Systematic Biology (2020), **0**, 0, *pp.* 1–20
doi:10.1093/sysbio/DefineSpeciesGeneFlow

Defining Species When There is Gene Flow

XIYUN JIAO[1], AND ZIHENG YANG[1,*]

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

[*]*Ziheng Yang, Department of Genetics, Evolution and Environment, University College London, Gower Street, London WC1E 6BT, UK.*
Phone: +44 20 76794379 (z.yang@ucl.ac.uk) (ORCID: 0000-0003-3351-7981)

ABSTRACT

Whatever one's definition of species, it is generally expected that individuals of the same species should be genetically more similar to each other than they are to individuals of another species. Here we show that in the presence of cross-species gene flow, this expectation may be incorrect. We use the multispecies coalescent model with continuous-time migration or episodic introgression to study the impact of gene flow on genetic differences within and between species and highlight a surprising but plausible scenario in which different population sizes and asymmetrical migration rates cause a genetic sequence to be on average more closely related to a sequence from another species than to a sequence from the same species. Our results highlight the extraordinary impact that even a small amount of gene flow may have on the genetic history of the species. We suggest that contrasting long-term migration rate and short-term hybridization rate, both of which can be estimated using genetic data, may be a powerful approach to detecting the presence of reproductive barriers and to define species boundaries.

Key words: Gene flow, introgression, migration, species concept, species delimitation, multispecies coalescent

INTRODUCTION

The concept of species is a controversial one, with a number of definitions proposed in the literature (Mallet, 2013; Zachos 2016). The biological species concept emphasizes reproductive isolation, although low levels of cross-species gene flow are tolerated in modern versions of the concept (Coyne and Orr, 2004). The lineage species concept considers species as independently evolving lineages (De Queiroz, 2007). Despite the differences in species definitions, it is generally expected that an individual is genetically more closely related to an individual of the same species than to an individual of a different species. Here we may measure genetic relatedness in two ways. First, if we sample two sequences a_1 and a_2 from species A and one sequence b from species B, we expect the average sequence distances to satisfy $\mathbb{E}(t_{aa}) < \mathbb{E}(t_{ab})$. Second, we expect gene tree $G_1 = ((a_1, a_2), b)$ to occur with a higher probability than gene trees $G_2 = ((b, a_1), a_2)$ or $G_3 = ((b, a_2), a_1)$.

Two approaches to identifying and delimiting species make use of those expectations explicitly. First, DNA barcoding is a fast approach to species identification and is occasionally applied to species delimitation as well (Hebert *et al.*, 2003). A genetic-distance threshold or 'barcoding gap' based on a universal marker (such as mitochondrial cytochrome oxidase 1 or

cytochrome b) is used to distinguish within- and between-species divergences. A query specimen is assigned to an existing species in the database if the sequence distance between the query and the sequences in the library is smaller than the threshold. Otherwise the specimen is considered a new species not yet represented in the library. The threshold may be arbitrary (Hebert *et al.*, 2003) or estimated from a database by minimizing assignment errors (Meyer and Paulay, 2005; Puillandre *et al.*, 2012). The use of one barcoding threshold for different species in the database may lead to errors of identification when different species have very different population sizes and divergence times (e.g., Hudson and Turelli, 2003; Meyer and Paulay, 2005; Dasmahapatra *et al.*, 2010; Yang and Rannala, 2017). Here we emphasize the fact that barcoding methods rely on a distance threshold, with the expectation that within-species sequence divergence is smaller than between-species divergence, $\mathbb{E}(t_{aa}) < \mathbb{E}(t_{ab})$. Second, the recently developed genealogical divergence index or *gdi* (Jackson *et al.*, 2017) is a simple method for fast species delimitation, useful for generating hypotheses of species status for systematic evaluations integrating different sources of information. The *gdi* is a linear transform on $\mathbb{P}(G_1)$. Two populations are considered distinct species if $\mathbb{P}(G_1) > 0.8$ or one single species if $\mathbb{P}(G_1) < 0.47$, with the species status undecided if $\mathbb{P}(G_1)$ falls between the two limits. It is expected that $\mathbb{P}(G_1) > \frac{1}{3} > \mathbb{P}(G_2) = \mathbb{P}(G_3)$.

The two expectations, $\mathbb{E}(t_{aa}) < \mathbb{E}(t_{ab})$ and $\mathbb{P}(G_1) > \mathbb{P}(G_2)$, are correct if isolation is complete and there is no cross-species gene flow (Fig. 1A). When we trace the genealogical history of sequences a_1, a_2, and b backwards in time, there is the possibility that sequences a_1 and a_2 coalesce before reaching the common ancestor R (Fig. 1A). If this happens, the gene tree will be G_1; otherwise the three possible gene trees will occur with equal probability. Thus $\mathbb{P}(G_1) > \mathbb{P}(G_2) = \mathbb{P}(G_3)$. Similarly, if sequences a_1 and a_2 coalesce in species A, they will have a shorter expected distance than between species; otherwise their distance will be the same as between species. Thus we expect $\mathbb{E}(t_{aa}) < \mathbb{E}(t_{ab})$.

However, it is not so clear whether the expectations are correct when there is cross-species gene flow. In this paper we study the impact of gene flow on genetic divergences within and between species, by using the Markov chain characterization of the process of coalescent and migration developed in the structured coalescent framework in population genetics (Notohara, 1990; Wilkinson-Herbots, 2008). We demonstrate that with different population sizes (and thus coalescent rates) and asymmetrical migration rates, it is possible for a gene sequence to be on average more distant from another sequence of the same species than from a sequence randomly sampled from another species. We refer to the region of the parameter space in which $\mathbb{P}(G_1) < \mathbb{P}(G_2)$ or $\mathbb{E}(t_{aa}) > \mathbb{E}(t_{ab})$ as the *species-definition anomaly zone*, similar to the *species-tree anomaly zone* discussed by Degnan and Rosenberg (2006). Our results highlight the complexity of defining and delimiting species in the presence of gene flow: for example, in the anomaly zone, application of any barcoding criterion or the *gdi* index may lead to incorrect inference of one species when two exist.

We note that in the past decade, analyses of genomic sequence data have detected cross-species gene flow in a variety of species including *Arabidopsis* (Arnold *et al.*, 2016), corals (Mao *et al.*, 2018), mosquitoes (Fontaine *et al.*, 2015; Thawornwattana *et al.*, 2018), butterflies (Martin *et al.*, 2013), birds (Ellegren *et al.*, 2012), cats (Li *et al.*, 2019), bears (Liu *et al.*, 2014), cattle (Wu *et al.*, 2018), gibbons (Chan *et al.*, 2013), and hominins (Nielsen *et al.*, 2017). Empirical studies suggest very high proportions of species that hybridize with at least one other species (Mallet, 2005, 2008). It is thus of great importance to examine the impact of cross-species gene flow on the definition and identification of species. Here we formulate our results in the context of using genomic sequence data to infer the history of species divergences and gene flow and to delimit species boundaries. We focus on the continuous-time migration model (the IM model, Hey, 2010) (Fig. 1B), but will show that the same behavior occurs under the episodic introgression model or the multispecies coalescent with introgression (MSci) model (Yu *et al.*, 2014; Flouri *et al.*, 2020) (Fig. 1C), which may be more realistic for some biological systems.

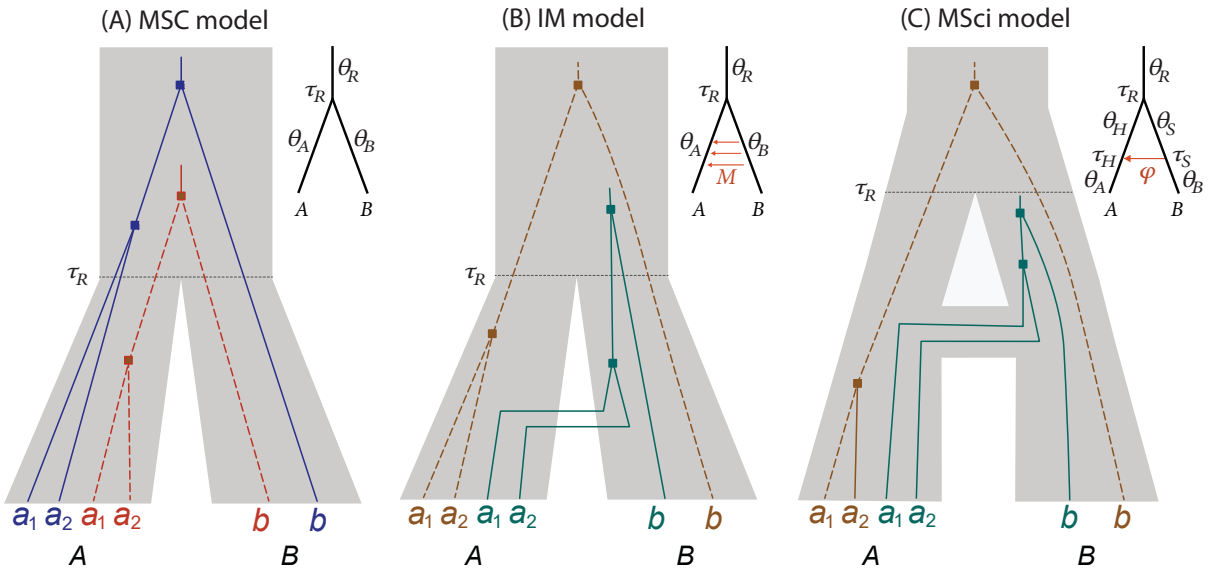

Fig. 1. (A) The multispecies coalescent (MSC) model for two species (A and B) with four parameters shown in the inset (species divergence time $\tau_R = \tau_{AB} = \tau$ and three population size parameters: θ_A, θ_B, and θ_R). Both θ and τ are measured in the number of mutations per site. Two gene trees for three sequences (a_1 and a_2 from A and b from B) are shown inside the species tree. If sequences a_1 and a_2 coalesce in species A, the gene tree will be $G_1 = ((a_1, a_2), b)$; otherwise all three sequences will enter species R and the three gene trees G_1, $G_2 = ((b, a_1), a_2)$ and $G_3 = ((b, a_2), a_1)$ will occur with equal probability ($\frac{1}{3}$ each). (B) The MSC model with migration (the IM model) and (C) the MSC model with introgression (the MSci model), with 5 and 8 parameters, respectively, shown in the inset. Under the IM model, the migration rate $M = M_{BA} = N_A m_{BA}$ is the expected number of $B \to A$ migrants in species A per generation, with $m_{BA} = m$ to be the proportion of immigrants (from species B) in species A. Under the MSci model, $\tau_H = \tau_S$ while φ is the introgression probability. Under both the IM and MSci models, there are multiple scenarios under which gene tree G_1 may occur. For example, in the red tree, a_1 and a_2 coalesce in species A, while in the green tree, a_1 and a_2 migrate (backwards in time) into species B and then coalesce in species B.

THE IM MODEL FOR TWO SPECIES AND THREE SEQUENCES

Consider two diploid species A and B, which diverged time $\tau = \tau_R$ ago and have since been undergoing migration from species B to species A, at the rate of $m = m_{BA}$ per generation (Fig. 1B). We formulate our theory in the context of analyzing genomic sequence data, so that time is scaled by mutations and both θ and τ are measured in the number of mutations per site. For each species, the population size parameter is defined as $\theta = 4N\mu$, where N is the effective population size and μ is the mutation rate per site per generation. We define the migration rate m_{BA} as the proportion of $B \to A$ immigrants in the receiving population A, so that $M_{BA} = m_{BA} N_A$ is the expected number of $B \to A$ migrants per generation. The parameters in the IM model (Hey, 2010) include τ_R, θ_A, θ_B, θ_R, and M_{BA} (Fig. 1B). The IM model of Figure 1B is a special case of the model of Long and Kubatko (2018, Fig. 1b), which allows migration in both directions.

