on average internal branch lengths in estimated triplet gene trees, while MSci network models were inferred using PHYLONET/MCMC_SEQ (Wen and Nakhleh 2018). Summary methods such as *D* statistics and QuIBL are based on genome-wide averages and ignore information from genealogical variation across the genome in multi-locus sequence data; they ignore gene flow between sister species and have limited ability to characterize the direction, timing, and magnitude of gene flow (as measured by the migration rate or introgression probability). PHYLONET/MCMC_SEQ is a Bayesian Markov chain Monte Carlo (MCMC) method based on sequence alignments, but is computationally applicable only to a small number of loci. In Edelman et al. (2019), PHYLONET produced inferences with considerable uncertainty of species tree topology and timing and directions of introgression. Finally, Massardo et al. (2020) re-analyzed genome data from the *erato-sara* clade after adding two additional genomes using approaches similar to those of Edelman et al. (2019).

The sliding-window analysis is a useful descriptive tool for exploratory analysis of genomic data (Martin and Van Belleghem (2017) but could run into difficulties when used in inference. The proportions of gene trees among the nonoverlapping sliding windows do not have a clear biological interpretation, as they depend on window size and reflect phylogenetic reconstruction errors, stochastic coalescent fluctuations, as well as population-level processes such as species divergences and cross-species introgression. To use estimated gene trees from the windows for inference, one must consider the likelihood for the observed gene tree distributions from the sliding windows under alternative models (for example, a complete isolation model versus a model of continuous migration after species divergence). Even under a complete-isolation model, large fluctuations in gene tree topologies and sequence divergences are expected due to the natural coalescent fluctuations (Barton 2006), and the gene-tree discordance can range anywhere from near 0 to near 1, depending on the parameters in the model. Introgression and phylogenetic reconstruction errors add further variation. Furthermore, estimated gene-tree proportions are sensitive to the choice of window size: with small windows, the results may be affected by phylogenetic errors, whereas with large windows, one or two gene trees will dominate due to the averaging effects, even though the average gene tree for the sliding window may differ from the species tree (Roch and Steel 2015). Indeed, a previous sliding-window analysis appeared to have produced incorrect species trees for the *Anopheles gambiae* group of African mosquitoes (Fontaine et al. 2015; Thawornwattana et al. 2018). It is thus worth asking whether similar systematic biases affected the sliding-window analysis in Edelman et al. (2019).

A standard practice in genome assembly has been to "haploidify" the diploid sequence to a single haploid sequence for each genome, with each heterozygous site represented by only one nucleotide base, e.g., the one with more reads. This loses half of the information, and, worse still, produces chimeric haplotypes that do not exist in nature because the genotypic phase at heterozygous sites is resolved effectively at random. Recent simulation studies (Andermann et al. 2019; Huang et al. 2021) found that such haploid consensus sequences can lead to serious biases in downstream phylogenomic analyses if the species tree is shallow, with species divergence times comparable to coalescent times. The impact of phasing errors in haploidified sequences on estimation of species trees and inference of gene flow in *Heliconius* is unknown. In this paper, we processed the raw reads for the genomic data of Edelman et al. (2019) for six species in the *erato-sara* clade to compile alignments of unphased diploid genomic sequences, and used them as input data for analysis using BPP. The BPP program (Yang 2015; Flouri et al. 2018) implements a method that analytically integrates over all possible phase resolutions of heterozygotes, weighting them according to their likelihood based on the sequence alignment at the locus (Gronau et al. 2011). In simulations, this approach performed nearly as well as analysis of fully phased diploid genomes (which could be generated, for instance, by costly single-molecule cloning and sequencing) (Huang et al. 2021).

We infer the species phylogeny and introgression history of the six species. Unlike previous

attempts, we used two coalescent-based full-likelihood phylogenetic approaches that explicitly account for deep coalescence and introgression as sources of genealogical variation across the genome. One approach is based on a multispecies coalescent model with introgression (MSci) (Degnan 2018; Wen and Nakhleh 2018; Zhang et al. 2018), implemented in the program BPP (Flouri et al., 2020). In this approach, introgression is modelled as discrete events that occur at particular time points in the past. Another approach is based on an isolation-with-migration (IM) model (Hey and Nielsen 2004; Hey 2010) implemented in the program 3s (Zhu and Yang 2012; Dalquen et al. 2017), which allows for continuous migration at a constant rate per generation after species divergence. Advantages of using full likelihood methods over approximate coalescent methods or summary statistics include making full use of information in the sequence data and properly accounting for uncertainty in gene trees (Xu and Yang 2016). These methods allow us not only to infer the presence of gene flow, but also to estimate its direction, timing and magnitude, along with species divergence times and effective population sizes. Estimation of such important evolutionary parameters from genome-scale sequence data can provide powerful insights into the divergence history of species, and a basis for further investigations of the evolution of adaptive traits of interest.

METHODS

Genome sequence data and genotyping

We generated new alignments of unphased diploid sequences by processing the raw sequencing read data of Edelman et al. (2019). The base-calling error rate (or per-base sequencing error rate) was estimated from the genomic data. We developed a procedure to estimate the genotype-calling error as a function of base-calling error and read depth, and used it to determine cutoffs on read depth for data filtering.

We used the raw sequencing data from six *Heliconius* species in the *erato-sara* clade and an outgroup (*H. melpomene*) (**Table S1**), previously generated by Edelman et al. (2019). Sequencing reads were filtered for Illumina adapters using cutadapt v1.8.1 (Martin, 2011) and then mapped to the chromosome-level genome assembly of *H. erato demophoon* v1 available from lepbase.org, using BWA mem v0.7.15 (Li 2013) with default parameters and marking short split hits as secondary. Mapped reads were sorted and duplicate reads removed using sambamba v0.6.8 (Tarasov et al. 2015). Realignment around indels was performed with the Genome Analysis Toolkit (GATK) v3.8 *RealignerTargetCreator* and *IndelRealigner* modules (McKenna et al. 2010; DePristo et al. 2011), in order to reduce the number of indel miscalls. Read depth and other relevant read alignment quality control metrics were computed using QualiMap v2.2.1 (Okonechnikov et al. 2016).

Genotype calling was performed on each individual separately with bcftools v1.5 (Li et al. 2009) *mpileup* and *call* modules (Li 2011), using the multiallelic-caller model (*call -m*) and requiring a minimum base and mapping quality of 20. Genotype calls were filtered using the bcftools *filter* module. Both invariant and variant sites were required to have a minimum quality score (QUAL) of 20. Furthermore, we required that each genotype had a genotype quality score (GQ) ≥20 and a read depth (DP) satisfying max(1/2*meanDP, 20) ≤ DP ≤ 2*meanDP where meanDP is the average read depth of the sample. The meanDP-based filters were used to reduce erroneous calls when the read depths were too low or too high from the sample average, accounting for variation of sequencing depths across individuals. The DP ≥ 20 filter was chosen to minimize genotyping error rates (see next section) while retaining a sufficiently large number of loci across the genome. At our estimated base-calling error (~0.1%), this filter achieved the genotype-calling error of <0.05% (see next section) (**Fig. 1**, **Tables S2–S3**). For the female Z chromosome, we used DP ≥ 10 instead since only one chromosome copy was present. All genotypes that did not fulfill these requirements or were located within 5-bp of an indel were recoded as missing data.

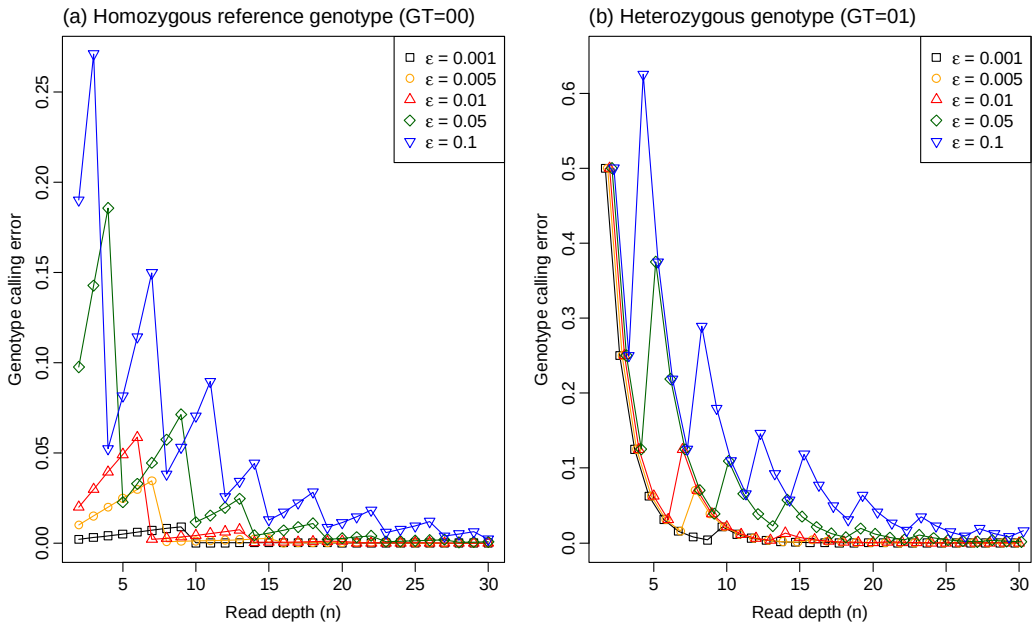


Figure 1. The genotype-calling error rate for a given base-calling error (ε) and read depth (n) when the true genotype is (**a**) a homozygote for the reference allele (GT = 00) or (**b**) a heterozygote (GT = 01). Note that the genotype-calling error does not decrease monotonically with the increase in n when ε is fixed, or with the reduction in ε when n is fixed. The calculation is performed using a C program written by Z.Y. that implements the ML method of Li (2011) and calculates its error rate.

Analysis of base-calling and genotype-calling error rates to guide data compilation

The base-calling error rate was estimated from the proportion of non-matching bases in homozygous genotype calls from reads mapped to the *H. melpomene* reference genome, which was more complete than the *H. erato demophoon* reference. Only positions with homozygous genotype calls with read depth (DP) ≥ 50 were retained. Because mapping errors can bias our estimates of base-calling error rate, sites overlapping repetitive regions were excluded. For sites passing those filters, we recorded the read depth (DP) and the number of reads supporting each of the reported alleles (AD), with the difference to be the number of erroneous base calls. The base-calling error was calculated as the ratio of the number of erroneous base calls over the read depth, both summed across all sites passing filters. The calculation was done for homozygous reference genotype (GT = 00) and homozygous alternative genotype (GT = 11).

We then analyzed the genotype-calling error rate to determine a suitable cutoff for the read depth, following the maximum likelihood (ML) method for genotype-calling of Li (2011). The results were used to guide our choice of the filter (DP ≥ 20), to achieve a genotype-calling error of <0.05%.

