Estimation of Cross-Species Introgression Rates using Genomic Data Despite Model Unidentifiability

Ziheng Yang[a,1] **and Tomáš Flouri**[a]

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

This manuscript was compiled on August 14, 2021

The multispecies coalescent with introgression (MSci) model accommodates both the coalescent process and cross-species introgression/hybridization events, two major processes that create genealogical fluctuations across the genome and gene-tree–species-tree discordance. Full likelihood implementations of the MSci model take such fluctuations as a major source of information about the history of species divergence and gene flow, and provide a powerful tool for estimating the direction, timing and strength of cross-species introgression using multilocus sequence data. However, introgression models, in particular those that accommodate bidirectional introgression (BDI), are known to cause unidentifiability issues of the label-switching type, whereby different models or parameters make the same predictions about the genomic data and thus cannot be distinguished by the data. Nevertheless, there has been no systematic study of unidentifiability when full likelihood methods are applied. Here we characterize the unidentifiability of arbitrary BDI models and derive simple rules for its identification. In general, an MSci model with k BDI events has 2^k unidentifiable towers in the posterior, with each BDI event between sister species creating within-model unidentifiability and each BDI between non-sister species creating cross-model unidentifiability. We develop novel algorithms for processing Markov chain Monte Carlo (MCMC) samples to remove label switching and implement them in the BPP program. We analyze genomic sequence data from *Heliconius* butterflies as well as synthetic data to illustrate the utility of the BDI models and the new algorithms.

Multispecies coalescent | introgression | unidentifiability | BPP | MSci | label-switching

Genomic sequences sampled from modern species contain rich historical information concerning species divergences and cross-species gene flow. In the past two decades, analysis of genomic sequence data has demonstrated the widespread nature of cross-species hybridization or introgression (1, 2). A number of statistical methods have been developed to infer introgression using genomic data, most of which use data summaries such as the estimated gene trees (3–5). Full-likelihood methods applied directly to multi-locus sequence alignments (6–8) allow estimation of evolutionary parameters including the timing and strength of introgression, as well as species divergence times and population sizes for modern and extinct ancestral species. These have moved the field far beyond simply testing for the presence of cross-species gene flow.

Models of cross-species introgression are known to cause unidentifiability issues, whereby different introgression models make the same probabilistic predictions about multilocus sequence data, and cannot be distinguished by such data (9–12). If the probability distributions of the data are identical under model m with parameters Θ and under model m' with parameters Θ', with

$$f(X|m,\Theta) = f(X|m',\Theta') \qquad [1]$$

for essentially all possible data X, the models are unidentifiable by data X. Here we use the term *within-model unidentifiability* if $m = m'$ and $\Theta \neq \Theta'$, or *cross-model unidentifiability* if $m \neq m'$. In the former case, two sets of parameter values in the same model are unidentifiable, whereas in the latter, two distinct models are unidentifiable. There have been very limited studies of unidentifiability of introgression models, which examined heuristic methods that use gene tree topologies (either rooted or unrooted) as data (10–12), but the issue has not been studied when full-likelihood methods are applied. Note that unidentifiability depends on the data and the method of analysis. An introgression model unidentifiable given gene tree topologies alone may be identifiable given gene trees with coalescent times. Similarly, a model unidentifiable using heuristic methods may be identifiable when full likelihood methods are applied to the same data. It is thus important to study the problem when full likelihood methods are applied, because unidentifiability by a heuristic method may reflect its inefficient use of information in the data rather than the intrinsic difficulty of the inference problem (13).

Among the different types of MSci models developed (6–8), the bidirectional-introgression (BDI) model (or model D in (8), fig. 1a) is one of the most useful in real data analysis (e.g., 14, 15). The basic BDI model for two species involves nine parameters, with $\Theta = (\theta_A, \theta_B, \theta_X, \theta_Y, \theta_R, \tau_R, \tau_X, \varphi_X, \varphi_Y)$ (fig. 1a). Note that an introgression model is similar to a species tree except that there are hybridization nodes representing cross-species introgressions, besides speciation nodes representing species divergences. While a speciation node has one parent and two daughters, a hybridization node has two parents and one daughter. The introgression probabilities (φ and $1-\varphi$) describe the contributions of the two parental populations to the hybrid species. When we trace the genealogical history of a sample of sequences from the modern species backwards in time and reach a hybridization node, each of the sequences takes the two parental paths with probabilities φ and $1-\varphi$. There are thus three types of parameters in an introgression (or MSci) model: the times of species divergence and introgression (τs), the (effective) population sizes of modern and ancestral species (θs), and the introgression probabilities (φs). Both the divergence times (τs) and population sizes (θs) are measured in the expected number of mutations per site.

The BDI model, in the case of two species (fig. 1), is noted to have an unidentifiability issue (8). Let Θ' be a set of parameters with the same values as Θ except that $\varphi'_X = 1-\varphi_X$, $\varphi'_Y = 1-\varphi_Y$, $\theta'_X = \theta_Y$, and $\theta'_Y = \theta_X$. Then $f(G|\Theta) = f(G|\Theta')$ for any gene tree G (fig. 1b&c). Thus for every point Θ in the parameter space, there is a 'mirror' point Θ' with exactly the same likelihood. With Θ, the A

ZY and TF conceived the research, designed and conducted the experiment, and wrote the paper.

No competing interests are declared.

[2]To whom correspondence should be addressed. E-mail: z.yangucl.ac.uk

sequences take the left (upper) path at X and enter population RX with probability $1-\varphi_X$, coalescing at the rate $\frac{2}{\theta_X}$, while with Θ', the same A sequences may take the right (horizontal) path and enter population RY with probability $\varphi'_X = 1-\varphi_X$, coalescing at the rate $\frac{2}{\theta'_Y} = \frac{2}{\theta_X}$. The differences between Θ and Θ' are in the labelling, with 'left' and X under Θ corresponding to 'right' and Y under Θ', but the probabilities involved are the same. The same argument applies to sequences from B going through node Y, and to any numbers of sequences from A and B considered jointly. Thus $f(G|\Theta) = f(G|\Theta')$ for essentially all G. If the priors on φ_X and φ_Y are symmetrical, say $\varphi \sim \text{beta}(\alpha,\alpha)$, the posterior density will satisfy $f(\Theta|X) = f(\Theta'|X)$ for all X. Otherwise the "twin towers" may not have exactly the same height.

The situation is very similar to the label-switching problem in Bayesian clustering (16–19). Consider data $X = \{x_i\}$ as a sample from a mixture of two normal distributions, $\mathbb{N}(\mu_1,1)$ and $\mathbb{N}(\mu_2,1)$ with the mixing proportions p_1 and $1-p_1$. Let $\Theta = (p_1, \mu_1, \mu_2)$ be the parameter vector. Then $\Theta' = (1-p_1, \mu_2, \mu_1)$ will have exactly the same likelihood, with $f(X|\Theta) = f(X|\Theta')$ for essentially all data X. In effect, the labels 'group 1' and 'group 2' are switched between Θ and Θ'.

As an example, we fit the BDI model of figure 2a to the first 500 noncoding loci on chromosome 1 in the genomic data from three *Heliconius* butterfly species: *H. melpomene*, *H. timareta*, and *H. numata* (14, 20). Figure 3a shows the trace plots for parameters φ_X and φ_Y from a Markov chain Monte Carlo (MCMC) run. The Markov chain moves between two peaks, centered around $(\varphi_X, \varphi_Y) = (0.35, 0.1)$ and $(0.65, 0.9)$, respectively. In effect, the algorithm is switching between Θ and Θ' and changing the definition of parameters. This is a label-switching problem, as occurs in Bayesian clustering. The usual practice of estimating parameters by their posterior means (which are 0.54 for φ_X and 0.62 for φ_Y in fig. 3a) and constructing the credibility intervals is inappropriate. Indeed the posterior distribution of Θ is exactly symmetrical with twin towers, and if the chain is run long enough, the posterior means of φ_X and φ_Y will be exactly $\frac{1}{2}$. The results are similar when the first 500 exonic loci are analyzed, in which the Markov chain moves between two towers centered around $(0.3, 0.1)$ and $(0.7, 0.9)$ (fig. S1a).

Unidentifiable models cannot be applied to real data as they are trying to "distinguish the indistinguishable" (10). Results such as those of figures 3a & S1a raise two questions. First, are BDI models with more than two species or two BDI events unidentifiable, and what are the rules? Second how do we deal with the problem of label-switching and make the models useful for real data analyses? Those two problems are addressed in this paper. We study the unidentifiability issue of BDI models for an arbitrary number of species with an arbitrary species tree, when a full-likelihood method is applied to multilocus sequence data. It has been conjectured that an MSci model is identifiable by full likelihood methods using data of multilocus sequence alignments if and only if it is identifiable when the data consist of gene trees with coalescent times (8). We make use of this conjecture and consider two BDI models to be unidentifiable if and only if they generate the same distribution of gene trees with coalescent times. We identify general rules for the unidentifiability of the BDI models. We then develop new algorithms for post-processing the MCMC samples generated from a Bayesian analysis under the BDI model to remove the label-switching. Those advancements make the BDI models usable for real data analysis despite their unidentifiability. We use the BPP program to analyze synthetic datasets as well as genomic data from *Heliconius* butterflies to demonstrate the utility of the BDI models and the new algorithms.