We consider the genealogical relationships among three sequences: a_1 and a_2 from A and b from B. In this setting, the gene trees and coalescent times are random variables, with distributions specified by the parameters in the model. The backward process of coalescence and migration during time interval $(0, \tau_R)$ is described by a Markov chain (Notohara, 1990), where the state of the chain is specified by the number of sequences remaining in the sample, the populations in which they reside, the population IDs (A and B) and the sequence IDs (a_1, a_2 and b) (Zhu and Yang, 2012; Andersen *et al.*, 2014; Tian and Kubatko, 2016; Dalquen *et al.*, 2017). For example, in the state $A_{a_1} A_{a_2} B_b$, abbreviated AAB, there are three sequences in the sample, and sequences a_1, a_2 and b are in populations A, A and B, respectively. This is the initial state for the

Markov chain as our sample consists of sequences a_1 and a_2 from A and b from B. Similarly ABB is the state reached when sequence a_2 migrates (backwards in time) into population B. The state $A_{aa}B_b$, abbreviated AB_b, means that two sequences remain in the sample, with the ancestor of a_1 and a_2 in population A, and sequence b in population B. This is the state reached when sequences a_1 and a_2 coalesce in species A. The generator matrix $Q^{①}$ for the Markov chain is

	AAB	ABB	BAB	BBB	AB_b	$A_{a_1}B$	$A_{a_2}B$	BB_b	$B_{a_1}B$	$B_{a_2}B$	B
AAB	$-2w_{BA}-c_A$	w_{BA}	w_{BA}	0	c_A	0	0	0	0	0	0
ABB	0	$-w_{BA}-c_B$	0	w_{BA}	0	c_B	0	0	0	0	0
BAB	0	0	$-w_{BA}-c_B$	w_{BA}	0	0	c_B	0	0	0	0
BBB	0	0	0	$-3c_B$	0	0	0	c_B	c_B	c_B	0
AB_b	0	0	0	0	$-w_{BA}$	0	0	w_{BA}	0	0	0
$A_{a_1}B$	0	0	0	0	0	$-w_{BA}$	0	0	w_{BA}	0	0
$A_{a_2}B$	0	0	0	0	0	0	$-w_{BA}$	0	0	w_{BA}	0
BB_b	0	0	0	0	0	0	0	$-c_B$	0	0	c_B
$B_{a_1}B$	0	0	0	0	0	0	0	0	$-c_B$	0	c_B
$B_{a_2}B$	0	0	0	0	0	0	0	0	0	$-c_B$	c_B
B	0	0	0	0	0	0	0	0	0	0	0

where $w_{BA} = \frac{m_{BA}}{\mu} = \frac{4M_{BA}}{\theta_A}$ is the mutation-scaled migration rate, and $c_A = \frac{2}{\theta_A}$ and $c_B = \frac{2}{\theta_B}$ are the coalescent rates. Here one time unit is the expected time taken to accumulate one mutation per site. In a species with a scaled population size $\theta = 4N\mu$, each pair of sequences coalesce at the rate $\frac{2}{\theta}$, with the average coalescent time to be $\frac{\theta}{2}$.

Probabilities of gene trees

We calculate the probabilities for the three gene trees: $G_1 = ((a_1,a_2),b)$; $G_2 = ((b,a_1),a_2)$; and $G_3 = ((b,a_2),a_1)$, as functions of the parameters in the IM model (Fig. 1B). Note that the gene tree topology is determined by the first coalescent event, so that there is no need to follow the Markov chain any further after the first coalescent has occurred. Thus we construct a simplified Markov chain in which all two-sequence states (such as AB_b and $A_{a_1}B$) are changed into absorbing states, and state B is unreachable and thus removed from the chain. Similarly as soon as the chain enters the state BBB, with all three sequences in species B, the three gene trees occur with equal probabilities. Thus we make BBB an absorbing state as well, with BB_b, $B_{a_1}B$, and $B_{a_2}B$ unreachable and removed. The modified Markov chain then has 7 states, with the generator $Q^{②}$

	AAB	ABB	BAB	BBB	AB_b	$A_{a_1}B$	$A_{a_2}B$
(1) AAB	$-2w_{BA}-c_A$	w_{BA}	w_{BA}	0	c_A	0	0
(2) ABB	0	$-w_{BA}-c_B$	0	w_{BA}	0	c_B	0
(3) BAB	0	0	$-w_{BA}-c_B$	w_{BA}	0	0	c_B
(4) BBB	0	0	0	0	0	0	0
(5) AB_b	0	0	0	0	0	0	0
(6) $A_{a_1}B$	0	0	0	0	0	0	0
(7) $A_{a_2}B$	0	0	0	0	0	0	0

Let $P^{②}(\tau) = \exp(Q^{②}\tau)$ be the matrix of transition probabilities over time $\tau = \tau_R$. We

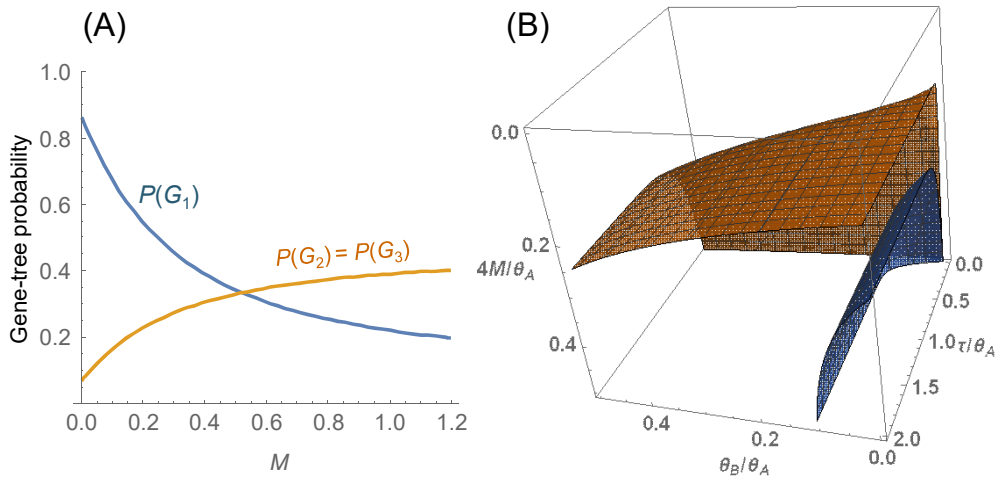

Fig. 2. (A) Probabilities of gene trees $G_1 = ((a_1,a_2),b)$ and $G_2 = ((b,a_1),a_2)$ as functions of the migration rate M when the other parameters in the IM model (Fig. 1B) are fixed: $\tau = 0.02$, $\theta_A = 0.025$, and $\theta_B = 0.001$. Note that when $M > M^* = 0.521361$ immigrants per generation, $\mathbb{P}(G_1) < \mathbb{P}(G_2)$. (B) Partition of the parameter space for the IM model (Fig. 1B) according to probabilities for gene trees. Below the outer tent, $\mathbb{P}(G_1) < \mathbb{P}(G_2)$, while below the inner tent $\mathbb{P}(G_1) < 0.2$ and $\mathbb{P}(G_2) = \mathbb{P}(G_3) > 0.4$.

have the probability for gene tree G_1 to be

$$\mathbb{P}(G_1) = P_{15}^{②}(\tau) + [P_{11}^{②}(\tau) + P_{12}^{②}(\tau) + P_{13}^{②}(\tau) + P_{14}^{②}(\tau)]/3. \tag{1}$$

The different terms account for different scenarios that lead to gene tree G_1. First, sequences a_1 and a_2 may coalesce in population A, before reaching time τ: this occurs with probability $P_{15}^{②}(\tau)$. Second if both sequences a_1 and a_2 enter species B any time during the time interval $(0,\tau)$, the chain will be in state 4 (BBB): each gene tree will then have probability $\frac{1}{3}$ when the coalescent events occur at random in species B or R (Fig. 1B). Finally if no coalescent occurs over the time interval $(0,\tau)$ and if at most one of the A sequences enters species B, the chain will be in states 1, 2, or 3 (for AAB, ABB or BAB) at time τ: then all three sequences will enter the common ancestor R and each gene tree occurs with probability $\frac{1}{3}$.

The eigenvalues of $Q^{②}$ are on the diagonal: $\lambda_1 = -\frac{2+8M}{\theta_A}$, $\lambda_2 = \lambda_3 = -\frac{4M}{\theta_A} - \frac{2}{\theta_B}$, and $\lambda_4 = \cdots = \lambda_7 = 0$. These are all real, as are the eigenvectors. We derive $P^{②}(\tau)$ using Mathematica, but the expression is tedious and not presented here. Let $e_1 = \mathrm{e}^{\lambda_1 \tau}$ and $e_2 = \mathrm{e}^{\lambda_2 \tau}$. Then equation 1 can be simplified, to give

$$\begin{aligned}\mathbb{P}(G_1) = \big[&((2-4M)e_1 - 3)\theta_A^2 + (3 - 8M^2 - (2+8M^2)e_1 + 4M(1+4M)e_2)\theta_A\theta_B \\ &+2M(1+2M)(3+4M-2e_1)\theta_B^2\big] \big/ [3(1+4M)(\theta_A+2M\theta_B)(-\theta_A+\theta_B+2M\theta_B)]\,.\end{aligned} \tag{2}$$

Similarly the probability for gene tree G_2 is

$$\mathbb{P}(G_2) = P_{17}^{②}(\tau) + [P_{11}^{②}(\tau) + P_{12}^{②}(\tau) + P_{13}^{②}(\tau) + P_{14}^{②}(\tau)]/3. \tag{3}$$

From equations 1 and 3 we can see that $\mathbb{P}(G_2) > \mathbb{P}(G_1)$ if and only if $P_{17}^{②}(\tau) > P_{15}^{②}(\tau)$. Indeed which of gene trees G_1, G_2 and G_3 is more probable depends on the relative likelihoods of three scenarios (Fig. 1B):

(i) a_1 and a_2 coalesce in A, which occurs with probability $P_{15}^{②}(\tau)$ and leads to G_1;

(ii) a_1 migrates (backwards in time) to B and coalesces with b, with probability $P_{17}^{②}(\tau)$ for G_2;

(iii) a_2 migrates (backwards in time) to B and coalesces with b, with probability $P_{16}^{②}(\tau)$ for G_3.

In all other scenarios, the three gene trees occur with equal probability. When the coalescence rate is much lower in species A than in B (or when $\theta_A \gg \theta_B$) and the migration rate from B to A is high, case (i) may be less probable than (ii) or (iii).

The anomaly in gene tree probabilities identified here is similar to the species-tree anomaly analyzed by Long and Kubatko (2018). The assumption of unidirectional migration in our model allows us to obtain simpler or more expressive analytical results than is possible under the model of bidirectional migration of Long and Kubatko (2018).