Given the base-calling error rate (ε) and the read depth (n), the genotype-calling error rate (e) can be calculated by following the ML procedure of genotype calling of Li (2011). Here we assume that ε is the same among the reads and is independent of the true base. Given the data of k 1s and ($n - k$) 0s among the n reads, the likelihoods for the three genotypes (GT = 00, 01, and 11) are given by the binomial probabilities as

$$L(00|k) = \Pr(k|\mathrm{GT}=00) = \binom{n}{k}(1-\varepsilon)^{n-k}\varepsilon^{k},$$
$$L(01|k) = \Pr(k|\mathrm{GT}=01) = \binom{n}{k}\left(\tfrac{1}{2}\right)^{n}, \quad (1)$$
$$L(11|k) = \Pr(k|\mathrm{GT}=11) = \binom{n}{k}(1-\varepsilon)^{k}\varepsilon^{n-k}$$

(Li 2011). The genotype achieving the highest likelihood is the called (inferred) genotype. The genotype-calling error is an average over the possible read outcomes (i.e., over k = 0, ..., n)

$$e_{\mathrm{GT=00}} = \sum_{k=0}^{n} \Pr(k|\mathrm{GT}=00) \times I_{GT\neq 00}, \quad (2)$$
$$e_{\mathrm{GT=01}} = \sum_{k=0}^{n} \Pr(k|\mathrm{GT}=01) \times I_{GT\neq 01},$$

where the indicator $I_{GT\neq 00}$ is 1 if the called genotype (from data k) is not 00 (in error) and 0 otherwise, and $I_{GT\neq 01}$ is defined similarly.

Multilocus datasets for BPP and 3S analyses

We prepared two sets of data, one for BPP analyses (under the MSC and MSci models) and the other for 3S analysis (under the IM model). The 'BPP dataset' has six species in the *erato-sara* clade. The '3S dataset' has an extra species, *H. melpomene*, as an outgroup (**Table S1**). Both datasets were prepared in the same way. For each dataset, we defined coding and noncoding regions based on the gene annotation of *H. erato demophoon* v1 reference assembly. For the noncoding regions, we extracted the genotype calls for small genomic segments, referred to as loci, with sizes between 100 and 2,000 bp, and with the further requirement that any two consecutive loci must be at least 2,000 bp apart. This minimum spacing was used in previous analyses of *Drosophila* genomic datasets based on estimated recombination rates and was found to produce similar results to the use of larger spacing such as 10 kb (Wang and Hey 2010; Dalquen et al. 2017). For each locus, we produced sequence alignments from the genotype calls, with heterozygotes represented using IUPAC codes (with Y for T/C, say). We removed sites with missing genotypes and sites overlapping with repetitive regions (based on the repeat annotation of *H. erato demophoon* v1 reference assembly, available at http://download.lepbase.org/v4/repeatmasker/). Loci with ≤10 sites after processing or with >50% gaps were excluded. For the coding regions, the same filters were applied except that there was no maximum locus size. There were 74,999 coding loci (median length of 165 and median informative sites of 3) and 92,966 noncoding loci (median length of 237, median number of informative sites of 5).

Inferring species divergence history across the genome using multispecies coalescent model (BPP A01 analysis)

We inferred species trees using Bayesian inference under the MSC model implemented in BPP v4.3.0 (Yang and Rannala 2014; Rannala and Yang 2017; Flouri et al. 2018). This A01 analysis under the MSC accounts for deep coalescence and the resulting gene tree heterogeneity along the genome but assumes no gene flow. We analyzed autosomes and the Z chromosome separately and coding and noncoding regions of autosomes separately, with four sets of data in total (**Table S4**). Two inversion regions in chromosomes 2 and 15 were analyzed separately, denoted 2b and 15b, respectively. The 2b region included part of the Herato0211 scaffold (from position 1434133) and the whole scaffolds Herato0212, Herato0213 and Herato0215 (a total of 1.95 Mb). The 15b region corresponded to Herato1505:1977997–2558395 (580 kb), which included the *cortex* gene (Herato1505:2074108–

2087841, ~13.7 kb). Thus, there were 25 chromosomal regions in total (for 21 chromosomes, with chromosomes 2 and 15 split into three regions).

For each of the four sets, we grouped the loci into blocks of 100 loci and inferred the species tree for the block (**Table S4**). This block size was large enough for the inferred species tree to achieve high posterior support and small enough to allow for the local introgression history to be reflected in the inferred species tree history. The choice also allowed for at least two blocks for each chromosome region. Any final blocks of each chromosomal region with fewer than 40 loci were discarded due to limited information. To assess the impact of block size, we repeated the analysis using blocks of 200 loci. While the neutral coalescent process may be expected to be largely homogeneous across the genome, the introgression rate is expected to be highly variable, due to selective removal of introgressed alleles, affected by the strength of selection, local recombination rate, etc. The analysis using blocks of loci can thus capture the variation across the genome due to differential rates of gene flow.

We assigned diffuse inverse gamma priors to the root age τ_0 and population size parameters (θ) with the means to be close to rough estimates from the data in preliminary runs. We used $\theta \sim$ InvG(3, 0.04), with the mean 0.04 / (3 – 1) = 0.02 for all populations. For divergence times, we used $\tau_0 \sim$ InvG(3, 0.06) for the coding loci and $\tau_0 \sim$ InvG(3, 0.12) for the noncoding loci. Given τ_0, divergence times at other nodes of the species tree were generated from a uniform-Dirichlet distribution (Yang and Rannala 2010, eq. 2). Population-size parameters (θ) are integrated out analytically to improve mixing of the MCMC in the A01 analysis of species tree estimation (Yang 2015).

The MCMC was run for 2×10^6 iterations after a burn-in of 10^5 iterations, with samples taken every 20 iterations. The same analysis of each block is repeated ten times using different starting species trees, and consistency among runs is used to assess convergence of the MCMC. Non-convergent runs were discarded. The samples were then combined to produce the posterior summary such as the maximum a posteriori (MAP) tree.

Exploring gene flow scenarios using the IM model for species triplets (3S analysis)

In order to formulate hypotheses about gene flow between species, we attempted to use several heuristic methods, including PHYLONET/MCMC_SEQ using sequences (Wen and Nakhleh 2018), PHYLONET/ML (Yu et al. 2014), and SNaQ in PHYLONETworks (Solís-Lemus et al. 2017), the latter two using estimated gene trees as input data. Our attempts were not successful. Different runs of those programs inferred different network models, which in general did not appear to be reliable, with apparently spurious introgression events around the root of the species tree.

Instead we used the maximum likelihood program 3S (Dalquen et al. 2017) to estimate migration rates between all pairs of species. This implements the IM model for three species S_1, S_2 and S_3, assuming the species tree ((S_1, S_2), S_3), and accommodates both deep coalescence and migration. While limited to three species and three sequences per locus, it can be applied to large datasets with >10,000 loci. Gene flow is allowed between the two ingroup species only (S_1 and S_2) while the third species (S_3) is used as an outgroup to improve parameter estimation and the power of the test (Dalquen et al. 2017). *H. melpomene* was used as an outgroup for all pairs. The outgroup species was added to the dataset by mapping it to the *H. erato demophoon* reference genome.

We analyzed coding and noncoding regions separately and the autosomal loci and the Z chromosome separately, but treated all loci in each chromosomal region as one dataset (rather than breaking them into blocks of 100 or 200 loci) (**Table S4**). We also performed an analysis that used all autosomal loci. Since the sequences were unphased and 3S requires phased haploid sequences, we first phased the loci using PHASE v2.1.1 (Stephens et al. 2001), resulting in two phased haploid

sequences per individual. In the simulations of Huang et al. (2021), which examined the Bayesian method BPP, this approach of computational phasing produced very small biases, and the same may be expected to apply to the ML method implemented in 3s. The 3s program uses three sequences but allows multiple sequences per species. For each locus, we sampled three sequences of configurations 123, 113 and 223 with probabilities 0.5, 0.25 and 0.25, respectively. Here, 123 means one sequence from each species, etc.

Each dataset was analyzed using 3s to fit two models: MSC without migration (M0) and MSC with migration (M2; isolation-with-migration or IM). Model M0 involves two divergence time parameters (τ_1 and τ_0) and four effective population sizes (θ_1 and θ_2 for species S_1 and S_2, θ_4 for the root, and θ_5 for the ancestor of S_1 and S_2), while M2 involves in addition two migration rates M_{12} and M_{21}. Here M_{12} is the expected number of migrants from species S_1 to S_2 per generation. Each ML analysis was repeated ten times, and the run with the highest log-likelihood value was used. The two models (M0 and M2) were compared using a likelihood ratio test (LRT), with χ^2_2 used as the null distribution. Only gene flow scenarios that passed the LRT at the 1% level were considered later.

Evidence for pairwise gene flow from the 3s analysis was used together with the inferred species trees from the previous BPP A01 analysis of blocks of loci to generate a species tree model with introgression (MSci model). There were two difficulties in this approach. First, IM (3s) and MSci (BPP) are two extreme versions of models of gene flow and both may be too simplistic to fit the real data. The IM model assumes continuous gene flow between species S_1 and S_2 since their split, and may be unrealistic for many of the species pairs, for example, if gene flow occurred immediately after species split but then stopped as the two species became more diverged. This is particularly the case as a branch in the 3-species tree used in the 3s analysis may represent multiple branches in the full tree for all six species. The MSci model instead assumes episodic introgression events at specific time points. Our expectation is that if introgression is episodic, the LRT based on the continuous IM model will still detect gene flow, but with distorted parameter estimates. The opposite may be true as well: if gene flow is continuous as in the IM model, fitting the MSci model will give distorted parameter estimates (Jiao et al. 2020). Second, the LRT suggested significant evidence for gene flow between most pairs of species. We prioritized introgression edges that could reconcile the different species trees across the genome from the previous BPP analysis under the MSC model. We also took a parsimonious approach of minimizing the number of introgression events by assuming gene flow only if the model of no gene flow cannot explain the data: if the LRT suggested gene flow between most of their descendant species, we placed introgression edges between the ancestral populations.

Estimation of parameters for species divergence and cross-species introgression (BPP A00 analysis under MSci)

Given a species tree model with introgression, we ran BPP v4.3.0 (Flouri et al. 2020) to estimate the population sizes (θ), species divergence times (τ) and introgression probabilities (φ). An inverse-gamma prior, InvG(3, 0.01), with mean 0.005, was assigned to θ. The prior for τ_0 was InvG(3, 0.04). The priors for φ were beta(4, 2), with mean 0.75. Initial values of φ were set to 0.8 or 0.9. These prior settings were based on rough estimates from preliminary runs. We also considered alternative priors for θ (e.g. gamma distributions) and φ (e.g. beta(1,1)) to assess the stability of the posterior estimates when mixing and convergence were of concern. We performed inference for each of the 25 chromosomal regions, using either all coding or noncoding loci in each region (**Table S4**).