Theory

The rule of unidentifiability of BDI models. Suppose species A and B exchange migrants at time $\tau_X = \tau_Y$ through bidirectional introgression (fig. 4). To study the backwards-in-time process of coalescent and introgression, we can treat nodes X and Y as one node, XY. When sequences from A reach node XY, each of them has probability $1-\varphi_X$ of taking the left parental path (RX) and probability φ_X of taking the right parental path (SY). Similarly when sequences from B reach node XY, they have probabilities φ_Y and $1-\varphi_Y$ of taking the left (RX) and right (YS) parental paths, respectively. If we swap branches A and B, carrying their population size parameters (θ) and introgression probabilities (φ) in the process, the probability density of the gene-trees remains unchanged. Thus the species tree-parameter combinations (S,Θ) and (S',Θ') of figure 4b&c are unidentifiable. The processes of coalescent and introgression before reaching nodes A and B are identical between Θ and Θ', as are the processes past nodes X and Y. For example, the rule still applies if each of A and B is a subtree, with introgression events inside, or if there are introgression events involving a descendant of A and a descendant of B.

In the case of two species, the parental species R and S (fig. 4) are one node, and the species trees (A,B) and (B,A) are the same. As a result, Θ and Θ' in figure 4 correspond to two sets of parameter values in the same model, so this is a case of within-model unidentifiability. Otherwise the unidentifiability will be cross-model.

Canonical cases of BDI models. Here we study major BDI models to illustrate the rule of unidentifiability and to provide reference for researchers who may apply those models to analyze genomic datasets.

If we add subtrees onto branches XA, YB, or the root branch R in the two-species tree of figure 1a, so that the BDI event remains to be between two sister species, the model will exhibit within-model parameter unidentifiability (fig. S2), just like the basic model of figure 1a.

If the BDI event is between non-sister species, the model exhibits cross-model unidentifiability. Figures S3a&a′ show a model with a BDI event between cousins, while in figures S3b&b′, the two species involved in the BDI event are more distantly related.

Figures S4a, b &c show three models each with a BDI event between non-sister species. In figure S4a, X and Y are non-sister species on the original binary species tree. In figure S4b&c, X and Y are non-sister species because there are introgression events involving branches RX and/or RY. In all three cases, there is cross-model unidentifiability, with the twin towers shown in S4a′, b′, &c′.

The case of two non-sister BDI events for three species is illustrated in figure S5. According to our rule, there are four unidentifiable models in the posterior, with parameter mappings shown in figure S5. One way of seeing that the four models are equivalent or unidentifiable is to assume that the introgression probabilities (φ_X, φ_Y, φ_Z, and φ_W) are all $< \frac{1}{2}$, and then work out the major routes taken when we trace the genealogical history of sequences sampled from modern species. In such cases, all four models of figure S5 predict the following: most sequences from A will take the route ZR at node ZW with probability $1-\gamma$; most sequences from B will take the route X-W at node XY (with probability $1-\alpha$), then the route WS at node ZW (with probability $1-\delta$), before reaching SR; and most sequences from C will take the route YS at node XY (with probability $1-\beta$, before reaching SR. Of course the four models are unidentifiable whatever values the introgression probabilities take. Those models have been used to analyze genomic data from Texas horned lizard (*Phrynosoma cornutum*) (15).

Figure 5 shows two models for five species, each model involving

three BDI events. In figure 5a, all three BDI events involve sister species, so that there are $2^3 = 8$ unidentifiable within-model towers in the posterior. In figure 5b, one BDI event involves non-sister species while two involve sister species, so that there are two unidentifiable models, each of which has four unidentifiable within-model towers in the posterior.

In general, if there are m BDI events between sister species and n BDI events between non-sister species, there will be 2^m unidentifiable models, each having 2^n within-model unidentifiable towers.

Unidentifiability of double-DBI models. Figure 6a shows two BDI events between species A and B, which occurred at times $\tau_X = \tau_Y$ and $\tau_Z = \tau_W$, respectively. To apply the rule of figure 4, we treat Z and W as one node so that X and Y are considered sister species. There are then four unidentifiable within-model towers in the posterior space, shown as Θ_1-Θ_4 in fig. 6. The parameter mappings are

Θ	φ_X	φ_Y	θ_X	θ_Y	φ_Z	φ_W	θ_Z	θ_W
$\Theta_1 : \varphi_X < \frac{1}{2}, \varphi_Z < \frac{1}{2}$	α	β	θ_X	θ_Y	γ	δ	θ_Z	θ_W
$\Theta_2 : \varphi_X < \frac{1}{2}, \varphi_Z > \frac{1}{2}$	α	β	θ_X	θ_Y	$1-\gamma$	$1-\delta$	θ_W	θ_Z
$\Theta_3 : \varphi_X > \frac{1}{2}, \varphi_W < \frac{1}{2}$	$1-\alpha$	$1-\beta$	θ_Y	θ_X	δ	γ	θ_W	θ_Z
$\Theta_4 : \varphi_X > \frac{1}{2}, \varphi_W > \frac{1}{2}$	$1-\alpha$	$1-\beta$	θ_Y	θ_X	$1-\delta$	$1-\gamma$	θ_Z	θ_W

[2]

In general, with k BDI events between two species, which occurred at different time points in the past, there will be 2^k unidentifiable within-model towers in the posterior. There may be little information in practical datasets to estimate so many parameters: if all sequences have coalesced before they reach the ancient introgression events near the root of the species tree, the introgression probabilities (φs) and the associated population sizes (θs) will be nearly impossible to estimate. Thus we do not consider more than two BDI events between two species. Note that even the model with one BDI event is not identifiable by heuristic methods that use gene tree topologies only. A small simulation is conducted to illustrate the feasibility of applying the double-BDI model (fig. 6) to genomic datasets; see Results.

Addressing unidentifiability issues and difficulties with identifiability constraints. According to our rule, MSci models with BDI events can create both within-model and cross-model unidentifiability. Cross-model unidentifiability is relatively simple to identify and deal with. If the MCMC is run with the MSci model fixed (8), only one of the models (e.g., model S_1 with parameters Θ_1 in fig. S5) is visited in the chain. One can then summarize the posterior distribution for parameters under that model (which may be smooth and single-moded), and the posterior summary may be mapped onto the other unidentifiable models according to the rule. See ref. (15) for such an application of BDI models of figure S5. If the MCMC is trans-model and visits different models in the chain (6, 7), the posterior space is symmetrical between the unidentifiable models (such as models S_1–S_4 of fig. S5). However, such symmetry is unlikely to be achieved in the MCMC sample due to well-known mixing difficulties of trans-model MCMC algorithms. One may choose to focus on one of the models (e.g., S_1 of fig. S5) and post-process the MCMC sample to map the sample onto the chosen model before producing the within-model posterior summary. Oftentimes the MCMC may explore the within-model posterior space very well, despite difficulties of moving from one model to another. In all cases, the researcher has to be aware of the unidentifiable models which are equally good explanations of the data (see Discussion).

Our focus here is on within-model unidentifiability created by BDI events between sister species. When there are multiple modes in the posterior, each mode may offer a sensible interpretation of the data, but it is inappropriate to merge MCMC samples from different modes, or to construct posterior summaries such as the posterior means and CIs using MCMC samples that traverse different modes. It is instead more appropriate to summarize the samples for each mode.

A common strategy for removing label-switching is to apply so-called *identifiability constraints*. In the simple BDI model of figure 1, any of the following constraints may be applicable: $\varphi_X < \frac{1}{2}$, $\varphi_Y < \frac{1}{2}$, and $\theta_X < \theta_Y$. Such identifiability constraints may be imposed during the MCMC or during post-processing of the MCMC samples. As discussed previously (17, 18), such a constraint may be adequate if the posterior modes are well separated, but may not work well otherwise. For example, when φ_X is far away from $\frac{1}{2}$ in all MCMC samples, it is simple to process the MCMC sample to impose the constraint $\varphi_X < \frac{1}{2}$. This is the case in analyses of the large datasets in this paper, for example, when all noncoding and exonic loci from chromosome 1 of the *Heliconius* data are analyzed (table 1). However, when the posterior modes are not well-separated (either because the true parameter value is close to the boundary defined by the inequality or because the data lack information so that the CIs are wide), different identifiability constraints can lead to very different parameter posteriors (16), and an appropriate constraint may not exist. A serious problem in such cases is that imposing an identifiability constraint may generate posterior distributions over-represented near the boundary, with seriously biased posterior means (17, 18). For example, φ_X may have substantial density mass both below and above $\frac{1}{2}$, and imposing the constraint $\varphi_X < \frac{1}{2}$ will artificially generate high density mass close to $\varphi_X = \frac{1}{2}$. Similarly the posterior distributions of θ_X and θ_Y may overlap, so that the constraint $\theta_X < \theta_Y$ may not be appropriate.

New algorithms to process MCMC samples from the BDI model to remove label switching. One approach to dealing with label-switching problems in Bayesian clustering is *relabelling*. The MCMC is run without any constrain, and the MCMC sample is then post-processed to remove label-switching, by attempting to move each point in the MCMC sample to its alternative unidentifiable positions in order to, as far as possible, make the marginal posterior distributions smooth and unimodal (17, 18). The processed sample is then summarized to generate the posterior of the parameters. Here we follow this strategy and implement three relabelling algorithms for use with the BDI model.

Let $\Theta = (\varphi_X, \varphi_Y, \theta_X, \theta_Y)$, which has a mirror point $\Theta' = (\varphi'_X, \varphi'_Y, \theta'_X, \theta'_Y) = (1-\varphi_X, 1-\varphi_Y, \theta_Y, \theta_X)$ (fig. 1). The other parameters in the model are not involved in the unidentifiability and are simply copied along. Let Θ_t, $t = 1, \cdots, N$, be the N samples of parameters generated by the MCMC algorithm. Each sample is a point in the 4-D space. Let σ_t be a transform for point t, with $\sigma_t(\Theta_t) = \Theta_t$ to be the original point, and $\sigma_t(\Theta_t) = \Theta'$ to be the transformed or mirror point (fig. 1b&c). With a slight abuse of notation, we also treat σ_t as an indicator, with $\sigma_t = 0$ and 1 representing Θ_t and Θ'_t, respectively. For each sample t, we choose either the original point or its mirror point, to make the posterior of the parameters look smooth and single-moded as far as possible. The first two algorithms, called center-of-gravity algorithms CoG_0 and CoG_N, loop through two steps.

