RESEARCH ARTICLE

Determining the probability of hemiplasy in the presence of incomplete lineage sorting and introgression

Mark S Hibbins[1]*, Matthew JS Gibson[1], Matthew W Hahn[1,2]

[1]Department of Biology, Indiana University, Bloomington, United States;
[2]Department of Computer Science, Indiana University, Bloomington, United States

Abstract The incongruence of character states with phylogenetic relationships is often interpreted as evidence of convergent evolution. However, trait evolution along discordant gene trees can also generate these incongruences – a phenomenon known as hemiplasy. Classic comparative methods do not account for discordance, resulting in incorrect inferences about the number, timing, and direction of trait transitions. Biological sources of discordance include incomplete lineage sorting (ILS) and introgression, but only ILS has received theoretical consideration in the context of hemiplasy. Here, we present a model that shows introgression makes hemiplasy more likely, such that methods that account for ILS alone will be conservative. We also present a method and software (*HeIST*) for making statistical inferences about the probability of hemiplasy and homoplasy in large datasets that contain both ILS and introgression. We apply our methods to two empirical datasets, finding that hemiplasy is likely to contribute to the observed trait incongruences in both.

***For correspondence:** mhibbins@iu.edu

Competing interests: The authors declare that no competing interests exist.

Funding: See page 21

Received: 06 October 2020
Accepted: 18 December 2020
Published: 21 December 2020

Reviewing editor: Antonis Rokas, Vanderbilt University, United States

Introduction

Convergent evolution of the same phenotype in distantly related species provides some of the most compelling evidence for natural selection. Comparative inferences of convergence require that the species history is known (***Felsenstein, 1985***). Comparative methods applied to such histories implicitly assume that the loci underlying convergent traits also follow the species tree. However, gene trees at individual loci can disagree with each other and with the species tree, a phenomenon known as gene tree discordance. While genomic data allow us to overcome many technical sources of discordance (***Delsuc et al., 2005***; ***Dunn et al., 2008***; ***Misof et al., 2014***), discordance also has biological causes (***Degnan and Rosenberg, 2009***), and remains a common feature of phylogenomic datasets (***Pollard et al., 2006***; ***Fontaine et al., 2015***; ***Pease et al., 2016***; ***Novikova et al., 2016***; ***Wu et al., 2018***; ***Vanderpool et al., 2020***).

Gene tree discordance can have multiple sources, including biological causes such as incomplete lineage sorting (ILS), introgression, and horizontal gene transfer, and technical causes such as hidden paralogy or errors in gene tree inference (***Schrempf and Szöllősi, 2020***). Here, we focus primarily on the first two biological causes: ILS and introgression. Looking backwards in time, ILS is the failure of lineages to coalesce within a population before reaching the next most recent ancestral population. The probability of discordance due to ILS is a classic result of the multispecies coalescent, and depends on the population size and the length of time in which coalescence can occur (***Hudson, 1983***; ***Pamilo and Nei, 1988***). More recently, the classic multispecies coalescent model has been extended to include introgression (a term we use to encompass hybridization and subsequent gene flow), in a framework called the 'multispecies network coalescent' (***Yu et al., 2012***; ***Yu et al., 2014***; ***Wen et al., 2016***). In this model, species relationships are modeled as a network, with introgression represented by horizontal reticulation edges. Individual loci probabilistically follow or do

not follow the reticulation edge, after which they sort according to the multispecies coalescent process (i.e. with ILS). A major advantage of this approach is that ILS and introgression can be modeled simultaneously (reviewed in **Degnan, 2018**), allowing for more detailed study of the consequences of discordance.

Importantly, discordant gene trees can lead to the appearance of apparently convergent traits. This is because discordant gene trees have internal branches that do not exist in the species tree. If a mutation occurs along such a branch at a locus controlling trait variation, it may produce a pattern of character states that is incongruent with the species tree. Incongruent trait patterns are the basis for inferences of convergent evolution ('homoplasy'), and thus this phenomenon has become known as hemiplasy (**Avise and Robinson, 2008**). Since hemiplasy can produce the same kinds of trait incongruence as homoplasy, failing to account for gene tree discordance can generate misleading inferences about convergence (**Mendes and Hahn, 2016**; **Mendes et al., 2016**). Studies in systems with widespread discordance have found that hemiplasy is a likely explanation for many patterns of incongruence (**Copetti et al., 2017**; **Wu et al., 2018**; **Guerrero and Hahn, 2018**).

The problem of hemiplasy makes it clear that robust inferences about the evolution of traits must account for gene tree discordance (**Hahn and Nakhleh, 2016**). Recent work has provided expressions for the probabilities of hemiplasy and homoplasy (**Guerrero and Hahn, 2018**), allowing for an assessment of whether a single transition (hemiplasy) or two transitions (homoplasy) is more likely to explain trait incongruence. This model shows that the most important factors contributing to a high risk of hemiplasy relative to homoplasy are a short internal branch on the species tree (which increases the rate of gene tree discordance), and a low mutation rate (which reduces the probability of the multiple independent transitions needed for homoplasy). However, applying this model in present form to empirical phylogenetic data faces two major limitations. First, incomplete lineage sorting is the only source of gene tree discordance considered, excluding scenarios with gene flow. Second, the model is limited to evolution along a three-taxon tree, restricting calculations for the exact probability of hemiplasy in larger clades.

With genomic data now available for many species, it has become clear that introgression is a common phenomenon (**Mallet et al., 2016**). Introgression leads to different patterns of gene tree discordance than expected under ILS alone – specifically, minority gene tree topologies supporting a history of introgression are expected to become more common than those produced via ILS alone. These differences form the conceptual basis for common tests of introgression using genomic data (**Reich et al., 2009**; **Green et al., 2010**; **Durand et al., 2011**; **Patterson et al., 2012**; **Pease and Hahn, 2015**). Introgression also affects the expected coalescence times between pairs of species (**Joly et al., 2009**; **Brandvain et al., 2014**; **Hibbins and Hahn, 2019**; **Hahn and Hibbins, 2019**). Pairs of species that have exchanged genes will have lower levels of sequence divergence, and therefore longer shared internal branches, at introgressed loci than expected under ILS alone. These differences in the frequency and branch lengths of genealogies produced by introgression should meaningfully affect the probability of hemiplasy. Therefore, it is important that both sources of gene tree discordance be accounted for in models of trait evolution.

For trees with more than three taxa, the number of possible gene trees and mutational configurations that could explain a particular pattern of trait incongruence increases dramatically. To illustrate this problem, we consider two cases of empirical incongruence of a binary trait. First, consider the case of New Guinea lizards that have evolved green blood from a red-blooded ancestor (**Figure 1A**; **Rodriguez et al., 2018**). A clade of 15 taxa contains both the green-blooded species and red-blooded species (the ancestral state). Given the phylogenetic distribution of the six green-blooded species—and no consideration of gene tree discordance—four independent transitions are necessary to explain this incongruence (**Figure 1**). However, the internal branches on this tree are short and discordance is likely. Individual loci could therefore group the green-blooded taxa into as few as one and as many as six separate clades. Depending on the history at loci affecting blood color, the distribution of green-blooded taxa could therefore be explained by anywhere from one to six mutations, and even more if we consider back-mutations. The one-mutation case represents a single transition due to hemiplasy along a branch that does not exist in the species tree, while the two- and three-mutation cases represent a combination of hemiplasy and homoplasy. The problem becomes even more complex when introgression occurs in the phylogeny, because each reticulation event introduces a new set of gene trees formed from the coalescent process at introgressed loci (**Hibbins and Hahn, 2019**). One such example is the origin of a chromosomal inversion spanning a

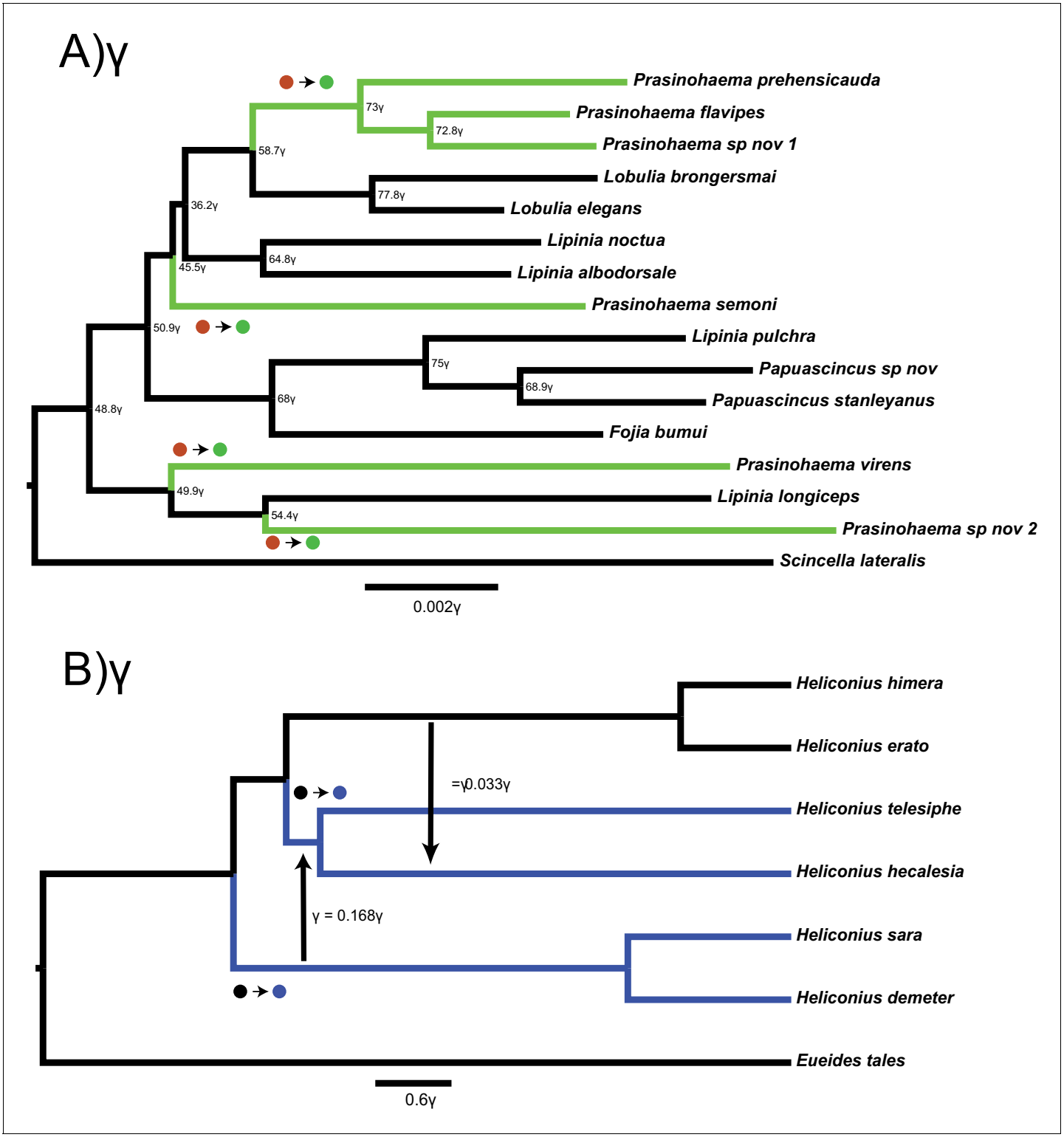

Figure 1. Two empirical examples of apparent convergence in character states that could potentially be explained by hemiplasy. (**A**) Maximum-likelihood species tree of the clade including green-blooded lizards and an outgroup, constructed from the concatenation of 3220 ultra-conserved elements (data from ***Rodriguez et al., 2018***). Branch lengths in substitutions per site; nodes labeled with site concordance factors. (**B**) Coalescent network of *Heliconius erato/sara* clade, processed from the network constructed for the clade in ***Edelman et al., 2019***. Branch lengths in units of 2*N* generations; rate, direction, and approximate timing of introgression events indicated by vertical arrows. In both trees, taxa with derived characters are colored, and the most parsimonious transitions from ancestral to derived states are labeled with circles.

Figure 1 continued on next page

Figure 1 continued

The online version of this article includes the following source data and figure supplement(s) for figure 1:

Source data 1. Input file given to *HeIST* for the lizard analysis, using the tree and character states shown in ***Figure 1A***.
Source data 2. Input file given to *HeIST* for the butterfly analysis, using the tree, character states, and introgression events shown in ***Figure 1B***.
Figure supplement 1. Full 43-species maximum-likelihood lizard phylogeny constructed by *RAxML*, with branch lengths in units of substitutions per site.
Figure supplement 2. Phylogeny of green-blooded lizards and an outgroup inferred from 3220 UCE gene trees using *ASTRAL* (data from ***Rodriguez et al., 2018***).

gene involved in wing coloration in the *Heliconius erato/sara* clade of butterflies (***Figure 1B***; ***Edelman et al., 2019***). Overall, the huge number of possible gene trees (>213 trillion for 15 lizard species; ***Felsenstein, 2004***) and the large number of possible mutational events on these trees makes it infeasible to derive an explicit mathematical solution to address questions about hemiplasy in many empirical systems.

Here, we make two steps toward addressing these problems. First, we derive expressions for the probabilities of hemiplasy and homoplasy under the multispecies network coalescent for three taxa. Our results show that hemiplasy becomes increasingly likely relative to homoplasy as introgression occurs at a higher rate and at a more recent time relative to speciation. We also show how this pattern is influenced by the direction of introgression. These results highlight the need to account for both ILS and introgression in order to understand the origins of a trait incongruence. Second, we present a tool called *HeIST* (*He*miplasy *I*nference *S*imulation *T*ool) that uses coalescent simulation to dissect patterns of hemiplasy and homoplasy in larger phylogenies. This tool provides an estimate of the most likely number of transitions giving rise to observed incongruence of binary traits, and accounts for both ILS and introgression. Lastly, we apply *HeIST* to two empirical cases of apparent convergence in a binary trait, finding that hemiplasy is likely to contribute to the observed trait incongruences.

Results

A model for the probability of hemiplasy under the multispecies network coalescent

To study the effects of introgression on the probability of hemiplasy, we combine concepts from two previously published models: the 'parent tree' framework of ***Hibbins and Hahn, 2019***, and the model of binary-trait evolution presented in ***Guerrero and Hahn, 2018*** (see ***Wang et al., 2020*** for an alternative way to extend the model to incorporate introgression). Consider a rooted three-taxon tree with the topology ((A,B), C). We define t_1 as the time of speciation between lineages A and B in units of $2N$ generations, and t_2 as the time of speciation between C and the ancestor of A and B. We also imagine an instantaneous introgression event between species B and C at time t_m, which can be in either direction (C → B or B → C). We define the total probability of a locus following an introgressed history as δ, with δ_2 denoting the probability of C → B introgression, and δ_3 the probability of B → C introgression. Introgression in both directions at an individual locus is not allowed in our model. However, a single introgression event in both directions can be modeled by allowing different directions at different loci. The history described here is represented by the phylogenetic network shown in ***Figure 2*** (top). Other introgression scenarios can be accommodated by our model (see Discussion), but will not be considered here.