As an example, consider $\mathbb{P}(G_1)$ as a function of M, with other parameters fixed at $\tau = 0.02, \theta_A = 0.025$, and $\theta_B = 0.001$ (Fig. 2B). When $M = 0$, the IM model of Figure 1B reduces to the simple MSC model of Figure 1A, and the gene tree probabilities are $\mathbb{P}(G_1) = 1 - \frac{2}{3}e^{-2\tau/\theta_A} = 0.865402$ and $\mathbb{P}(G_2) = \mathbb{P}(G_3) = \frac{1}{3}e^{-2\tau/\theta_A} = 0.067299$. Here $\tau/\frac{\theta_A}{2}$ is the branch length of branch A in coalescent units (as the average coalescent time in population A is $\frac{1}{2}\theta_A$ mutations per site), and $e^{-2\tau/\theta_A}$ is the probability that two sequences (a_1 and a_2) do not coalesce along branch A or over the time interval $(0, \tau)$. At the threshold value $M^* = 0.521361$, $\mathbb{P}(G_1) = \mathbb{P}(G_2) = \frac{1}{3}$. When $M = 0.8$, the probabilities for the three scenarios described above are $P_{15}^{②}(\tau) = 0.23781$ and $P_{16}^{②}(\tau) = P_{17}^{②}(\tau) = 0.35753$, with $\mathbb{P}(G_1) = 0.25352 < \mathbb{P}(G_2) = \mathbb{P}(G_3) = 0.37324$. In the limit of $M = \infty$, sequences a_1 and a_2 will immediately migrate (backwards in time) into B and the three sequences will coalesce at random, with $\mathbb{P}(G_1) \to \frac{1}{3}$.

To verify the analytical results, we used the *simulate* option of BPP (Flouri *et al.*, 2018) to generate 10^7 gene trees at those parameter values. The estimates of $\mathbb{P}(G_1)$ are 0.865374, 0.333393, and 0.253542, for $M = 0$, 0.521361, and 0.8, respectively, which differ from the above analytical calculations by less than 10^{-4}.

Figure S1A&B examines the impact of the divergence time (τ_R) and the ratio of the population sizes (θ_A/θ_B) on gene tree probabilities, when other parameters are fixed at the values of Figure 2. Those two parameters similarly partition the parameter space into two zones, with the anomaly $\mathbb{P}(G_1) < \mathbb{P}(G_2)$ occurring for large τ_R (with $\tau_R > 0.00119461$) and for very different population sizes (with $\theta_A/\theta_B > 2.66667$). Nevertheless, τ_R and θ_A/θ_B appear to have less impact than the migration rate M (Fig. 2A).

Figure 2B shows a partition of the 3-D parameter space into two zones: when the parameters are inside the outer tent, we have the anomaly $\mathbb{P}(G_1) < \mathbb{P}(G_2)$.

Average coalescent times

Next we consider the average coalescent times or sequence distances within and between species. One could in principle use the Markov chain $Q^{①}$ constructed earlier for the process of coalescence and migration for the three sequences in the sample (a_1, a_2, and b). However, it is far simpler to use a reduced Markov chain with fewer states for two sequences only. To derive the density of the coalescent time between sequences a_1 and a_2, i.e., t_{aa}, we consider a Markov chain with 4 states. We abbreviate states like $A_{a_1}A_{a_2}$ as AA, and merge states A and B (which are states reached when the two sequences have coalesced with the ancestral sequence in either A or B) into one absorbing state, $A|B$ (for 'A or B') (Andersen *et al.*, 2014). The generator matrix $Q^{③}$ is

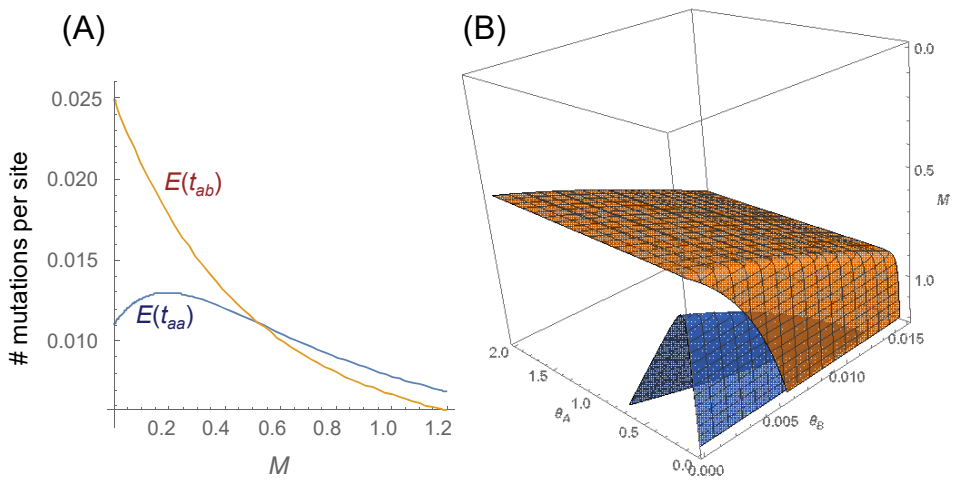

Fig. 3. (A) The expected coalescent times, $\mathbb{E}(t_{aa})$ and $\mathbb{E}(t_{ab})$, as functions of the migration rate M under the IM model (Fig. 1B). The two curves cross at $M^* = 0.5254101$, and when $M > M^*$, $\mathbb{E}(t_{aa}) > \mathbb{E}(t_{ab})$. Other parameters are fixed at $\tau = 0.02$, $\theta_A = 0.025, \theta_B = 0.001$, and $\theta_R = 0.01$. (B) Partition of the parameter space defined by θ_A, θ_B, and M under the IM model according to whether $\mathbb{E}(t_{aa}) > \mathbb{E}(t_{ab})$. Inside the outer tent, $\mathbb{E}(t_{aa}) > \mathbb{E}(t_{ab})$. Inside the inner tent $\mathbb{E}(t_{aa}) > \mathbb{E}(t_{ab}) + 0.001$. Other parameters are fixed at $\tau = 0.02$ and $\theta_R = 0.01$.

	AA	AB	BB	$A\|B$
AA	$-(2w_{BA} + c_A)$	$2w_{BA}$	0	c_A
AB	0	$-w_{BA}$	w_{BA}	0
BB	0	0	$-c_B$	c_B
$A\|B$	0	0	0	0

The eigenvalues of $Q^{③}$ are on the diagonal: $\lambda_1 = -\frac{8M}{\theta_A} - \frac{2}{\theta_A}$, $\lambda_2 = -\frac{4M}{\theta_A}$, $\lambda_3 = -\frac{2}{\theta_B}$, and $\lambda_4 = 0$. Let the transition probability matrix over time t be $P^{③}(t) = \exp(Q^{③}t)$, which is a function of $e^{\lambda_k t}$, $k = 1, 2, 3$. Like τ, time t is measured in the expected number of mutations per site. Thus

$$f(t_{aa}) = \begin{cases} P^{③}_{AA,AA}(t_{aa}) \cdot \frac{2}{\theta_A} + P^{③}_{AA,BB}(t_{aa}) \cdot \frac{2}{\theta_B}, & \text{if } 0 < t_{aa} < \tau_R, \\ \left[1 - P^{③}_{AA,A|B}(\tau_R)\right] \frac{2}{\theta_R} e^{-\frac{2}{\theta_R}(t_{aa} - \tau_R)}, & \text{if } t_{aa} > \tau_R. \end{cases} \quad (4)$$

Note that according to the definition of the probability density function, $f(t_{aa})\Delta t$ is the probability that the coalescent time falls in the interval $(t_{aa}, t_{aa} + \Delta t)$. When $t_{aa} < \tau_R$, this is the sum of two terms, as the coalescent event can occur in either species A or B. The first term, $P^{③}_{AA,AA}(t_{aa}) \cdot \frac{2}{\theta_A}\Delta t$, is the probability that sequences a_1 and a_2 are both in species A right before time t_{aa}, multiplied by the probability, $\frac{2}{\theta_A}\Delta t$, that they coalesce during the time interval $(t_{aa}, t_{aa} + \Delta t)$. The second term, $P^{③}_{AA,BB}(t_{aa}) \cdot \frac{2}{\theta_B}\Delta t$, is the probability for the coalescent to occur in species B. Similarly, in the case $t_{aa} > \tau_R$, both a_1 and a_2 enter R, with probability $1 - P^{③}_{AA,A|B}(\tau_R)$, and then coalesce in R in the time interval $(t_{aa}, t_{aa} + \Delta t)$, with probability $\frac{2}{\theta_R} e^{-\frac{2}{\theta_R}(t_{aa} - \tau_R)}\Delta t$.

The expectation of t_{aa} is given by averaging over the three cases of equation 4 in which a_1

and a_2 coalesce in A, B, and R:

$$\begin{aligned}
\mathbb{E}(t_{aa}) &= \int_0^{\tau_R} tP^{③}_{AA,AA}(t)\mathrm{d}t \cdot \frac{2}{\theta_A} + \int_0^{\tau_R} tP^{③}_{AA,BB}(t)\mathrm{d}t \cdot \frac{2}{\theta_B} + \left[1 - P^{③}_{AA,A|B}(\tau_R)\right]\left(\tau_R + \frac{\theta_R}{2}\right) \\
&= \left[e_1 - \frac{4M(e_1-e_2)}{1+2M} - \frac{8e_3M^2\theta_B^2}{(\theta_A-\theta_B-4M\theta_B)(2M\theta_B-\theta_A)}\right. \\
&\quad \left. - \frac{8e_2M^2\theta_B}{(1+2M)(2M\theta_B-\theta_A)} - \frac{8e_1M^2\theta_B}{(1+2M)(\theta_A-\theta_B-4M\theta_B)}\right]\left(\tau_R + \frac{\theta_R}{2}\right) \\
&\quad + \frac{\theta_A[1-e_1(1-\lambda_1\tau_R)]}{2(1+4M)^2} - \frac{\theta_A^2[1-e_2(1-\lambda_2\tau_R)]}{(1+2M)(2M\theta_B-\theta_A)} \\
&\quad + \frac{4M^2\theta_A^2}{(1+4M)^2(2M\theta_B-\theta_A)}\left[\frac{1}{1+2M} + \frac{\theta_B}{\theta_A-\theta_B-4M\theta_B}\right][1-e_1(1-\lambda_1\tau_R)] \\
&\quad - \frac{4M^2\theta_B^3[1-e_3(1-\lambda_3\tau_R)]}{(2M\theta_B-\theta_A)(\theta_A-\theta_B-4M\theta_B)},
\end{aligned} \tag{5}$$

where $e_k = \mathrm{e}^{\lambda_k\tau_R}$, $k = 1,2,3$.