Algorithms CoG_0 and CoG_N. Initialize. For each point t, $t = 1, \cdots, N$, pick either the original point (Θ_t) or its mirror point (Θ'_t). We set σ_t to 0 (for Θ_t) if $\varphi_X < \frac{1}{2}$ or $\varphi_Y < \frac{1}{2}$, or to 1 (for Θ'_t) otherwise.

- Step 1. Determine the center of gravity, given by the sample means of the parameters, $\mu = (\bar{\varphi}_X, \bar{\varphi}_Y, \bar{\theta}_X, \bar{\theta}_Y)$.
- Step 2. For each point $t = 1, \cdots, N$, compare the current and its mirror positions and choose the one closer to the center of

gravity (μ).

In step 2, we use the Euclidean distance

$$d_0(\Theta,\mu) = \left[\sum_j^4 (\phi_j - \mu_j)^2\right]^{1/2}, \quad [3]$$

where ϕ_j are the four parameters in Θ: $\varphi_X, \varphi_Y, \theta_X, \theta_Y$. This is algorithm CoG_0.

If we consider different scales in the different dimensions (for example, φ_X and θ_X may have very different posterior variances), we can calculate the sample variances v (in addition to the sample means μ) in step 1 and use them as weights to normalize the differences in step 2, with

$$d_N(\Theta,\mu) = \left[\sum_j^4 \frac{1}{v_j}(\phi_j - \mu_j)^2\right]^{1/2}. \quad [4]$$

We refer to this as algorithm CoG_N.

The third algorithm, called the β–γ algorithm, follows the relabelling algorithm in ref. (18) for Bayesian clustering. We use maximum likelihood (ML) to fit the sample $\{\Theta_t\}$ to independent beta distributions for φ_X and φ_Y and gamma distributions for θ_X and θ_Y:

$$f(\Theta;\omega) = b(\varphi_X;p_X,q_X)\cdot b(\varphi_Y;p_Y,q_Y) \times g(\theta_X;a_X,b_X)\cdot g(\theta_Y;a_Y,b_Y), \quad [5]$$

where

$$b(\phi;p,q) = \frac{1}{B(p,q)}\phi^{p-1}(1-\phi)^{q-1},$$
$$g(\phi;a,b) = \frac{b^a}{\Gamma(a)}\phi^{a-1}\,\mathrm{e}^{-b\phi} \quad [6]$$

are the beta and gamma densities and where $\omega = (p_X, q_X, p_Y, q_Y, a_X, b_X, a_Y, b_Y)$ is the vector of hyper-parameters.

The log likelihood, as a function of the hyper-parameters ω and the transforms $\sigma = \{\sigma_t\}$, is

$$\ell(\omega,\sigma) = \sum_t^N \ell_t(\omega,\sigma_t(\Theta_t)) = \sum_t^N \log f(\sigma_t(\Theta_t);\omega). \quad [7]$$

We have implemented the following iterative algorithm to estimate ω and σ by maximizing ℓ.

Algorithm β–γ. Initialize $\sigma_t, t = 1,\cdots,N$. As before, we set σ_t to 0 (for Θ_t) if $\varphi_X < \frac{1}{2}$ or $\varphi_Y < \frac{1}{2}$, or to 1 (for Θ'_t) otherwise.

- Step 1. Choose $\hat{\omega}$ to maximize the log likelihood ℓ (eq. 7) with the transforms σ fixed.
- Step 2. For $t = 1,\cdots,N$, choose σ_t to maximize $\ell_t(\hat{\omega},\sigma_t(\Theta_t))$ with the hyper-parameters ω fixed. In other words compare Θ_t and Θ'_t and choose the one that better fits the beta and gamma distributions.

Step 1 fits two beta and two gamma distributions by ML and involves four separate 2-D optimization problems. The maximum likelihood estimates (MLEs) of p and q for the beta distribution $b(\phi;p,q)$ are functions of $\sum_t \log\phi_t$ and $\sum_t \log(1-\phi_t)$, whereas the MLEs of a and b for the gamma distribution $g(\phi;a,b)$ are functions of $\sum_t \phi_t$ and $\sum_t \log\phi_t$. These are simple optimization problems, which we solve using the BFGS algorithm in the PAML program (21). Step 2 involves N independent optimization problems, each comparing two points ($\sigma_t = 0$ and 1), with ω fixed. It is easy to see that the algorithm is nondecreasing (that is, the log likelihood ℓ never decreases) and converges, as step 1 involves ML estimation of parameters in the beta and gamma distributions, and step 2 involves comparing two points.

Note that algorithm β–γ becomes algorithm CoG_N if the beta and gamma densities are replaced by normal densities.

Algorithms CoG_0, CoG_N, and β–γ for the double-BDI model (fig. 6a). Under the double-BDI model, there are four within-model unidentifiable towers, specified by eight parameters (eq. 2). Thus σ_t takes four values $(0,1,2,3)$. Let $\Theta = (\varphi_X,\varphi_Y,\varphi_Z,\varphi_W,\theta_X,\theta_Y,\theta_Z,\theta_W)$. We use the same strategy and fit four beta distributions to the φs and four gamma distributions to the θs, with 16 hyper-parameters in ω. We implement the three algorithms (β–γ, CoG_N, and CoG_0) as before. We prefer the tower in which the introgression probabilities are small and initialize the algorithm accordingly. The transforms (σ_t) are as follows (eq. 2)

- $\sigma_t = 0$: if the parameters are in Θ_1, do nothing.
- $\sigma_t = 1$: if in Θ_2, let $\varphi_Z = 1-\varphi_Z$, $\varphi_W = 1-\varphi_W$, and swap θ_Z and θ_W.
- $\sigma_t = 2$: if in Θ_3, let $\varphi_X = 1-\varphi_X$, $\varphi_Y = 1-\varphi_Y$, swap θ_X and θ_Y, swap φ_Z and φ_W, swap θ_Z and θ_W;
- $\sigma_t = 3$: if in Θ_4, let $\varphi_X = 1-\varphi_X$, $\varphi_Y = 1-\varphi_Y$, swap θ_X and θ_Y, and let $\varphi_Z = 1-\varphi_W$ and $\varphi_W = 1-\varphi_Z$.

The algorithms are implemented in C and require minimal computation and storage. Processing 5×10^5 samples takes several seconds, mostly spent on reading and writing files. The algorithms are integrated into the BPP program (22) so that MCMC samples from various BDI models are post-processed and summarized automatically. We also provide a stand-alone program in the github repository `abacus-gene/bpp-msci-D-process-mcmc`.

Results

Introgression between *Heliconius melpomene* and *H. timareta*. We fitted the BDI model of figure 2 to the genomic sequence data from three species of *Heliconius* butterflies: *H. melpomene*, *H. timareta*, and *H. numata* (14, 20). When we used the first 500 loci, either noncoding or exonic, there was substantial uncertainty in the posterior of φ_X and φ_Y, and the MCMC jumped between the twin towers, and the marginal posteriors had multiple modes, due to label switching (figs. 3a & S1a). Post processing of the MCMC sample using the new algorithms led to single-moded marginal posterior distributions (figs. 3b–d & S1b–d). The three algorithms produced extremely similar results for both datasets. For example, the posterior mean and 95% CI for φ_X from the noncoding data were 0.356 (0.026, 0.671) by CoG_0, 0.357 (0.026, 0.674) by CoG_N, and 0.354 (0.022, 0.664) by β–γ, while those for φ_Y were 0.103 (0.000, 0.304) by CoG_0 and CoG_N, and 0.104 (0.000, 0.306) by β–γ.

We then analyzed all the 2592 noncoding and 3023 exonic loci on chromosome 1. With the large datasets, the parameters were better estimated with narrower CIs and the unidentifiable towers were well-separated. In fact, the MCMC visited only one of the two towers, but that tower was well explored so that multiple runs produced highly consistent results. We started the MCMC with small values for φ_X and φ_Y, and post-processing the MCMC samples had no effect.

Estimates of all parameters from the small (with $L = 500$) and large datasets are summarized in table 1. In the small datasets, the introgression probabilities were $\varphi_X \approx 0.354$ (with the CI 0.022–0.664) for the noncoding data and 0.280 (with CI 0.002–0.547) for the coding loci, while φ_Y was 0.104 (CI 0.000–0.306) for the coding data and 0.116 (CI 0.000–0.318) for the exonic data. When all loci from chromosome 1 were used, φ_X was 0.124 (with the CI 0.007–0.243)

for the noncoding data and 0.161 (with CI 0.070–0.264) for the coding loci, while φ_Y was 0.048 (CI 0.000–0.139) for the coding data and 0.019 (CI 0.000–0.056) for the exonic data. The estimates were similar between the coding and noncoding data, with greater proportions of migrants in *H. timareta* from *H. melpomene* than in the opposite direction. This was so despite the fact that *H. melpomene* had a smaller effective population size than *H. timareta*. Note that *H. melpomene* has a widespread geographical distribution whereas *H. timareta* is restricted to the Eastern Andes; the small θ_M estimates are most likely due to the fact that the *H. melpomene* sample was from a partially inbred strain to avoid difficulties with genome assembly. Estimates of θs and τs were smaller for the coding loci than for the noncoding loci, due to selective constraint on nonsynonymous mutations.

Estimates of φ_X and φ_Y showed large differences between the small and large datasets, but they involved large uncertainties, with the CIs for large datasets mostly inside the CIs for the small datasets. One reason for the differences may be the variable rate of gene flow across the genome or chromosome. Note that φ in the MSci model reflects the long-term effects of gene flow and selection purging introgressed alleles, influenced by linkage to gene loci under natural selection.