To make it easier to track the history of different gene trees, we imagine that a phylogenetic network can be split into a set of 'parent trees' which describe the history at individual loci (***Meng and Kubatko, 2009***; ***Liu et al., 2014***; ***Hibbins and Hahn, 2019***; ***Figure 2***, bottom). Within each of these parent trees, which describe either the species history or the history of introgression, gene trees sort under the multispecies coalescent process. Loci follow the species history, referred to as parent tree 1, with probability $1 - (\delta_2 + \delta_3)$. With C → B introgression, some loci will follow the alternative history within parent tree 2, with probability δ_2. In parent tree 2, species B and C are sister and share a 'speciation' time of t_m. B → C introgression causes loci to follow parent tree 3 with probability δ_3; in this history, lineages B and C are sister and split at time t_m, while the split time of A and the ancestor of

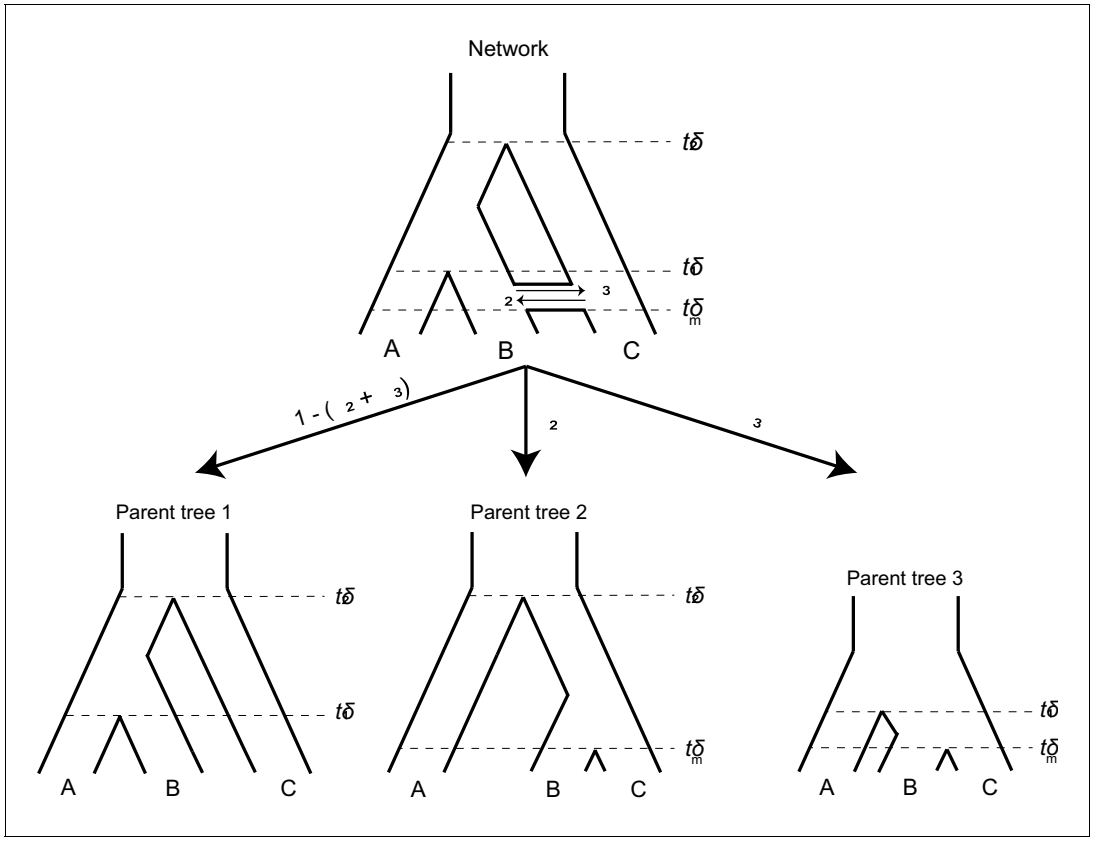

Figure 2. A phylogenetic network (top) can be split into a set of parent trees (bottom) representing the possible histories at individual loci. The probability that a locus is described by a particular parent tree depends on the probability of introgression (arrow labels). The horizontal 'tube' shown in the phylogenetic network does not depict introgression over a continuous time interval, but rather shows the timing of introgression (t_m) in an instantaneous pulse, while allowing for coalescence to be visualized for loci that follow a history of introgression.
The online version of this article includes the following figure supplement(s) for figure 2:

Figure supplement 1. Each parent tree in our model generates four gene trees: one generated from lineage sorting (Panels A and E), and three equally likely trees generated from incomplete lineage sorting (panels B-D, F–H).

BC is reduced t_1 to This reduction in the second split time in parent tree 3 occurs becaus ence of loci from lineage B in lineage C allows C to trace its ancestry through B g Since B is more closely related to A than C this allows C to coalesce wi*Fig- ure 2*). Each introgression event is modeled as a discrete and instantaneous pulse that its own parent tree and in our model we consider a single introgression event for sim ever multiple events or introgression over a continuous time interval can be modeled by multiple pulses with different directions timings or probabilities. Each such event introduces parent tree and set of gene trees.

Each parent tree can produce four gene trees under the multispecies coalescent pro tree from lineage sorting and three equally probable trees from incomplete lineage sorting *ure 2—figure supplement 1*). In other words introgression always involves ILS as these a mutually exclusive histories. Each of these possible gene trees has five branches along tions can occur three tip branches an internal branch and an ancestral branch. A subs possible gene trees within each parent tree can lead to hemiplasy while homoplasy can any gene tree (*Figure 3*). *Guerrero and Hahn, 2018* provide exact expectations for the probability of a mutation on each branch of each genealogy in an ILS-only model. Before extendin work to incorporate introgression the ILS-only model will be briefly described here using updated notation that will make it easier to include the effects of introgression.

Consider a binary trait that is incongruent with the described species tree where specie have the derived state and A has the ancestral state. λ_1 λ_2 and λ_3 denote the tip branches in

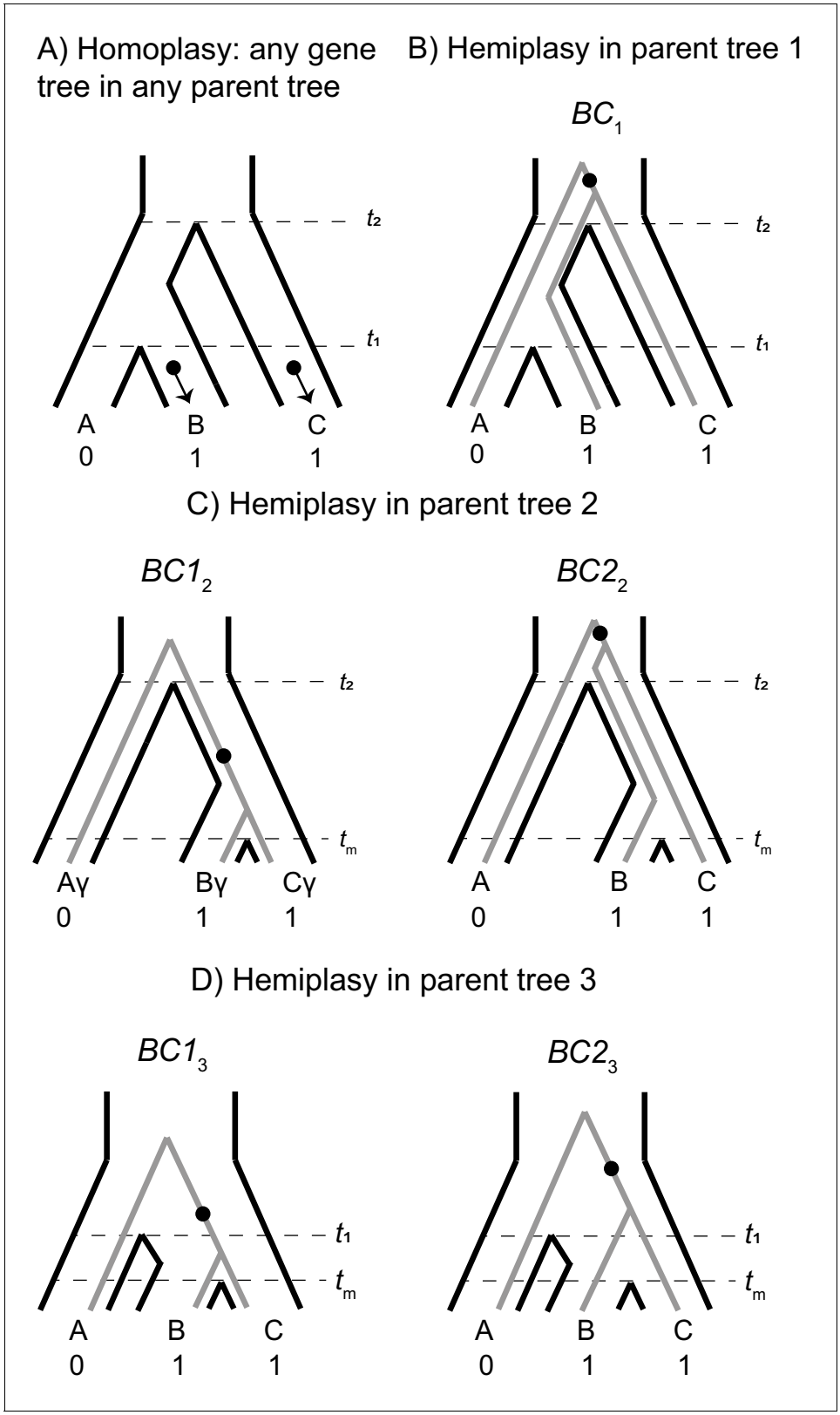

Figure 3. The possible paths to homoplasy and hemiplasy under the multispecies network coalescent. Homoplasy can happen on any gene tree, as long as there are two independent mutations on tip branches (panel **A**). Homoplasy can also happen via a mutation in the ancestor of all three species, followed by a reversal (not shown). All cases of hemiplasy require a transition on the internal branch of a gene tree with the topology ((B,C),A). In parent tree 1 (panel **B**), only one such possible gene tree exists (shown in gray; BC_1). In both parent trees 2 and 3 (panels **C** and **D** respectively), there are two

Figure 3 continued on next page

Figure 3 continued

possible gene trees with this topology. These gene trees differ in internal branch lengths, depending on the parent tree of origin and whether the tree is the result of lineage sorting ($BC1_2$ and $BC1_3$) or incomplete lineage sorting ($BC2_2$ and $BC2_3$) within introgressed histories.

any topology leading to species A, B, and C respectively; λ_4 denotes the internal branch of any topology, and λ_5 the branch subtending the root. The notation $\nu(\lambda, \tau)$ represents the probability of a mutation on branch λ_i in genealogy τ, where τ represents any of the four gene trees from any of the three parent trees. The rates of 0 → 1 and 1 → 0 mutations are assumed to be equal, and the rate among lineages is assumed to be constant. Finally, to describe individual genealogies, we use the notation $XY_{i=1,2,3}$, where X and Y denote the sister taxa, and the subscript i denotes the parent-tree of origin. In cases where a tree topology can be produced by either lineage sorting or ILS, a non-subscripted 1 or 2 is used, respectively. Under the ILS-only model, hemiplasy can only occur through a substitution on branch λ_4 of genealogy BC_1 (***Figure 2—figure supplement 1C***, ***Figure 3B***). This occurs with the following probability:

$$P_e[BC_1 = \left(\frac{1}{3}e^{-(t_2-t_1)}\right)\nu(\lambda_4, BC_1)\prod_{i\neq 4}(1-\nu(\lambda_i, BC_1)) \quad (1)$$

(***Guerrero and Hahn, 2018***). ***Equation 1*** has three components: the probability of observing genealogy BC_1, the probability that a mutation happens on the internal branch of that genealogy, and the probability that no other mutations occur. See section 1 of the Appendix for the full expressions for each mutation probability.

Now consider the phylogenetic network described earlier and shown in ***Figure 2***. At an introgressed locus, the parent tree topology is ((B,C), A), but could be either parent tree 2 or 3. Within each of these parent trees, there are two possible gene trees that share this topology: one produced by lineage sorting (***Figure 2—figure supplement 1E***, ***Figure 3C***) and one produced by ILS where B and C are still the first to coalesce (***Figure 2—figure supplement 1F***, ***Figure 3C***). While these trees have the same topology, their expected frequencies and internal branch lengths differ. These quantities also differ depending on the direction of introgression at the locus, that is whether the history follows parent tree 2 or 3.

We first consider the C → B direction of introgression, and genealogy $BC1_2$, which is the result of lineage sorting within parent tree 2. This gives:

$$P_e[BC1_2 = 1 - e^{-(t_2-t_m)}\ \nu(\lambda_4, BC1_2)\prod_{i\neq 4}(1-\nu(\lambda_i, BC1_2)) \quad (2)$$

While ***Equation 2*** has the same three core components as ***Equation 1***, there are several important differences. First, the gene tree probability is the probability of lineage sorting within parent tree 2, which differs from the probability of ILS within parent tree 1. Second, the lower bound of coalescence is t_m rather than t_1, resulting in a higher probability of lineage sorting in parent tree 2 as compared to parent tree 1. Third, because B and C coalesce more quickly in this tree, they share a longer internal branch, which means the probability of mutation on that branch is higher (see section 1 of the Appendix).

ILS within parent tree two produces gene tree $BC2_2$, in which B and C are the first to coalesce in the common ancestor of all three species. The probability of hemiplasy in this case is:

$$P_e[BC2_2 = \left(\frac{1}{3}e^{-(t_2-t_m)}\right)\nu(\lambda_4, BC2_2)\prod_{i\neq 4}(1-\nu(\lambda_i, BC2_2)) \quad (3)$$

In ***Equation 3***, the gene tree probability represents ILS in parent tree 2. This probability is lower than its equivalent in parent tree 1, again because t_m is the lower bound for coalescence. Since the upper bound to coalescence is the same (t_2), the probability of a mutation on the internal branch of this gene tree is the same as for BC_1 (the ILS topology within parent tree 1). To get the overall probability of hemiplasy due to both ILS and introgression when there is gene flow from C → B, we weight the probability from each gene tree (***Equations 1-3)*** by the admixture proportion, giving the following:

$$P_e[ILS,\ \mathrm{C} \rightarrow \mathrm{B} = (1 - \delta_2)P_e[BC_1 + \delta_2(P_e[BC1_2 + P_e[BC2_2) \quad (4)$$

From ***Equation 4***, we can see that introgression will increase the probability of hemiplasy over ILS alone (***Equation 1***) whenever the probability of hemiplasy from parent tree two is higher than from parent tree 1 (i.e. $P_e[BC1_2 + P_e[BC2_2 > P_e[BC_1)$. This is true whenever $t_2 > t_m$ (see section 2 of the Appendix), which is by definition always true in this model.