The MCMC was run for 10^6 iterations, sampling every 100 iterations, after a burn-in of 10^6 iterations. Ten independent runs were performed and convergence was assessed by examining consistency between runs. Non-convergent runs were discarded. For models with a label-switching

unidentifiability issue (Flouri et al. 2020), the MCMC samples were post-processed before they were combined for posterior summaries.

RESULTS

Analysis of base-calling and genotype-calling error rates to guide data compilation

We processed the raw reads from the whole-genome re-sequencing data of Edelman et al. (2019) for six *Heliconius* species from the *erato-sara* clade: *H. erato demophoon*, *H. himera*, *H. hecalesia formosus*, *H. telesiphe telesiphe*, *H. demeter* and *H. sara magdalena* (**Table S1**). To guide our compilation of the unphased loci, we estimated the base-calling error rate in the genomic data and analyzed the genotype-calling error. The base-calling error was calculated to be ~0.08% for the homozygous reference allele and ~0.20% for the homozygous alternative allele, with variation among individual genomes (**Tables S2–S3**).

Given a base-calling error rate (ε), we calculated the genotype-calling error rate (e) for the homozygous reference genotype (GT = 00) and the heterozygous genotype (GT = 01) by using equation (2) (**Fig. 1**). Note that even at a very low base-calling error rate, the genotype-calling error can be very high, especially at low read depth. Furthermore, the error rate for heterozygotes is much higher than for homozygotes.

Given the read depth n, the genotype-calling error does not necessarily decrease when ε decreases. For example, when $n = 7$, the genotype-calling error for a homozygote (GT = 00) is 0.0020 at $\varepsilon = 0.01$, but rises to 0.0345 when ε is reduced to 0.005 (**Fig. 1**). This is due to the discrete nature of the read outcome. At $\varepsilon = 0.01$, the called genotype is 00 (the truth) when $k = 0$ or 1, is 01 when $k = 2$–5 and is 11 when $k = 6$ or 7, and the genotype-calling error is a sum over $k = 2, \ldots, 7$ (i.e., six out of the eight cases are in error). In contrast, at $\varepsilon = 0.005$, the called genotype is 00 when $k = 0$, is 01 when $k = 1$–6 and is 11 when $k = 7$, so that the genotype-calling error is a sum over $k = 1, \ldots, 7$ (i.e., seven out of the eight cases are in error), and is higher than at $\varepsilon = 0.01$. Similarly given ε, the genotype-calling error does not necessarily decrease with the increase of the read depth n (**Fig. 1**). For example, when $\varepsilon = 0.01$ the genotype-calling error for a homozygote is 0.0199 at $n = 2$, but rises to 0.0585 when n increases from 2 to 6; for those values of n, genotype-calling error is a sum over $k = 1, \ldots, n$. Then when $n = 7$, error drops to 0.00269 as the error is a sum over $k = 2, \ldots, n$. The strong periodicity in the genotype-calling error is similarly due to the discrete nature of the problem.

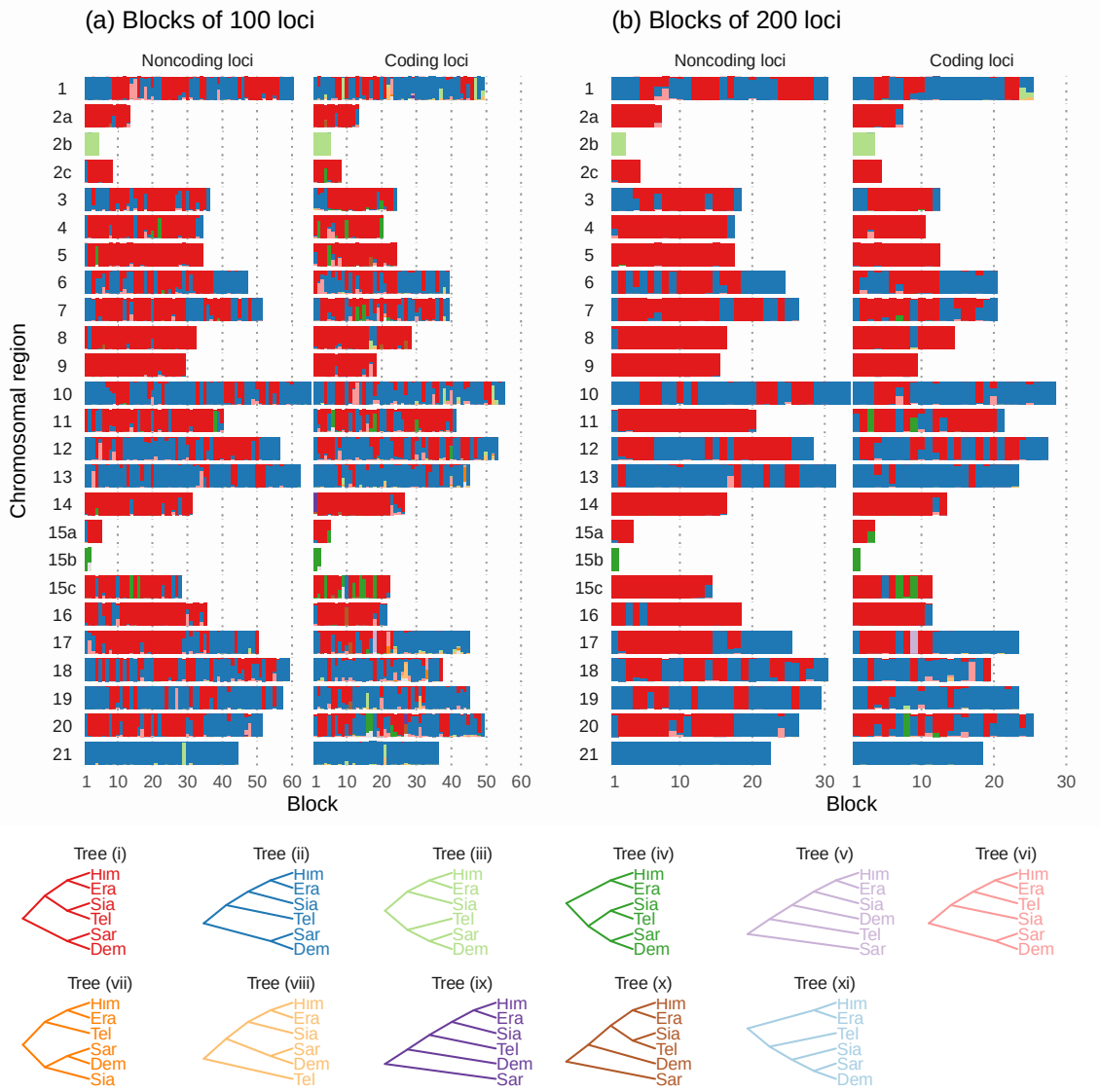
With the base-calling error estimated at $\varepsilon \approx 0.001$, we estimate the genotype-calling error rate to be 1.1×10^{-6} for homozygotes and 4.0×10^{-4} for heterozygotes at the read depth $n = 20$ (**Fig. 1**). We therefore filtered our data for coverage $n \geq 20$. Since the read coverage was >60x (Edelman et al. 2019), this caused relatively little dropout.

Genealogical variation across the genome

We first used BPP to infer species trees under the multispecies coalescent (MSC) model using blocks of 100 or 200 loci along the genome. There were 749 coding blocks and 933 noncoding blocks over the 25 chromosomal regions (21 chromosomes with chromosomes 2 and 15 each split into three regions by an inversion) (**Table S4**). This is the BPP A01 analysis of Yang (2015), which explicitly accounts for deep coalescence but ignores gene flow. Ten species trees were the best estimate (i.e., the maximum *a posteriori* probability tree or MAP tree) in at least one block (**Fig. 2**, **Tables S5–S7**). These are referred to as trees i-x. Trees i and ii accounted for over 95% of the blocks. These two trees differ only in the position of *H. hecalesia* within the *erato* clade: in tree i, *H. hecalesia* is sister

385 to *H. telesiphe* while in tree ii, it is sister to the clade (*H. erato*, *H. himera*).

386



387

388 **Figure 2**. Posterior probabilities of species trees from BPP analysis of (**a**) blocks of 100 loci and (**b**)
389 blocks of 200 loci across the genome (Table S2) under the MSC model without gene flow. The y-axis
390 represents the posterior probability and ranges from 0 to 1. Trees (i)–(x) in the legend are MAP
391 trees in at least one block. Tree (xi) appeared as one of the top eight trees in the sliding-window
392 analysis of Edelman et al. (2019) but not in our analysis. Tree colours match those in Edelman et al.
393 (2019) (their Fig. 2). Era: *H. erato*, Him *H. himera*, Sia: *H. hecalesia*, Tel: *H. telesiphe*, Dem: *H.*
394 *demeter*, Sar: *H. sara*.

395

396 The species trees estimated in this analysis are expected to reflect the history of species divergences
397 as well as cross-species introgression, and the major cause of the differences among the blocks is the
398 variable rate of gene flow along the genome. The use of 100 or 200 loci in each block helps to filter
399 out stochastic fluctuations in the coalescent process. The results were highly consistent between

the two choices of the block size and between coding and noncoding loci (**Fig. 2**). Such consistencies suggest that the patterns revealed in the analysis are real rather than a result of analytical artefacts.

There was some variation in the estimated species trees across the genomic regions. Three regions in particular had species tree distributions that differed from the rest of the genome: chromosome 21 (Z chromosome), the inversions on chromosomes 2 and 15 (2b and 15b). Chromosome 21 was the only chromosome for which tree ii is a MAP tree for almost all blocks (**Fig. 2**, **Table S5**). Among the autosomal regions excluding inversions, tree ii was the MAP tree in ~40% and 46% of noncoding and coding blocks, respectively, while the corresponding proportions for tree i were 58% and 47% (**Table S5**). Thus tree i was the autosomal majority tree. Inversion 2b had an unusual history in which *H. telesiphe* is more closely related to the *sara* clade (tree iii). In this inversion region, *H. erato*, *H. himera* and *H. hecalesia* share a derived inverted rearrangement relative to *H. melpomene, H. sara* and *H. demeter* (Van Belleghem et al. 2017; Davey et al. 2017; Edelman et al. 2019), consistent with trees ii and iii where these species are clustered together. Inversion 15b supported tree iv in both coding and noncoding regions, in which the (*H. telesiphe*, *H. hecalesia*) clade was sister to the *sara* clade instead of other members of the *erato* clade as in tree i or ii. This grouping strongly suggests that *H. telesiphe*, *H. hecalesia*, *H. demeter* and *H. sara* share the derived inverted rearrangement of this region relative to *H. erato* (Edelman et al. 2019). The 15b inversion contains the *cortex* gene that controls mimetic wing colour patterning across *Heliconius* species (Nadeau et al. 2016; Van Belleghem et al. 2017). Tree iv also appeared as a MAP tree sporadically in other parts of the autosomes, sometimes with high posterior probabilities (**Fig. 2**, **Table S6**). These include regions on chromosome 15 outside the inversion as well as on chromosomes 4, 5 and 11.