Analysis of data simulated under the double-BDI model of figure 6a. We conducted a small simulation to illustrate the feasibility of the double-BDI model (fig. 6), simulating 10 replicate datasets of $L = 500$, 2000, and 8000 loci. The three algorithms were used to process the MCMC samples, before they were summarized. A typical case is shown in figure 7 for the case of $L = 500$. While there are four unidentifiable towers in the 8-D posterior space (eq. 2), the MCMC visited only two of them, with different values for parameters around the *ZW* BDI event. The dataset of $L = 500$ loci are very informative about the parameters for the recent BDI event at node *XY* (φ_X, φ_Y, θ_X, θ_Y), so that these had highly concentrated posteriors with well separated towers. We started the Markov chains with small values (e.g., 0.1 and 0.2) for φ_X and φ_Y, so that the sampled points were all around the correct tower for those parameters. If the chain started with large φ_X and φ_Y, it would visit a 'mirror' tower. Thus post-processing of the MCMC samples in the case of $L = 500$ mostly affected parameters around the BDI event at *ZW* (φ_Z, φ_W, θ_Z, θ_W). Figure 7 shows the effects on parameters φ_Z and φ_W using the β–γ algorithm. The CoG_0 and CoG_N algorithms produced nearly identical results, and all algorithms were effective in removing label switching. The post-processed samples were summarized to calculate the posterior means and the HPD CIs (fig. 8).

At $L = 2000$ or 8000 loci, the four towers were well-separated and the MCMC visited only one of them. Applying the post-processing algorithms either had no effect or, in rare occasions, moved all sampled points from another tower.

Posterior means and the 95% highest-probability-density (HPD) credibility intervals (CI) for all parameters were summarized in figure 8. Parameters around the BDI event at *ZW* (φ_Z, φ_W, θ_Z, θ_W) are the most difficult to estimate. Nevertheless, with the increase of data size, the CIs for all parameters become smaller, and the posterior means are converging to the true values. Note that while the simulation is conducted using one set of correct parameter values (say, Θ_1 of fig. 6), we consider the estimates to be good if they are close to any of the four towers (say, Θ_2, Θ_3, or Θ_4).

Analysis of data simulated with one BDI event with poorly separated modes. We simulated a more challenging dataset for the relabelling algorithms, with $L = 500$ loci under the BDI model of figure 1a with parameter values $(\varphi_X, \varphi_Y) = (0.7, 0.2)$ (see table S1). As φ_X and φ_Y are not too far away from $\frac{1}{2}$ and the dataset is small, the posterior modes are poorly separated, with considerable mass near $(\frac{1}{2}, \frac{1}{2})$. The unprocessed sample from BPP shows two modes for φ_Y, and one mode around $\frac{1}{2}$ for φ_X, with the posterior means at 0.51 for φ_X and 0.50 for φ_Y, very close to $\frac{1}{2}$ (fig. S6. These are misleading summaries, as the sample is affected by label switching. The three algorithms (β–γ, CoG_N, and CoG_0) produce similar results, with single-moded posterior, around the mirror tower $\Theta' = (0.3, 0.8)$. The posterior means for φ_X are 0.245, 0.236, and 0.235, for the three algorithms (β–γ, CoG_N, and CoG_0), and those for φ_Y are 0.553, 0.539, and 0.538 (table S1). The three algorithms have worked well even when the posterior modes are poorly separated.

The parameters involved in the label switching, φ_X, φ_Y, θ_X, θ_Y, are poorly estimated, due to the difficulty of separating the towers and to influence from the priors. The estimates should improve if more loci are used in the data. Other parameters in the model are all well estimated (table S1).

Discussion

Identifiability and low information content of MSci models. The identifiability of other MSci models implemented in BPP are simpler. MSci model A is consistent with three different biological scenarios (fig. 9a-c). In scenario A_1, two species *SH* and *TH* merge to form a hybrid species *HC*, but the two parental species become extinct after the merge. This scenario may be rare. In scenario A_2, species *SUX* contributes migrants to species *THC* at time τ_H and has since become extinct or is unsampled in the data. In scenario A_3, *TUX* is the ghost species. The three scenarios are unidentifiable using genomic data. Model B_1 assumes introgression from species *RA* to *TC* at time $\tau_S = \tau_H$ (fig. 9d). This is distinguishable using genetic data from the alternative model B_2 in which there is introgression from *RB* to *SC* (B_2, fig. 9e). Note that models B_1 and B_2 are both special cases of model A_1 with different constraints.

We note that there are many parameter settings and data configurations in which some parameters are hard to estimate, because the data lack information about them. For example, ancestral population sizes for short and deep branches in the species tree are hard to estimate, because most sequences sampled from modern species may have coalesced before reaching that population when we trace the genealogy of the sample backwards in time. Similarly, if not many sequences reach a hybridization node, there will be little information in the data about the introgression probabilities at that node. In such case, even if the model is identifiable mathematically, it may be nearly impossible to estimate the parameters with any precision even with large datasets.

In some cases, certain parameters may be very near the boundary of the parameter space, and this may create near unidentifiability with multiple modes in the posterior. For example, the introgression probability may be close to $\varphi = 0$ or 1, or speciation events may have occurred in rapid succession so that the mother and daughter nodes on the species tree have nearly the same age) (see (15) for an example). The MCMC samples around different modes should be summarized separately.

Estimation of introgression probabilities despite unidentifiability. The three algorithms for post-processing MCMC samples under the BDI model produced very similar results in our applications. In particular the simple center-of-gravity algorithms produced results as good as the more elaborate β–γ algorithm, despite the fact the normal distribution is a poor approximation to the posterior of the introgression probabilities (φ_X and φ_Y). This may be due to the fact that the distributions (or the distance in the CoG algorithms) are used to compare the sampled points with their unidentifiable mirror

points only, and are not used to directly approximate the posterior distribution of those parameters, which are estimated by using the processed samples. Similarly, while we fit independent distributions for parameters in the algorithms (eq. 6), there is no need to assume independence in the posterior for the algorithms to work. Furthermore, if there exist multiple modes in the posterior that are not due to label-switching, such genuine multimodality will not be removed by the algorithms (18).

A model with a label-switching type of unidentifiability can still be applied in real data analysis. In the clustering problem, the Bayesian analysis may reveal the existence of two groups, in proportions p_1 and $1-p_1$ with means μ_1 and μ_2, and it may not matter if it cannot decide which group should be called 'group 1'. The twin towers Θ and Θ' under the BDI model of figure 1 constitute a mathematically similar label-switching problem. However, Θ and Θ' may represent different biological scenarios or hypotheses. Suppose that species A and B are distributed in different habitats (dry for A and wet for B, say), and suppose the ecological conditions have changed little throughout the history of the species. Θ may mean that species A has been in the dry habitat over the whole time period since species divergence at time τ_R, while species B has been in the wet habitat, and they came into contact and exchanged migrants at time τ_X. In contrast, Θ' may mean that species A was in the wet habitat since species divergence while species B was in the dry habitat, but when they came into contact (at time τ_X) they nearly replaced each other, switching places, so that today species A is found in the dry habitat while species B in the wet habitat. The two sets of parameters Θ and Θ' may thus mean different biological hypotheses. The scenario of total replacement may be implausible for most systems, and in our algorithms, we start with the initial conditions $\varphi_X < \frac{1}{2}$ and/or $\varphi_Y < \frac{1}{2}$ as much as possible. When the introgression probabilities are intermediate, the biological interpretations may not be so clear-cut, but unidentifiability exists nevertheless. In the example of figure S6 and table S1, the choice between the two unidentifiable towers $\Theta = (\varphi_X, \varphi_Y) = (0.7, 0.2)$ and $\Theta' = (0.3, 0.8)$ may not be easy. Ultimately, genomic data from modern species provide information about the order and timings of species divergences and cross-species introgressions, but not about the geographical locations and ecological conditions in which the divergences and introgressions occurred. Unidentifiable models discussed in this paper are all of this nature. The algorithms we developed in this paper remove label switching in the MCMC sample, but do not remove the unidentifiability of the BDI models. The researcher has to be aware of the unidentifiability or the equally supported explanations of the genomic data.

In the current implementation of BDI models in BPP, each branch in the species tree is assigned its own population size parameter (8). We note that if all species on the species tree are assumed to have the same population size (θ), unidentifiability persists. However, if we assume that the population size remains unchanged by the introgression event: e.g., $\theta_X = \theta_A$ and $\theta_Y = \theta_B$ in figure 1, the model becomes identifiable. The assumption of the same population size before and after a migration event appears to be plausible biologically. It reduces the number of parameters by two for each BDI event, and removes unidentifiability. It may be worthwhile to implement such models. At any rate, the relabelling algorithms we have implemented makes it possible to apply the BDI models to genomic sequence data despite their unidentifiability.

Materials and Methods

Introgression in *Heliconius* butterflies. We fitted the BDI model to the genomic sequence data for three species of *Heliconius* butterflies: *H. melpomene*, *H. timareta*, and *H. numata* (23, 24). The species tree or MSci model assumed is shown in figure 2, with introgression between *H. melpomene* and *H. timareta*. The two species are known to hybridize, although no attempt has yet been made to infer the direction or magnitude of introgression (except for colour-pattern genes) (24). There are 31,166 autosomal noncoding loci and 36,138 autosomal exonic loci, with 2592 noncoding and 3023 exonic loci on chromosome 1. We conducted two sets of analysis, using either the first 500 loci or all the loci on chromosome 1.