To derive the density of the coalescent time t_{ab} between sequences a_1 and b, we consider a Markov chain with three states describing the backward process of coalescence and migration during time interval $(0,\tau_R)$. We abbreviate states like $A_{a_1}B_b$ as AB here. The generator matrix $Q^{④}$ is

	AB	BB	B
AB	$-w_{BA}$	w_{BA}	0
BB	0	$-c_B$	c_B
B	0	0	0

Thus the transition probability matrix is $P^{④}(t) = \exp(Q^{④}t)$, and

$$f(t_{ab}) = \begin{cases} P^{④}_{AB,BB}(t_{ab}) \cdot \frac{2}{\theta_B}, & \text{if } 0 < t_{ab} < \tau_R, \\ \left[1 - P^{④}_{AB,B}(\tau_R)\right]\frac{2}{\theta_R}\mathrm{e}^{-\frac{2}{\theta_R}(t_{ab}-\tau_R)}, & \text{if } t_{ab} > \tau_R. \end{cases} \tag{6}$$

The expectation of t_{ab} is given by averaging over the two cases:

$$\begin{aligned}
\mathbb{E}(t_{ab}) &= \int_0^{\tau_R} tP^{④}_{AB,BB}(t)dt \cdot \frac{2}{\theta_B} + \left[1 - P^{④}_{AB,B}(\tau_R)\right]\left(\tau_R + \frac{\theta_R}{2}\right) \\
&= \frac{4M^2\theta_B^2[1-e_3(1-\lambda_3\tau_R)] - \theta_A^2[1-e_2(1-\lambda_2\tau_R)]}{4M(2M\theta_B-\theta_A)} \\
&\quad + \left[e_2 - \frac{2M\theta_B(e_2-e_3)}{2M\theta_B-\theta_A}\right]\left(\tau_R + \frac{\theta_R}{2}\right).
\end{aligned} \tag{7}$$

We plot $\mathbb{E}(t_{aa})$ and $\mathbb{E}(t_{ab})$ against the migration rate M in Figure 3A, with other parameters in the model fixed at $\tau = 0.02$, $\theta_A = 0.025$, $\theta_B = 0.001$, and $\theta_R = 0.01$. In the extreme case of $M = 0$, the IM model becomes a model of complete isolation (or the MSC model, Fig. 1A), in which case $\mathbb{E}(t_{ab}) = \tau_R + \frac{1}{2}\theta_R = 0.025$ and $\mathbb{E}(t_{aa}) = \frac{1}{2}\theta_A + P_A \cdot \frac{1}{2}(\theta_R - \theta_A) = 0.01099$,

with $P_A = \exp(-\frac{2}{\theta_A}\tau) = 0.2019$ to be the probability that a_1 and a_2 do not coalesce in species A. Here $\mathbb{E}(t_{aa})$ is given by the approach of "iterated corrections", since the coalescent process between a_1 and a_2 occurs at different rates (determined by θ_A and θ_R) before and after τ_R (Burgess and Yang, 2008, eq. 7). If θ_A and θ_R were equal, the mean coalescent time would be $\frac{\theta_A}{2}$. Thus applying a correction for different population sizes, which affects a proportion P_A of the coalescent times, leads to $\mathbb{E}(t_{aa}) = \frac{1}{2}\theta_A + P_A \cdot \frac{1}{2}(\theta_R - \theta_A)$. In the other extreme case with $M \to \infty$, both a_1 and a_2 will migrate (backwards in time) into B immediately and then coalesce with b at random, so that $\mathbb{E}(t_{aa}) = \mathbb{E}(t_{ab}) = \frac{1}{2}\theta_B + P_B \cdot \frac{1}{2}(\theta_R - \theta_B) = 0.0005$, where $P_B = \exp(-\frac{2}{\theta_B}\tau)$. When M is greater than a threshold value, $M^* = 0.5254101$, $\mathbb{E}(t_{aa}) > \mathbb{E}(t_{ab})$.

We used BPP to simulate 10^7 gene trees at the parameter values of Figure 3A to verify the equations. At $M^* = 0.5254101$, the estimates of $\mathbb{E}(t_{aa})$ and $\mathbb{E}(t_{ab})$ are 0.0110556 and 0.0110550, in comparison with 0.0110557 and 0.0110557 from equations 5 and 7. At $M = 0.8$ they are 0.00902285 and 0.00808002, in comparison with 0.00902284 and 0.00808021 from equations 5 and 7.

Figure S1C&D examines the impact of the divergence time τ_R and the ratio of population sizes θ_A/θ_B on the average coalescent times, with other parameters fixed at the values of Figure 3. The anomaly $\mathbb{E}(t_{aa}) > \mathbb{E}(t_{ab})$ occurs when τ_R is large and when θ_A is much greater than θ_B.

Figure 3B shows a partition of a 3-D parameter space. The anomaly $\mathbb{E}(t_{aa}) > \mathbb{E}(t_{ab})$ occurs more easily for large M and when θ_A is much greater than θ_B.

THE MSCI MODEL FOR TWO SPECIES WITH THREE SEQUENCES

Consider the introgression (MSci) model for two species A and B, with $B \to A$ introgression at time $\tau_H = \tau_S$ and introgression probability φ (Fig. 1C). Again consider a sample of three sequences, a_1 and a_2 from A and b from B. We derive the probabilities for the three gene trees: $G_1 = ((a_1, a_2), b)$, $G_2 = ((b, a_1), a_2)$, and $G_3 = ((b, a_2), a_1)$, as well as the expected within-species and between-species coalescent times: $\mathbb{E}(t_{aa})$ and $\mathbb{E}(t_{ab})$.

Probabilities of gene trees

The gene tree topology depends on whether sequences a_1 and a_2 coalesce in species A (i.e., over the time interval 0–τ_H), and, if they do not, on whether they migrate into population S, and so on (Fig. 1C). Note that in population A, sequences a_1 and a_2 coalesce according to a Poisson process at the rate $\frac{2}{\theta_A}$. Thus the probability that a_1 and a_2 do not coalesce in A before reaching time τ_H is

$$P_A = \mathrm{e}^{-\frac{2}{\theta_A}\tau_H}. \tag{8}$$

Similarly we define

$$P_H = \mathrm{e}^{-\frac{2}{\theta_H}(\tau_R - \tau_H)} \text{ and } P_S = \mathrm{e}^{-\frac{2}{\theta_S}(\tau_R - \tau_H)} \tag{9}$$

to be the probabilities that two sequences entering populations H or S, respectively, do not coalesce in that population (Fig. 1C). Then the probabilities for the three gene trees are

$$\begin{aligned}
\mathbb{P}(G_1) &= (1 - P_A) + P_A(1-\varphi)^2(1 - P_H) + \frac{1}{3}P_A(1-\varphi)^2 P_H \\
&\quad + \frac{2}{3}P_A\varphi(1-\varphi)P_S + \frac{1}{3}P_A\varphi^2, \\
\mathbb{P}(G_2) &= \mathbb{P}(G_3) = \frac{1}{3}P_A(1-\varphi)^2 P_H + P_A\varphi(1-\varphi)\left(1 - \frac{1}{3}P_S\right) + \frac{1}{3}P_A\varphi^2,
\end{aligned} \tag{10}$$

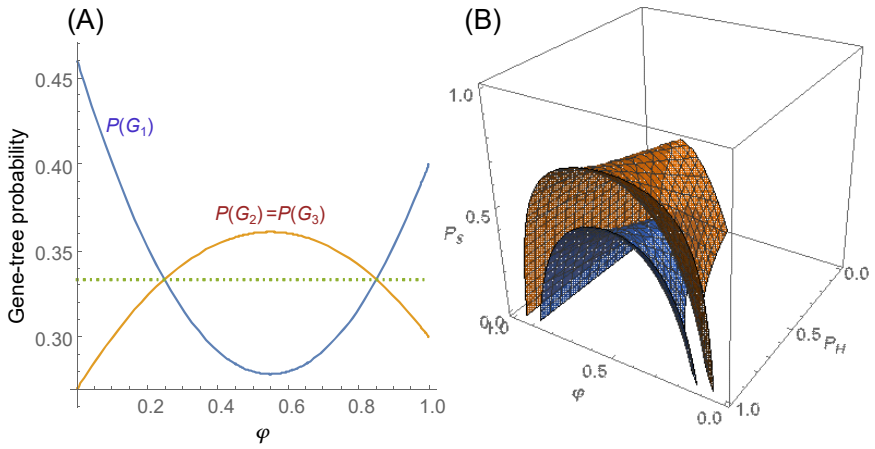

Fig. 4. (A) Probabilities of gene trees G_1 and G_2 as functions of the introgression probability φ in the MSci model (Fig. 1C) when $P_A = P_H = 0.9$ and $P_S = 0.1$ are fixed. (B) Partition of the parameter space according to gene tree probabilities: $\mathbb{P}(G_1) < \mathbb{P}(G_2)$ if and only if the parameter values are inside the tent. The inner and outer tents correspond to $P_A = 0.90$ and 0.95, respectively.

with $\mathbb{P}(G_1) + 2\mathbb{P}(G_2) = 1$. Here $\mathbb{P}(G_1)$ is a sum of five terms, corresponding to different scenarios in which the first coalescent event is between a_1 and a_2. The first term, $1 - P_A$, is the probability that a_1 and a_2 coalesce in population A. The second term, $P_A(1-\varphi)^2(1-P_H)$, is the probability that a_1 and a_2 do not coalesce in population A, they both enter population H (branch RH in the species tree, Figure 1C) and coalesce in H. The third term, $P_A(1-\varphi)^2 P_H \cdot \frac{1}{3}$, is the probability that a_1 and a_2 do not coalesce in A, and they both enter H and then R, where the three sequences coalesce in random order. The fourth term, $P_A \cdot 2\varphi(1-\varphi)P_S \cdot \frac{1}{3}$, is the probability that a_1 and a_2 do not coalesce in A and one of them enters S but does not coalesce with b in S, so that all three sequences enter R and coalesce in random order. The fifth term, $P_A\varphi^2 \cdot \frac{1}{3}$, is the probability that a_1 and a_2 do not coalesce in A and they both enter S, so that all three sequences enter S and coalesce at random in S or R.

Similarly $\mathbb{P}(G_2)$ is a sum of three terms, corresponding to three different scenarios in which sequences a_1 and b coalesce first. The first term, $P_A(1-\varphi)^2 P_H \cdot \frac{1}{3}$, is for a_1 and a_2 not to coalesce in A but to enter H and then R, and then for a_1 and b to coalesce in R. The second term, $P_A\varphi(1-\varphi) \cdot \left(P_S \cdot \frac{1}{3} + (1-P_S) + P_S \cdot \frac{1}{3}\right)$, is for one of a_1 and a_2 to enter H and the other to enter S. If a_1 enters H, and a_2 enters S and does not coalesce with b in S, then a_1 and b can coalesce in R. If a_2 enters H and a_1 enters S, then a_1 and b may coalesce in S or R. Lastly the third term, $P_A\varphi^2 \cdot \frac{1}{3}$, is for both a_1 and a_2 to enter S and then for the three sequences to coalesce at random in S or R.

We have $\mathbb{P}(G_1) < \mathbb{P}(G_2) = \mathbb{P}(G_3)$ if and only if

$$P_A\varphi(1-\varphi)(1-P_S) > 1 - P_A + P_A(1-\varphi)^2(1-P_H) \tag{11}$$

or

$$P_A(2-P_H-P_S)\varphi^2 - P_A(3-2P_H-P_S)\varphi + 1 - P_AP_H < 0. \tag{12}$$

While the MSci model of Figure 1C has seven parameters (we do not count θ_B since it is not needed to simulate sequence data of a_1, a_2, and b), the gene tree probabilities depend on only four: the introgression probability φ and the three branch lengths in coalescent units for branches A, H and S (Fig. 1C). Note that P_A, P_H and P_S are simply functions of the respective branch

lengths (in coalescent units). We plot $\mathbb{P}(G_1)$ and $\mathbb{P}(G_2)$ against φ in Figure 4A, with $P_A = P_H = 0.9$ and $P_S = 0.1$ fixed. Note that when $\varphi = 0$ or 1, the MSci model reduces to the simple MSC model for two species with changing population sizes but without introgression. At $\varphi = 0$, we have $\mathbb{P}(G_1) = 1 - \frac{2}{3}P_A P_H = 0.46$ while at $\varphi = 1$, $\mathbb{P}(G_1) = 1 - \frac{2}{3}P_A = 0.4$. The anomaly $\mathbb{P}(G_1) < \mathbb{P}(G_2)$ occurs in the zone $0.247694 < \varphi < 0.852306$. When φ is close to 1 (or $>$ 0.852306), a_1 and a_2 either coalesce in A or both will very likely enter species S and coalesce with b at random, so that $\mathbb{P}(G_1) > \mathbb{P}(G_2)$. Note that $\mathbb{P}(G_1)$ is not a monotonic function of φ: when introgression is either very rare or virtually guaranteed there is an increased chance for a_1 and a_2 to be in the same population and coalesce. Figure S2A, B & C examines the impact of τ_R, θ_H/θ_S, and τ_H (Fig. 1C) on gene tree probabilities. The anomaly $\mathbb{P}(G_1) < \mathbb{P}(G_2)$ occurs when τ_R is in a certain range, when θ_H is much greater than θ_S, and when τ_H is small.