Pairwise gene flow rates

Gene flow among species could reconcile different species trees from the BPP A01 analysis reported above. We investigated this possibility by explicitly estimating the migration rates between each species pair under the IM model implemented for species triplets in the program 3s (Zhu and Yang 2012; Dalquen et al. 2017). The model allows migration between the two ingroup species at constant rates per generation in the two directions, while migration involving the outgroup is not allowed. We used *H. melpomene* as an outgroup in all triplets for parameter estimation.

We found evidence of bidirectional gene flow between *H. telesiphe* and *H. hecalesia*, consistent with a scenario in which tree i (autosome-majority tree) and tree ii (Z chromosome tree) are related through introgression between the two species (**Figs. 3, 4, 5**). However, gene flow between these species was detected only for autosomal loci (both coding and noncoding) and not on the Z chromosome (**Fig. 3**). The results suggest that the prevailing topology on the Z chromosome, tree ii rather than tree i, is the true species tree. In particular, we found no evidence of gene flow in 3s analyses from *H. erato* and *H. himera* towards *H. hecalesia*-(**Fig. 3**), which is necessary to explain tree ii as a result of introgression if tree i were the true species tree. We note that in other systems, such as *Heliconius melpomene* and *H. cydno* (Martin et al. 2019) and *Anopheles* mosquitoes (Thawornwattana et al. 2018), the sex chromosome was also less prone to gene flow. In addition, in the BPP analysis under the MSci model to be discussed in detail later, there is only very weak introgression ($1 - \varphi_c < 10\%$) from *H. telesiphe* to *H. hecalesia* on the Z and almost none for the 2b inversion (**Figs. 5, 6**). Together, these results suggest that tree ii most likely reflects the true species history while tree i is the result of cross-species introgression.

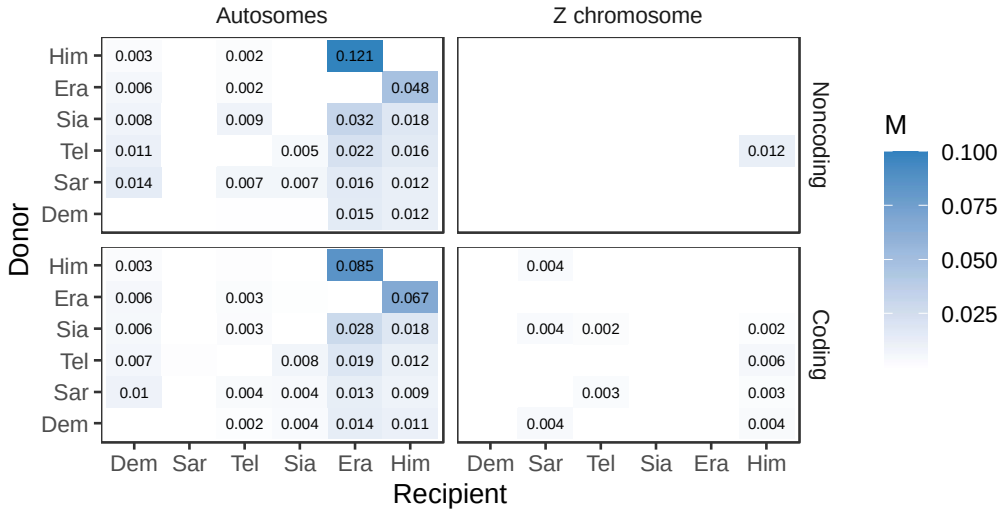
Given tree ii, the Z chromosome, particularly among noncoding loci, was almost devoid of gene flow, in sharp contrast with the autosomes (**Fig. 3**, **Table S8**). For the autosomes, the highest rates of migration were between the two sister species *H. erato* and *H. himera*, with gene flow occurring in both directions (at the rate of 0.085–0.121 migrants per generation from *H. himera*

448 to *H. erato* and 0.048–0.067 in the opposite direction) (**Fig. 3**). According to 3s results, there was
449 gene flow into *H. erato*, *H. himera* and *H. demeter* from every other species, and gene flow from
450 *H. sara* to all other species. These results were largely consistent among the individual
451 autosomes (**Fig. S1**), across pairs of species (**Fig. S2**) and with an independent Bayesian analysis of
452 the 3s datasets under the same model using GPhocs (Gronau et al. 2011) (**Fig. S3**).

453 The pattern of gene flow inferred using 3s may reflect complex introgression in this group of
454 species as well as the difficulty of using pairwise migration rates to reconstruct the full migration
455 history for all species. If gene flow involved ancestral branches on the tree for all six species, we
456 would expect the LRT to detect it in multiple pairs of species, although with distorted estimates of
457 times and rates of migration. One scenario is extensive introgression involving common
458 ancestors of the *sara* group and the *erato* group, which should show up as initial migration after
459 species divergence that ceased after a certain time.

460 We used the results from the previous two analyses (BPP A01 and 3s) to formulate a plausible
461 history of species divergences and cross-species introgression for the *erato-sara* clade. The Z
462 chromosome tree from the BPP A01 analysis was used as the backbone, onto which 3s-supported
463 introgression events were added to reconcile other species trees from the bpp A01 analysis. The
464 result is shown in **Figure 4**.

465



466

467 **Figure 3**. Maximum-likelihood estimates of the pairwise migration rate ($M = Nm$) under the IM
468 model from 3s using all coding and noncoding loci in the autosomes (left column) and in the Z
469 chromosome (right column; see **Table S4** for the number of loci) for each pair of species. The donor
470 and recipient species of gene flow are given in the y- and x-axis, respectively. *H. melpomene* was
471 used as an outgroup. Only significant migration rate estimates (by likelihood ratio test at $p < 0.01$)
472 larger than 0.001 are shown. See **Figure S2** and **Table S8** for estimates of other parameters.

473

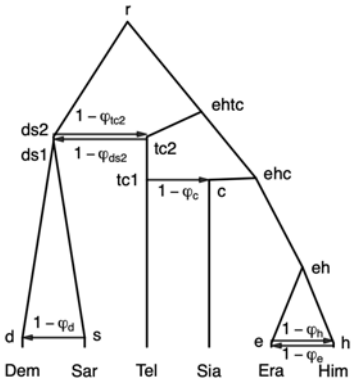


Figure 4. Proposed introgression model based on the BPP species tree estimation under MSC and 3s analysis under the IM model. Each horizontal arrow represents a unidirectional introgression event. Branch lengths represent posterior means of divergence/introgression times in the BPP analysis of all 6,030 noncoding loci in chromosome 1. Estimates for other chromosomal regions are in **Figures 5** and **S4**. Note that the introgression probability φ_d from *H. sara* to *H. demeter* was estimated to be zero in all chromosomal regions under this model (**Fig. 5**).

Construction of a full history of species divergences and cross-species introgression

We then used the species-tree model with introgression of **Figure 4** to fit the MSci model to estimate introgression probabilities ($1 - \varphi$), population size parameters (θ) and species divergence/introgression times (τ). (Note: in the version of BPP used, φ is the backbone inheritance probability and $1 - \varphi$ is the corresponding introgression probability). All coding loci are analyzed as one dataset for each chromosomal region, as are all noncoding loci. These correspond to A00 analysis under MSci in BPP (Flouri et al. 2020).

Estimates of introgression probability from *H. telesiphe* into *H. hecalesia* were consistently high ($1 - \varphi_c > 0.5$, ~0.8 on average) across the genome, except for the 2b inversion region and Z chromosome (**Figs. 5–6**). The time estimates suggest that this introgression occurred almost immediately after *H. hecalesia* split from the common ancestor of *H. erato* and *H. himera* (**Fig. 6**), supporting the hypothesis that *H. hecalesia* is a hybrid species that was generated during a single catastrophic event. Even though our model assumed different times for species divergence (τ_{ehc}) and for introgression (τ_c), with $\tau_{ehc} > \tau_c$ (equivalent to model B in Flouri et al. (2020), posterior estimates strongly suggest that those two times actually coincided, with $\tau_{ehc} \approx \tau_c$ (equivalent to model C in Flouri et al. (2020). This pattern was consistent with our estimates of the species trees from the blocks of loci (BPP A01 analysis; **Fig. 2**) where the autosomes, with an exception of the 2b region, were dominated by tree i as a result of *H. telesiphe* → *H. hecalesia* introgression on tree ii.

For the more ancient introgression between *H. telesiphe* and the common ancestor of *H. demeter* and *H. sara*, the estimated introgression probability from *H. telesiphe* ($1 - \varphi_{ds2}$) was substantial across the entire genome, ~0.3–0.4 on average, with $1 - \varphi_{tc2} \approx 0.1$ in the reverse direction (**Fig. 5**, **Table S9**). This suggests genome-wide flow between the *erato* clade and the *sara* clade prior to the *H. telesiphe* → *H. hecalesia* introgression/hybridization. The estimates of φ_{ds2} and φ_{tc2} in the two small inversion regions, 2b and 15b, were more extreme, either very close to 0 or 1, with complex identifiability issues. We discuss the introgression history of the two inversion regions below.

Other introgression events had relatively low probabilities across the genome despite evidence from the 3s analysis under the IM model. The bidirectional gene flow between *H. erato* and *H. himera* had probabilities of $1 - \varphi_h \approx 15\%$ for *H. erato* → *H. himera* and $1 - \varphi_e \approx 5\%$ for *H. himera* → *H. erato*,

consistently throughout the genome (**Fig. 5**). Note that the introgression probability in the MSci model is expressed as the proportion of immigrants in the receiving population at the time of hybridization, so the smaller rate in the *H. himera* → *H. erato* direction may reflect the large population size for *H. erato*. In contrast, the migration rate in the IM model is estimated to be in the range of 0.07–0.12 migrants per generation, and is larger in the *H. himera* → *H. erato* direction (**Fig. 3**). We note that there may not be a simple correspondence between the introgression probability in the MSci model (BPP) and the migration rate in the IM model (3s); even a small migration rate over a prolonged time period may result in a high introgression probability (Jiao et al. 2020). For sister species that hybridize today, the IM model of continuous migration may be more realistic than the MSci model assuming introgression at isolated time points in the past. The present-day range of *H. himera* is confined to a small area surrounded by several races of *H. erato*, and it is known that hybridization between the two species is ongoing (Mallet et al. 1998). Furthermore, migration rate estimates vary considerably across the chromosomes according to both 3s (**Figs. S1** and **S2**, **Table S8**) and GPhoCS (**Figure S3**).