Finally, we consider the probability of hemiplasy when introgression is in the direction B → C (represented by admixture fraction δ_3). As mentioned previously, this direction of introgression results in an upper bound to coalescence of t_1 rather than t_2. This is the primary difference between the directions of introgression, affecting both the expected gene tree frequencies and mutation probabilities (compare to ***Equations 2 and 3***):

$$P_e[BC1_3 = 1 - e^{-(t_1 - t_m)}\ \nu(\lambda_4, BC1_3)\prod_{i \neq 4}(1 - \nu(\lambda_i, BC1_3)) \quad (5)$$

and

$$P_e[BC2_3 = \left(\frac{1}{3}e^{-(t_1 - t_m)}\right)\nu(\lambda_4, BC2_3)\prod_{i \neq 4}(1 - \nu(\lambda_i, BC2_3)) \quad (6)$$

For the general probability of hemiplasy, including both directions of introgression, we now have:

$$\begin{aligned} P_e[ILS,\ \mathrm{C} \rightarrow \mathrm{B},\ \mathrm{B} \rightarrow \mathrm{C} = {} & (1 - (\delta_2 + \delta_3))P_e[BC_1 + \delta_2(P_e[BC1_2 + P_e[BC2_2) + \\ & \delta_3(P_e[BC1_3 + P_e[BC2_3) \end{aligned} \quad (7)$$

Finally, we consider the probability of homoplasy. As described in ***Guerrero and Hahn, 2018***, there are two possible paths to homoplasy for a three-taxon tree where taxa B and C carry the derived state. The first is parallel 0 → one mutations on branches λ_2 and λ_3 (***Figure 3A***), and the second is a 0 → one mutation on branch λ_5 followed by a 1 → 0 reversal on branch λ_1. Both these paths to homoplasy can happen on any possible genealogy, because every topology contains independent tip branches leading to species B and C, as well as an internal branch ancestral to all three species. This gives the following:

$$P_o = \sum_{\tau} p(\tau)\left[\nu(\lambda_2, \tau)\nu(\lambda_3, \tau)\prod_{i \neq 2,3}(1 - \nu(\lambda_i, \tau)) + \nu(\lambda_5, \tau)\prod_{i \neq 1,5}(1 - \nu(\lambda_i, \tau))\right] \quad (8)$$

where τ denotes the set of all possible gene trees. (Note that the sum inside ***Equation 8*** is multiplied by $\frac{1}{p(\tau)}$ in the main text of ***Guerrero and Hahn, 2018***. This is a typo in that paper, but the results presented from their model use the correct expression, $p(\tau)$.) This formulation can also be applied to the extended model with introgression, with the understanding that τ now also includes the gene trees produced by parent trees 2 and 3. Each gene tree used in this summation will have a different set of mutation probabilities, which are detailed in section 1 of the Appendix.

To understand the analytical effect of introgression on the relative risks of hemiplasy and homoplasy, we plotted the ratio P_e/P_o over a realistic range of admixture proportions, timings, and directions (***Figure 4***). The values of t_1 and t_2 were held constant at 1 and 3.5 coalescent units, respectively, with a population-scaled mutation rate of $\theta = 0.002$. These settings ensured a constant contribution of incomplete lineage sorting to the risk of hemiplasy, leading to a baseline ratio of hemiplasy to homoplasy, P_e/P_o, of 0.818 with no introgression. We varied the admixture proportion from 0 to 10%, and the value of t_m from 0.99 (just after the most recent speciation) to 0.01, for three different direction conditions: C → B only, B → C only, and equal rates in both directions.

Introgression makes hemiplasy more likely than incomplete lineage sorting alone

Using our model for the probability of hemiplasy and of homoplasy, we examined the ratio P_e/P_o over a range of different introgression scenarios. This ratio summarizes how much more probable hemiplasy is than homoplasy for a given area of parameter space; for example, a value of $P_e/P_o = 2$ means hemiplasy is twice as likely as homoplasy. We find that the probability of hemiplasy relative to

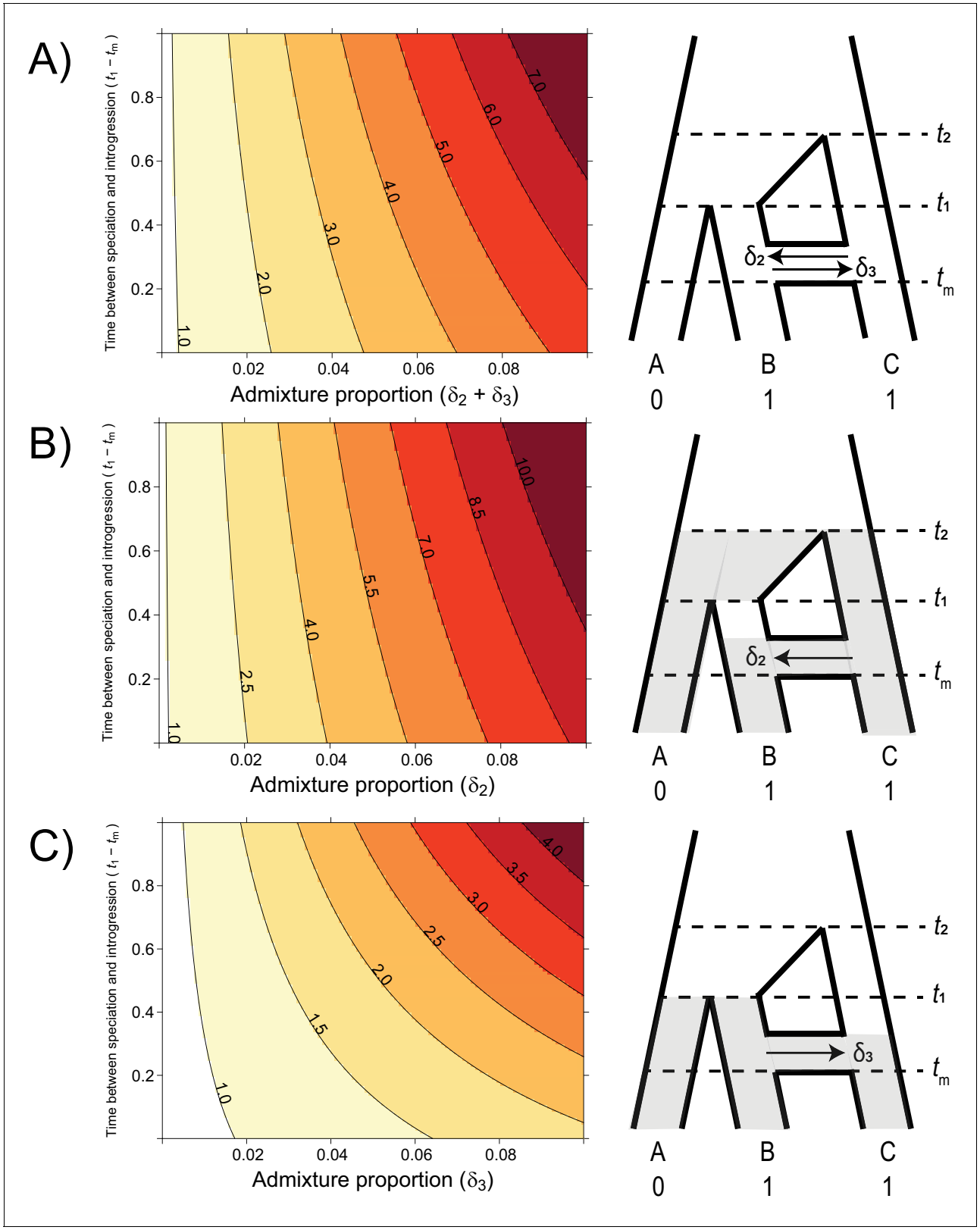

Figure 4. The probability of hemiplasy relative to homoplasy (contours) as a function of the admixture proportion (x-axis), the time between speciation and introgression (y-axis), and the direction of introgression (panels). The contours delineate the factor difference between hemiplasy and homoplasy; for instance, a contour value of 2.0 means hemiplasy is twice as probable as homoplasy in that area of parameter space. At x = 0 in each panel, P_e/

Figure 4 continued on next page

Figure 4 continued

P_o = 0.818. (**A**) Equal rates of introgression in both directions. (**B**) Introgression in only the C → B direction. (**C**) Introgression in only the B → C direction.

The online version of this article includes the following source data for figure 4:

Source data 1. Data used to generate the contour plots in ***Figure 4***.

homoplasy increases as a function of the admixture proportion and how recently introgression occurs relative to speciation (***Figure 4***). As mentioned in the Introduction, there are several possible reasons for these observed trends. The strongest effect on P_e/P_o comes from the admixture proportion: a higher proportion means more loci evolving under parent trees 2 and 3, which means higher frequencies of the genealogies where hemiplasy is possible (i.e. $BC1_2$, $BC2_2$, $BC1_3$,$BC2_3$). The range of simulated admixture proportions from 0 to 10% was meant to capture a biologically plausible range of values, although rates of introgression can sometimes be much higher than this (e.g. ***Fontaine et al., 2015***). Even in this modest range, the effect on the probability of hemiplasy can be substantial. We found that an admixture proportion of 5% results in hemiplasy being anywhere from 1.5 to 4 times more likely than homoplasy (depending on the timing and direction of introgression; ***Figure 4***). Given the baseline value of P_e/P_o with no introgression for our chosen parameters (0.818), this represents at minimum a doubling of the probability of hemiplasy relative to homoplasy.

The effect of the timing of introgression is more complicated, as it manifests in multiple ways. First, more recent introgression increases the values of $t_2 - t_m$ and $t_1 - t_m$, which in turn increases the degree of lineage sorting in parent trees 2 and 3, respectively. This leads to a higher frequency of gene trees where hemiplasy is possible. Second, the expected length of the internal branches in these two genealogies increases as introgression becomes more recent, which leads to a higher probability of mutations occurring on these branches. Third, since the total height of each tree is being held constant, more recent introgression reduces the lengths of the tip branches leading to species B and C. This reduces the probability of homoplasy due to parallel substitutions, again making hemiplasy relatively more likely. Finally, the strength of the effect of the timing of introgression increases with the admixture proportion, since it is a property of introgressed loci; in other words, the values of of $t_2 - t_m$ and $t_1 - t_m$ do not matter unless loci follow a history of introgression.

The direction of introgression affects the relationship between the admixture proportion, the timing of introgression, and hemiplasy risk (***Figure 4B and C***). While hemiplasy becomes more likely than homoplasy with increased admixture in either direction, P_e/P_o is lower in any given part of parameter space for B → C introgression (***Figure 4C***). This is because the bounds of coalescence for parent tree 3 are t_1 and t_m, which are always closer in time than t_2 and t_m (***Figure 2***). The smaller internal branch in parent tree 3 leads to a higher rate of ILS, in addition to a shorter internal gene tree branch (and lower mutation probability) on genealogies that undergo lineage sorting in these histories. Finally, the timing of introgression has a stronger effect on P_e/P_o in the B → C direction (***Figure 4C***). This is likely because parent tree 3 is truncated relative to parent tree 2 (see ***Figure 2***), and so the difference $t_1 - t_m$ makes up a proportionally larger part of the tree height.

HeIST: *He*miplasy *I*nference *S*imulation *T*ool

As described above, it is possible to infer the most likely number of transitions for an incongruent trait while accounting for discordance in a rooted tree with three taxa. However, similar calculations are computationally difficult for larger numbers of taxa. Here, we present a tool built on top of the coalescent simulator *ms* (***Hudson, 2002***) and sequence simulator *Seq-Gen* (***Rambaut and Grassly, 1997***) that provides an intuitive way to interrogate the parameter space of larger trees. Our tool, dubbed *HeIST*, takes a phylogenetic tree (including an option to specify introgression events) with observed character states as input and returns a simulated distribution of the number of transitions necessary to explain those character states. Introgression events must be specified as an instantaneous 'pulse' from one lineage to another, but we allow flexibility with respect to the timing of that pulse, as well as the rate, direction, and the lineages involved. The input phylogeny must be in coalescent units, but we also include a tool for converting trees given in units of substitutions per site to coalescent units, as long as branches are also associated with concordance factors (see section entitled 'Inferring the tip branch lengths of a phylogeny in coalescent units' below).

HeIST uses *ms* to simulate a large number of gene trees from the specified species tree or species network, and then simulates the evolution of a single nucleotide site along each of these gene trees using *Seq-Gen*. Loci where the simulated nucleotide states (transformed into 0/1 characters representing ancestral and derived states) match the character states observed on the species tree are taken as replicate simulations of the evolution of the trait being studied. In these 'focal' cases, *HeIST* counts the number of mutations that occurred along the gene tree in each simulation. It also returns information on the frequency of tip vs. internal branch mutations, transition vs. reversal mutations, the distribution of gene tree topologies, and whether gene trees originate from the species branching history or introgression history. Finally, it returns a summary of how much hemiplasy is likely to contribute to observed character states, using Fitch parsimony (***Fitch, 1971***) to obtain a homoplasy-only baseline for comparison. *HeIST* is implemented in Python 3 and the package/source code are freely available from https://github.com/mhibbins/HeIST.

HeIST effectively captures the effects of ILS and introgression on hemiplasy risk

To evaluate the performance of *HeIST*, we simulated across nine conditions with increasing expected probabilities of hemiplasy, across five different trait mutation rates. The results, shown in ***Figure 5***, confirm the theoretical predictions shown in ***Figure 4***: the probability of hemiplasy increases as a function of decreasing internal branch length (ILS1-ILS3), increasing probability of introgression (INT1-INT3), and more recent introgression (INT4-INT6). The effect of the timing of introgression is weaker than the effect of the introgression rate, also in line with theoretical expectations. These results held true for both the probability conditional on observing the specified trait pattern (***Figure 5A***) and the raw probability (***Figure 5B***).

While the change in the probability of hemiplasy is broadly consistent with theoretical expectations, the probabilities estimated from *HeIST* consistently underestimated the exact values predicted from theory by a small amount (***Figure 5—figure supplement 2***). We suspect this is due to the occurrence of multiple hits on the same branch of a gene tree, which are not accounted for in our theoretical model. Reversals on branches where hemiplasy can occur would slightly reduce the number of observed hemiplasy cases, leading to the observed underestimation. Consistent with our hypothesis, the mean-normalized mean squared error between simulated and expected values is lower for both lower mutation rates and simulated conditions with a shorter internal branch (***Figure 5—figure supplement 3***). Overall, the mismatch between simulations and theory appears to be negligible for lower, more realistic trait mutation rates, so we do not believe this will be a concern for most empirical applications.

When the parameters of a phylogenetic network are estimated from empirical data, it is possible that many different parameter combinations may be equally likely, especially when only a subset of features are used to fit the model. However, these combinations may differentially affect the probability of hemiplasy: for instance, if the frequency of gene trees is used to fit the network model, but the length of gene tree branches is ignored. To investigate this, we applied *HeIST* to five simulated conditions in which the probability and recency of introgression were increased, while the frequency of the discordant gene tree that could cause hemiplasy was held constant (***Figure 5—figure supplement 4***). We found that, despite a constant gene tree probability, the conditional probability of hemiplasy increased in each successive condition as introgression became more recent and frequent (***Figure 5—figure supplement 5***). These results to some extent merely serve to reinforce the notion that introgression has an effect on hemiplasy above and beyond the effect of ILS alone: by lengthening the branch on the discordant tree that hemiplastic mutations can occur on, introgression has a larger effect than ILS alone. But even when network models that include introgression are used, the estimated effects on hemiplasy will be conservative if parameters are estimated using gene tree frequencies alone.