We used gamma priors for the population sizes (θ) and for the age of the root (τ_0): $\theta \sim G(2, 400)$ with the mean 0.005 substitution per site, and $\tau \sim G(2, 400)$ with mean 0.005. The introgression probabilities were assigned beta priors $\varphi \sim B(1, 1)$, which is the uniform $\mathbb{U}(0, 1)$. We used a burn-in of 16000 iterations, and then took 2×10^5 samples, sampling every 5 iterations. Running time on a server with 9 threads of Intel Xeon Gold 6154 CPU (3.0GHz) was about 1 hour for the small datasets and 10 hours for the large ones.

Convergence of the MCMC algorithms was assessed by checking for consistency between independent runs, taking into account possible label-switching issues. In the large datasets analyzed in this paper, the MCMC typically visits only one of the unidentifiable towers, but that tower is well-explored, with the different runs producing highly consistent posterior after label switching is removed. In such cases, reliable inference is possible (cf.:(19)).

Simulation under the double-BDI model. We simulated and analyzed data to under the double-BDI model of figure 6. We generated gene trees with branch lengths (coalescent times) and sequences under the JC model (25). The parameters used are $\varphi_X = 0.1, \varphi_Y = 0.2, \varphi_Z = 0.2, \varphi_W = 0.3$, $\tau_R = 0.005$, $\tau_Z = \tau_W = 0.0025$, $\tau_X = \tau_Y = 0.00125$, $\theta_R = \theta_Z = \theta_X = \theta_A = 0.005$, and $\theta_W = \theta_Y = \theta_B = 0.02$. Each dataset consists of $L = 500, 2000$ and 8000 loci, with $S = 16$ sequences per species per locus, and with the sequence length to be 500 sites. The numbre of replicate datasets is 10.

The data were then analyzed using BPP under the double-BDI model (fig. 6) to estimate the 14 parameters. We use gamma priors $\tau_0 \sim G(2, 400)$ for the root age with the mean to be the true value (0.005), and $\theta \sim G(2, 200)$ with the mean 0.01 (true values are 0.005 and 0.02). We used a burn-in of 32,000 iterations, and then took 5×10^5 samples, sampling every 2 iterations. Analysis of each dataset took ~10hrs for $L = 500$ and ~ 130hrs for $L = 8000$, using 8 threads on a server. The MCMC samples were processed to remove label-switching before they are summarized to approximate the posterior distribution.

ACKNOWLEDGMENTS. We thank James Mallet and Fernando Seixas for providing the genomic datasets for the *Heliconius* butterflies, and James Mallet for comments on an earlier draft of the paper. This study has been supported by Biotechnology and Biological Sciences Research Council grant (BB/T003502/1) and a BBSRC equipment grant (BB/R01356X/1).

1. RG Harrison, EL Larson, Hybridization, introgression, and the nature of species boundaries. *J. Hered.* **105 (S1)**, 795–809 (2014).
2. J Mallet, N Besansky, MW Hahn, How reticulated are species? *BioEssays* **38**, 140–149 (2016).
3. JH Degnan, Modeling hybridization under the network multispecies coalescent. *Syst. Biol.* **67**, 786–799 (2018).
4. RAL Elworth, HA Ogilvie, J Zhu, L Nakhleh, Advances in computational methods for phylogenetic networks in the presence of hybridization in *Bioinformatics and Phylogenetics*. (Springer) Vol. 29, pp. 317–360 (2019).
5. X Jiao, T Flouri, Z Yang, Multispecies coalescent and its applications to infer species phylogenies and cross-species gene flow. *Nat. Sci. Rev.* (2021) DOI:10.1093/nsr/nwab127.
6. D Wen, L Nakhleh, Coestimating reticulate phylogenies and gene trees from multilocus sequence data. *Syst. Biol.* **67**, 439–457 (2018).
7. C Zhang, HA Ogilvie, AJ Drummond, T Stadler, Bayesian inference of species networks from multilocus sequence data. *Mol. Biol. Evol.* **35**, 504–517 (2018).
8. T Flouri, X Jiao, B Rannala, Z Yang, A Bayesian implementation of the multispecies coalescent model with introgression for phylogenomic analysis. *Mol. Biol. Evol.* **37**, 1211–1223 (2020).
9. Y Yu, JH Degnan, L Nakhleh, The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. *PLoS Genet.* **8**, e1002660 (2012).
10. F Pardi, C Scornavacca, Reconstructible phylogenetic networks: do not distinguish the indistinguishable. *PLoS Comput. Biol.* **11**, e1004135 (2015).
11. S Zhu, JH Degnan, Displayed trees do not determine distinguishability under the network multispecies coalescent. *Syst. Biol.* **66**, 283–298 (2017).
12. C Solis-Lemus, A Coen, C Ane, On the identifiability of phylogenetic networks under a pseudo-likelihood model. *ArXive* (2020).
13. T Zhu, Z Yang, Complexity of the simplest species tree problem. *Mol. Biol. Evol.* (2021) DOI: 10.1093/molbev/msab009.

14. Y Thawornwattana, J Mallet, Z Yang, Complex introgression history of the erato-sara clade of heliconius butterflies. *bioRxiv* (2021).
15. N Finger, et al., Genome-scale data reveal deep lineage divergence and a complex demographic history in the texas horned lizard (*Phrynosoma cornutum*) throughout the southwestern and central usa. *Genome Biol. Evol.* (2021).
16. S Richardson, P Green, On Bayesian analysis of mixtures with an unknown number of components (with discussions). *J. R. Stat. Soc. B* **59**, 731–792 (1997).
17. G Celeux, M Hurn, C Robert, Bayesian inference for mixture: the label switching problem in *COMPSTAT*, eds. R Payne, PJ Green. (Physica, Heidelberg), pp. 227–232 (1998).
18. M Stephens, Dealing with label switching in mixture models. *J. R. Stat. Soc. B.* **62**, 795–809 (2000).
19. A Jasra, CC Holmes, DA Stephens, Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Stat. Sci.* **1**, 50–67 (2005).
20. NB Edelman, et al., Genomic architecture and introgression shape a butterfly radiation. *Science* **366**, 594–599 (2019).
21. Z Yang, Paml 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
22. T Flouri, X Jiao, B Rannala, Z Yang, Species tree inference with BPP using genomic sequences and the multispecies coalescent. *Mol. Biol. Evol.* **35**, 2585–2593 (2018).
23. GC Heliconius, Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* **487**, 94–98 (2012).
24. SH Martin, et al., Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Res.* **23**, 1817–1828 (2013).
25. T Jukes, C Cantor, Evolution of protein molecules in *Munro, H.N., ed. Mammalian Protein Metabolism.* (Academic Press, New York), pp. 21–123 (1969).

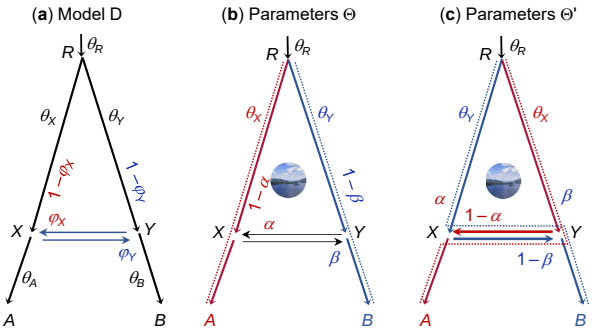

Fig. 1. (**a**) Bidirectional introgression model or model D (8) assumes introgression in both directions between species A and B at time $\tau_X = \tau_Y$. (**b**) and (**c**) Two sets of parameters Θ and Θ', with the same parameter values except that $\varphi'_X = 1 - \varphi_X$, $\varphi'_Y = 1 - \varphi_Y$, $\theta'_X = \theta_Y$, and $\theta'_Y = \theta_X$. The dotted lines indicate the main routes taken by sequences sampled from species A and B, if both introgression probabilities α and β are $\ll \frac{1}{2}$.

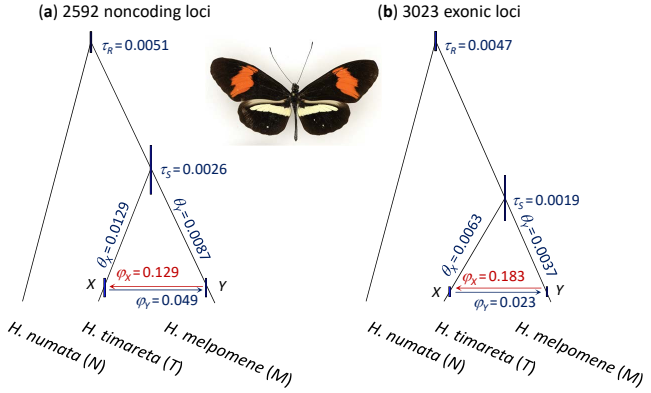

Fig. 2. Species tree and BDI model for *Heliconius melpomene*, *H. timareta*, and *H. numata*. The branch lengths are drawn to represent the estimated species divergence times (posterior means) using the noncoding and exonic loci from chromosome 1, while the node bars represent the 95% HPD CIs. See table 1 for estimates of other parameters. Photo of *H. timareta* courtesy of James Mallet.