Lastly, our BPP analysis under the MSci model did not support the *H. sara* → *H. demeter* introgression suggested by 3s analysis. The introgression probability ($1 - \varphi_d$) was either small (<1%) or had a large posterior interval in all chromosomal regions (**Fig. 5**), in either case providing no strong support for such introgression.

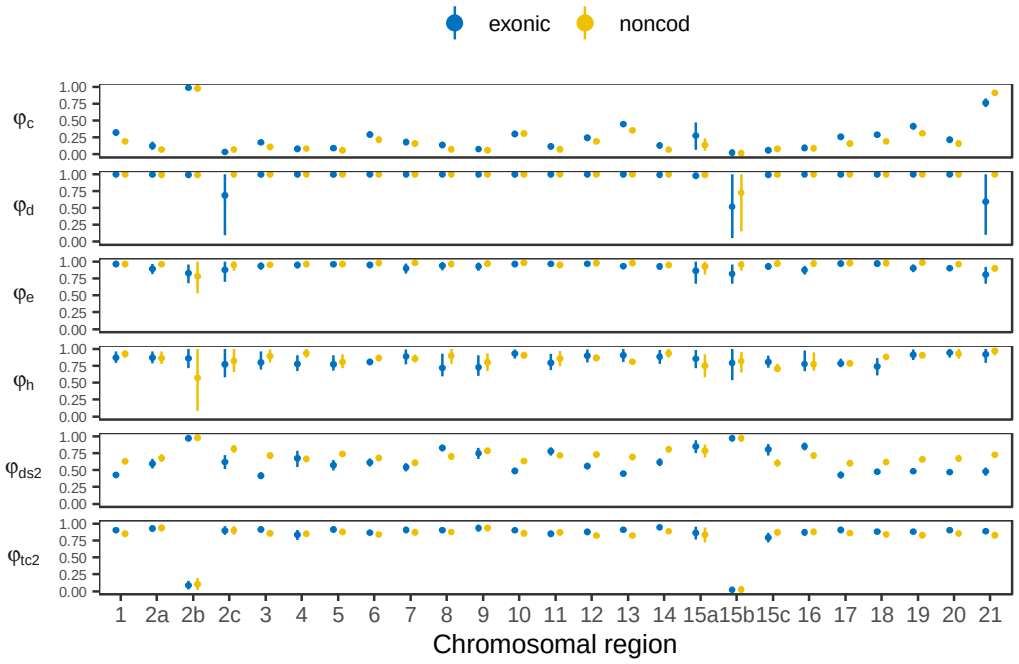


Figure 5. Posterior means (dots) and 95% HPD intervals (bars) of introgression probabilities (φ) under the MSci model of **Figure 4** obtained using BPP for the 25 chromosomal regions (see **Table S4** for the number of loci). Results for other parameters in the model are in **Figure S4**. For 2b and 15b, there is an alternative set of posterior estimates resulting from within-model unidentifiability, shown in **Figure 7**.

Overall, estimates of population sizes (θ) and species divergence/introgression times (τ) were broadly consistent across most chromosomal regions, as well as between coding and noncoding loci (**Figs. 6, S4, S5**). Each chromosomal region with a sufficient number of loci yielded parameter

estimates with high precision, and this consistency is an indication of the reliability of the estimates. The estimates of divergence times were consistently proportional between noncoding and coding loci, with regression coefficient ~0.6 ($r^2 \geq 0.99$) (**Fig. S6**), suggesting that purifying selection on coding loci reduced the neutral mutation rate to about 0.6 x that for noncoding loci.

The age of the base of the *erato-sara* clade (τ_r) was estimated to be about 0.027 substitutions per site for noncoding loci on average (**Fig. S4**, **Table S9**). If we use the mutation rate of 2.9×10^{-9} neutral mutations per site per generation and 4 generations per year (Keightley et al. 2015), this translates to about 2.3 million years (Ma) of divergence.

Population size parameters (θ) for populations corresponding to short branches on the species tree (**Fig. 4**) were difficult to estimate reliably, with considerable variation among chromosomes and with large posterior intervals (**Fig. S4**, **Table S9**), reflecting both low information content for those parameters and the impact of the heavy-tailed inverse gamma prior. Most populations had θ in the order of 0.01 while a few had more extreme estimates. *H. erato* (θ_{Era}) had small values, in the order of 0.0001–0.001, for most chromosomes, except in the 2b inversion where the θ_{Era} estimates were unreliable due to lack of data. In the current implementation of the MSci model, the same species before and after an introgression event are considered distinct species and are assigned independent θ parameters. It appears more sensible to assign the same θ, in particular when the introgression event is inferred to be nonexistent (as in the case of the *H. sara* $\rightarrow$ *H. demeter* introgression) and/or to be extremely recent. Thus we believe the estimates of θ for the parent populations prior to introgression events as the population size estimates for the four modern species: *H. demeter* ($\theta_{\text{d}} \approx 0.01$, better than $\theta_{\text{Dem}} \approx 0.002$-$0.004$), *H. sara* ($\theta_{\text{s}} \approx 0.01$, better than $\theta_{\text{Sar}} \approx 0.06$-$0.13$), *H. erato* ($\theta_{\text{e}} \approx 0.06$-$0.30$, better than $\theta_{\text{Era}} \approx 0.0001$-$0.0166$) and *H. himera* ($\theta_{\text{h}} \approx 0.005$-$0.017$, better than $\theta_{\text{Him}} \approx 0.0009$-$0.0022$) all for noncoding autosomal loci (**Table S9**). The genome sequences of *H. erato* and *H. himera* were obtained from partially inbred individuals to aid assembly, which likely explains their apparent low effective population sizes and large variation among chromosomes. The *H. erato* genome in particular has long stretches of homozygosity on some chromosomes due to inbreeding (**Fig. S5**) corresponding approximately to the per chromosome estimates of θ_{Era}.

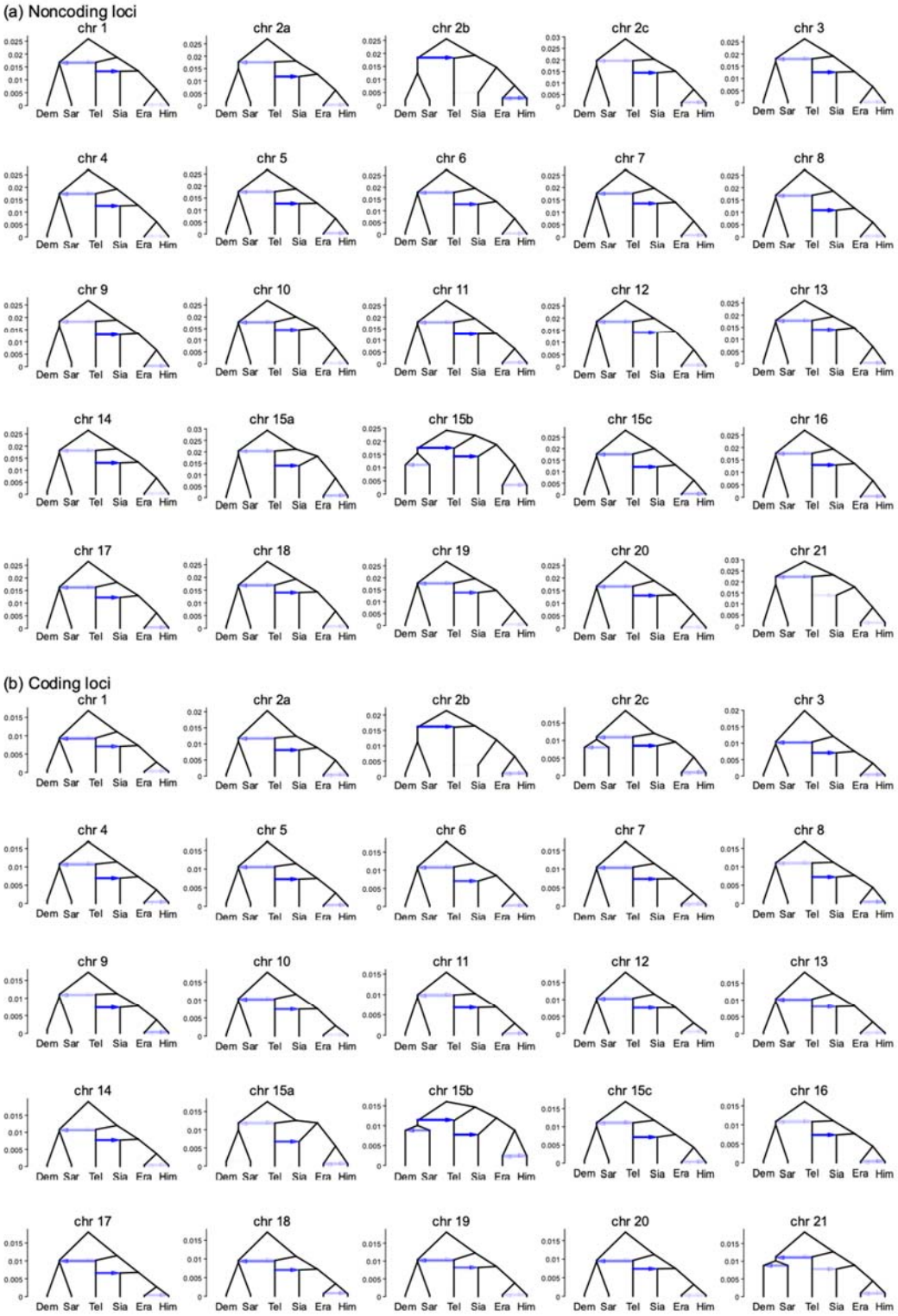


Figure 6. Estimated introgression history in each chromosomal region obtained from BPP analysis under the MSci model of **Figure 4** using (**a**) noncoding and (**b**) coding loci. Intensity of the horizontal edges represents the posterior mean of introgression probability, while the y-axis represents the divergence times in the units of the expected number of mutation per site.

DISCUSSION

Major features of the species tree and introgression: comparison with previous studies

In this study, we processed genomic sequences from the *erato-sara* clade of *Heliconius* to compile unphased diploid sequence alignments, and accommodate unknown phase resolution in the likelihood calculation. We also assessed the probable genotype calling error rate based on our high coverage reads and used a high coverage filter to exclude almost all genotype calling errors. We then used the data to infer the history of speciation and introgression under the MSci model in a full-likelihood framework. To our knowledge, this is the first phylogenomic study that treats unphased diploid genomic data properly. The coalescent-based full-likelihood method is an improvement over summary methods or sliding-window analysis, which do not make proper and full use of information in the genomic sequence data.