To evaluate the effects of using *HeIST* on real data, we compared results using the 'true' species tree (***Figure 5—figure supplement 6A***) to those obtained from estimating the species tree with branch lengths using simulated DNA data. This data was run through a pipeline involving estimating a phylogeny using *RAxML*, converting branch lengths to coalescent units, and smoothing (***Figure 5—figure supplement 6B and C***). In all cases, the tree was comprised of eight taxa with no introgression, with three incongruent taxa sharing a hypothetical derived character with a mutation rate of 0.05 per $2N$ generations. Regardless of whether the 'extend' or 'redistribute' method was used for

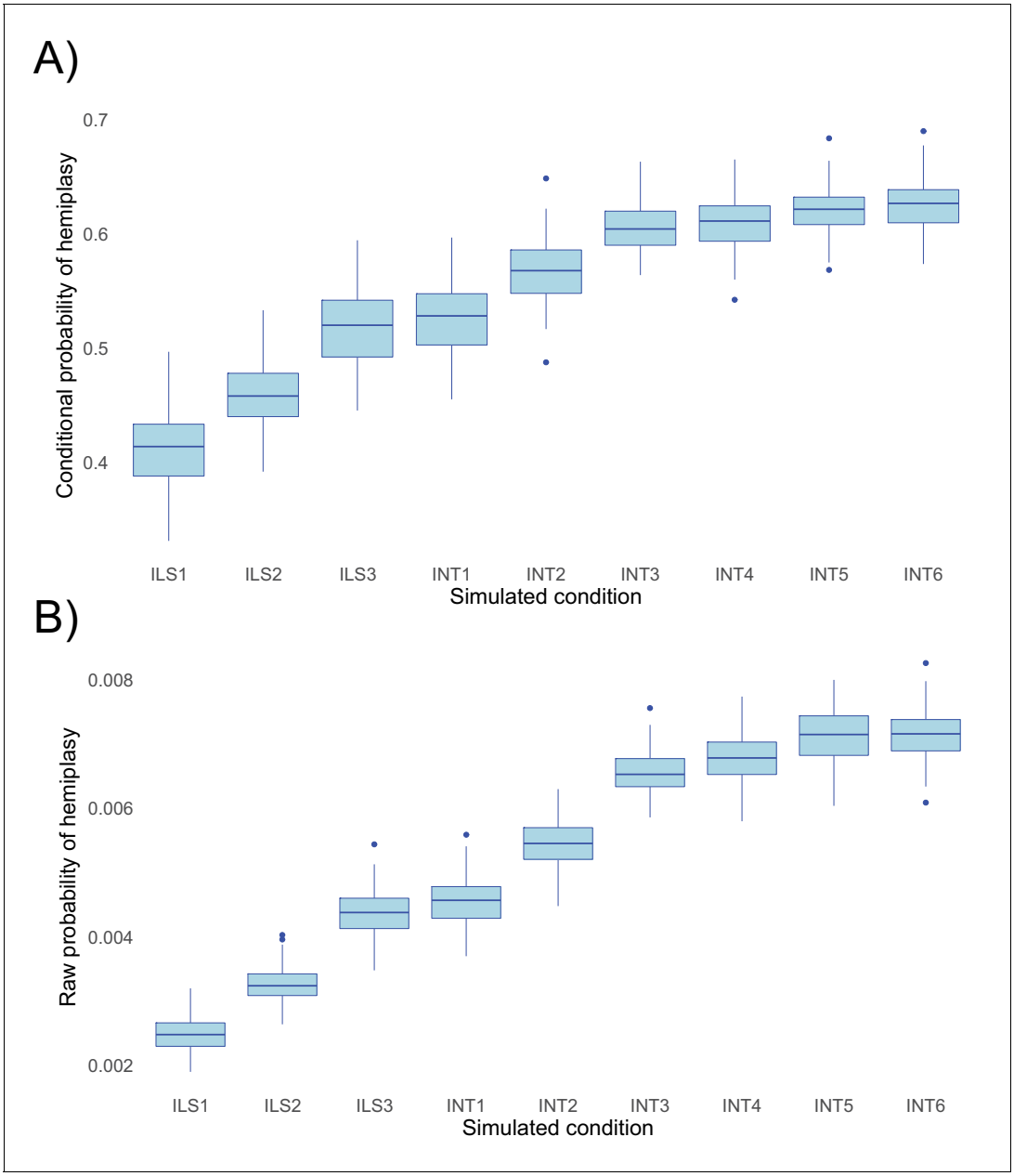

Figure 5. Probabilities of hemiplasy estimated from *HeIST* across nine simulated conditions. ILS1-ILS3 decrease the internal branch length of the species tree; INT1-INT3 introduce introgression between derived taxa with increasing probability; INT4-INT6 make introgression more recent while holding the probability constant. See ***Figure 5—figure supplement 1*** for the exact parameters used in each condition. Panel **A** shows the probability conditional on observing the trait pattern, whereas panel **B** shows the raw probability out of 100,000 simulations.

The online version of this article includes the following source data and figure supplement(s) for figure 5:

Source data 1. Data used to generate ***Figure 5***.

Figure supplement 1. Parameters used for benchmarking simulations in HeIST.

Figure supplement 2. Mismatch between simulated (blue boxes) and theoretical (red diamonds) raw probabilities of hemiplasy across simulated conditions, using a mutation rate per 2N generations of 0.05.

Figure supplement 3. Degree of mismatch between simulated and theoretical values of the raw probability of hemiplasy for our nine simulated conditions (colors) across five mutation rates (x-axis).

Figure supplement 4. Parameters used for simulations demonstrating trade-offs in introgression parameters in HeIST.

Figure supplement 5. Trade-offs of different network parameters in HeIST.

Figure supplement 6. Effect of phylogenetic inference, branch length unit conversion, and smoothing on estimated probabilities of hemiplasy in HeIST.

smoothing the overall effect of estimating the tree from sequence data was to lengthen b nal and tip branch lengths reducing the conditional probability of hemiplasy relative to true tree was used (see ***Figure 5—figure supplement 6*** for exact probabilities). These results sug- gest that when our unit-conversion approach and smoothing are applied to empirical data resulting probability estimates will be conservative with respect to the hypothesis of hemipla

The distribution of green-blooded New Guinea lizards is likely to have arisen from fewer than four transitions

e investigated the most likely number of transitions to green blood from a red-blooded in New inea iards of the genus *Prasinohaema* (***Rodriguez et al., 2018***). Phylogenies constructed using *RAxML* (***Figure 1***–***Figure 1—figure supplement 1***) and *ASTRAL* (***Figure 1—figure supplement 2***) recover the phylogeny published by ***Rodriguez et al., 2018*** including the placement of green-blooded species and also confirm the existence of very short internal branches. In this observation site concordance factors estimated from CEs indicate very high rates of dance in this clade with some approaching a star tree (i.e. all topologies having freuenc (***Figure 1***–***Supplementary file 1***). This strongly suggests that the apparent convergent evolution the green blood phenotype has been affected by hemiplasy.

e used *HeIST* with the 15-taxon subclade containing six green-blooded species to determ the most likely number of trait transitions. sing our branch length unit conversion tool we obtained a best fit line $y = 0.3038 + 157.03x$ with an adusted R^2 of 0.554 (***Figure 6—figure supplement 1A***). This formula was used to predict the tip branch lengths of the iard p coalescent units for input to *HeIST* (***Figure 6—figure supplement 2***). This analysis was repeated using two different outgroups which differed in their distance from the focal subclade. Th were essentially the same using both outgroups here we present probabilities using the c group *Scincella lateralis*. After simulating 10^{10} loci from the iard phylogeny using *HeIST* we obtained 2042 loci with a distribution of derived states that matched the empirical distri green-blooded species. It is important to note that this number is expected to be a ve portion of the total number of simulated loci. This occurs because it is necessary to sim tories randomly but we use only the ones that match the observed distribution. ue enormous space of possibilities the probability of any single trait distribution will be very cially with large numbers of taxa and high rates of trait incongruence.

ith four independent transitions reuired without discordance there are three possible scen ios that involve at least one hemiplastic transition (***Figure 6A***). The first is a hemiplasy-only scenario in which all green-blooded species are grouped into a single monophyletic clade in a ge a single transition in the ancestor of this clade explains the observed distribution (***Figure 6A***, left). Out of 2042 focal loci 726 (35.5) correspond to this hemiplasy-only case. In the secon green-blooded species may be grouped into one or two clades in a gene tree and t independent transitions at least one of which must involve a discordant ancestral (***Figure 6A***, mid-left). Since there are still multiple independent transitions this case repres combination of hemiplasy and homoplasy but exactly which mutations on which branches a plasy vs. homoplasy will depend on the gene tree topology. Of 2042 focal cases 1316 (spond to this scenario. In the third case the green-blooded species are grouped into three clades with three independent transitions at least one of which must be he (***Figure 6A***, mid-right). inally the green-blooded species may be grouped into four clades four independent transitions as in the species tree (***Figure 6A***, right). e observed no instances of the latter two cases out of 2042 focal loci. These results strongly support the conclusion hemiplasy the green-blooded phenotype arose from one or two independent transitions r than four.

In all 2042 simulated focal cases the gene trees on which mutations arose grouped blooded species as monophyletic regardless of the number of mutations that occurred on In addition almost all these monophyletic clades share the same structure containing t clades one containing *P. semoni*, *P. prehensicauda*, *P. flavipes* and *P. sp nov 1* another containing *P. virens* and *P. sp nov 2*. It is important to note that the freuency of monophyletic groupings expected to reflect the overall distribution of gene trees but rather the distribution condit observing the trait incongruence of interest. These observations make our estimated proba

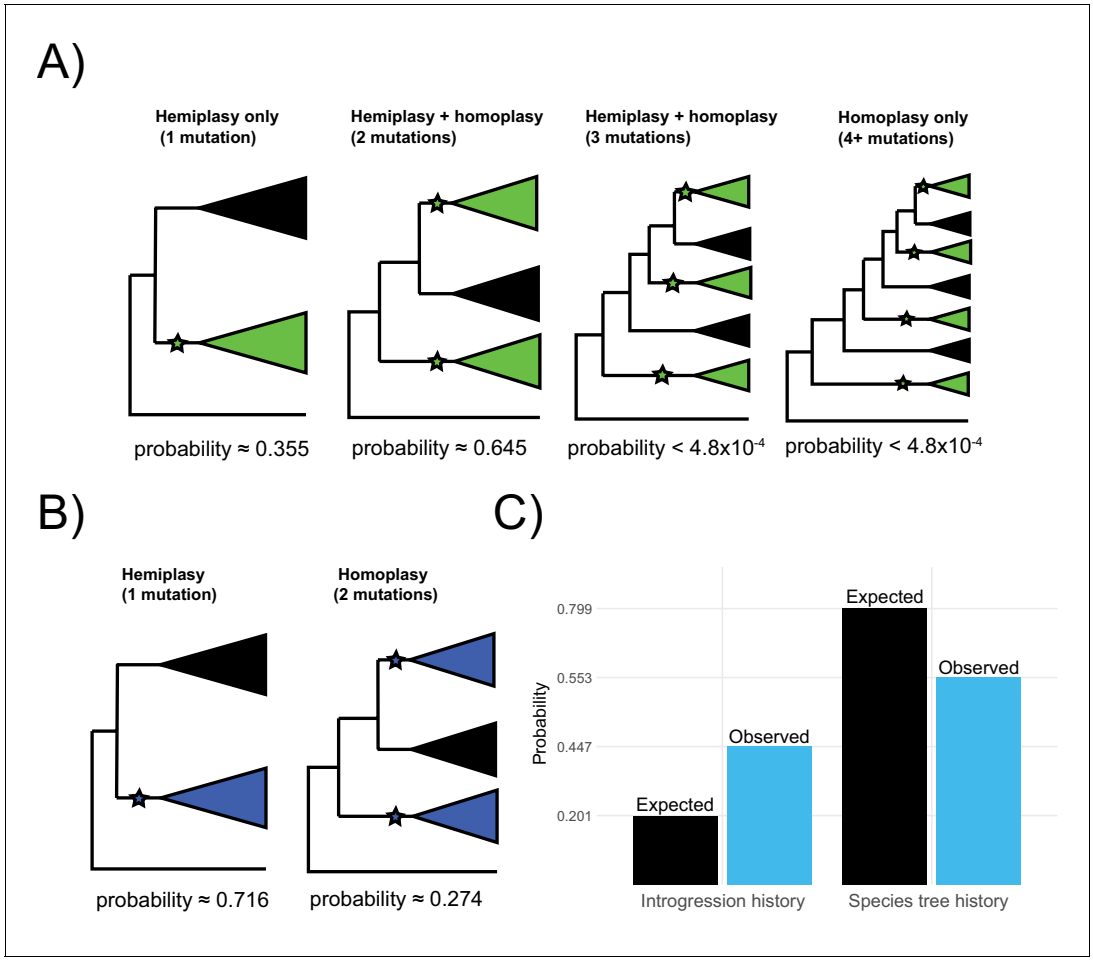

Figure 6. Probable histories for (**A**) the origin of green blood in New Guinea lizards and (**B**) the chromosomal inversion spanning the gene *cortex* in *Heliconius*, calculated using *HeIST*. Trees depict the maximum number of clades expected for gene tree topologies under each scenario, with green-blooded clades in green and inversion clades in blue. Branches with proposed ancestral-to-derived transitions are labeled with stars. Exactly which species are sorted into these clades can vary, meaning many possible gene trees exist for each of the depicted scenarios. Correspondingly, any of the labeled hypothetical mutations could represent hemiplasy or homoplasy (except in the case of a single mutation, which must be hemiplasy), depending on the gene tree topology. Reported probabilities are based on 10^{10} simulated trees for New Guinea lizards and 10^{7} trees for *Heliconius*, with probabilities conditional on matching the empirical trait distributions. Panel **C** shows what proportion of gene trees originate from a history of introgression vs. the species tree for the results summarized in panel B (blue) as compared to what would be expected based on the inferred network in ***Figure 1*** (black).

The online version of this article includes the following source data and figure supplement(s) for figure 6:

Source data 1. Output file from the *HeIST* lizard analysis, from which the probabilities are reported in ***Figure 6A***.
Source data 2. Output file from the *HeIST* butterfly analysis, from which probabilities are reported in ***Figure 6B and C***.
Figure supplement 1. Regression of internal branch lengths in substitutions per site (x-axis) against the same branch estimated in coalescent units using concordance factors (y-axis) for the 43-species lizard phylogeny (panel **A**) and the 6-species Heliconius phylogeny (panel **B**).
Figure supplement 2. Ultrametric phylogeny of green-blooded lizards, with branch lengths in units of 2N generations and green-blooded taxa labeled in green.

easy to interpret□ if there was one mutation□ it happened in the ancestor of the green-blo
if there were two mutations□ they most likely occurred in the ancestors of the two subcla
ōllowing the logic of phyloĀS□ (***Pease et al., 2016***)□ we checked biallelic sites in the CE
alignment and topologies from the CE gene trees for a monophyletic clade of green-bloo
ards in order to identify regions potentially associated with variation in blood color. ōwev
the gene tree and CE datasets contained missing samples□ which made it difficult to
identify truly monophyletic clades. On average□ approximately nine taxa were unrepresented
tips of individual gene trees□ and approximately 10 were not assigned a base at indivi

CEs. The identity of the missing taxa varied across sites and trees but often included s the 15-taxon subclade containing the green-blooded species which made it more difficult to tently polarie and compare patterns of monophyly. In the small proportion of gene trees sites where information was available for all taxa we did not find any monophyletic g green-blooded species.

A chromosomal inversion in the *Heliconius erato/sara* clade likely has a single origin

In addition to the analysis of green-blooded liards we also investigated the origins of somal inversion in the *Heliconius erato/sara* clade (*Edelman et al., 2019*). This inversion spans the gene *cortex* which is known to influence wing patterning and coloration across but (*Joron et al., 2006*; *Nadeau et al., 2016*). hile parsimony applied to the species phylogeny would suggest two independent origins of the inversion (*Figure 1B*) there is clear evidence in *Edelman et al., 2019* of both incomplete lineage sorting and introgression among the clades ing the inversion implicating a role for hemiplasy.

e inferred branch lengths in coalescent units for the phylogenetic network of these given in *Edelman et al., 2019*. sing our unit conversion tool we obtained a best-fit line $y = -1.815 + 302.49x$ with an adusted R^2 of 0.98 (though as a note of caution R^2 with this regression contained only five data points; see *Figure 6—figure supplement 1B*). The predicted branch leading to the outgroup was extremely long (generations) so the tree was smoothed using the extend method without the outgroup and the outgroup was re-added post-smoo at a length proportional to the original network. The two most highly supported introgressio in the inferred phylogenetic network were then added to the coalescent tree with their inferred direction rate and approximate timing before being given as input to *HelST* (*Figure 1B*).

sing *HelST* we found that a single origin of the inversion was most likely representi 923 (71.5) focal cases (*Figure 6B* left). The scenario involving two mutations was less likely was still found in 253 of 923 cases (27.4) (*Figure 6B* right). e also observed a small number (3 923 0.32) of focal cases with three independent transitions. Overall our results support th nal findings of *Edelman et al., 2019* that the inversion likely arose once and then was s between lineages via introgression.