Figure 4B shows the zone of parameters in which $\mathbb{P}(G_1) < \mathbb{P}(G_2)$ in a 3-D space. When P_A and P_H are large and P_S is small (or when θ_A and θ_H are large and θ_S is small), the anomaly $\mathbb{P}(G_1) < \mathbb{P}(G_2)$ may occur even with $\varphi < 0.5$.

Average coalescent times

The density of the coalescent time between sequences a_1 and a_2 is

$$f(t_{aa}) = \begin{cases} \frac{2}{\theta_A} e^{-\frac{2}{\theta_A} t_{aa}}, & \text{if } 0 < t_{aa} < \tau_H, \\ P_A \left[(1-\varphi)^2 \frac{2}{\theta_H} e^{-\frac{2}{\theta_H}(t_{aa}-\tau_H)} + \varphi^2 \frac{2}{\theta_S} e^{-\frac{2}{\theta_S}(t_{aa}-\tau_H)} \right], & \text{if } \tau_H < t_{aa} < \tau_R, \\ P_A[(1-\varphi)^2 P_H + \varphi^2 P_S + 2\varphi(1-\varphi)] \frac{2}{\theta_R} e^{-\frac{2}{\theta_R}(t_{aa}-\tau_R)}, & \text{if } t_{aa} > \tau_R. \end{cases} \quad (13)$$

First, the probability, $f(t_{aa})\Delta t$, that sequences a_1 and a_2 coalesce during the time interval $(t_{aa}, t_{aa}+\Delta t)$, with $t_{aa} < \tau_H$, is given by the probability, $e^{-\frac{2}{\theta_A} t_{aa}}$, that they do not coalesce before time t_{aa}, multiplied by the probability, $\frac{2}{\theta_A}\Delta t$, that they coalesce during the time interval (t_{aa}, $t_{aa}+\Delta t$). Second, for $\tau_H < t_{aa} < \tau_R$, $f(t_{aa})\Delta t$ is the sum of two terms, as the coalescent event can occur in either species H or S. The first term, $P_A(1-\varphi)^2 \frac{2}{\theta_H} e^{-\frac{2}{\theta_H}(t_{aa}-\tau_H)}\Delta t$, is the probability that a_1 and a_2 do not coalesce in species A, but both enter species H and coalesce there. Similarly, the second term, $P_A \varphi^2 \frac{2}{\theta_S} e^{-\frac{2}{\theta_S}(t_{aa}-\tau_H)}\Delta t$, is the probability that a_1 and a_2 do not coalesce in species A, but both enter species S and coalesce there. Finally, in the case $t_{aa} > \tau_R$, both a_1 and a_2 enter R with probability $P_A[(1-\varphi)^2 P_H + \varphi^2 P_S + 2\varphi(1-\varphi)]$, and coalesce in R in the time interval (t_{aa}, $t_{aa}+\Delta t$) with probability $\frac{2}{\theta_R} e^{-\frac{2}{\theta_R}(t_{aa}-\tau_R)}\Delta t$.

The expectation of t_{aa} is given by averaging over the four cases of equation 13 in which a_1 and a_2 coalesce in A, H, S, and R.

$$\begin{aligned} \mathbb{E}(t_{aa}) = {} & \frac{\theta_A}{2} - P_A(\tau_H + \frac{\theta_A}{2}) + P_A(1-\varphi)^2 \left[\tau_H + \frac{\theta_H}{2} - P_H(\tau_R + \frac{\theta_H}{2}) \right] \\ & + P_A \varphi^2 \left[\tau_H + \frac{\theta_S}{2} - P_S(\tau_R + \frac{\theta_S}{2}) \right] + P_A[P_H(1-\varphi)^2 + P_S\varphi^2 + 2\varphi(1-\varphi)](\tau_R + \frac{\theta_R}{2}). \quad (14) \end{aligned}$$

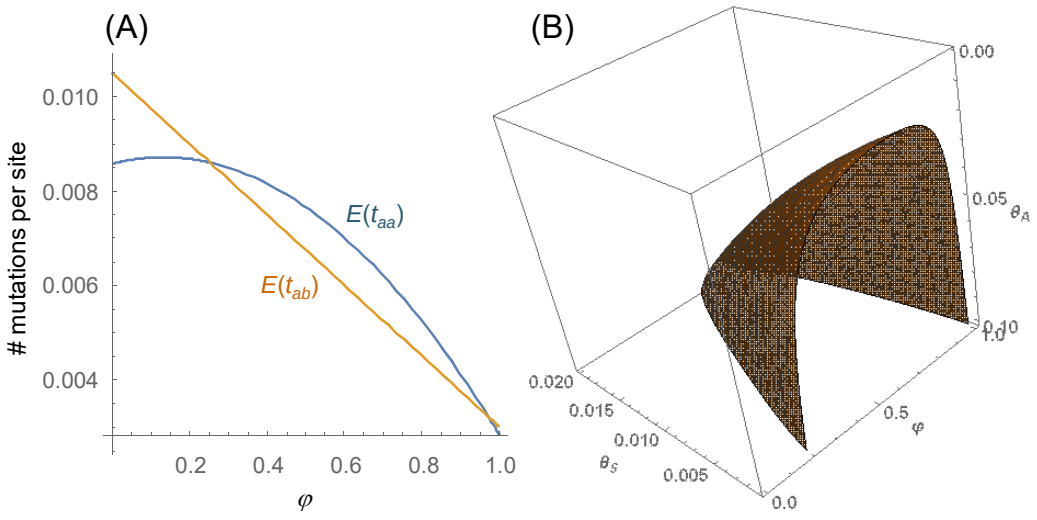

Fig. 5. (A) The expected coalescent times, $\mathbb{E}(t_{aa})$ and $\mathbb{E}(t_{ab})$, as functions of the introgression probability φ under the MSci model (Fig.1C). Other parameters are fixed at $\theta_A = \theta_H = 0.05$, $\theta_S = \theta_R = 0.001$, $\tau_R = 0.01$, and $\tau_H = 0.0025$. (B) Partition of the parameter space defined by θ_A, θ_S, and φ under the MSci model (Fig. 1C): inside the tent, $\mathbb{E}(t_{aa}) > \mathbb{E}(t_{ab})$ while outside it the opposite is true. Other parameters are fixed at $\theta_H = 0.05$, $\theta_R = 0.001$, $\tau_R = 0.01$, and $\tau_H = 0.0025$.

Similarly the density of the coalescent time between sequences a_1 and b is

$$f(t_{ab}) = \begin{cases} \varphi \frac{2}{\theta_S} e^{-\frac{2}{\theta_S}(t_{ab}-\tau_H)}, & \text{if } \tau_H < t_{ab} < \tau_R, \\ [(1-\varphi)+P_S\varphi]\frac{2}{\theta_R} e^{-\frac{2}{\theta_R}(t_{ab}-\tau_R)}, & \text{if } t_{ab} > \tau_R. \end{cases} \tag{15}$$

When $\tau_H < t_{ab} < \tau_R$, the coalescent occurs in species S, and $f(t_{ab})\Delta t$ is given by the probability, φ, that sequence a_1 enters species S times the probability, $\frac{2}{\theta_S} e^{-\frac{2}{\theta_S}(t_{ab}-\tau_H)}\Delta t$, that a_1 and b coalesce in S in the time interval $(t_{ab}, t_{ab}+\Delta t)$. In the case of $t_{ab} > \tau_R$, the coalescent occurs in species R. The probability that both a_1 and b enter R is $(1-\varphi)+\varphi P_S$, and the probability that they coalesce in R in the time interval $(t_{ab}, t_{ab}+\Delta t)$ is $\frac{2}{\theta_R} e^{-\frac{2}{\theta_R}(t_{ab}-\tau_R)}\Delta t$.

The expectation of t_{ab} is given by

$$\mathbb{E}(t_{ab}) = \varphi\left[\tau_H + \frac{\theta_S}{2} + P_S\left(\frac{\theta_R}{2} - \frac{\theta_S}{2}\right)\right] + (1-\varphi)\left(\tau_R + \frac{\theta_R}{2}\right). \tag{16}$$

This is a weighted average depending on whether sequence a enters S (with probability φ) or H (with probability $1-\varphi$). If a enters S, the mean coalescent time is $\tau_H + \frac{\theta_S}{2} + P_S\left(\frac{\theta_R}{2} - \frac{\theta_S}{2}\right)$ by the argument of iterated corrections. Similarly with probability $1-\varphi$ sequence a enters H and coalesces with b in R, with the mean coalescent time to be $\tau_R + \frac{\theta_R}{2}$.

Thus $\mathbb{E}(t_{aa}) > \mathbb{E}(t_{ab})$ if and only if

$$\begin{aligned}&\frac{\theta_A}{2} - P_A(\tau_H + \frac{\theta_A}{2}) + P_A(1-\varphi)^2\left[\tau_H + \frac{\theta_H}{2} - P_H(\tau_R + \frac{\theta_H}{2})\right]\\&+ (P_A\varphi^2 - \varphi)\left[\tau_H + \frac{\theta_S}{2} - P_S(\tau_R + \frac{\theta_S}{2})\right] + [P_A P_H(1-\varphi)^2\\&+ P_S\varphi(P_A\varphi - 1) + (2P_A\varphi - 1)(1-\varphi)](\tau_R + \frac{\theta_R}{2}) > 0.\end{aligned} \tag{17}$$

We plot $\mathbb{E}(t_{aa})$ and $\mathbb{E}(t_{ab})$ against φ in Figure 5A, with other parameters in the MSci model fixed: $\theta_A = \theta_H = 0.05$, $\theta_S = \theta_R = 0.001$, $\tau_R = 0.01$, and $\tau_H = 0.0025$. Note that the coalescent times depend on all seven parameters of the MSci model except θ_B (Fig. 1C). The cases $\varphi = 0$ and 1 correspond to MSC (complete-isolation) models for two species with changing population sizes. With $\varphi = 0$, sequences a_1 and a_2 coalesce at different rates determined by population sizes θ_A, θ_H, and θ_R, so the approach of iterated corrections gives $\mathbb{E}(t_{aa}) = \frac{\theta_A}{2} + P_A[(\frac{\theta_H}{2} - \frac{\theta_A}{2}) + P_H(\frac{\theta_R}{2} - \frac{\theta_H}{2})] = 0.00857716$. Also at $\varphi = 0$, sequences a_1 and b can coalesce in R only, with $\mathbb{E}(t_{ab}) = \tau_R + \frac{\theta_R}{2} = 0.0105$. At $\varphi = 1$, sequences a_1 and a_2 can coalesce in A, S, or R, so that $\mathbb{E}(t_{aa}) = \frac{\theta_A}{2} + P_A[(\frac{\theta_S}{2} - \frac{\theta_A}{2}) + P_S(\frac{\theta_R}{2} - \frac{\theta_S}{2})] = 0.00283148$ while sequences a_1 and b can coalesce in S or R, with $\mathbb{E}(t_{ab}) = \tau_H + \frac{\theta_S}{2} + P_S(\frac{\theta_R}{2} - \frac{\theta_S}{2}) = 0.003$. When φ is close to 1, either a_1 and a_2 coalesce in A or they both enter S and coalesce with b at random, so that $\mathbb{E}(t_{aa}) < \mathbb{E}(t_{ab})$. When $0.252962 < \varphi < 0.971179$, we have the anomaly $\mathbb{E}(t_{aa}) > \mathbb{E}(t_{ab})$. If species A has a much larger population size than S, it may be more likely for sequence a_1 or a_2 to migrate into species S and coalesce with b than for a_1 to coalesce with a_2, causing $\mathbb{E}(t_{aa}) > \mathbb{E}(t_{ab})$. Such anomaly may occur even if φ is much smaller than $\frac{1}{2}$.