Table 1. Posterior means and 95% HPD CIs (in parenthees) for parameters in the BDI model of figure 2 for the *Heliconius* data

	Noncoding, $L = 500$	Noncoding, $L = 2592$	Exonic, $L = 500$	Exonic, $L = 3023$
τ_R	4.73 (4.33, 5.13)	5.10 (4.89, 5.30)	4.39 (3.98, 4.81)	4.71 (4.54, 4.88)
τ_S	3.12 (2.05, 4.19)	2.58 (2.12, 3.05)	1.95 (1.07, 2.82)	1.78 (1.38, 2.19)
$\tau_X = \tau_Y$	0.62 (0.21, 1.02)	0.25 (0.09, 0.40)	0.20 (0.03, 0.37)	0.13 (0.05, 0.24)
θ_M	1.50 (0.62, 2.34)	0.69 (0.35, 1.10)	0.38 (0.08, 0.70)	0.32 (0.14, 0.52)
θ_T	2.55 (1.40, 3.74)	1.23 (0.65, 1.84)	0.79 (0.13, 1.28)	0.63 (0.32, 0.94)
θ_N	15.1 (12.0, 18.5)	23.0 (20.3, 25.7)	11.2 (9.11, 13.5)	12.4 (11.4, 13.4)
θ_R	5.08 (4.12, 6.05)	5.74 (5.23, 6.24)	5.76 (4.83, 6.70)	6.68 (6.24, 7.11)
θ_S	4.62 (1.85, 7.40)	6.92 (5.48, 8.37)	5.31 (3.38, 7.36)	7.50 (6.51, 8.49)
θ_X	11.4 (2.83, 21.2)	12.9 (7.35, 19.6)	8.04 (1.67, 15.4)	5.80 (3.60, 8.36)
θ_Y	6.78 (2.42, 11.6)	8.74 (5.69, 12.0)	4.03 (0.60, 7.51)	3.49 (2.56, 4.50)
φ_X	0.354 (0.022, 0.664)	0.124 (0.007, 0.243)	0.280 (0.002, 0.547)	0.161 (0.070, 0.264)
φ_Y	0.104 (0.000, 0.306)	0.048 (0.000, 0.139)	0.116 (0.000, 0.318)	0.019 (0.000, 0.056)

Note.— Estimates of τs and θs are multiplied by 10^3. MCMC samples are processed using the β–γ algorithm before they are summarized.

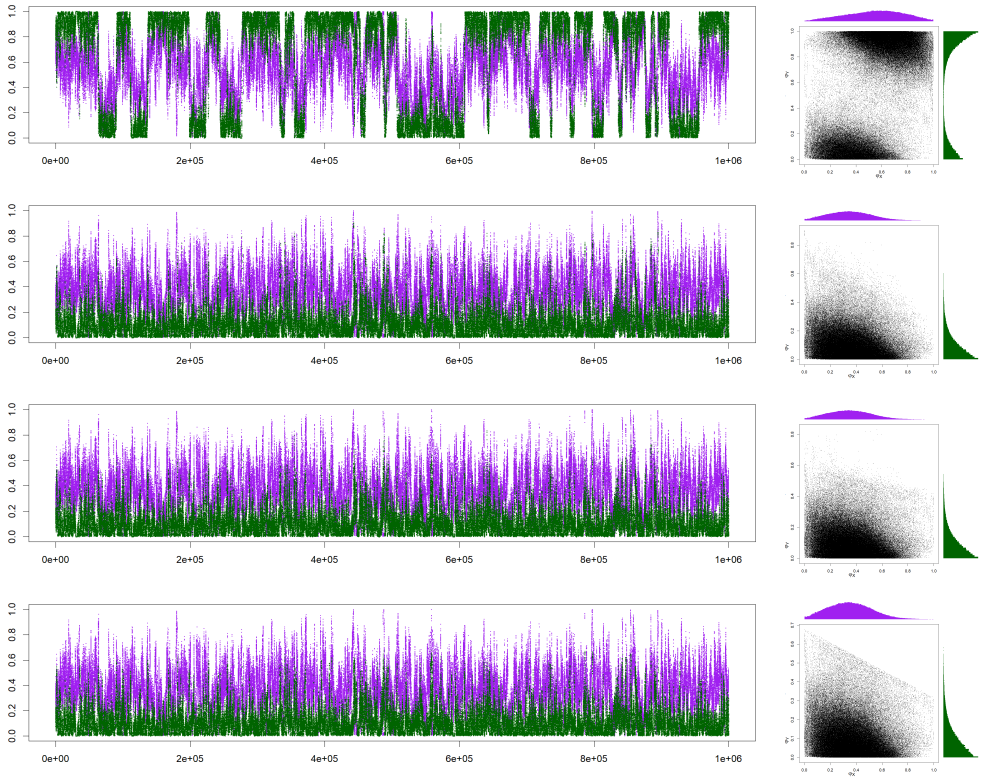

Fig. 3. Trace plots of MCMC samples and 2-D scatter plots for parameters φ_X (purple) and φ_Y (green) before (top) and after (bottom three) the post-processing of the MCMC sample in the BPP analysis of the first 500 noncoding loci from chromosome 1 of the *Heliconius* data under the MSci model of figure 2. The three algorithms used are β–γ, CoG_N, and CoG_0.

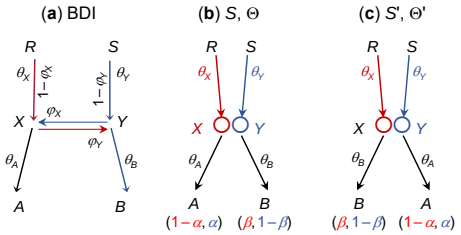

Fig. 4. The rule of BDI unidentifiability. (**a**) In the BDI model, species RA and SB exchange migrants at time $\tau_X = \tau_Y$. Treat X and Y as one node with left parent RX with population size θ_X and right parent SY with population size θ_Y. When a sequence from A reaches XY, it takes the left and right parental paths with probabilities $1-\varphi_X$ and φ_X, respectively. When a sequence from B reaches XY, it goes left and right with probabilities φ_Y and $1-\varphi_Y$, respectively. (**b** & **c**) Placing the two daughters in the order (A,B) as in Θ or (B,A) as in Θ' does not affect the distribution of gene trees, and constitutes unidentifiable towers in the posterior space. If X and Y are sister species and have the same mother node (with R and S to be the same node), the unidentifiability is within-model; otherwise it is cross-model.

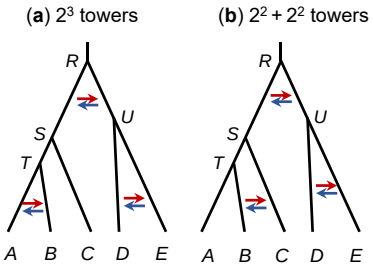

Fig. 5. (**a**) Three BDI events between sister species creating $2^3 = 8$ within-model towers in the posterior. (**b**) Two BDI events between sister species and one BDI event between non-sister species creating two unidentifiable models each with four within-model unidentifiable towers in the posterior space.

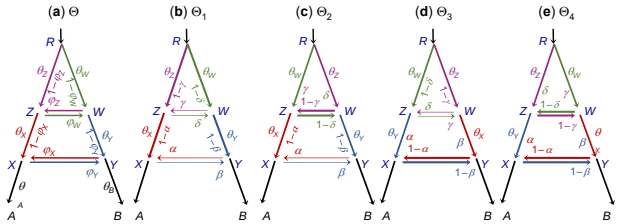

Fig. 6. Double-BDI model between two species A and B, with four within-model towers (Θ_1, Θ_2, Θ_3, and Θ_4). (**a**) The parameters in the model include 7 θs, 3 τs, and 4 φs, with 14 parameters in total. (**b**)-(**e**) Four unidentifiable towers showing the mappings of parameters (eq. 2). To apply the rule of figure 4, we treat each pair of BDI nodes as one node, so that X and Y have the same node ZW as the parent, and the unidentifiability caused by the BDI event at nodes X-Y is within-model, as is the unidentifiability for the BDI event at nodes Z-W .

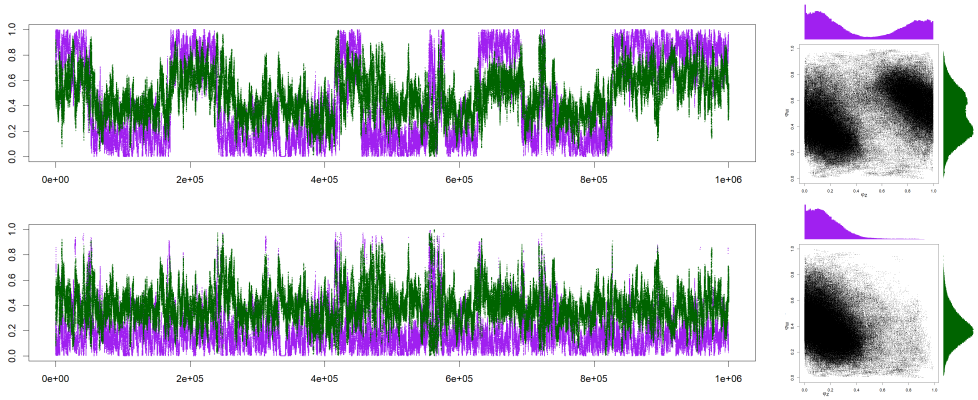

Fig. 7. Trace plots of MCMC samples and 2-D scatter plots for parameters φ_Z (purple) and φ_W (green) before (top) and after (bottom) the post-processing of the MCMC samples for the double-DBI model of figure 6a. Post processing used the β–γ algorithm, while CoG_N and CoG_0 produced nearly identical results (not shown). This is for replicate 2 for $L = 500$ loci.