In general, our analysis confirms some major conclusions of previous analyses but reveals far more detail of the evolutionary history of the clade. As in Edelman et al. (2019), we find that the Z chromosome supports tree ii, while autosomes favour tree i. We weigh the evidence and find that that the Z chromosomal tree is almost certainly the true species tree (see 'Pairwise gene flow rates'), even though large-scale introgression has distorted genealogical histories of autosomes so that the most common gene tree in the genome has a different topology (mostly tree i). Major genealogical conflicts between the sex chromosome and the autosomes were inferred also in other species groups including *Anopheles gambiae* (Fontaine et al. 2015; Thawornwattana et al. 2018), felids (Li et al. 2019) and small finches (Stryjewski and Sorenson 2017). In the *Anopheles gambiae* group of African mosquitoes, the X chromosomal tree was inferred to be the species tree while the autosomal tree reflects rampant gene flow (Fontaine et al. 2015; Thawornwattana et al. 2018). However, the sliding-window analsyis of genomic data from the *Anopheles gambiae* group (Fontaine et al. 2015) led to an incorrect species tree while coalescent-based full-likelihood methods produced a tree more consistent with chromosomal inversion data and with established patterns of cross-species hybridization and gene flow (Thawornwattana et al. 2018). In our *Heliconius* dataset, concatenated windows and MSci analyses led to similar species trees. Concatenation may produce anomalous gene trees (Roch and Steel 2015) so that the estimated gene tree from large sliding windows may differ from the species phylogeny, but this was not the case for the *Heliconius* data. In the *Anopheles gambiae* group, all species arose very rapidly, creating very short internal branches at the base of the species tree, while the *Heliconius* species studied here are well separated with longer internal branches (**Fig. 4**). Thus the sliding-window gene tree approach was more successful in *Heliconius* than in the *Anopheles gambiae* group. Nevertheless, given the well-known caveats of sliding-window analysis, it was important to test the previous results.

Our improved data and analytical methods also made it possible to conduct more powerful analyses, answering additional questions. For example, *D* statistics and internal branch length tests (Edelman et al. 2019) can be used to test for the presence of gene flow, but not to identify gene flow between sister species, to infer the direction of gene flow, or to estimate its magnitude. In contrast, migration rate $M = Nm$ and introgression probabilities $(1 - \varphi)$ are parameters in the IM and MSci models that provide direct estimates of the magnitude of gene flow. Note that the proportions of gene trees in sliding-window analyses are not directly comparable with introgression probabilities or with proportions of block-based species trees across the genome in the BPP A01 analysis.

For example, the frequencies of estimated gene trees i and ii in Edelman et al. (2019) are 24.3% and 25.8% among 10kb sliding windows (their Fig. S78), but are 35.5% and 34.8% among the 50kb windows (mapped to the *H. erato* reference, their Fig. 2B), and 36.4% and 45.4% among the 50 kb windows (mapped to the more distant *H. melpomene* reference, their Fig. S77). It is easy to see why the proportion of windows with the dominant gene tree will increase when the window becomes larger due to the averaging effect of concatenation. Even though the objective of the analysis is to

infer the population-level process of species divergences and cross-species introgression, the gene-tree proportions calculated using sliding windows are affected by this averaging effect as well as random phylogenetic errors due to a limited window size and coalescent fluctuations in the genealogical trees. In the BPP A01 analysis under the MSC model, the variation in the rate of cross-species gene flow (affected by linkage to selected loci and local recombination rate) is expected to be the most important factor for fluctuations among blocks. Gene-tree conflicts due to ancestral polymorphisms are filtered out by the MSC model while they contribute to fluctuations among gene trees in sliding-window analysis. In our BPP A01 analysis, the frequencies of (species) trees i and ii are 54.8% and 42.8% among the 100-loci noncoding blocks, 44.2% and 48.3% among the 100-loci coding blocks, 58.9% and 40.2% among the 200-loci noncoding blocks, and 46.9% and 49.5% among the 200-loci coding blocks. These are reasonably consistent between coding and noncoding loci and relatively insensitive to the number of loci in each block. There was a somewhat greater representation of the presumed introgression tree i in noncoding loci (55-59%) than coding loci (44-47%) and a moderate excess of introgression *H. telesiphe* $\rightarrow$ *H. hecalesia* ($1 - \varphi_c$) in noncoding versus coding loci on each chromosome (**Fig. 5**). This may reflect greater constraint on introgression for coding sequences as found in other systems, for instance for Neanderthal ancestry in modern humans (Sankararaman et al. 2014). In effect, the BPP-estimated species tree variation across the genome (**Fig. 2**) approximates history while the concatenated sliding-window analysis includes both noise and history, which are hard to separate without a proper inference framework.

Previous estimates of migration rates and timings in *Heliconius* under models of gene flow such as the IM model have been limited to a few loci (Bull et al. 2006; Kronforst et al. 2006; Kronforst 2008; Salazar et al. 2008; Pardo-Diaz et al. 2012) or a few species (Kronforst et al. 2013; Van Belleghem et al. 2020), most of which have focused on the *melpomene* clade. Using joint site-frequency spectrum and a secondary contact model, Van Belleghem et al. (2020) estimated migration rates ($M = Nm$) between *H. himera* and two races of *H. erato* (*H. erato favorinus* and *H. erato emma*) to be around 0.5–0.6 migrants per generation from *H. erato* to *H. himera*, and 0.07–0.13 in the opposite direction. This asymmetric gene flow accords with our MSci results, in which the probability of flow is 15% from *H. erato* to *H. himera*, and 5% in the opposite direction. Our estimates from the 3s analysis of the genomic data are 0.05–0.07 migrants per generation from *H. erato* to *H. himera*, and 0.09–0.12 in the opposite direction. The differences may be due to different data and methods used. No estimates of the other introgression probabilities in **Fig. 4** have previously been reported.

Limitations of our analyses

Implementation of the MSci model in BPP (Flouri et al. 2020) provides a powerful tool for estimating the timing and magnitude of introgression on a species phylogeny from genome-scale data while accounting for deep coalescence and genealogical heterogeneity across the genome. The implementation also retains heterozygous sites in the genomic data and averages over all possible heterozygote phase resolutions (Gronau et al. 2011), avoiding systematic errors caused by the use of haploid consensus sequences (Huang et al. 2021). Other likelihood-based methods incorporating cross-species gene flow are under active development (Hey et al. 2018; Wen and Nakhleh 2018; Zhang et al. 2018; Jones 2019), but they are computationally demanding and currently impractical for data of >100 loci. We have also tested for the effects of sequencing error on genotype calls; erroneous genotype calls may also potentially obscure the genealogical signal in the data. We find that genotype call error should be very low in our study due to high read coverage (>60x).

Here we discuss several limitations of our approach. First, the introgression model we formulated (**Fig. 4**) relies on two main sources of information: the variation in the species tree estimated from blocks of loci across the genome and pairwise migration rate estimates from 3s. Integrating such information into a single species tree model with introgression is nontrivial. Ideally, we would be

able to infer the species phylogeny and introgression events simultaneously with associated parameters such as introgression and divergence times, but this has yet to be implemented.

Second, even with a fixed MSci model (**Fig. 4**), inference of demographic parameters is computationally challenging. Some parameters were more difficult to estimate reliably, such as population sizes associated with short branches in the species tree (e.g., θ_{Sar}, θ_{e}, and θ_{ds1}; see **Fig. S4**). Contributing factors include unidentifiability of the model, multimodal posterior distribution, poor MCMC mixing, and potential model misspecification. The unidentifiability problem arises because of the symmetry in the likelihood induced by bidirectional introgression (such as the one between *H. erato* and *H. himera*, or between the ancestor of *H. sara* + *H. demeter* and the ancestor of *H. telesiphe*) (Flouri et al. 2020). The nature of the unidentifiability in models with many species and many introgression events can be complex, involving both within-model and cross-model unidentifiability. The MSci model of **Figure 4** involves two bidirectional introgression events. The early bidirectional introgression between the ancestor of *H. telesiphe* and the ancestor of *H. sara* and *H. demeter* yields a pair of unidentifiable models (**Fig. S9**). The bidirectional introgression between *H. himera* and *H. erato* also leads to a pair of unidentifiable likelihood peaks (not shown). We prefer species tree S_1 of **Figure S9** because the introgression probabilities $1 - \varphi_{ds2} < ½$ and $1 - \varphi_{tc2} \approx 0.1$, which appear to be biologically plausible, while the unidentifiable alternative model S_2 has introgression probabilities $\varphi_{ds2} > ½$ and $\varphi_{tc2} \approx 0.9$ for the horizontal introgression branches, implying mutual replacements of *H. telesiphe* by the ancestor of *H. sara* and *H. demeter* and the ancestor of *H. sara* and *H. demeter* by *H. telesiphe*, which does not appear plausible biologically. Similarly the bidirectional introgression event between *H. erato* and *H. himera* creates a within-model unidentifiability, with two peaks in the posterior (Flouri et al. 2020), and we focus on the peak with small introgression probabilities (with $1 - \varphi_{e} < ½$ and $1 - \varphi_{h} < ½$) (**Fig. 5**) instead of the alternative peak of mutual population replacements. We suspect that introgression in those species is likely to be occurring over extended periods of time rather than at isolated time points due to a certain catastrophic event (Mallet et al. 2016). The unidentifiability caused by the bidirectional introgression events is of the label-switching type and is resolved by applying simple constraints such as introgression probability $1 - \varphi < ½$ (Flouri et al. 2020). An example of MCMC outputs and processing is provided in **Figure S10**. We also note that unidentifiability can occur when the introgression probability is estimated to be 0 or 1. This is the case for the 2b and 15b inversion regions; see our discussions of the different inversion scenarios below.

Third, we avoided explicit modeling of recombination by analyzing short genomic loci (~200 bp on average) that are far apart (at least 2 kb), assuming that sites within each locus have zero recombination, whereas different loci are free to recombine. Sites are expected to be approximately independent at physical distances of ~10 kb apart, at least based on linkage disequilibrium decay estimated in the *H. melpomene* group species (Dasmahapatra et al. 2012). Thus, some correlation of gene genealogies between consecutive loci may be expected. The unrealistic assumption of free recombination is expected to have a slight effect of exaggerating information content within each block of loci, leading potentially to artificially narrow credibility intervals for parameter estimates. The assumption of no recombination among sites of the same locus may be of more concern since recombination causes different parts of the sequence to have different histories while the model assumes one history. However, the loci used in this study are so short that the assumption of no recombination within a locus is expected to hold approximately. Previously Lanier and Knowles (2012) found that species tree estimation under MSC is robust to realistic levels of recombination. The impact of recombination on estimates of the migration rate or introgression probability is not well understood, in particular when recombination rate varies across the genome. Variable recombination rates as well as variable selective pressures across the genome will differentially affect the probability that introgressed alleles are accepted in the recipient population, that is, the rate of gene flow, as found in *Heliconius* (Edelman et al. 2019; Martin et al. 2019) and other organisms (Burri et al. 2015; Schumer et al. 2018).