We confirmed our derivations by simulating 10^7 gene trees using BPP (Flouri *et al.*, 2018). With $\varphi = 0.4$ in Figure 5A, the estimates are 0.00815726 for $\mathbb{E}(t_{aa})$ and 0.007499884 for $\mathbb{E}(t_{ab})$, compared with 0.008157341 and 0.0075 from equations 14 and 16.

Figure S2D, E & F examines the impact of τ_R, θ_H/θ_S, and τ_H (Fig. 1C) on the average coalescent times, when other parameters are fixed at the values of Figure 5. $\mathbb{E}(t_{aa}) > \mathbb{E}(t_{ab})$ when τ_R is in a certain range, when θ_H is much greater than θ_S, and when τ_H is small.

Figure 5B shows the anomaly zone with $\mathbb{E}(t_{aa}) > \mathbb{E}(t_{ab})$ in the 3-D space of parameters θ_A, θ_S, and φ, with other parameters fixed.

DISCUSSION

The nature of the anomaly

The species-definition anomaly zone, in which the within-species divergence is greater than the between-species divergence, with divergence measured by either the gene tree probability or the average genetic distance, is very similar to the species-tree anomaly zone (Degnan and Rosenberg, 2006). In the species-tree anomaly zone, the use of the most common gene tree topology as the species tree estimate will be statistically inconsistent, although it should be emphasized that the problem disappears if one takes a likelihood approach and uses the likelihood (that is, the probability of the gene trees) to compare different species trees (Xu and Yang, 2016). The models considered in this paper (Fig. 1B&C) involve only two species with only one simple species tree: (A, B). However, one may consider the gene tree $G_1 = ((a_1, a_2), b)$ to match the species tree, and gene trees $G_2 = ((b, a_1), a_2)$ and $G_3 = ((b, a_2), a_1)$ to be the mismatching trees. Then the anomaly $\mathbb{P}(G_1) < \mathbb{P}(G_2)$ means that the matching gene tree has a smaller probability than either mismatching gene tree, a situation very similar to the anomaly

zone in species tree estimation. Nevertheless, the anomaly zone for species tree estimation is due to polymorphism in ancestral species and the resulting deep coalescence, while the anomaly discussed in this paper is due to cross-species gene flow and different population sizes. In the context of phylogenetic network (i.e., MSci) models, Zhu *et al.* (2016) defined an anomalous gene tree as one that has a higher probability than any gene tree that matches a *displayed species tree* — displayed species trees are binary trees that remain when one of the two parental branches at each hybridization node in the species network is removed (Zhu *et al.*, 2016; Zhu and Degnan, 2017). All such anomalies share the feature that the most probable gene tree under the data-generating model does not match one's intuitive expectation. Here we stress that the "counter-intuitive" results do not imply that genetic sequence data contain misleading information about the history of species divergences.

The species-definition anomaly does not occur in the MSC model without gene flow (Fig. 1A). Nor does it occur in simple models of population subdivision in population genetics. For example, under the islands and stepping-stones models, the expected coalescent time between sequences sampled from the same population must be smaller than that between sequences sampled from two different populations (Li, 1976; Strobeck, 1987; Slatkin, 1987, 1991). Those models assume symmetry in the population size and migration rate: the different populations are assumed to have the same size and the migration rate is assumed to be the same between any two populations in the islands model or between any two adjacent populations in the stepping-stones model. In the IM and MSci models considered here, cross-species gene flow and large differences in population size are the main causes for the anomaly. We note that the anomaly described in this paper can occur in more general settings than we considered. For example, we have assumed unidirectional migration (from B to A only) in the IM model, but the same behavior should occur in a more general model of bidirectional migration (Long and Kubatko, 2018), as long as there is sufficient asymmetry in the population size and in the migration rate.

In our analysis we have assumed a simple neutral coalescent model (with and without gene flow) and have not considered the effects of natural selection or population structure. Selection may distort the distribution of the gene tree topologies and coalescent times, especially when the population sizes and thus the efficacy of purifying selection differs between species (He *et al.*, 2020). Previously coding loci were found to produce highly consistent species-tree and parameter estimates with the noncoding parts of the genome (Shi and Yang, 2018; Thawornwattana *et al.*, 2018; Flouri *et al.*, 2020), suggesting that the effects may be minor if purifying selection operates in similar ways in different species. However, species-specific selection, as expected for gene loci responsible for ecological adaptation of the species (Turner *et al.*, 2005; Pardo-Diaz *et al.*, 2012), will likely have major impacts on the gene tree distribution. Furthermore our analysis has assumed that each species is a population of panmixia. Population subdivision may lead to an inflated effective population size for the species, and may create a scenario that is similar to the model studied here. Suppose species A has a wide-ranging geographical distribution with population subdivision, while species B has a very limited distribution and is close to one of the geographical populations of species A. Our analysis suggests that such gene flow can easily create a species-definition anomaly zone, with two sequences randomly sampled from species A to be on average more distantly related than two sequences from the two different species.

How common is the species-definition anomaly?

While our theoretical calculations suggest that the species-definition anomaly is possible in large zones of the parameter space, it is not known how often it occurs in nature. This empirical question can be addressed by estimating the relevant parameters (in particular the migration rate M and the introgression probability φ) under the IM and MSci models using genomic sequence data. Currently such estimates are rare and mostly based on small datasets, while it may be

necessary to use hundreds or thousands of loci to get reliable estimates. Nevertheless, available estimates (e.g., Pinho and Hey, 2010, table S1) suggest that population sizes can differ by orders of magnitude even between closely related species, and migration is often asymmetrical, providing opportunities for the anomaly to occur.

Here we briefly review a few recent studies which generated estimates of migration rates from genomic data, from fruit flies, mosquitoes, butterflies, and gibbons. Several studies have found significant evidence for gene flow from *Drosophila simulans* to *D. melanogaster*, at the rate of $M_{S\to M} = 0.02$–0.04 migrant individuals per generation, but no migration in the opposite direction ($M_{M\to S} \approx 0$) (Wang and Hey, 2010; Dalquen *et al.*, 2017, tables 9 & 10). Population sizes were around $\theta_S = 0.013$ and $\theta_M = 0.005$, with the divergence time $\tau_{SM} \approx 0.012$–0.014 (Dalquen *et al.*, 2017, tables 9 & 10). In the *Anopheles gambiae* species complex of African mosquitoes, hybridization occurs between several pairs of non-sister species. Gene flow from *A. arabiensis* to *A. gambiae* (or *A. coluzzii*) occurs so frequently for the autosomes that the gene trees reflect the migration history rather than the history of species divergences (Fontaine *et al.*, 2015; Thawornwattana *et al.*, 2018). Estimates from the genomic data are in the order of $M_{A\to G} \approx 0.2$ migrants per generation while $M_{G\to A} = 0$ (Thawornwattana *et al.*, 2018, table S3; Flouri *et al.*, 2020, table 1), in agreement with crossing experiments, which showed that introgressed alleles from *A. arabiensis* to *A. gambiae* persisted over many generations, while it was not possible to maintain an introgression colony in the opposite $G \to A$ direction (Slotman *et al.*, 2005). Other parameters were around $\theta_A = 0.014$, $\theta_G = 0.02$–0.03, and $\tau_{AG} = 0.007$ (Thawornwattana *et al.*, 2018, table S3). *Heliconius* butterflies constitute one of the best studied groups for cross-species hybridization/introgression, involving many sister- and nonsister-species pairs, and involving both recent and ancient gene flow (Bull *et al.*, 2006; Kronforst *et al.*, 2006; Mallet *et al.*, 2007; Salazar *et al.*, 2008; Pardo-Diaz *et al.*, 2012; Martin *et al.*, 2013). A recent study (Van Belleghem *et al.*, 2020) applied coalescent-based simulation to joint site-frequency spectrum data to estimate the migration rates and population sizes between two incipient species: *H. erato* and *H. himera*, finding strong evidence for highly asymmetrical introgression, predominantly from *H. erato favorinus* to *H. himera*, at the rate of $M = 0.5$–0.6 migrants per generation, with $\tau \approx 0.002$, $\theta_E = 0.01$, and $\theta_H = 0.0008$. In an analysis of genomic sequences from five species of gibbons (which belong to four different genera), gene flow was inferred between two species of the same genus: *Hylobates moloch* and *H. pileatus*, but not between species of different genera. The migration rates were estimated to be $M_{M\to P} \approx 0.008$ migrants per generation, while $M_{P\to M} \approx 0$, with $\theta_M = 0.0014$, $\theta_P = 0.0005$, and $\tau = 0.0017$ (Shi and Yang, 2018, Fig. 1).

The parameter estimates suggest that those species pairs are not in the species-definition anomaly zone as discussed in this paper. Nevertheless, they do suggest large differences in population size and in the migration rate in the two directions. They also indicate that the parameter values used in our example calculations (figs. 2, 3, 4, 5) are representative of real biological systems. We leave it to future genomic analyses to determine how common the anomaly is in the real world. As more and more genomes are sequenced, and as analytical methods are improved to handle large datasets, we see exciting opportunities for using genomic data to infer the evolutionary history of species divergence and gene flow.

The impact of gene flow on the definition and identification of species

It is noteworthy that the migration rate required for the species-definition anomaly to occur may be much less than one migrant per generation. For a species like the mosquitoes the population size may well be larger than a million, which means that a proportion of migrants less than one in a million is sufficient to change the apparent genetic history of the species. In population genetic models of population subdivision, migration rates of $M \ll 1$ are low enough so that the populations will be differentiated or isolated (as measured by $F_{\rm st}$) (Wright, 1931). However, in the IM model, such low levels of gene flow can have a dramatic impact on the history of the

species as represented in gene genealogies or genetic distances. Similarly Jiao *et al.* (2020) found that even a small amount of migration per generation can have a huge impact on species tree estimation under the simple MSC model (see also Long and Kubatko, 2018).

The dramatic impact of gene flow on the genetic history of the species suggests that one has to consider this effect when defining and identifying species. In the species-definition anomaly zone, simple application of DNA barcoding or the *gdi* will lump genuinely distinct species into the same species. Thus if those methods suggest one species but there is evidence for asymmetrical gene flow between the populations and drastically different population sizes, the results from those methods should be re-examined for the impact of gene flow. We suggest that estimating and contrasting the long-term migration rate and the short-term hybridization rate as an effective approach to establishing the existence of reproductive barriers and evidence for species status. Note that genomic sequence data may contain rich information concerning evolutionary parameters such as species divergence times, population sizes, and migration rates or introgression probabilities, which may be invaluable for delimiting species boundaries (Fujita *et al.*, 2012; Leaché *et al.*, 2019). The migration rate estimated from genomic data under the IM model reflects the long-term impact of gene flow and genetic drift, as well as natural selection against introgressed alleles (Martin and Jiggins, 2017). Genomic sequence data can also be used to identify recent hybridization/admixture events (Anderson and Thompson, 2002; Anderson, 2008; Veller *et al.*, 2019). A greatly reduced migration rate relative to the hybridization rate (e.g., a migration rate of $m = 10^{-6}$ per generation relative to a proportion of F_1 hybrids of 0.1%) may be strong evidence that introgressed alleles are deleterious and removed from the receiving population by natural selection and that reproductive barriers exist between the species. While genomic data may be currently lacking for many species groups, this approach may become feasible in the near future with advancements in genome sequencing technologies and development of reduced-representation datasets (Lemmon *et al.*, 2012; Edwards *et al.*, 2017), as well as advancements of analytical methods that accommodate both coalescent and gene flow (Dalquen *et al.*, 2017; Hey *et al.*, 2018; Wen and Nakhleh, 2018; Zhang *et al.*, 2018; Flouri *et al.*, 2020).