Out of 923 simulated loci matching the trait pattern we found that 413 originated fro gressed history. This proportion (0.447) is substantially higher than the sum of introgressio bilities specified in the input (0.201) which suggests that introgression contributes more probability of observing the trait incongruence than would be expected by chance. In addi the liard simulations we found that almost every simulated focal tree (913923 98.9) groupe *Heliconius* species that share the inversion as monophyletic. owever there is more variation structure of the subclades than there was in the liards. Nevertheless we can infer fro two-mutation cases are most likely to arise as independent mutations in the ancestors clades that are part of a larger monophyletic group.

Discussion

Phenotypic convergence among species can provide important evidence for natural selectio molecular variation underlying this convergence can arise through independent mutations molecular level (*Storz, 2016*). owever it has recently become clear that such cases of true gence need to be distinguished from cases of apparent convergence due to introgression and hemiplasy (*Hahn and Nakhleh, 2016*). Some effort has been made in this regard through the use of coalesc tion summary statistics and updated comparative approaches (*Pease et al., 2016*; *Copetti et al., 2017*; *Guerrero and Hahn, 2018*; *Wu et al., 2018*). owever these approaches often assume incomplete lineage sorting as the only source of discordance and cannot explicitly resolve ber of transitions required to explain a trait distribution while accounting for discordance recently *Bastide et al., 2018* and *Karimi et al., 2020* developed extensions to comparative meth- ods that allow antitative trait likelihoods to be calculated on phylogenetic networks. ow while phylogenetic network inference methods are often robust to the effects of IL (*Lemus and Ané, 2016*; *Wen et al., 2018*) the estimated networks themselves do not contain t

necessary information to simultaneously capture the effects of ILS and introgression on tr
bilities (*Mendes et al., 2018*).

ere we take two important steps toward addressing these problems by (1) studying th
of introgression on the risk of hemiplasy under the multispecies network coalescent mode
providing a tool that can infer the most probable number of transitions given a phyloge
bution of binary traits. e find that introgression increases the risk of hemiplasy over ILS
uncover likely hemiplastic origins for the evolution of green blood from a red-blooded a
New inea liards and a chromosomal inversion spanning a gene important for wing color
Heliconius. hile our work has important implications for studies of trait evolution it also
numerous limitations and simplifying assumptions which suggest logical next steps for furthe
Below we discuss these implications assumptions and future directions.

The probability of hemiplasy due to introgression

A multitude of studies have revealed the potential role of introgression in shaping pheno
vergence and adaptation (e.g. *Heliconius Genome Consortium, 2012*; *Huerta-Sánchez et al., 2014*; *Jones et al., 2018*; *Mullen et al., 2020*). owever such studies rarely consider how introgres-
sion could lead to false inferences of convergence due to hemiplasy at both the molecu
notypic levels if left unaccounted for. Our model results show that both ILS and introgr
be accounted for in order to make robust inferences of convergent evolution.

Our model for the probability of hemiplasy with introgression combining concepts from tw
viously published models also shares most of their assumptions. irst we have assumed th
possible introgression scenario involving a single pair of species and with introgression
instantaneously at some point in the past. owever much more complex introgression scena
possible including introgression between multiple species pairs involving ancestral populatio
(and internal branches) at multiple time points in the past or continuously over a peri
oriontal gene transfer which is more common in prokaryotes would also reuire networks
contain reticulation edges spanning very long periods of time. It is not always clear hov
bility of hemiplasy would be affected under these alternative introgression scenarios. or e
we assume that the taxa sharing the derived state are also the ones involved in intr
introgression between other species pairs could alter patterns of discordance and therefore
the hemiplasy risk albeit less directly. any of these scenarios could be incorporated into
eral SNC framework as additional parent trees but with more complex histories this may
mathematically intractable even in the three-taxon case our hemiplasy inference tool HeIST is
designed to ameliorate this issue. espite these limitations we can generally expect that i
sion will increase the overall risk of hemiplasy whenever rates of introgression are high
pairs of species that also share the derived state for an incongruent trait. This is beca
matters is the generation of gene tree topologies with internal branches where hemiplast
tions can occur the increased variance in coalescence times under more complex introgres
narios while affecting mutation probabilities should have a comparatively minor effect (*Figure 4*).

e also assume that the coalescence times and gene tree freencies of loci underlying
ation follow neutral expectations even though alleles controlling trait variation are often
some form of selection. irectional selection on such variation will relative to neutral
expectations which will decrease the rate of incomplete lineage sorting and conseuently he
due to ILS. Of course the amount of ILS used in our simulations is not taken direc
expectations but rather is estimated from real data. Therefore the effects of selection or
interest will only be manifest if they are greater than the general effects of linked se
the regions used to estimate discordance (*Kern and Hahn, 2018*). On the other hand introgressed
alleles can lead to hemiplasy even in cases where there is no ILS. In fact directional
also make it more likely that introgressed loci have a discordant topology as it reduce
parent trees 2 and 3. Alternatively balancing selection can maintain ancestral polymorphis
increase rates of discordance due to ILS. This will also increase the risk of h
(*Fontaine et al., 2015*; *Lamichhaney et al., 2016*; *Palesch et al., 2018*).

Considerations for the inference tool *HeIST*

hile the software we introduce here allows for multiple novel types of inferences it also limitations that are important to address. Errors common to all phylogenetic methods can duced into the user-specified species treenetwork at several steps including errors in identification tree topology concordance factors and branch lengths (via both the conversi coalescent units and tree smoothing). The process of smoothing the coalescent tree sho duce predictable biases in branch length estimates. here is a method for redistributing branch lengths internal branches that are very short may have their length increased long external branches may be shortened. The lengthening of internal branches decreases all rate of discordance and makes inferences about hemiplasy from *HeIST* conservative. Similarly when smoothing is done using our function for extending tip branch lengths the probability pendent mutations on those tip branches (i.e. homoplasy) is increased again making h inferences conservative. The results presented in *Figure 5—figure supplement 6* capture the overall effects of errors in phylogeny estimation branch length prediction and smoothing.

Errors in inference may affect our approach to branch length unit conversion in seve concordance factors are underestimates for instance due to errors in gene tree reconstruc then the branch lengths in coalescent units will also be underestimates of their true value would be simulations with more ILS and discordance than actually occurred. In cases whe concerns about branch length estimates we suggest *HeIST* running across multiple values for tip branches the option exists when *HeIST* to use the lower and upper bounds of the prediction val in addition to the predictions themselves. In addition if there are tip branch lengths nal tree that fall outside the range of internal branch length values the predicted value branches in coalescent units may be less reliable since it reires extrapolation beyond the datapoints used to fit the regression. Lastly we note that *Bastide et al., 2018* propose an approach to estimating coalescent tip branch lengths on a network using the method of least-sares pairwise genetic distances and network pairwise distances. e expect this approach to ha similar performance to ours since linear regression is done using least-sares and pairwise distances should be highly correlated with concordance factors.

There are also several practical points to consider when applying *HeIST* to empirical data. hen researchers have estions about hemiplasy involving either very large phylogenies or ve mutation rates only a small number of simulated trees may match the incongruent patter real data. The large number of simulations reired may not be computationally feasible careful pruning of species that do not affect inferences of hemiplasy may greatly reduce tion. By default *HeIST* will prune the input phylogeny to include the smallest subclade that all the taxa with the derived state plus a specified outgroup. In addition while *HeIST* can simulate phylogenies with introgression it reires that the timing direction and rate of each introgr event is provided. To obtain this information we recommend using a phylogenetic networ approach such as PhyloNet (*Wen et al., 2018*) SNaQ (*Solís-Lemus and Ané, 2016*) or the Species-Network (*Zhang et al., 2018a*) package within BEAST2 (*Bouckaert et al., 2019*).

inally an issue that concerns both our theoretical work and *HeIST* is the specification of the mutation rate. In both cases we assume that the rates of 1 0 transitions are eivalent and that these rates are constant across the tree under study. iolations of these assu certainly influence the probabilities of hemiplasy and homoplasy although it is unlikely that ing mutation rates will vary substantially among closely related lineages (*Lynch, 2010*). ore important these rates represent the mutation rate among character states and may not alwa same as nucleotide mutation rates. e have assumed in the results presented here that between character states are controlled by a single site and therefore that the nucleotide rate is a good approximation of the trait mutation rate. owever the degree to which will depend on the genetic architecture underlying a trait. or example transitions in floral often underlain by loss-of-function mutations and many mutational targets can potentially l the same phenotypic changes (*Rausher, 2008*; *Smith and Rausher, 2011*). In such cases the rate o trait transitions can potentially be many times higher than the nucleotide mutation rate wi plasy becoming more probable as a result. In contrast trait transitions can also reire molecular changes the order of which may be constrained by pleiotropy and epistasis. Suc underlie for instance high-altitude adaptation of hemoglobin in mammals (*Storz et al., 2009*

Tufts et al., 2015). In these cases, the rate of trait transitions may be many times low nucleotide mutation rate, with hemiplasy becoming more probable as a result.

Evolution of green-blooded lizards and the *Heliconius* inversion

In our analysis of lizards in the genus *Prasinohaema*, we found strong support for one or two in pendent origins of green blood from a red-blooded ancestor, with two origins being the m This contrasts with analyses that do not account for gene tree discordance, in which fou is the best explanation. In *Heliconius*, we found support for a single origin of a chromosomal sion, in contrast to methods that do not account for discordance. Both these results stron that hemiplasy has played a role in the evolution of these traits.

Applications of *HeIST* to these clades involves some system-specific assumptions, the first which relates to the genetic architecture of the traits under study. For the lizard analysis, the potentially strong assumption that the green-blooded phenotype is achievable by a sing tion. While the physiological mechanism for this phenotype is well-understood (*Hantzsche and Lessing, 1994*), the genetic architecture underlying the transition from a red-blooded ancestor is not cussed in the previous section, this architecture will affect the choice of the trait evolution- ary rate in our simulations. Since the genetic architecture is unknown, we based our θ on what is typically observed for nucleotide mutations in vertebrate systems (*Lynch, 2010*). For the *Heli- conius* inversion, the architecture is more clear-cut, since chromosomal inversions are a singl tional event by definition. While the per-generation rate of de novo chromosomal inversions known for many systems, it is certain to be lower than the rate for nucleotide mutat Nucleotide θ is estimated at 0.02–0.08 in *H. melpomene* (*Martin et al., 2016*), and averages around 0.01 in invertebrates (*Lynch, 2010*). Our choice of θ for the inversion was one order of magnitud lower than these estimates.

Another key assumption is that the estimated gene trees and concordance factors are as is the regression approach for converting branch length units. The R^2 of 0.554 observed for the unit-conversion in the lizard dataset might be interpreted as surprisingly low given that it i sion of the same quantity measured in two different units. This value likely reflects uncerta ated in several steps of our analysis, including the estimation of branch lengths in th likelihood species tree, and the procedure of randomly sampling quartets to estimate sCs, u *IQ-TREE*. In *Heliconius*, the R^2 was much higher at 0.98, but with only five data points there ited information about the true relationship. Nonetheless, we observed the expected positive lation in both cases, and a sufficient amount of variation is explained to ensure that estimated in coalescent units are proportionally similar to those in the maximum-likelihood t gesting that the regression approach works well as an approximation. In addition, the r line on the lizard data appears to slightly over-estimate very short branch lengths in coale making our inferences of hemiplasy conservative.

Conclusions

A major question in the study of convergent evolution is whether phenotypic convergence i lain by convergent changes at the molecular level (*Storz, 2016*). The work presented here is con cerned primarily with such molecular changes, and the results of our empirical analyses how apparently convergent phenotypes can arise from a single molecular change. Such changes come about as a result of gene tree discordance due to ILS, introgression, or nation of the two. Given that these phenomena are common in phylogenomic d (*Pollard et al., 2006*; *Fontaine et al., 2015*; *Pease et al., 2016*; *Novikova et al., 2016*; *Wu et al., 2018*), perhaps it should be less surprising that phylogenetically incongruent traits often hav mon genetic basis.

Finally, while the tools presented here may help to rule out cases of molecular conver observation of a single molecular origin for a trait does not rule out the occurrence o adaptation in general. Parallel selective pressures from the environment on the same mole ation may be regarded as one of many possible modes of convergent evolution (*Losos, 2017*). In studying novel phenotypes such as green blood or wing patterning and coloratio still tremendous interest in understanding the ecological pressures that may have led to th dent fixation of single, ancestral changes along multiple lineages. In general, integrative app

combining modern phylogenomics with an ecological context will pave the way toward an understanding of the nature of convergent evolution.

Materials and methods

Accuracy of *HeIST*

To confirm that *HeIST* accurately counts mutation events and is consistent with our theoretical ings we evaluated its performance under nine simulated conditions with increasing le expected hemiplasy. All simulated conditions involve a four-taxon tree with the topology (((4 Species 4 and 2 carry the derived state for a hypothetical binary character. The split from the ancestor of 4 3 and 2 occurred *N* generations in the past. The first three simulated con tions contain no introgression and progressively decrease the length of the internal branch ing species 4 and 3. The total tree height was held constant. The simulated internal were *N* 1.5*N* and *N* generations for conditions *ILS1* *ILS2* and *ILS3* respectively. The subsequent six conditions maintain *ILS3* condition for branch lengths with the addition of an introgress event from species 2 into species 4. For conditions *INT1* *INT2* and *INT3* the timing of introgression was held constant at generations while the introgression probability was set to 0.01 0.05 0.1 respectively. For conditions *INT4* *INT5* and *INT6* the introgression probability was held constant at 0.1 while the timing of introgression was reduced to and *N* respectively. The parameters used for each condition are summarized in *Figure 5—figure supplement 1*.

We performed two sets of simulations (1) 100 replicates of each condition consisting of gene trees each with a constant mutation rate of per generations (2) 20 replicates of each condition for each of five different mutation rates per generations (0.0005 0.0025 0.005 0.025 0.05) each consisting of 1000000 gene trees. For each combination of parameters we es the probability of hemiplasy conditional on observing the specified trait pattern and the ra bility of hemiplasy out of the total number of replicates. For the latter simulation set we the mean-squared error (SE) using the simulated values as observations and the expecte from theory as the true mean. These SE values were divided by the simulated mean error across conditions with different ranges of expected values.

Trade-offs among parameters mean that many combinations of estimated network param may be equally consistent with patterns in subsets of the observed data. To investigate effects on the probability of hemiplasy we evaluated the performance of *HeIST* under five additional simulation conditions (*Figure 5—figure supplement 4*). In each successive condition the probability of introgression was increased while the timing of introgression was made more recent. T of the internal branch in the species tree was also increased such that the expected fr discordant gene tree that causes hemiplasy remained approximately constant (*Figure 5—figure supplement 5A*). These simulations used the same tree topology derived taxa split time of t tral population and mutation rate as the first set of benchmarking simulations. Condition 1 same parameters as *ILS1*. Conditions 2 5 used the following sets of parameters respectively 2 2.32*N* 2.6*N* 2.8*N* generations for the length of the internal branch 0.01 0.025 0.04 0.05 probability of introgression 0.4*N* 0.3*N* 0.2*N* 0.1*N* generations for the timing of introgression (*Figure 5—figure supplement 4*). For each condition we performed 100 replicate simulations 100000 gene trees each *HeIST* and estimated the probability of hemiplasy conditional on obse ing the trait pattern.