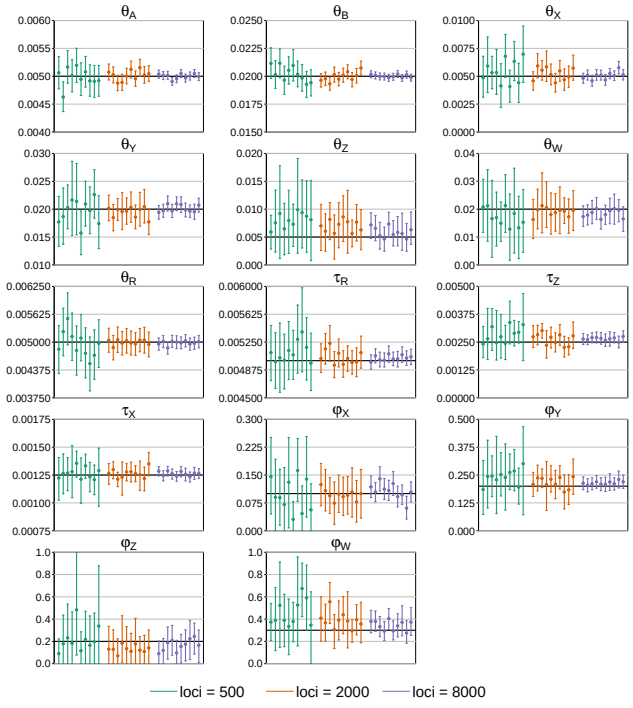

Fig. 8. Posterior means and the 95% HPD CIs in 10 replicate datasets of $L = 500, 2000$, and 8000 loci, simulated and analyzed under the double-BDI model of figure 6a. The MCMC samples are post-processed using the β–γ algorithm before they are summarized.

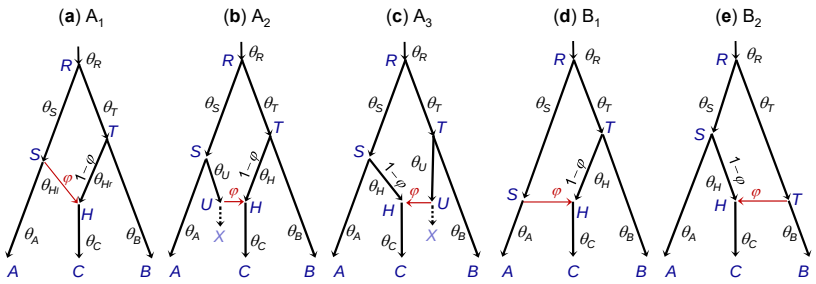

Fig. 9. (**a**-**c**)Three interpretations of MSci model A are indistinguishable/unidentifiable. (**d**, **e**) Two versions of MSci model B are identifiable.

Supporting Information (SI).

- Figure S1: Analysis of the first 500 exonic loci of the *Heliconius* data.
- Figure S2: Three models with a BDI event between sister species.
- Figure S3: Two models with a BDI event between nonsister species.
- Figure S4: Three models with a BDI event between nonsister species.
- Figure S5: Two BDI events between non-sister species creating four unidentifiable models.
- Figure S6: Trace plots for ϕ_X and ϕ_Y in analysis of a dataset of $L = 500$ loci simulated under the BDI model of figure 1.
- Table S1: Posterior means and 95% HPD CIs for parameters in the BDI model from a simulated data of $L = 500$ loci.

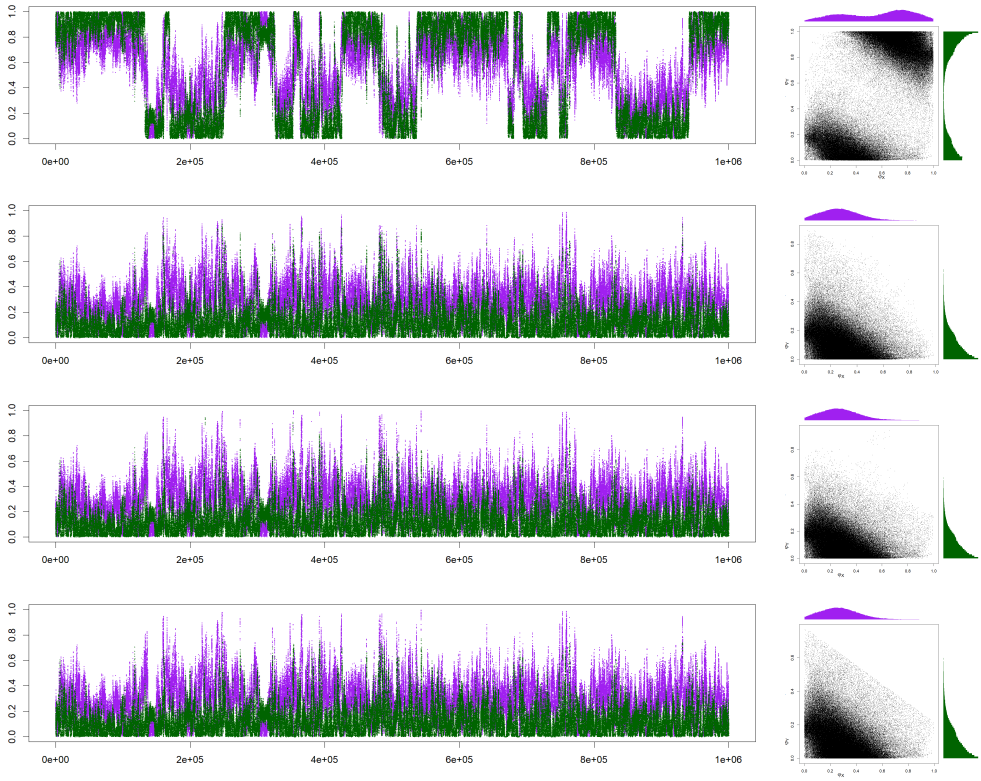

Fig. S1. Analysis of the first 500 exonic loci on chromosome 1 from the *Heliconius* data. See legend to figure 3.

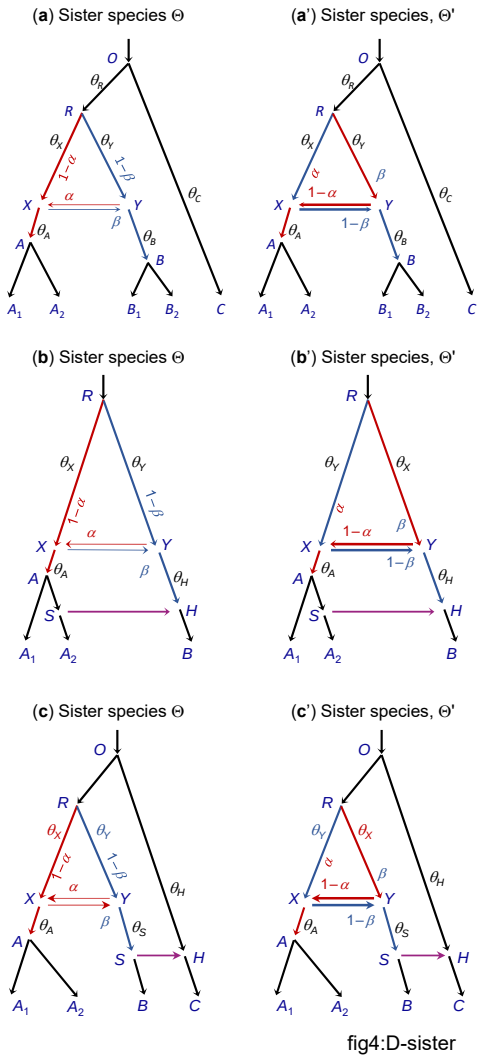

Fig. S2. Three models, each involving a BDI event between sister species, creating within-model unidentifiability. (**a**) & (**a′**) Subtrees are added to branches A, B, and R in the basic model of figure 1a. (**b**) & (**b′**) A BDI event between sister species X and Y with a unidirectional introgression involving descendant branches of X and Y. (**c**) and (**c′**) A BDI event between sister species X and Y with a unidirectional introgression involving one descendant branch and another branch that is not a descendant of X or Y. In all three cases, the parameter mapping is $\varphi'_X = 1 - \varphi_X$, $\varphi'_Y = 1 - \varphi_Y$, $\theta'_X = \theta_Y$, and $\theta'_Y = \theta_X$.

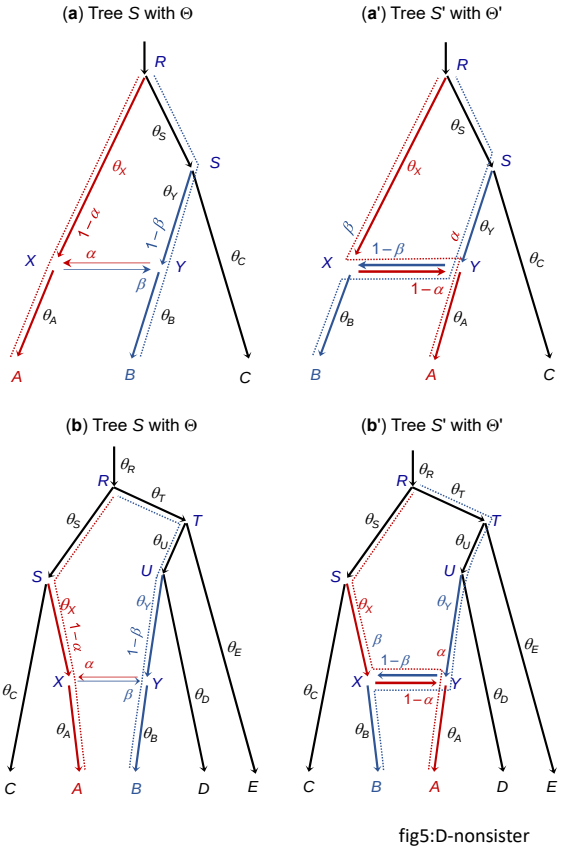

Fig. S3. BDI between non-sister species creates cross-model unidentiability. (**a** & **a′**) A pair of unidentifiable models with a BDI event between non-sister species. The dotted lines indicate the main routes taken by sequences sampled from species A and B, if the introgression probabilities α and β are $< \frac{1}{2}$. (**b** & **b′**) Another pair of unidentifiable models with a BDI event between non-sister species. The parameter mapping from Θ to Θ' in both cases is $\varphi'_X = 1 - \varphi_Y$ and $\varphi'_Y = 1 - \varphi_X$, with all other parameters (such as θ_X, θ_Y, θ_A, and θ_B) to be identical between Θ and Θ'.