Fourth, the migration (IM) models in 3s and the MSci model in BPP are necessarily simplified. Currently we lack a good understanding of the behaviour of our inference methods when the model of gene flow is misspecified, for example, when the MSci model is fitted to data generated under the IM model or vice versa (Jiao et al. 2020, Fig. 7). In reality one may expect the magnitude of gene flow to vary across the genome and over time. It will be interesting to examine the performance of estimation methods such as 3s and BPP under such complex scenarios of gene flow, and to develop methods that account for such variation.

Evidence of introgression from natural populations

Using IM and MSci models, our analyses support three introgression events (**Fig. 4**): (1) between *H. telesiphe* and the common ancestor of the *sara* clade, (2) from *H. telesiphe* into *H. hecalesia*, and (3) between *H. erato* and *H. himera*. The first two are consistent with previous genomic studies (Edelman et al. 2019; Kozak et al. 2021). Ancient introgression such as (1) will be difficult to confirm using evidence from natural populations today. The recent introgression (3) is well-documented in natural populations and in mating experiments (Jiggins et al. 1997; Mcmillan et al. 1997; Mallet et al. 2007).

However, the six species included in this study constitute but a fraction of the species and geographic races in the *erato-sara* clade, so caution should be exercised in interpretation of inferred introgression. In particular, signals of introgression may be indirect, involving related species or subspecies unsampled in the data.

For example, the *H. telesiphe* → *H. hecalesia* introgression inferred from our data may not represent direct gene flow between the two species. *H. telesiphe* and *H. hecalesia* do not currently overlap geographically (Rosser et al. 2012). *H. hecalesia* does overlap with *H. clysonymus* and its sister species *H. hortense* in West Ecuador, the Colombian Andes, and Central America (Rosser et al. 2012), and there are documented natural hybrids between *H. hecalesia* and both *H. clysonymus* and *H. hortense* (Mallet et al. 2007). Furthermore, *H. telesiphe* forms a well-supported clade with *H. clysonymus* and *H. hortense*, nested within the *erato* clade (Kozak et al. 2015; Massardo et al. 2020; Kozak et al. 2021). Therefore, it is likely that introgressed loci in *H. hecalesia* actually came from *H. clysonymus*, *H. hortense*, or their common ancestor. Introgression from *H. clysonymus* or *H. hortense* is also supported by a phylogenetic network analysis and *D* statistics in Kozak et al. (2021).

Another case is the possible indirect gene flow signal between *H. erato* and *H. himera*. *H. himera* is considered an incipient species within *H. erato* (*sensu lato*). It is restricted to middle elevations (800–2,000 metres) of the Andes in South America; in contrast, the subspecies of *H. erato* characterized here, *H. erato demophoon*, is found in Central America (Rosser et al. 2012). However, *H. himera* is parapatric with several subspecies of *H. erato* (such as *cyrbia*, *favourinus*, *lativitta* and *emma*) (Mallet 1993; Rosser et al. 2012), with narrow contact zones where natural hybrids can be found at frequencies of 5–10% (Jiggins et al. 1996; Jiggins et al. 1997; Mallet et al. 1998). Thus the introgression signal almost certainly arose from these adjacent subspecies of *H. erato* rather than from *H. erato demophoon*.

Introgression history of the two inversion regions

Our analyses consistently suggest that two inversion regions, 2b and 15b, have genetic histories distinct from the rest of the genome (**Figs. 5–6**). All introgression probabilities for those two regions were close to either 0 or 1, except for estimates of φ_h and φ_d which involve large uncertainties due to lack of data (**Fig. 5**). This suggests that loci within each of the two inversion regions were inherited

almost without recombination as a single locus. This is likely if an inverted region introgresses and becomes fixed in a population that previously lacked the inversion (or vice-versa), but that during the period of polymorphism, recombination in the inverted region was strongly suppressed, which is a known biological effect of inversions. Although the MSC framework explicitly models ancestral polymorphism, inversion polymorphism is not accounted for since it creates substructure within the ancestral population while the model assumes random mating within species and free recombination among loci. Nonetheless, the result that the probability of introgression is either 0 or 1 is a sensible result. In the MSci model, when introgression probabilities are near the boundary of the parameter space [0,1], several parameters can become unidentifiable, and different parameter values corresponding to different introgression histories can explain the data equally well. For each of the two inversion regions, we identified distinct peaks in the posterior distribution which correspond to different historical scenarios with opposite introgression directions between (*H. sara*, *H. demeter*) and *H. telesiphe* (**Fig. 7**). Both scenarios are consistent with the known inversion orientations in modern species. Here, we discuss evidence in favour of each introgression scenario, although it is difficult to rule out alternatives.

In the chromosome 2b region (1.95 Mb, with 411 noncoding loci and 516 coding loci), introgression between the common ancestor of the *sara* clade and *H. telesiphe* was inferred to be unidirectional with a high introgression probability ($1 - \varphi_{tc2} \approx 0.9$, $1 - \varphi_{ds2} \approx 0$) in scenario 1 (**Fig. 7a**). In scenario 2, introgression was mostly unidirectional but in the opposite direction ($1 - \varphi_{tc2} \approx 0$, $1 - \varphi_{ds2} \approx 0.8$). Introgression in either direction is consistent with tree iii (**Fig. 2**), where *H. telesiphe* clusters with the *sara* clade and the *H. telesiphe* → *H. hecalesia* introgression was absent ($1 - \varphi_c \approx 0$). The two scenarios make similar predictions about the gene-tree histories and the genetic data (for example, both scenarios predict the origin of the inversion in the branch leading to ehc), but the effective root of the species tree for the 2b region is at node r in scenario 1 and at node ehtc in scenario 2.

To determine which scenario is more consistent with other autosomal regions of the genome, we compared the root age of the inversion region with τ_r and τ_{ehtc} on other chromosomes. Divergence times (τ) for the 2b region are highly similar to those in other regions of the genome, with regression slopes ~1.01–1.08 for the noncoding loci (**Fig. S7**), suggesting that the 2b region has a neutral mutation rate similar to other genomic regions. Scenario 1 is supported if the root age (τ_r in scenario 1, τ_{ehtc} in scenario 2) for inversion 2b is close to estimates of τ_r in other regions, while scenario 2 is supported if it is closer to τ_{ehtc} from the rest of the genome. For non-coding loci, the root age of 2b (posterior mean 0.026) was closer to the root age or the rest of the genome expected under scenario 1 (τ_r = 0.26–0.29) than that for scenario 2 (τ_{ehtc} = 0.018–0.023, **Fig. S4**, **Table S9**). Although we do not know whether *H. telesiphe* has the inversion, its ancestor likely had the inversion, but introgression from the ancestor of *H. demeter* and *H. sara* (which lack the inversion) under scenario 1 caused replacement without recombination. We therefore predict that *H. telesiphe* does not today carry the inversion, but that its ancestor did. In contrast, scenario 2 predicts that *H. telesiphe* always lacked the inversion, but it would be unclear why the non-inverted arrangement from *H. telesiphe* replaced the original non-inverted sequence in the *H. demeter/H. sara* ancestor with an almost complete lack of recombination. Thus this secondary evidence also supports scenario 1 (**Fig. 7a**).

An alternative to an introgression hypothesis requires ancestral polymorphism of the 2b inversion, where *H. sara*, *H. demeter* and *H. telesiphe* retained the ancestral non-inverted orientation through incomplete lineage sorting after divergence of the *erato* and *sara* clades (Edelman et al. 2019). Compared with the introgression scenario, ancestral inversion polymorphism predicts that the divergence time τ_{tc2} for the common ancestor of *H. telesiphe*, *H. sara* and *H. demeter* must pre-date or overlap the divergence time of the entire clade found for the rest of the genome. Our results did not support this hypothesis. Estimated τ_{tc2} in the 2b region (for non-coding regions, 0.018) was always much closer to the introgression time in the rest of the genome (τ_{tc2} = 0.016–0.022) than to the root in the rest of the genome (τ_r = 0.026–0.029) (**Fig. S4**, **Table S9**), suggesting that the

introgression hypothesis is more likely than the inversion polymorphism hypothesis in the ancestral population prior to the divergence of the entire *erato-sara* clade.

Based on pairwise sequence distances (D_{XY}) Edelman et al. (2019) argued that ancestral inversion polymorphism could explain the data as well as or better than a (*H. sara*, *H. demeter*) → *H. telesiphe* introgression for inversion 2b (their Fig. S95). The introgression scenario was rejected because the *H. sara*–*H. telesiphe* divergence of 2b was not much lower than the genome-wide average. However, this conclusion assumes that that all other chromosomes were free of such introgression. The prevalence of bidirectional introgression between (*H. sara*, *H. demeter*) and *H. telesiphe* across the genome, as suggested by our results (**Figs. 5–6**), could explain the patterns of D_{XY} in Edelman et al. (2019). The possibility of the opposite introgression with *H. telesiphe* as the source species (scenario 2 in **Fig. 7a**) was also not considered by Edelman et al. (2019).

For the chromosome 15b inversion region (580 kb, with only 149 noncoding loci and 167 coding loci), two sets of estimates were always obtained from independent MCMC runs, each occurring about 50% of the time. This inversion contains the *cortex* gene known to be involved in mimicry in the *H. melpomene*/silvaniform group as well as the *H. erato*/*sara* group of *Heliconius* (Nadeau et al. 2016), although the precise role of the inversion in mimicry is unknown for the species studied here (Edelman et al. 2019). As for the 2b region, a pair of scenarios correspond to alternative histories differing in direction of introgression between the common ancestor of the *sara* clade and *H. telesiphe* (**Fig. 7b**). Again, both scenarios are consistent with the pattern of shared derived orientation of the inversion region among *H. telesiphe*, *H. hecalesia*, *H. demeter* and *H. sara*, as represented by tree iv (**Fig. 2**). In both scenarios, the inversion originated along the branch between the root and the introgression between the ancestral populations of *H. telesiphe* and *H. demeter*/*H. sara* (**Fig. 7b**). As before, we used the root age (τ_r under scenario 1 and τ_{ehtc} under scenario 2) to distinguish between these two possibilities. Note that the divergence time estimates in the 15b region are also highly comparable to those in other genomic regions, with regression slopes ~1.00–1.07 for the noncoding loci (**Fig. S8**). The comparison suggests that introgression from the common ancestor of *H. sara* and *H. demeter* into the ancestor of *H. telesiphe* was more compatible with the other regions of the genome than the opposite direction (**Fig. S4**, **Table S9**), even though the opposite direction of introgression was marginally more prevalent across the rest of the genome (**Fig. 6**). Our result for 15b is concordant with the findings of Edelman et al. (2019), who used internal branch lengths and D_{XY} to suggest suggest introgression as a likely scenario, although they did not consider the alternative scenario 2 (**Fig. 7b**) of introgression.