Inferring the tip branch lengths of a phylogeny in coalescent units

Inferences made under the multispecies coalescent require branch lengths specified in coa units. However most standard methods for building phylogenies infer branches in units of tions per site. Units of absolute time inferred from substitution rates using molecula approaches can be converted into coalescent units provided that the generation time and population size are known. However these parameters are sometimes not available or accurat given system. As an alternative estimates of gene tree discordance can be used to esti branch lengths in coalescent units but these provide no information about the length branches. For example the species tree inference software *ASTRAL* (*Zhang et al., 2018b*) does not infer tip branch lengths while the software *MP-EST* (*Liu et al., 2010*) adds branches of length nine

for every tip. These tip lengths are necessary to make accurate inferences about hem homoplasy from empirical data since they affect the probability of mutation on tip branches

To ameliorate this problem we have applied a simple regression approach for infe lengths in coalescent units (see *Bastide et al., 2018* for an alternative method). Our approach makes use of concordance factors estimates of the fraction of concordant loci with respect to branch in a species tree. Concordance factors come in two flavors gene concordance fac (*Gadagkar et al., 2005*; *Ané et al., 2007*) which estimate the concordance of gene tree topologie and site concordance factors (*Minh et al., 2020a*) which do the same for parsimony-informa- tive sites. In general concordance factors estimated from quartets provide an estimate of $1 - \frac{2}{3}\frac{T}{}$ where T is the length of the internal branch in coalescent units. ith concordance factors the internal branches of a tree that has lengths in substitutions per site the aforementio can be used to obtain estimates of those same branch lengths in coalescent units. Mgeneralized of 2 sion of the internal branch length estimates in both units can then be used to obtain unit conversion between the *HeIST* uses this formula to predict the tip branch lengths of th in coalescent units. To partially account for uncertainty introduced during tip branch length tion *HeIST* can also be run using the lower or upper bounds of the prediction 95 confide val as the inferred tip lengths in addition to the predictions themselves. As a final process the tree in coalescent units is smoothed to ensure the input tree to be ultrametric. *HeIST* has two options for how to perform this smoothing. The first redistributes the tr lengths so that the distance from the root to each tip is the same this is done usi the ultrametric() function in the Python library *Phylo3* (*Huerta-Cepas et al., 2016*). The second extends the lengths of tip branches while preserving internal branch lengths this function is coo *HeIST* but was borrowed from a commented out task on the source code.

To investigate the potential bias introduced to results by *HeIST* from either phylogenetic infer- ence the branch regression approach or subsequent smoothing we compared the *HeIST* outputs of run from an eight-taxon test tree with known branch lengths. To generate realistic dataset simulated 3000 gene trees from the known species tree and using then simulated 1 b of sequence from each locus with $\theta = 0.001$ using *Seq-Gen*. These loci were concatenated into a sing 3 b alignment which was given to *RAxML* version 8.2.12 (*Stamatakis, 2014*) using the TR substi- tution model with rate heterogeneity to infer a species tree in units of substitutions pe inferred tree and the concatenated alignment were given to *IQ-TREE* version 2.0 (*Minh et al., 2020b*) to infer site concordance factors. This substitution tree with nodes labeled with co factors was given as input to *HeIST* where our branch regression approach was applied. e test both methods for tree smoothing from this input. e estimated the probability of each nu mutations conditional on observing the incongruent site pattern and results from this analy compared to those obtained using the true test tree (*Figure 5—figure supplement 6*).

Our regression approach is implemented in the *HeIST* and can be run as part of the overall hemi analysis or separately using the *mbsuite coal*.

Empirical applications of *HeIST*

e applied *HeIST* to two empirical case studies where hemiplasy appeared to be a plausibl nation for observed trait incongruences. The first is a dataset of New inea (*Rodriguez et al., 2018*). As described in the Introduction the genus *Prasinohaema* contains six spe- cies that have evolved green blood from a red-blooded ancestor (*Figure 1A*). Previous analyses of the species tree built from thousands of loci inferred that four independent transitions are to explain the phylogenetic distribution of green-blooded (*Rodriguez et al., 2018*). This conclusion is the same using any standard phylogenetic comparative method whether a state reconstruction is carried out using maximum likelihood (*Rodriguez et al., 2018*) or itch parsi- mony (this study). owever the phylogeny for this clade contains many short branches (*Figure 1A*), suggesting that a scenario involving at least some hemiplasy (in this case 1 3 mutatio preferred over homoplasy-only scenarios when discordance is accounted for.

To address this uestion we used the original dataset of (*Rodriguez et al., 2018*) consisting of 3220 ultra-conserved elements (CEs) totalling approximately 1.3 b for 43 species. e down-sampled these species to 15 taxa in the clade including the green-blooded specie outgroup (*Figure 1A*). e constructed a concatenated maximum likelihood species tree in addi

to gene trees for each CE□RAxML version 8.2.12 (*Stamatakis, 2014*). To verify the species tree topology for the 15-taxon subclade□ we also constructed a *ASTRAL-III* tree with version 5.6.3 (*Zhang et al., 2018a*). Site and gene concordance factors were calculated for this tree u *IQ-TREE* version 2.0 (*Minh et al., 2020a*□*Minh et al., 2020b*). To obtain the phylogeny in coalescent units□ we employed the regression approach described above for unit conversion as implem *HeIST*. The extend□ method was used for tree smoothing. e *HeIST* to simulate 10 loci from the liard subclade containing green-blooded species□ with a population-scaled mutation (θ) of 0.0005 p generations. hile specific parameter estimates are not available for this tem□ our choice θ reflects broad estimates N_e of and μ on the order of 10 (*Lynch, 2006*) and 10^{-8} □ 10^{-9} per-base per-generation (*Lynch, 2010*)□ respectively□ in vertebrates (see iscussion). This analysis was performed for each of two *Lygosoma* subgroups *sp*□ which is sister to 40 species the 43-species phylogeny□ and *Scincella lateralis*□ which is sister to the 15-taxon clade containing green-blooded species. e also calculated statistics (*Green et al., 2010*) for 12 trios involving green-blooded taxa□ finding no strong evidence of introgression (block bootstrap significance *Supplementary file 1*-Tables 2 and 3). Therefore□ our simulations did not include any introg events.

The second empirical case study involves the origins of a chromosomal inversion spann important for wing coloration in *Heliconius* butterflies (*Edelman et al., 2019*). The derived inversion arrangement is shared by four taxa□ grouped into two subclades (*H. timareta* and the group of *Helico-nius*. itch parsimony suggests two independent origins□ but a combination of short i branches and introgression between the ancestral populations sharing the inversion (*Edelman et al., 2019*) strongly suggests a role for hemiplasy. e obtained the phylogenetic network□ that species tree with reticulation edges□ inferred in units of substitutions per site□ in addition concordance factors□ from the authors. As our regression approach for conversion to co units cannot be used on phylogenetic networks directly□ we used the species tree embed network with concordance factors as input to *HeIST*. The two most strongly supported introgres-sion events were then added back onto the smoothed network in coalescent units□ using rates and directions□ with approximate timings based on the location of the events in network and our requirement that these events be instantaneous pulses. rom this input (sl *Figure 1B*)□ we simulated 10^7 gene trees in *HeIST* using a mutation rate of 0.0005 per gener-2 tions. That our choice of θ for this system is the same as in our liard analysis is ust a c reflects a trade-off between the generally higher effective population sie (*Lynch, 2007*) s and the lower mutation rate expected for chromosomal inversions□ see iscussion. e also formed the same simulations without specifying the introgression events to obtain an ILS-mate of the probability of hemiplasy.

Acknowledgements

e thank Rafael errero□ Leonie oyle□ and Ben ulton for helpful comments and advice□ as as three anonymous reviewers□ the associate editor□ and Chris Austin for suggestions tha improve this work. e also thank achary Rodrigue□ for sharing the liard data and Nate for sharing the *Heliconius* phylogenetic network. This work was supported by National Science d dation grant EB-1936187.

Additional information

Funding

Funder	Grant reference number	Author
National Science Foundation	DEB-1936187	Matthew W Hahn

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

ark S ibbins Conceptualiation ata curation Software ormal analysis alidation Investigation isualiation ethodology riting - original draft Proect administration riting - review and editing atthew S ibson ata curation Software alidation ethodology riting - original draft riting - review and editing atthew ahn Conceptualiation Supervision unding acuisition alidation riting - original draft Proect administration riting - review and editing

Author ORCIDs

ark S ibbins https orcid.org0000-0002-4651-3704
atthew S ibson https orcid.org0000-0001-7855-1628
atthew ahn https orcid.org0000-0002-5731-8808

Decision letter and Author response

ecision letter https://doi.org/10.7554/eLife.63753.sa1
Author response https://doi.org/10.7554/eLife.63753.sa2

Additional files

Supplementary files

- Supplementary file 1. Table 1 Site concordance factors from Itree. I the I of the ir branch in the full liard phylogeny as labeled in *Figure 1—figure supplement 1*. sC the value of the site concordance factor for the branch averaged over 100 randomly sampled uartets and s2 the site discordance factors for the first and second most common discordant sit terns at each branch respectively. sN the average number of informative sites averaged acr sampled uartets at each branch. Length is the length of the internal branch in substitut site. -Table 2 Trios involving green-blooded species which were evaluated for evidence of sion using statistics. Species are listed in the order P1 P2 P3 where an excess of P2P3 alleles would indicate evidence of introgression. Lygosoma sp was used as the o each trio. Table 3 statistic results for the 12 trios listed in Table 1 estimate concatenated alignment of ultra-conserved elements (CEs). Significance was evaluated for eac by bootstrap-sampling the CEs to generate a null distribution of alignments and asking ho the bootstrap distribution of 1000 statistics was at least as extreme as the observed v
- Transparent reporting form

Data availability

Availability of the liard genomic data and eliconius phylogenetic network is detailed i Acknowledgements section of the source manuscript. Source code and test cases for ou eIST are freely available from the itub repository. Source data files have been provided ures 1 4 5 and 6. The Appendix details all the mutation rate parameters of our model.

The following previously published datasets were used

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Rodriguez ZB, Perkins SL, Austin CC	2018	Raw sequencing reads from ultraconserved elements of Australasian skinks	https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA420910	NCBI BioProject, PRJNA420910
Edelman NB, Frandsen PF, Miyagi M, Clavijo B, Davey J, Dikow RB, Van Belleghem SM, Patterson N, Neafsey DE, Richard C, Kumar S, Moreira GRP, Salazar C,	2019	Data from: Genomic architecture and introgression shape a butterfly radiation	http://dx.doi.org/10.5061/dryad.b7bj832	Dryad Digital Repository, 10.5061/dryad.b7bj832

Chouteau M, Counterman B, Papa R, Blaxter M, Reed RD, Dasmahapatra KK, Kronforst M, Joron M, Jiggins CD, McMillan WO, Di Palma F, Blumberg AJ, Wakeley J, Jaffe D, Mallet J

References

Ané C□ Larget B□ Baum A□ Smith S□ Rokas A. 2007. Bayesian estimation of concordance among gene *Molecular Biology and Evolution* **24**412□426. OI□ https□doi.org10.1093□molbev□msl170 □PI□ 17095535

Austin CC□ essing □ 1994. reen-blood pigmentation in liard*Comparative Biochemistry and Physiology Part A: Physiology* **109**619□626. OI□ https□doi.org10.10160300-9629(94)90201-1

Avise JC□ Robinson T□ 2008. emiplasy□ a new term in the lexicon of *phylogenetics*. *Systematic Biology* **57**503□507. OI□ https□doi.org10.10801063515080216458 7 □PI□ 18570042

Bastide P□ Sos-Lemus C□ riebel R□ illiam Sparks □ Ané C. 2018. Phylogenetic comparative methods on phylogenetic networks with reticulations. *Systematic Biology* **67**800□820. OI□ https□doi.org10.1093□sysbio□syy033□PI□ 29701821

Bouckaert R□ aughan T□ Barido-Sottani □ uche□ ne S□ ourment □ avryushkina A□ eled □ ones□hnen□ e aio N□ atschiner □ endes □ u □ller N□ Ogilvie A□ du Plessis L□ Popinga A□ Rambaut A□ Rasmussen □ Siveroni I□ Suchard A□ et al. 2019. BEAST 2.5□ an advanced software platform for Bay evolutionary analysis. *PLOS Computational Biology* **15**e1006650. OI□ https□doi.org10.1371□ournal.pcbi.1006650□PI□ 30958812

Brandvain Y□ enney A□ lagel L□ Coop □ Sweigart AL. 2014. Speciation and introgression *Mimulus*□between *nasutus* and *Mimulus guttatus*. *PLOS Genetics* **10**e1004410. OI□ https□doi.org10.1371□ournal.pgen.1004410□PI□ 24967630

Copetti D□ Burue□ A□ Bustamante E□ Charboneau L□ Childs L□ Eguiarte LE□ Lee S□ Liu TL□ ccahon □ hiteman N□ ing RA□ ociechowski □ Sanderson □ 2017. Extensive gene tree discordance and hemiplasy shaped the genomes of North American columbines. *PNAS* **114**8003□12008. OI□ https□doi.org10.1073□pnas.1706367114 □PI□ 29078296

Degnan JH. 2018. odeling hybridiation under the network multispecies coalescent. *Systematic Biology* **67**□786□799.OI□ https□doi.org10.1093□sysbio□syy040 □PI□ 29846734

Degnan JH□ Rosenberg NA. 2009. ene tree discordance□ phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution* **24**332□340. OI□ https□doi.org10.1016□tree.2009.01.009 □PI□ 19307040

Delsuc F□ Brinkmann □ Philippe □ 2005. Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics* **6**361□375. OI□ https□doi.org10.1038□nrg1603 □PI□ 15861208

Dunn CW□ enol A□ atus □ Pang □ Browne E□ Smith SA□ Seaver E□ Rouse □ Obst □ Edgecombe □ S□rensen □ addock S□ Schmidt-Rhaesa A□ Okusu A□ ristensen R□ heeler C□ artindale □ iribet □ 2008. Broad phylogenomic sampling improves resolution of the animal tree. *Nature* **452**745□749. OI□ https□doi.org10.1038□nature06614 □PI□ 18322464

Durand EY□ Patterson N□ Reich □ Slatkin □ 2011. Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution* **28**2239□2252. OI□ https□doi.org10.1093□molbev□msr048 □PI□ 21325092

Edelman NB□ randsen PB□ iyagi □ Clavio B□ avey □ ikow RB□ ae Accioli □ an Belleghem S□ Patterson N□ Neafsey E□ Challis R□ umar S□ oreira RP□ Salaar C□ Chouteau □ Counterman BA□ Papa Blaxter □ Reed R□ asmahapatra □ et al. 2019. enomic architecture and introgression shape a butterfly radiation. *Science* **366**594□599. OI□ https□doi.org10.1126□science.aaw2090 □PI□ 31672890

Felsenstein J. 1985. Phylogenies and the comparative method. *The American Naturalist* **125**1□15. OI□ https□doi.org10.1086284325

Felsenstein J. 2004. *Inferring Phylogenies*. Sinauer Associates.