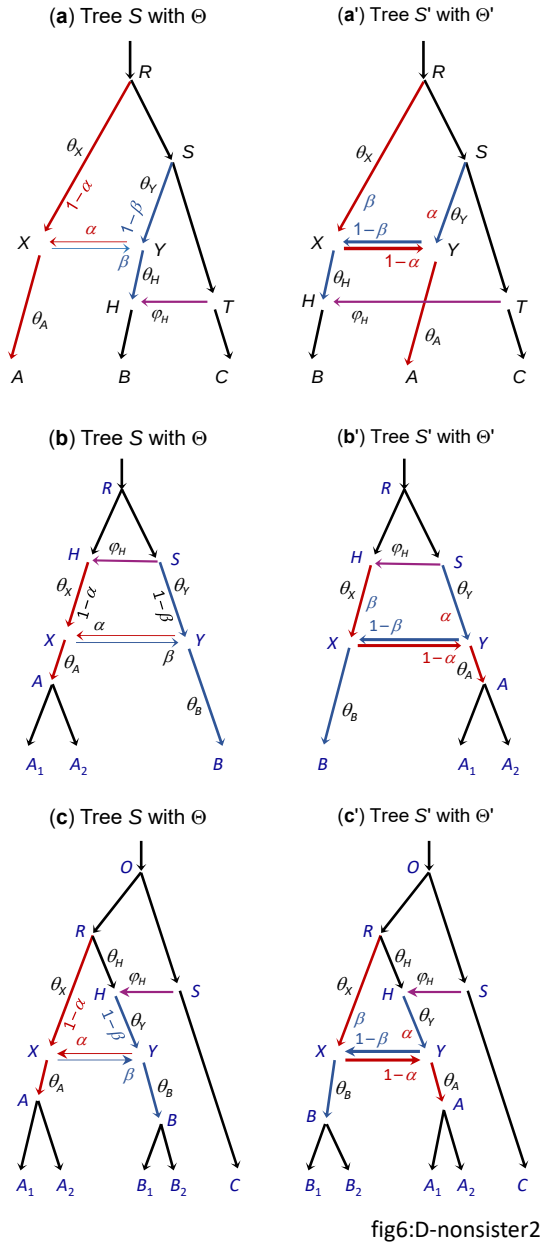

Fig. S4. Three pairs of unidentifiable models with one BDI event between non-sister species, illustrating the mapping of parameters (Θ and Θ'). In (**a**), RXA and SYH are non-sister species. In (**b** & **c**), nodes X and Y are non-sister species because of the unidirectional introgression event involving branches RX and/or RY. The mirror model (S' with Θ') is generated by pruning off branches AX at X and BY at Y, swapping places and reattaching, and applying the mapping $\varphi'_X = 1 - \varphi_Y$ and $\varphi'_Y = 1 - \varphi_X$.

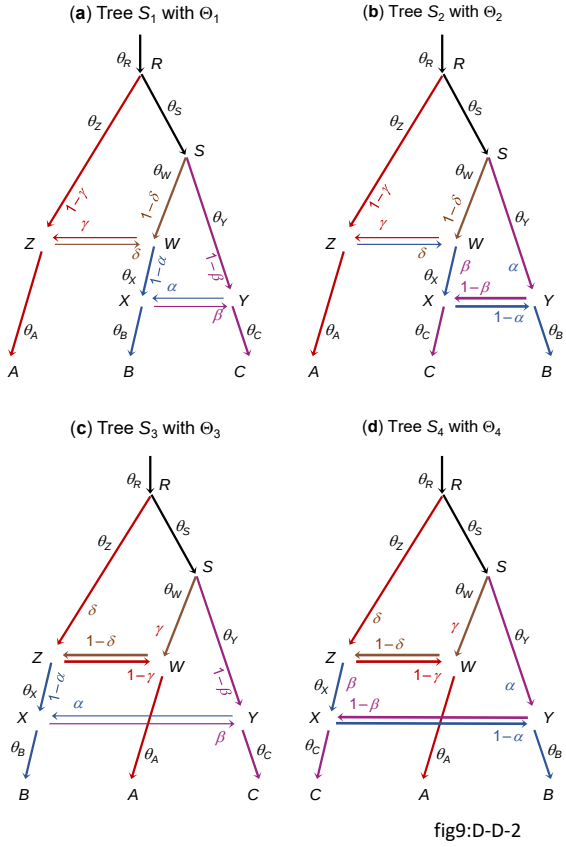

Fig. S5. Two BDI events involving non-sister species on a species tree for three species creating four unidentifiable models. The cross-model parameter mappings concern only the introgression probabilities $\varphi_X \equiv \alpha$, $\varphi_Y \equiv \beta$, $\varphi_Z \equiv \gamma$, and $\varphi_W \equiv \delta$, while all other parameters are the same among the models. The colored lines indicate the main routes taken by sequences sampled from A (red), B (blue), and C (purple), if the introgression probabilities α, β, γ, and δ are all $< \frac{1}{2}$, from which the unidentifiability of the four models can be seen easily.

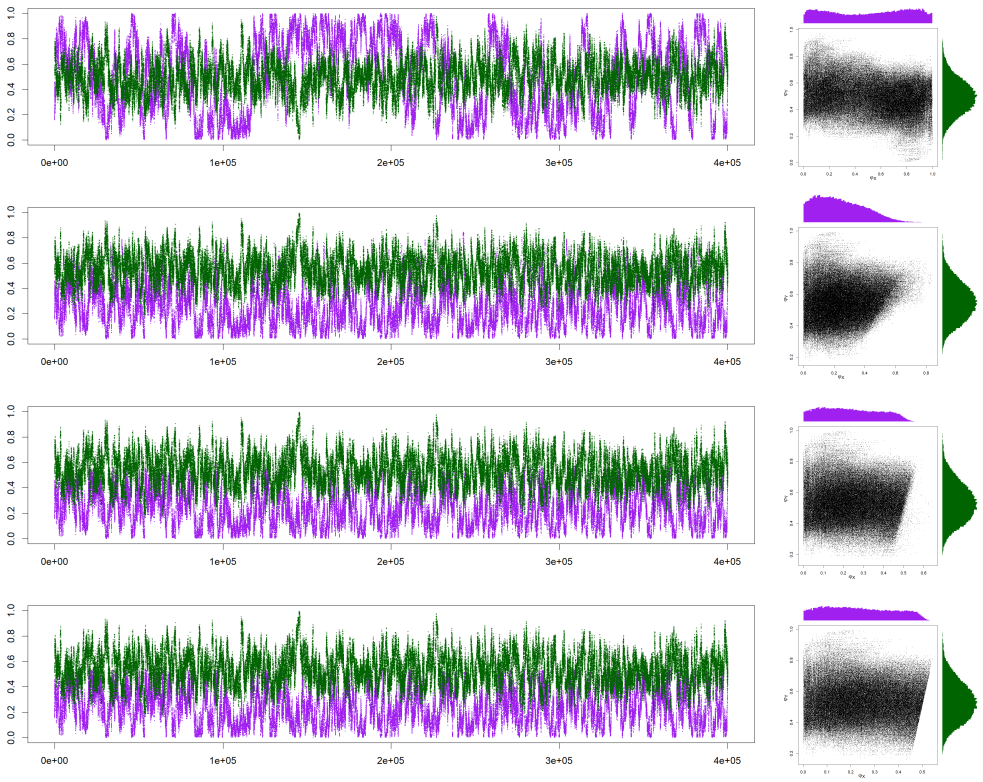

Fig. S6. Trace plots of MCMC samples for ϕ_X and ϕ_Y and 2-D scatter plots from BPP analysis of a dataset of $L = 500$ loci simulated under the BDI model of figure 1a. See table S1 for the true parameter values and posterior summaries. The plots are for, from top to bottom, unprocessed sample and processed samples using the β–γ, CoG$_N$, and CoG$_0$ algorithms. The true parameter values are $\Theta = (\varphi_X, \varphi_Y) = (0.7, 0.2)$, and the post-processing using all three algorithms mapped the samples to the mirror tower around $\Theta' = (0.3, 0.8)$.

Table S1. Posterior means and 95% HPD CIs (in parenthees) for parameters in the MSci model of figure 1a from a simulated dataset of $L = 500$ loci

	truth (Θ)	mirror (Θ')	beta-gamma	CoG$_N$	CoG$_0$
τ_R	0.01		0.0098 (0.0088, 0.0108)		
$\tau_X = \tau_Y$	0.005		0.0050 (0.0045, 0.0055)		
θ_A	0.002		0.0020 (0.0018, 0.0021)		
θ_B	0.01		0.0101 (0.0093, 0.0108)		
θ_R	0.002		0.0020 (0.0006, 0.0034)		
θ_X	0.002	0.01	0.0071 (0.0022, 0.0124)	0.0067 (0.0017, 0.0120)	0.0068 (0.0017, 0.0121)
θ_Y	0.01	0.002	0.0063 (0.0005, 0.0130)	0.0066 (0.0005, 0.0133)	0.0066 (0.0005, 0.0133)
ϕ_X	0.7	0.3	0.245 (0.001, 0.528)	0.236 (0.001, 0.472)	0.235 (0.001, 0.470)
ϕ_Y	0.2	0.8	0.553 (0.330, 0.791)	0.539 (0.305, 0.786)	0.538 (0.305, 0.788)

Note.— Empty cells mean the same values as on the left. MCMC samples are processed using the three algorithms and then summarized. See figure S6 for the tracecatter plots. The dataset of $L = 500$ loci, each consisting of four sequences per species (or eight sequences per locus) and 500 sites per sequence, is simulated using the true parameter values (Θ).