1. Acknowledgements

We thank James Mallet and Yuttapong Thawornwattana for many discussions and comments. We are grateful to three anonymous reviewers, Matthew Hahn and Bryan Carstens for many constructive comments. This study has been supported by Biotechnology and Biological Sciences Research Council grant (BB/P006493/1) to Z.Y. and a BBSRC equipment grant (BB/R01356X/1).

References

Andersen, L. N., Mailund, T., and Hobolth, A. 2014. Efficient computation in the IM model. *J. Math. Biol.*, 68: 1423–1451.

Anderson, E. and Thompson, E. 2002. A model-based method for identifying species hybrids using multilocus genetic data. *Genetics*, 160: 1217–1229.

Anderson, E. C. 2008. Bayesian inference of species hybrids using multilocus dominant genetic markers. *Phil. Trans. R. Soc. Lond. B: Biol. Sci.*, 363(1505): 2841–2850.

Arnold, B. J., Lahner, B., DaCosta, J. M., Weisman, C. M., Hollister, J. D., Salt, D. E., Bomblies, K., and Yant, L. 2016. Borrowed alleles and convergence in serpentine adaptation. *Proc. Natl. Acad. Sci. U.S.A.*, 113(29): 8320–8325.

Bull, V., Beltran, M., Jiggins, C., McMillan, W. O., Bermingham, E., and Mallet, J. 2006. Polyphyly and gene flow between non-sibling *Heliconius* species. *BMC Biology*, 4: 11.

Burgess, R. and Yang, Z. 2008. Estimation of Hominoid ancestral population sizes under Bayesian coalescent models incorporating mutation rate variation and sequencing errors. *Mol. Biol. Evol.*, 25: 1979–1994.

Chan, Y. C., Roos, C., Inoue-Murayama, M., Inoue, E., Shih, C. C., Pei, K. J., and Vigilant, L. 2013. Inferring the evolutionary histories of divergences in *Hylobates* and *Nomascus* gibbons through multilocus sequence data. *BMC Evol. Biol.*, 13: 82.

Coyne, J. A. and Orr, H. A. 2004. *Speciation*. Sinauer Assoc., Sunderland, Massachusetts.

Dalquen, D., Zhu, T., and Yang, Z. 2017. Maximum likelihood implementation of an isolation-with-migration model for three species. *Syst. Biol.*, 66: 379–398.

Dasmahapatra, K. K., Elias, M., Hill, R. I., Hoffman, J. I., and Mallet, J. 2010. Mitochondrial DNA barcoding detects some species that are real, and some that are not. *Mol. Ecol. Resour.*, 10(2): 264–273.

De Queiroz, K. 2007. Species concepts and species delimitation. *Syst. Biol.*, 56: 879–886.

Degnan, J. H. and Rosenberg, N. A. 2006. Discordance of species trees with their most likely gene trees. *PLoS Genet.*, 2: e68.

Edwards, S., Cloutier, A., and Baker, A. 2017. Conserved nonexonic elements: a novel class of marker for phylogenomics. *Syst. Biol.*, 66(6): 1028–1044.

Ellegren, H., Smeds, L., Burri, R., Olason, P. I., Backstrom, N., Kawakami, T., Kunstner, A., Makinen, H., Nadachowska-Brzyska, K., Qvarnstrom, A., Uebbing, S., and Wolf, J. B. W. 2012. The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature*, 491: 756–760.

Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., and Glenn, T. C. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Syst. Biol.*, 61(5): 717–726.

Flouri, T., Jiao, X., Rannala, B., and Yang, Z. 2018. Species tree inference with BPP using genomic sequences and the multispecies coalescent. *Mol. Biol. Evol.*, 35(10): 2585–2593.

Flouri, T., Jiao, X., Rannala, B., and Yang, Z. 2020. A Bayesian implementation of the multispecies coalescent model with introgression for phylogenomic analysis. *Mol. Biol. Evol.*, 37(4): 1211–1223.

Fontaine, M. C., Pease, J. B., Steele, A., Waterhouse, R. M., Neafsey, D. E., Sharakhov, I. V., Jiang, X., Hall, A. B., Catteruccia, F., Kakani, E., Mitchell, S. N., Wu, Y. C., Smith, H. A., Love, R. R., Lawniczak, M. K., Slotman, M. A., Emrich, S. J., Hahn, M. W., and Besansky, N. J. 2015. Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science*, 347(6217): 1258524.

Fujita, M. K., Leaché, A. D., Burbrink, F. T., McGuire, J. A., and Moritz, C. 2012. Coalescent-based species delimitation in an integrative taxonomy. *Trends Ecol. Evol.*, 27: 480–488.

He, C., Liang, D., and Zhang, P. 2020. Asymmetric distribution of gene trees can arise under purifying selection if differences in population size exist. *Mol. Biol. Evol.*, 37(3): 881–892.

Hebert, P. D., Cywinska, A., Ball, S. L., and deWaard, J. R. 2003. Biological identifications through DNA barcodes. *Proc. Biol. Sci.*, 270: 313–321.

Hey, J. 2010. Isolation with migration models for more than two populations. *Mol. Biol. Evol.*, 27: 905–920.

Hey, J., Chung, Y., Sethuraman, A., Lachance, J., Tishkoff, S., Sousa, V. C., and Wang, Y. 2018. Phylogeny estimation by integration over isolation with migration models. *Mol. Biol. Evol.*, 35(11): 2805–2818.

Hudson, R. R. and Turelli, M. 2003. Stochasticity overrules the "three-times rule": genetic drift, genetic draft, and coalescence times for nuclear loci versus mitochondrial DNA. *Evolution*, 57: 182–190.

Jackson, N. D., Carstens, B. C., Morales, A. E., and O'Meara, B. C. 2017. Species delimitation with gene flow. *Syst. Biol.*, 66(5): 799–812.

Jiao, X., Flouri, T., Rannala, B., and Yang, Z. 2020. The impact of cross-species gene flow on species tree estimation. *Syst. Biol.*, page in press.

Karin, B. R., Gamble, T., and Jackman, T. R. 2020. Optimizing phylogenomics with rapidly evolving long exons: Comparison with anchored hybrid enrichment and ultraconserved elements. *Mol. Biol. Evol.*, 37(3): 904–922.

Kronforst, M. R., Young, L. G., Blume, L. M., and Gilbert, L. E. 2006. Multilocus analyses of admixture and introgression among hybridizing *Heliconius* butterflies. *Evolution*, 60(6): 1254–68.

Leaché, A. D., Koo, M. S., Spencer, C. L., Papenfuss, T. J., Fisher, R. N., and McGuire, J. A. 2009. Quantifying ecological, morphological, and genetic variation to delimit species in the coast horned lizard species complex (Phrynosoma). *Proc. Natl. Acad. Sci. U.S.A.*, 106: 12418–12423.

Leaché, A. D., Zhu, T., Rannala, B., and Yang, Z. 2019. The spectre of too many species. *Syst. Biol.*, 68(1): 168–181.

Lemmon, A. R., Emme, S. A., and Lemmon, E. M. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. *Syst. Biol.*, 61(5): 727–744.

Li, G., Figueiro, H. V., Eizirik, E., and Murphy, W. J. 2019. Recombination-aware phylogenomics reveals the structured genomic landscape of hybridizing cat species. *Mol. Biol. Evol.*, 36(10): 2111–2126.

Li, W.-H. 1976. Distribution of nucleotide differences between two randomly chosen cistrons in a subdivided population: the finite island model. *Theor. Popul. Biol.*, 10: 303–308.

Liu, S., Lorenzen, E. D., Fumagalli, M., Li, B., Harris, K., Xiong, Z., Zhou, L., Korneliussen, T. S., Somel, M., Babbitt, C., Wray, G., Li, J., He, W., Wang, Z., Fu, W., Xiang, X., Morgan, C. C., Doherty, A., O'Connell, M. J., McInerney, J. O., Born, E. W., Dalen, L., Dietz, R., Orlando, L., Sonne, C., Zhang, G., Nielsen, R., Willerslev, E., and Wang, J. 2014. Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears. *Cell*, 157: 785–794.

Long, C. and Kubatko, L. 2018. The effect of gene flow on coalescent-based species-tree inference. *Syst. Biol.*, 67(5): 770–785.

Mallet, J. 2005. Hybridization as an invasion of the genome. *Trends Ecol. Evol.*, 20: 229–237.

Mallet, J. 2008. Hybridization, ecological races, and the nature of species: empirical evidence for the ease of speciation. *Phil. Trans. R. Soc. B: Biol. Sci.*, 363: 2971–2986.

Mallet, J. 2013. Concepts of species. In S. Levin, editor, *Encyclopedia of Biodiversity*, volume 6, pages 679–691. Academic Press, Massachusetts.

Mallet, J., Beltran, M., Neukirchen, W., and Linares, M. 2007. Natural hybridization in heliconiine butterflies: the species boundary as a continuum. *BMC Evol. Biol.*, 7: 28.

Mao, Y., Economo, E. P., and Satoh, N. 2018. The roles of introgression and climate change in the rise to dominance of *Acropora* corals. *Curr. Biol.*, 28(21): 3373–3382 e5.

Martin, S. H. and Jiggins, C. D. 2017. Interpreting the genomic landscape of introgression. *Curr. Opin. Genet. Dev.*, 47: 69–74.

Martin, S. H., Dasmahapatra, K. K., Nadeau, N. J., Salazar, C., Walters, J. R., Simpson, F., Blaxter, M., Manica, A., Mallet, J., and Jiggins, C. D. 2013. Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Res.*, 23(11): 1817–1828.

Meyer, C. P. and Paulay, G. 2005. DNA barcoding: error rates based on comprehensive sampling. *PLoS Biol.*, 3(12): e422.

Nielsen, R., Akey, J. M., Jakobsson, M., Pritchard, J. K., Tishkoff, S., and Willerslev, E. 2017. Tracing the peopling of the world through genomics. *Nature*, 541: 302.

Notohara, M. 1990. The coalescent and the genealogical process in geographically structured populations. *J. Math. Biol.*, 29: 59–75.

Pardo-Diaz, C., Salazar, C., Baxter, S. W., Merot, C., Figueiredo-Ready, W., Joron, M., McMillan, W. O., and Jiggins, C. D. 2012. Adaptive introgression across species boundaries in *Heliconius* butterflies. *PLoS Genet.*, 8(6): e1002752.

Pinho, C. and Hey, J. 2010. Divergence with gene flow: models and data. *Ann. Rev. Ecol. Evol. Syst.*, 41: 215–230.

Puillandre, N., Lambert, A., Brouillet, S., and Achaz, G. 2012. Automatic barcode gap discovery for primary species delimitation. *Mol. Ecol.*, 21: 1864–1877.

Salazar, C., Jiggins, C., Taylor, J. E., Kronforst, M., and Linares, M. 2008. Gene flow and the genealogical history of *Heliconius heurippa*. *BMC Evol. Biol.*, 8: 132.

Shi, C. and Yang, Z. 2018. Coalescent-based analyses of genomic sequence data provide a robust resolution of phylogenetic relationships among major groups of gibbons. *Mol. Biol. Evol.*, 35: 159–179.

Slatkin, M. 1987. The average number of sites separating DNA sequences drawn from a subdivided population. *Theor. Popul. Biol.*, 32: 42–49.

Slatkin, M. 1991. Inbreeding coefficients and coalescence times. *Genet. Res.*, 58: 167–175.

Slotman, M. A., della Torre, A., Calzetta, M., and Powell, J. R. 2005. Differential introgression of chromosomal regions between *Anopheles gambiae* and *An. arabiensis*. *Am. J. Trop. Med. Hyg.*, 73(2): 326–335.