Our results provide evidence that introgression according to scenario 1 (**Fig. 7**) between (*H. sara*, *H. demeter*) and *H. telesiphe* rather than ancestral polymorphism is a more likely explanation for the history of both 2b and 15b inversion regions, and that the introgression was almost wholly unidirectional. Additional genomes with inversion status from diverse species from this *erato-sara* clade and better breakpoint characterization would be useful to test these conclusions about the alternative scenarios.

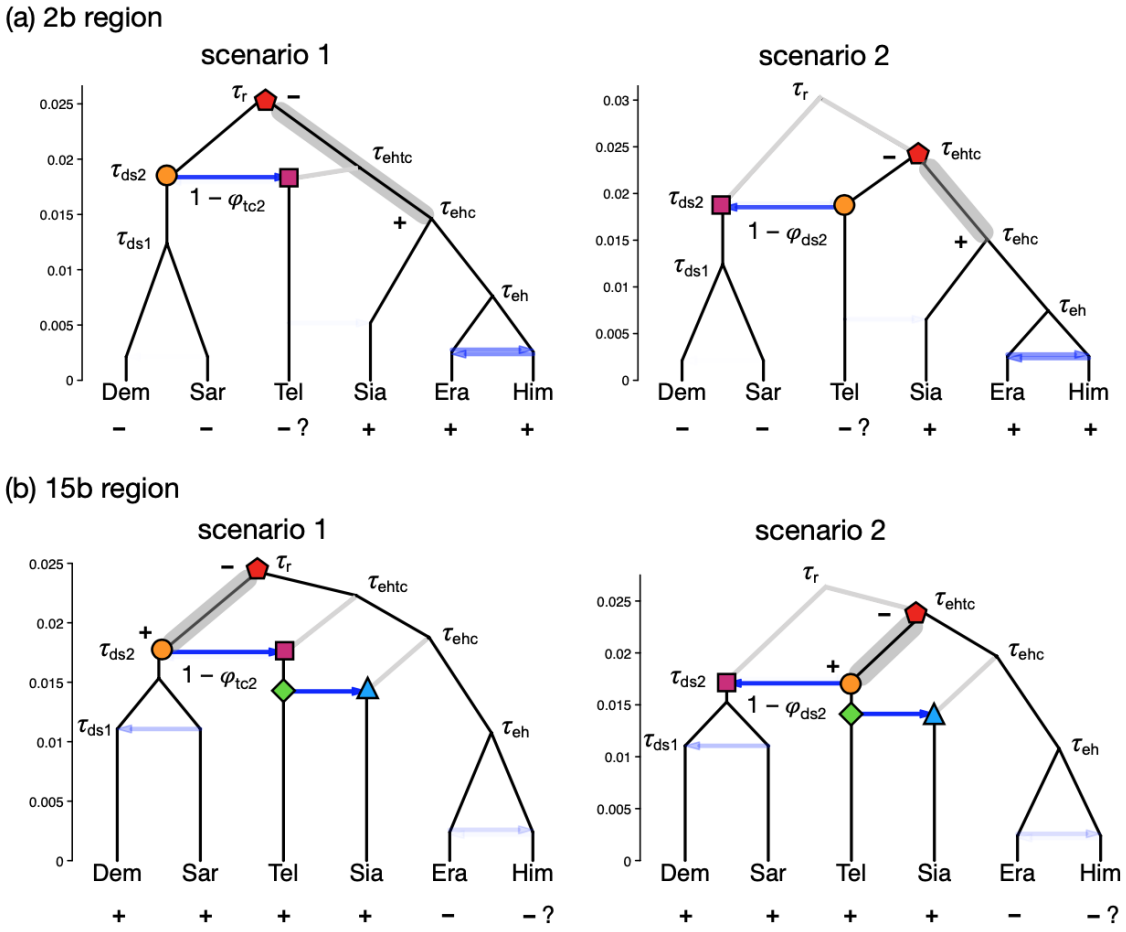


Figure 7. Within-model unidentifiability and possible introgression histories for inversion regions (**a**) 2b and (**b**) 15b, showing estimated divergence times from BPP analysis under the MSci model. Plus (+) represents the derived inverted orientation while minus (–) is the ancestral non-inverted orientation. Species with uncertain inversion orientation are marked with '?'. Node symbols and colours indicate matching nodes and times between the two scenarios. The estimates were based on posterior means from noncoding loci. Grey band indicates an ancestral population within which the derived inverted orientation (+) may have arisen from the non-inverted orientation (–). Note that parameters associated with nodes reached by very few sequences (indicated as grey branches) were expected to be poorly estimated, with a wide posterior interval. For example, in scenario 1 for 2b, τ_{ehtc}, θ_{ehtc} and θ_{ehc} were poorly estimated (**Fig. S4**, **Table S9**).

Conclusion

The full-likelihood approach employed here was able not only to detect gene flow among species, but also to estimate its direction, timing and magnitude by testing different models of species divergence and gene flow using genomic data. We found robust evidence of introgression across the genome involving distantly related species deep in the phylogeny as well as sister species in shallower parts of the tree. We confirm ancestral gene flow between the *sara* clade and an ancestral population of *H. telesiphe*, and infer a likely hybrid origin of *H. hecalesia* as well as gene flow between the sister species *H. erato* and *H. himera*. We clarify how introgression among ancestral species can explain the history of two chromosomal inversions deep in the phylogeny of the *erato-sara* group. For the first time, we estimate key population parameters such as species

874 divergence times and population sizes for modern and ancestral species, providing an opportunity to
875 understand the speciation and introgression history in finer detail than approximate/heuristic
876 approaches.

877

878 ## COMPETING INTEREST STATEMENT

879 The authors declare no competing interests.

880

881 ## SUPPLEMENTARY MATERIAL

882 The multilocus sequence datasets for bpp and 3s analyses are available in Zenodo at
883 https://doi.org/10.5281/zenodo.5078147. The C program for calculating the genotyping error rate
884 given the base-calling error and read depth is at https://github.com/abacus-gene/genotypecall.

885

886 ## ACKNOWLEDGEMENTS

887 We thank Nathaniel Edelman, Paul Frandsen, Michael Miyagi, and Bernardo Clavijo for their inputs
888 to the overall project.
889
890 ## FUNDING

891 The Broad Institute of Harvard and MIT funded the original sequencing of the *Heliconius* genomes
892 studied here. This study was supported by Harvard University and by Biotechnology and Biological
893 Sciences Research Council grant (BB/P006493/1 to Z.Y.) and a BBSRC equipment grant
894 (BB/R01356X/1).
895

REFERENCES

FIGURE LEGENDS

Figure 1. *The genotype-calling error rate for a given base-calling error (ε) and read depth (n) when the true genotype is (**a**) a homozygote for the reference allele (GT = 00) or (**b**) a heterozygote (GT = 01). Note that the genotype-calling error does not decrease monotonically with the increase in n when ε is fixed, or with the reduction in ε when n is fixed. The calculation is performed using a C program written by Z.Y. that implements the ML method of Li (2011) and calculates its error rate.*

Figure 2. *Posterior probabilities of species trees from BPP analysis of (**a**) blocks of 100 loci and (**b**) blocks of 200 loci across the genome (Table S2) under the MSC model without gene flow. The y-axis represents the posterior probability and ranges from 0 to 1. Trees (i)–(x) in the legend are MAP trees in at least one block. Tree (xi) appeared as one of the top eight trees in the sliding-window analysis of Edelman et al. (2019) but not in our analysis. Tree colours match those in Edelman et al. (2019) (their Fig. 2). Era: H. erato, Him H. himera, Sia: H. hecalesia, Tel: H. telesiphe, Dem: H. demeter, Sar: H. sara.*

Figure 3. *Maximum-likelihood estimates of the pairwise migration rate (M = Nm) under the IM model from 3s using all coding and noncoding loci in the autosomes (left column) and in the Z chromosome (right column; see **Table S4** for the number of loci) for each pair of species. The donor and recipient species of gene flow are given in the y- and x-axis, respectively. H. melpomene was used as an outgroup. Only significant migration rate estimates (by likelihood ratio test at $p < 0.01$) larger than 0.001 are shown. See **Figure S2** and **Table S8** for estimates of other parameters.*

Figure 4. *Proposed introgression model based on the BPP species tree estimation under MSC and 3s analysis under the IM model. Each horizontal arrow represents a unidirectional introgression event. Branch lengths represent posterior means of divergence/introgression times in the BPP analysis of all 6,030 noncoding loci in chromosome 1. Estimates for other chromosomal regions are in **Figures 5** and **S4**. Note that the introgression probability φ_d from H. Sara to H. demeter was estimated to be zero in all chromosomal regions under this model (**Fig. 5**).*

Figure 5. *Posterior means (dots) and 95% HPD intervals (bars) of introgression probabilities (φ) under the MSci model of **Figure 4** obtained using BPP for the 25 chromosomal regions (see **Table S4** for the number of loci). Results for other parameters in the model are in **Figure S4**. For 2b and 15b, there is an alternative set of posterior estimates resulting from within-model unidentifiability, shown in **Figure 7**.*

Figure 6. *Estimated introgression history in each chromosomal region obtained from BPP analysis under the MSci model of **Figure 4** using (**a**) noncoding and (**b**) coding loci. Intensity of the horizontal edges represents the posterior mean of introgression probability, while the y-axis represents the divergence times in the units of the expected number of mutation per site.*

Figure 7. *Within-model unidentifiability and possible introgression histories for inversion regions (**a**) 2b and (**b**) 15b, showing estimated divergence times from BPP analysis under the MSci model. Plus (+) represents the derived inverted orientation while minus (–) is the ancestral non-inverted orientation. Species with uncertain inversion orientation are marked with '?'. Node symbols and colours indicate matching nodes and times between the two scenarios. The estimates were based on posterior means from noncoding loci. Grey band indicates an ancestral population within which the derived inverted orientation (+) may have arisen from the non-inverted orientation (–). Note that parameters associated with nodes reached by very few sequences (indicated as grey branches) were expected to be poorly estimated, with a wide posterior interval. For example, in scenario 1 for 2b, τ_{ehtc}, θ_{ehtc} and θ_{ehc} were poorly estimated (**Fig. S4**, **Table S9**).*