Fitch WM. 1971. Toward defining the course of evolution□ minimum change for a specific tree topology *Systematic Zoology* **20**406□416. OI□ https□doi.org10.23072412116

Fontaine MC□ Pease B□ Steele A□ aterhouse R□ Neafsey E□ Sharakhov I□ iang □ all AB□ Catteruccia □ akani E□ itchell SN□ ü C□ Smith A□ Love RR□ Lawniczak □ Slotman A□ Emrich S□ ahn □ Besansky N□ 2015. Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science* **347**1258524. OI□ https□doi.org10.1126□science.1258524 □PI□ 25431491

Gadagkar SR□ Rosenberg S□ umar S. 2005. Inferring species phylogenies from multiple genes□ concatenate seuence tree versus consensus gene tree. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* **304B**64□74. OI□ https□doi.org10.1002□b.21026

Green RE□ rause □ Briggs A□ aricic T□ Stenel □ ircher □ Patterson N□ Li □ hai □ rit□ □ ansen N□ urand E□ alaspinas AS□ ensen □ arues-Bonet T□ Alkan C□ Fu □ eyer □ Burbano A□ ood □ et al. 2010. A draft seuence of the Neandertal genome. *Science* **328**710□722. OI□ https□doi.org10.1126□science.1188021□PI□ 20448178

Guerrero RF□ ahn □ 2018. uantifying the risk of hemiplasy in phylogenetic inference. *PNAS* **115**12787□12792. OI□ https□doi.org10.1073□pnas.1811268115 □PI□ 30482861

Hahn MW□ ibbins S. 2019. A three-sample test for introgression. *Molecular Biology and Evolution* **36**2878□2882. OI□ https□doi.org10.1093□molbevms178 □PI□ 31373630

Hahn MW□ Nakhleh L. 2016. Irrational exuberance for resolved species trees. *Evolution* **70**7□17. OI□ https□doi.org10.1111□evo.12832

Heliconius Genome Consortium. 2012. Butterfly genome reveals promiscuous exchange of mimicry adaptation among species. *Nature* **487**94□98. OI□ https□doi.org10.1038□nature11041 □PI□ 22722851

Hibbins MS□ ahn □ 2019. The timing and direction of introgression under the multispecies network coalescent. *Genetics* **211**1059□1073. OI□ https□doi.org10.1534□genetics.118.301831 □PI□ 30670542

Hudson RR. 1983. Testing the constant-rate neutral allele model with protein sequence data. *Evolution* **37**203□217. OI□ https□doi.org10.1111□1558-5646.1983.tb05528.x

Hudson RR. 2002. enerating samples under a right-isher neutral model of genetic variation. *Bioinformatics* **18**337□338. OI□ https□doi.org10.1093□bioinformatics18.2.337 □PI□ 11847089

Huerta-Cepas J□ Serra □ Bork P. 2016. ETE 3□ reconstruction□ analysis□ and visualiation of phylogenomic *Molecular Biology and Evolution* **33**1635□1638. OI□ https□doi.org10.1093□molbevmsw046 □PI□ 269213 90

Huerta-Sánchez E□ in □ Asan □ Bianba □ Peter B□ inckenbosch N□ Liang □ i □ e □ Somel □ Ni P□ ang B□ Ou □ uasang □ Luosang □ Cuo □ Li □ ao □ in □ ang □ et al. 2014. Altitude adaptation in Tibetans caused by introgression of enisovan-like DNA. *Nature* **512**194□197. OI□ https□doi.org10.1038□nature13408□PI□ 25043035

Joly S□ cLenachan PA□ Lockhart P□ 2009. A statistical approach for distinguishing hybridiation and incomplete lineage sorting. *The American Naturalist* **174**E54□E70. OI□ https□doi.org10.1086600082 □PI□ 19519219

Jones MR□ ills LS□ Alves PC□ Callahan C□ Alves □ Lafferty R□ iggins □ ensen □ elo-erreira □ ood □ 2018. Adaptive introgression underlies polymorphic seasonal camouflage in snowshoe hares. *Science* **360**1355□1358. OI□ https□doi.org10.1126□science.aar5273 □PI□ 29930138

Joron M□ Papa R□ Beltra n □ Chamberlain N□ ava re □ Baxter S□ Abanto □ Bermingham E□ umphray S□ Rogers □ Beasley □ Barlow □ ffrench-Constant R□ allet □ cillan O□ iggins C□ 2006. A conserved supergene locus controls colour pattern diversity in *Heliconius* butterflies. *PLOS Biology* **4**e303. OI□ https□doi.org10.1371□ournal.pbio.0040303 □PI□ 17002517

Karimi N□ rover CE□ allagher P□ endel □ Ane C□ Baum A. 2020. Reticulate evolution helps explain apparent homoplasy in floral biology and pollination in *Baobabs* (Bombacoideae□ alvaceae). *Systematic Biology* **69**462□478. OI□ https□doi.org10.1093□sysbiosy073 □PI□ 31693158

Kern AD□ ahn □ 2018. The neutral theory in light of natural selection. *Molecular Biology and Evolution* **35**1366□1371. OI□ https□doi.org10.1093□molbevmsy092 □PI□ 29722831

Lamichhaney S□ an □ Berglund □ ang C□ Alme n S□ ebster T□ rant BR□ rant PR□ Andersson L. 2016. A beak siе locus in arwins finches facilitated character displacement during a drought. *Science* **352**470□474. OI□ https□doi.org10.1126□science.aad8786 □PI□ 27102486

Lee KM□ Coop □ 2017. istinguishing among modes of convergent adaptation using population genomic da *Genetics* **207(4)** 15911619 OI□ https□doi.org10.1534□genetics.117.300417

Liu L□ u L□ Edwards S□ 2010. A maximum pseudo-likelihood approach for estimating species trees under coalescent model. *BMC Evolutionary Biology* **10**302. OI□ https□doi.org10.11861471-2148-10-302 □PI□ 20937096

Liu KJ□ ai □ Truong □ Song □ ohn □ Nakhleh L. 2014. An -based comparative genomic framework for detecting introgression in eukaryotes. *PLoS Computational Biology* **10**e1003649. OI□ https□doi.org10.1371□ournal.pcbi.1003649 □PI□ 24922281

Lynch M. 2006. The origins of eukaryotic gene structure. *Molecular Biology and Evolution* **23**450□468. OI□ https□doi.org10.1093□molbevms050 □PI□ 16280547

Lynch M. 2010. Evolution of the mutation rate. *Trends in Genetics* **26**345□352. OI□ https□doi.org10.1016□tig.2010.05.003□PI□ 20594608

Mallet J□ Besansky N□ ahn □ 2016. ow reticulated are species? *BioEssays* **38**140□149. OI□ https□doi.org□10.1002□bies.201500149 □PI□ 26709836

Martin SH□ ost □ Palmer □ Salaar C□ cillan O□ iggins □ iggins C□ 2016. Natural selection and genetic diversity in the butterfly *Heliconius melpomene*. *Genetics* **203**525□541. OI□ https□doi.org10.1534□genetics.115.183285□PI□ 27017626

Mendes FK□ ahn □ ahn □ 2016. ene tree discordance can generate patterns of diminishing convergence over time. *Molecular Biology and Evolution* **33**3299□3307. OI□ https□doi.org10.1093□molbevmsw197 □PI□ 27634870

Mendes FK□ uentes-ona le A□ Schraiber □ ahn □ 2018. A multispecies coalescent model for uantitative traits. *eLife* **7**e36482. OI□ https□doi.org10.7554□eLife.36482 □PI□ 29969096

Mendes FK□ ahn □ 2016. ene tree discordance causes apparent substitution rate variation. *Systematic Biology* **65**711□721. OI□ https□doi.org10.1093□sysbiosyw018 □PI□ 26927960

Meng C□ ubatko LS. 2009. etecting hybrid speciation in the presence of incomplete lineage sorting usir tree incongruence□ a mode*Theoretical Population Biology* **75**35□45. OI□ https□doi.org10.1016□tpb.2008.10.004□PI□ 19038278

Minh BQ□ ahn □ Lanfear R. 2020a. New methods to calculate concordance factors for phylogenomic datasets. *Molecular Biology and Evolution* **37**2727□2733. OI□ https□doi.org10.1093□molbev□msaa106 □PI□ 32365179

Minh BQ□ Schmidt A□ Chernomor O□ Schrempf □ oodhams □ von aeseler A□ Lanfear R. 2020b. I-TREE 2□ new models and efficient methods for phylogenetic inference in the*Molecular Biology and Evolution* **37**1530□1534. OI□ https□doi.org10.1093□molbev□msaa015□ PI□ 32011700

Misof B□ Liu S□ eusemann □ Peters RS□ onath A□ ayer C□ randsen PB□ are □ louri T□ Beutel R□ Niehuis O□ Petersen □ luerdo-Carrasco □ appler T□ Rust □ Aberer A□ Aspo□ Aspo□ Bartel □ Blanke A□ et al. 2014. Phylogenomics resolves the timing and pattern of insect evolution. *Science* **346**763□767. OI□ https□doi.org10.1126□science.1257570 □PI□ 25378627

Mullen SP□ anuren N□ hang □ Nallu S□ ristiansen EB□ uyun □ Liu □ ill RI□ Briscoe A□ ronforst R. 2020. isentangling population history and character evolution among hybridiing *Molecular Biology and Evolution* **37**1295□1305. OI□ https□doi.org10.1093□molbev□msaa004 □PI□ 31930401

Nadeau NJ□ Pardo-ia□ C□ hibley A□ Supple A□ Saenko S□ allbank R□ u C□ aroa L□ erguson L□ anly □ ines □ Salaar C□ errill R□ owling A□ ffrench-Constant R□ Llaurens □ oron □ cillan O□ iggins C□ 2016. The gene*Nature* **534**106□110. OI□ https□doi.org10.1038□nature17961 □PI□ 27251285

Novikova PY□ ohmann N□ Nihynska □ Tsuchimatsu T□ Ali □ uir □ uggisberg A□ Paape T□ Schmid □ edorenko O□ olm S□ Sal T□ Schtterer C□ arhold □ idmer A□ Sese □ Shimiu □ eigel □ ramer □ och A□ et al. 2016. Seuencing of the *Arabidopsis* identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. *Nature Genetics* **48**1077□1082. OI□ https□doi.org10.1038□ng.3617 □PI□ 27428747

Palesch D□ Bosinger SE□ Tharp □ anderford T□ Paiardini □ Chahroudi A□ ohnson P□ irchhoff □ ahn B□ Norgren RB□ Patel NB□ Sodora L□ awoud RA□ Stewart CB□ Seepo S□ arris RA□ Liu □ Raveendran □ an English A□ et al. 2018. Sooty mangabey genome seuence provides insight into AIS resistance in a host. *Nature* **553**77□81. OI□ https□doi.org10.1038□nature25140 □PI□ 29300007

Pamilo P□ Nei □ 1988. Relationships between gene trees and spe*Molecular Biology and Evolution* **5**□ 568□583. OI□ https□doi.org10.1093□oxfordournals.molbev.a040517 □PI□ 3193878

Patterson N□ oorani P□ Luo □ allick S□ Rohland N□ han □ enschoreck T□ ebster T□ Reich □ 2012. Ancier admixture in human histo*Genetics* **192**1065□1093. OI□ https□doi.org10.1534□genetics.112.145037 □PI□ 22960212

Pease JB□ aak C□ ahn □ oyle LC. 2016. Phylogenomics reveals three sources of adaptive variation during a rapid radiation. *PLOS Biology* **14**e1002379. OI□ https□doi.org10.1371□ournal.pbio.1002379 □PI□ 26871574

Pease JB□ ahn □ 2015. etection and polariation of introgression in a five-taxon phylogeny. *Systematic Biology* **64**651□662. OI□ https□doi.org10.1093□sysbio□syv023 □PI□ 25888025

Pollard DA□ Iyer N□ oses A□ Eisen B. 2006. idespread discordance of gene trees with species tree in *Drosophila*□ evidence for incomplete lineage sorting. *PLOS Genetics* **2**e173. OI□ https□doi.org10.1371□ournal.pgen.0020173 □PI□ 17132051

Rambaut A□ rassly NC. 1997. Se-en□ an application for the onte Carlo simulation of NA seuence evolution along phylogenetic trees. *Bioinformatics* **13**235□238. OI□ https□doi.org10.1093□bioinformatics13.3.235

Rausher MD. 2008. Evolutionary transitions in floral *International Journal of Plant Sciences* **169**7□21. OI□ https□doi.org10.1086523358

Reich D□ Thangara□ □ Patterson N□ Price AL□ Singh L. 2009. Reconstructing indian popu*Nature* **461**489□494. OI□ https□doi.org10.1038□nature08365 □PI□ 19779445

Rodriguez ZB□ Perkins SL□ Austin CC. 2018. ultiple origins of green blood in New *Scientific Reports* **Advances 4**eaao5017. OI□ https□doi.org10.1126□sciadv.aao5017 □PI□ 29774232

Schrempf D□ Sollo si □ 2020. The sources of phylogenetic conflicts. In□ Celine S□ elsuc □ Nicolas □ (Ed *Phylogenetics in the Genomic Era*. Self published. p. 3.11□ □3.123.

Smith SD□ Rausher □ 2011. ene loss and parallel evolution contribute to species difference in flower c *Molecular Biology and Evolution* **28**2799□2810. OI□ https□doi.org10.1093□molbev□msr109 □PI□ 21551271

Solís-Lemus C□ Ané C. 2016. Inferring phylogenetic networks with maximum pseudolikelihood under incom lineage sorting. *PLOS Genetics* **12**e1005896. OI□ https□doi.org10.1371□ournal.pgen.1005896 □PI□ 26950302

Stamatakis A. 2014. RAxL version 8□ a tool for phylogenetic analysis and post-analysis of large phylog *Bioinformatics* **30**1312□1313. OI□ https□doi.org10.1093□bioinformatics□btu033 □PI□ 24451623

Storz JF□ Runck A□ Sabatino S□ elly □ errand N□ oriyama □ eber RE□ ago A. 2009. Evolutionary and functional insights into the mechanism underlying high-altitude adaptation of deer mouse *PNAS* moglobin. **106**14450□14455. OI□ https□doi.org10.1073□pnas.0905224106 □PI□ 19667207

Storz JF. 2016. Causes of molecular convergence and parallelism in protein *Nature Reviews Genetics* **17**239□250. OI□ https□doi.org10.1038□nrg.2016.11 □PI□ 26972590

Tufts DM□ Nataraan C□ Revsbech I□ Proecto-arcia □ offmann □ eber RE□ ago A□ oriyama □ Stor□ □ 2015. Epistasis constrains mutational pathways of hemoglobin adaptation in high-altitude pikas. *Molecular Biology and Evolution* **32**287□298. OI□ https□doi.org10.1093□molbevmsu311 □PI□ 25415962

Vanderpool D□ inh B□ Lanfear R□ ughes □ urali S□ arris RA□ Raveendran □ uny □ ibbins S□ illiamson R□ ibbs RA□ orley C□ Rogers □ ahn □ 2020. Primate phylogenomics uncovers multiple rapid radiations and ancient interspecific introgression. *PLOS Biology* **18**e3000954. OI□ https□doi.org10.1371□urnal.pbio.3000954 □PI□ 33270638

Wang Y□ Cao □ Ogilvie □ Nakhleh L. 2020. Phylogenomic assessment of the role of hybridiation and introgression in trait evolution. *bioRxiv*. OI□ https□doi.org10.11012020.09.16.300343

Wen D□ ü □ Nakhleh L. 2016. Bayesian inference of reticulate phylogenies under the multispecies netwo coalescent. *PLOS Genetics* **12**e1006006. OI□ https□doi.org10.1371□urnal.pgen.1006006 □PI□ 27144273

Wen D□ ü □ hu □ Nakhleh L. 2018. Inferring phylogenetic networks using *PhyloNet*. *Systematic Biology* **67**□ 735□740.OI□ https□doi.org10.1093□sysbiosyy015 □PI□ 29514307