Strobeck, K. 1987. Average number of nucleotide differences in a sample from a single subpopulation: A test for population subdivision. *Genetics*, 117: 149–153.

Thawornwattana, Y., Dalquen, D., and Yang, Z. 2018. Coalescent analysis of phylogenomic data confidently resolves the species relationships in the *Anopheles gambiae* species complex. *Mol. Biol. Evol.*, 35(10): 2512–2527.

Tian, Y. and Kubatko, L. S. 2016. Distribution of coalescent histories under the coalescent model with gene flow. *Mol. Phylogenet. Evol.*, 105: 177–192.

Turner, T. L., Hahn, M. W., and Nuzhdin, S. V. 2005. Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biol.*, 3(9): e285.

Van Belleghem, S. M., Cole, J. M., Montejo-Kovacevich, G., Bacquet, C. N., McMillan, W. O., Papa, R., and Counterman, B. A. 2020. Selection and gene flow define polygenic barriers between incipient butterfly species. *bioRxiv*, page 2020.04.09.034470.

Veller, C., Edelman, N., Muralidhar, P., and Nowak, M. 2019. Recombination, variance in genetic relatedness, and selection against introgressed DNA. *bioRxiv:846147*.

Wang, Y. and Hey, J. 2010. Estimating divergence parameters with small samples from a large number of loci. *Genetics*, 184: 363–379.

Wen, D. and Nakhleh, L. 2018. Coestimating reticulate phylogenies and gene trees from multilocus sequence data. *Syst. Biol.*, 67(3): 439–457.

Wilkinson-Herbots, H. M. 2008. The distribution of the coalescence time and the number of pairwise nucleotide differences in the "isolation with migration" model. *Theor. Popul. Biol.*, 73: 277–288.

Wright, S. 1931. Evolution in Mendelian populations. *Genetics*, 16: 97–159.

Wu, D.-D., Ding, X.-D., Wang, S., Wojcik, J. M., Zhang, Y., Tokarska, M., Li, Y., Wang, M.-S., Faruque, O., Nielsen, R., Zhang, Q., and Zhang, Y.-P. 2018. Pervasive introgression facilitated domestication and adaptation in the bos species complex. *Nature Ecol. Evol.*, 2(7): 1139–1145.

Xu, B. and Yang, Z. 2016. Challenges in species tree estimation under the multispecies coalescent model. *Genetics*, 204: 1353–1368.

Yang, Z. and Rannala, B. 2017. Bayesian species identification under the multispecies coalescent provides significant improvements to DNA barcoding analyses. *Mol. Ecol.*, 26: 3028–3036.

Yu, Y., Dong, J., Liu, K. J., and Nakhleh, L. 2014. Maximum likelihood inference of reticulate evolutionary histories. *Proc. Natl. Acad. Sci. U.S.A.*, 111(46): 16448–16453.

Zachos, F. E. 2016. *Species Concepts in Biology*. Springer, New York.

Zhang, C., Ogilvie, H. A., Drummond, A. J., and Stadler, T. 2018. Bayesian inference of species networks from multilocus sequence data. *Mol. Biol. Evol.*, 35: 504–517.

Zhu, J., Yu, Y., and Nakhleh, L. 2016. In the light of deep coalescence: revisiting trees within networks. *BMC Bioinformatics*, 17: 415.

Zhu, S. and Degnan, J. H. 2017. Displayed trees do not determine distinguishability under the network multispecies coalescent. *Syst. Biol.*, 66(2): 283–298.

Zhu, T. and Yang, Z. 2012. Maximum likelihood implementation of an isolation-with-migration model with three species for testing speciation with gene flow. *Mol. Biol. Evol.*, 29: 3131–3142.

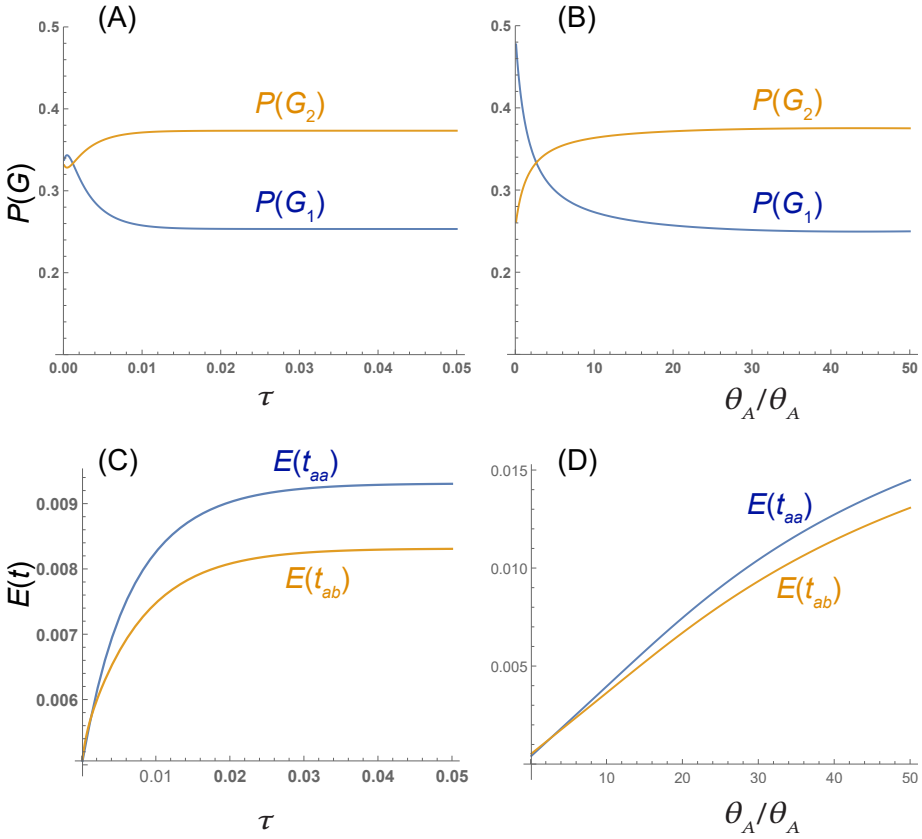

Figure S1: Characterization of the anomaly zone under the IM model. (A, B) $\mathbb{P}(G_1)$ (blue) and $\mathbb{P}(G_2)$ (orange) plotted against $\tau = \tau_R$ and θ_A/θ_B (with θ_B fixed), respectively. The anomaly $\mathbb{P}(G_1) < \mathbb{P}(G_2)$ occurs when $\tau > 0.00119461$ or when $\theta_A/\theta_B > 2.66667$, respectively. (C, D) $\mathbb{E}(t_{aa})$ (blue) and $\mathbb{E}(t_{ab})$ (orange) plotted against $\tau = \tau_R$ and θ_A/θ_B (with θ_B fixed), respectively. $\mathbb{E}(t_{aa}) > \mathbb{E}(t_{ab})$ if $\tau > 0.00123447$ or if $\theta_A/\theta_B > 2.66667$, respectively. Parameters that are not on the x-axis are fixed, at $\theta_A = 0.025$, $\theta_B = 0.001$, $\theta_R = 0.01$, $\tau_R = 0.02$ and $M = 0.8$.

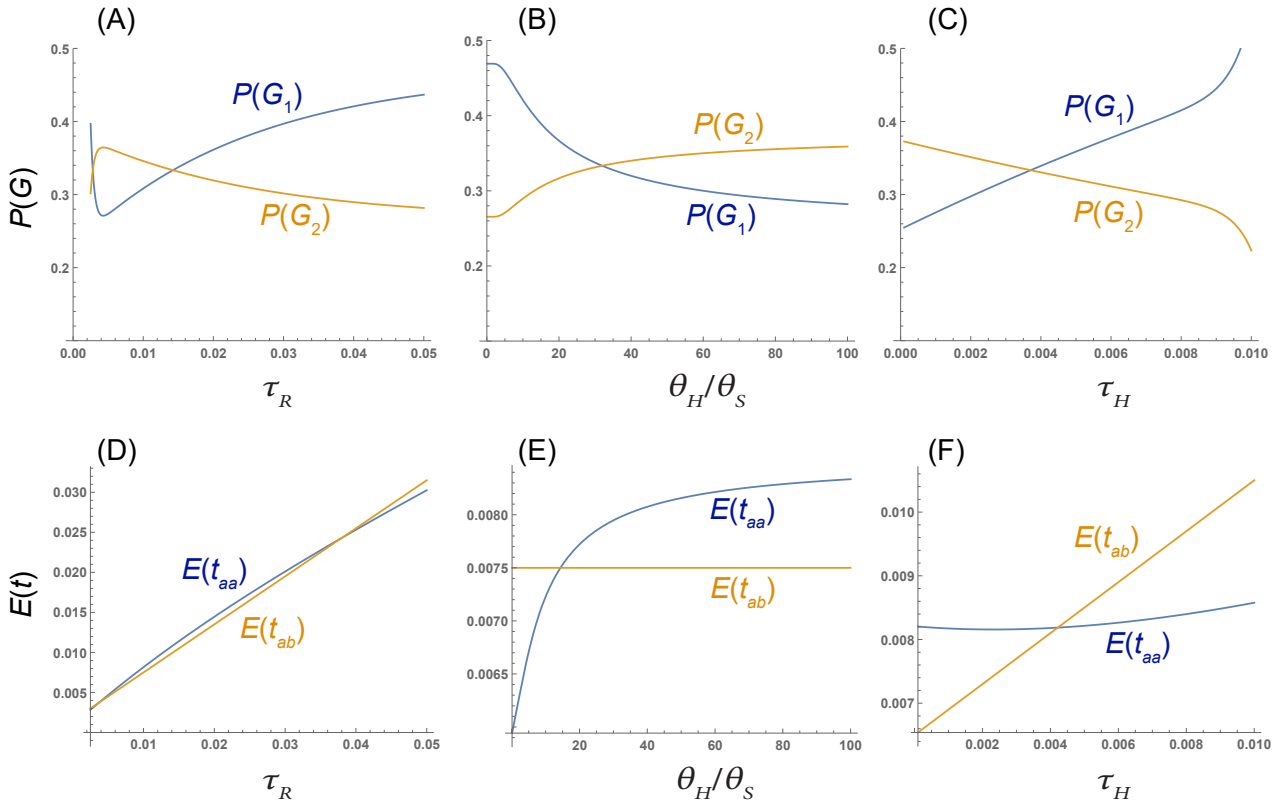

Figure S2: Characterization of the anomaly zone in the MSci model. (A, B, C) $\mathbb{P}(G_1)$ (blue) and $\mathbb{P}(G_2)$ (orange) plotted against τ_R, θ_H/θ_S (with θ_S fixed), and τ_H, respectively. The anomaly $\mathbb{P}(G_1) < \mathbb{P}(G_2)$ occurs if $0.00280476 < \tau_R < 0.0142311$, if $\theta_H/\theta_S > 31.9663$, or if $\tau_H < 0.00370836$. (D, E, F) $\mathbb{E}(t_{aa})$ (blue) and $\mathbb{E}(t_{ab})$ (orange) plotted against τ_R, θ_H/θ_S, and τ_H, respectively. The anomaly $\mathbb{E}(t_{aa}) > \mathbb{E}(t_{ab})$ occurs if $0.00365181 < \tau_R < 0.0380174$, if $\theta_H/\theta_S > 14.3751$, or if $\tau_H < 0.00421649$. Parameters that are not on the x-axis are fixed, at $\theta_A = \theta_H = 0.05$, $\theta_S = \theta_R = 0.001$, $\tau_R = 0.01$, $\tau_H = 0.0025$ and $\varphi = 0.4$.