Wu M□ ostyun L□ ahn □ oyle LC. 2018. issecting the basis of novel trait evolution in a radiation with widespread phylogenetic discordance. *Molecular Ecology* **27**3301□3316. OI□ https□doi.org10.1111□mec.14780

Yu Y□ egnan □ Nakhleh L. 2012. The probability of a gene tree topology within a phylogenetic networ applications to hybridiation detection. *PLOS Genetics* **8**e1002660. OI□ https□doi.org10.1371□urnal.pgen.1002660□PI□ 22536161

Yu Y□ ong □ Liu □ Nakhleh L. 2014. aximum likelihood inference of reticulate evolutionary histories. *PNAS* **111**16448□16453. OI□ https□doi.org10.1073□pnas.1407950111 □PI□ 25368173

Zhang C□ Ogilvie A□ rummond A□ Stadler T. 2018a. Bayesian inference of species networks from multiloc seuence data. *Molecular Biology and Evolution* **35**504□517. OI□ https□doi.org10.1093□molbevmsx307 □PI□ 29220490

Zhang C□ Rabiee □ Sayyari E□ irarab S. 2018b. ASTRAL-III□ polynomial time species tree reconstruction fr partially resolved gene trees. *BMC Bioinformatics* **19**153. OI□ https□doi.org10.1186□12859-018-2129-y □PI□ 29745866

Appendix 1

1 Mutation probabilities on genealogies

Each of the twelve possible genealogies under our parent tree model has a set of five along which mutations can occur. λ_1, λ_2, and λ_3 denote the tip branches leading to species A, B, and C respectively. λ_4 denotes the internal branch, and λ_5 denotes the ancestral branch. As described in the supplement of *Guerrero and Hahn, 2018*, the mutation probability on each of these branches has the general form $\int_0^\infty e^{-\mu x} f(x)dx$, where μ is the mutation probability per generations, x is the random variable for the branch length, and f(x) is the probability density function for . with the mutation probabilities for parent tree 1, which are found in the supplement of Guhahn, and will be re-written here to be consistent with notation. In the following, parent tree 1 will be denoted as $pt1$. Since many of the genealogies are identical in mutation probabilities on their branches can be written with general expressions. The first the genealogies $AB2_1$, BC_1, and AC_1, which are all produced via incomplete lineage sorting in p tree 1, and share the following set of mutation probabilities:

$$\nu_1[ILS,pt1] = \frac{1}{\Lambda}\int_0^{t_3-t_2}(1-e^{-\mu(t_1+(t_2-t_1)+x)})\frac{3}{2}(e^{-x}-e^{-3x})dx \quad (1)$$

$$\nu_2[ILS,pt1] = \frac{1}{\Lambda}\int_0^{t_3-t_2}(1-e^{-\mu(t_1+(t_2-t_1)+x)})3e^{-3x}(1-e^{-((t_3-t_2)-x)})dx \quad (2)$$

$$\nu_4[ILS,pt1] = \frac{1}{\Lambda_0}\int_0^{t_3-t_2}3e^{-3y}\left(\int_0^{(t_3-t_2)-y}e^{-x}(1-e^{-\mu x})dx\right)dy \quad (3)$$

$$\nu_5[ILS,pt1] = \frac{1}{\Lambda}\int_0^{t_3-t_2}(1-e^{-\mu((t_3-t_2)-x)})3e^{-3x}dx \quad (4)$$

In each of the above, $\Lambda = 1+\frac{1}{2}e^{-3(t_3-t_2)}-\frac{3}{2}e^{-(t_3-t_2)}$ is the probability of coalescence of A, B, and in their ancestral population, and t_3 denotes the total height of the tree, i.e. the time at the bas tree. The difference between t_3 and t_2 determines the duration of the ancestral population of three taxa, before speciation occurs. *Equations 1* through four each represent the mutation probabilities for multiple branches, which are as follows:

$$\nu_1[ILS,pt1] = \nu(\lambda_3,AB2_1) = \nu(\lambda_1,BC_1) = \nu(\lambda_2,AC_1) \quad (5)$$

$$\begin{aligned}\nu_2[ILS,pt1] &= \nu(\lambda_1,AB2_1) = \nu(\lambda_2,AB2_1) = \\ \nu(\lambda_2,BC_1) &= \nu(\lambda_3,BC_1) = \nu(\lambda_1,AC_1) = \nu(\lambda_3,AC_1)\end{aligned} \quad (6)$$

$$\nu_4[ILS,pt1] = \nu(\lambda_4,AB2_1) = \nu(\lambda_4,BC_1) = \nu(\lambda_4,AC_1) \quad (7)$$

$$\nu_5[ILS,pt1] = \nu(\lambda_5,AB2_1) = \nu(\lambda_5,BC_1) = \nu(\lambda_5,AC_1) \quad (8)$$

The gene tree produced by lineage sorting in parent tree 1 has a different set of mutation probabilities, since the branches have different expected lengths. These are:

$$\nu(\lambda_1,AB1_1) = \nu(\lambda_2,AB1_1) = \frac{1}{1-e^{-(t_2-t_1)}}\int_0^{t_2-t_1}(1-e^{-\mu(t_1+x)})e^{-x}dx \quad (9)$$

$$\nu(\lambda_3,AB1_1) = \frac{1}{1-e^{-(t_3-t_2)}}\int_0^{t_3-t_2}(1-e^{-\mu(t_1+(t_2-t_1)+x)})e^{-x}dx \quad (10)$$

$$\nu(\lambda_4,AB1_1) = \int_0^{t_2-t_1}\frac{e^{-y}}{1-e^{-(t_2-t_1)}}\left(\int_0^{t_3-t_2}(1-e^{-\mu((t_2-t_1)-y+x)})\frac{e^{-x}}{1-e^{-(t_3-t_2)}}dx\right)dy \quad (11)$$

$$\nu(\lambda_5, AB1_1) = \int_0^{t_2-t_1} \frac{1}{1-e^{-(t_3-t_2)}} \int_0^{t_3-t_2} (1-e^{-\mu((t_3-t_2)-x)}) e^{-x} dx \quad (12)$$

Now we consider introgression starting with parent tree 2. Many of the mutation probabilities are symmetrical with parent tree 1 and therefore remain the same and the remainder have a general form with different parameters. For the ILS genealogies $BC2_2$, AB_2 and AC_2 (**Equations 1 and 2**) have the time of A-B speciation (t_2) replaced with the timing of B-C introgression (t_m). This gives:

$$\nu_1[ILS, pt2 = \frac{1}{\Lambda} \int_0^{t_3-t_2} (1-e^{-\mu(t_m+(t_2-t_m)+x)}) \frac{3}{2} (e^{-x} - e^{-3x}) dx \quad (13)$$

$$\nu_2[ILS, pt2 = \frac{1}{\Lambda} \int_0^{t_3-t_2} (1-e^{-\mu(t_m+(t_2-t_m)+x)}) 3e^{-3x} (1-e^{-((t_3-t_2)-x)}) dx \quad (14)$$

$$\nu_4[ILS, pt2 = \nu_4[ILS, pt1 \quad (15)$$

$$\nu_5[ILS, pt2 = \nu_5[ILS, pt1 \quad (16)$$

These correspond to the following branch mutation probabilities:

$$\nu_1[ILS, pt2 = \nu(\lambda_1, BC2_2) = \nu(\lambda_3, AB_2) = \nu(\lambda_2, AC_2) \quad (17)$$

$$\begin{aligned} \nu_2[ILS, pt2 &= \nu(\lambda_2, BC2_2) = \nu(\lambda_3, BC2_2) = \\ \nu(\lambda_1, AB_2) &= \nu(\lambda_2, AB_2) = \nu(\lambda_1, AC_2) = \nu(\lambda_3, AC_2) \end{aligned} \quad (18)$$

$$\nu_4[ILS, pt2 = \nu(\lambda_4, BC2_2) = \nu(\lambda_4, AB_2) = \nu(\lambda_4, AC_2) \quad (19)$$

$$\nu_5[ILS, pt2 = \nu(\lambda_5, BC2_2) = \nu(\lambda_5, AB_2) = \nu(\lambda_5, AC_2) \quad (20)$$

For the genealogy produced by lineage sorting in parent tree 2 ($BC1_2$) we have:

$$\nu(\lambda_2, BC1_2) = \nu(\lambda_3, BC1_2) = \frac{1}{1-e^{-(t_2-t_m)}} \int_0^{t_2-t_m} (1-e^{-\mu(t_m+x)}) e^{-x} dx \quad (21)$$

$$\nu(\lambda_1, BC1_2) = \frac{1}{1-e^{-(t_3-t_2)}} \int_0^{t_3-t_2} (1-e^{-\mu(t_m+(t_2-t_m)+x)}) e^{-x} dx \quad (22)$$

$$\nu(\lambda_4, BC1_2) = \int_0^{t_2-t_m} \frac{e^{-y}}{1-e^{-(t_2-t_m)}} \left(\int_0^{t_3-t_2} (1-e^{-\mu((t_2-t_m)-y+x)}) \frac{e^{-x}}{1-e^{-(t_3-t_2)}} dx \right) dy \quad (23)$$

$$\nu(\lambda_5, BC1_2) = \nu(\lambda_5, AB1_1) \quad (24)$$

Finally we consider parent tree 3. The mutation probabilities have the same formulation as tree 2 with two key changes since parent tree 3 is shorter (t_2 in text) is replaced by t_1. This also applies to the value of Λ which we will denote for parent tree 3 $\Lambda_3 = 1 + \frac{1}{2}e^{-3(t_3-t_1)} - \frac{3}{2}e^{-(t_3-t_1)}$. For the ILS genealogies $BC2_3$, AB_3 and AC_3 this gives:

$$\nu_1[ILS, pt3 = \frac{1}{\Lambda_3} \int_0^{t_3-t_1} (1-e^{-\mu(t_m+(t_1-t_m)+x)}) \frac{3}{2} (e^{-x} - e^{-3x}) dx \quad (25)$$

$$\nu_2[ILS, pt3 = \frac{1}{\Lambda_3} \int_0^{t_3-t_1} (1-e^{-\mu(t_m+(t_1-t_m)+x)}) 3e^{-3x} (1-e^{-((t_3-t_1)-x)}) dx \quad (26)$$

$$\nu_4[ILS,pt3 = \frac{1}{\Lambda_3}\int_0^{t_3-t_1} 3e^{-3y}(\int_0^{(t_3-t_1)-y} e^{-x}(1-e^{-\mu x})dx)dy \tag{27}$$

$$\nu_5[ILS,pt3 = \frac{1}{\Lambda_3}\int_0^{t_3-t_1} (1-e^{-\mu((t_3-t_1)-x)})3e^{-3x}dx \tag{28}$$

here

$$\nu_1[ILS,pt3 = \nu(\lambda_1,BC2_3) = \nu(\lambda_3,AB_3) = \nu(\lambda_2,BC_3) \tag{29}$$

$$\begin{aligned}\nu_2[ILS,pt3 &= \nu(\lambda_2,BC2_3) = \nu(\lambda_3,BC2_3) = \\ \nu(\lambda_1,AB_3) &= \nu(\lambda_2,AB_3) = \nu(\lambda_1,AC_3) = \nu(\lambda_3,AC_3)\end{aligned} \tag{30}$$

$$\nu_4[ILS,pt3 = \nu(\lambda_4,BC2_3) = \nu(\lambda_4,AB_3) = \nu(\lambda_4,AC_3) \tag{31}$$

$$\nu_5[ILS,pt3 = \nu(\lambda_5,BC2_3) = \nu(\lambda_5,AB_3) = \nu(\lambda_5,AC_3) \tag{32}$$

inally for the genealogy $BC1_3$ the mutation probabilities are as follows

$$\nu(\lambda_2,BC1_3) = \nu(\lambda_3,BC1_3) = \frac{1}{1-e^{-(t_1-t_m)}}\int_0^{t_1-t_m}(1-e^{-\mu(t_m+x)})e^{-x}dx \tag{33}$$

$$\nu(\lambda_1,BC1_3) = \frac{1}{1-e^{-(t_3-t_1)}}\int_0^{t_3-t_1}(1-e^{-\mu(t_m+(t_1-t_m)+x)})e^{-x}dx \tag{34}$$

$$\nu(\lambda_4,BC1_3) = \int_0^{t_1-t_m}\frac{e^{-y}}{1-e^{-(t_1-t_m)}}(\int_0^{t_3-t_1}(1-e^{-\mu((t_1-t_m)-y+x)})\frac{e^{-x}}{1-e^{-(t_3-t_1)}}dx)dy \tag{35}$$

$$\nu(\lambda_5,BC1_3) = \frac{1}{1-e^{-(t_3-t_1)}}\int_0^{t_3-t_1}(1-e^{-\mu((t_3-t_1)-x)})e^{-x}dx \tag{36}$$

2 When does introgression makes hemiplasy more likely than ILS alone?

The probability of hemiplasy C with B introgression is

$$P_e = (1-\delta)P_e[BC_1 + \delta(P_e[BC1_2 + P_e[BC2_2]) \tag{37}$$

rom this it can be seen that introgression makes hemiplasy more likely than ILS alone

$$P_e[BC1_2 + P_e[BC2_2 \; P_e[BC_1 \tag{38}$$

hen is this true Substituting the relevant expressions from the main text gives

$$\begin{aligned}&(1-e^{-(t_2-t_m)})\nu(\lambda_4,BC1_2)\prod_{i\neq 4}(1-\nu(\lambda_i,BC1_2))+ \\ &(\frac{1}{3}e^{-(t_2-t_m)})\nu(\lambda_4,BC2_2)\prod_{i\neq}(1-\nu(\lambda_i,BC2_2)) \\ &(\frac{1}{3}e^{-(t_2-t_1)})\nu(\lambda_4,BC_1)\prod_{i\neq}(1-\nu(\lambda_i,BC_1))\end{aligned} \tag{39}$$

$$(1-e^{-(t_2-t_m)})v(\lambda_4,BC1_2)\prod_{i\neq 4}(1-v(\lambda_i,BC1_2))\square$$
$$(\frac{1}{3}e^{-(t_2-t_1)})v(\lambda_4,BC_1)\prod_{i\neq}(1-v(\lambda_i,BC_1))- \quad (40)$$
$$(\frac{1}{3}e^{-(t_2-t_m)})v(\lambda_4,BC2_2)\prod_{i\neq}(1-v(\lambda_i,BC2_2))$$

The mutation probabilities on the right side of the ineuality are eual since they are c topology with the same branch lengths. Therefore, Equation 40 can simplified

$$(1-e^{-(t_2-t_m)})v(\lambda_4,BC1_2)\prod_{i\neq 4}(1-v(\lambda_i,BC1_2))\square$$
$$v(\lambda_4,BC_1)\prod_{i\neq 4}(1-v(\lambda_i,BC_1))(\frac{1}{3}e^{-(t_2-t_1)}-\frac{1}{3}e^{-(t_2-t_m)}) \quad (41)$$

As the most conservative case, let us assume a hybrid speciation t_1 scenario. This where t_m represents the most conservative introgression scenario since Figure 4 in the main text shows that more recent introgression makes hemiplasy more likely. In this case, the right side of th simplifies to 0, leaving

$$(1-e^{-(t_2-t_m)})v(\lambda_4,BC1_2)\prod_{i\neq 4}(1-v(\lambda_i,BC1_2))\square 0 \quad (42)$$

This is true whenever $t_2>t_m$, which is true by definition in this model. Therefore, introgres always makes hemiplasy more likely than ILS alone.