doi:10.1002/evl3.77

Speciation genes are more likely to have discordant gene trees

Richard J. Wang[1,2] and Matthew W. Hahn[1,3]

[1]*Department of Biology, Indiana University, Bloomington, Indiana*

[2]*E-mail: rjwang@indiana.edu*

[3]*Department of Computer Science, Indiana University, Bloomington, Indiana*

Received March 29, 2018
Accepted July 6, 2018

Speciation genes are responsible for reproductive isolation between species. By directly participating in the process of speciation, the genealogies of isolating loci have been thought to more faithfully represent species trees. The unique properties of speciation genes may provide valuable evolutionary insights and help determine the true history of species divergence. Here, we formally analyze whether genealogies from loci participating in Dobzhansky–Muller (DM) incompatibilities are more likely to be concordant with the species tree under incomplete lineage sorting (ILS). Individual loci differ stochastically from the true history of divergence with a predictable frequency due to ILS, and these expectations—combined with the DM model of intrinsic reproductive isolation from epistatic interactions—can be used to examine the probability of concordance at isolating loci. Contrary to existing verbal models, we find that reproductively isolating loci that follow the DM model are often more likely to have discordant gene trees. These results are dependent on the pattern of isolation observed between three species, the time between speciation events, and the time since the last speciation event. Results supporting a higher probability of discordance are found for both derived–derived and derived–ancestral DM pairs, and regardless of whether incompatibilities are allowed or prohibited from segregating in the same population. Our overall results suggest that DM loci are unlikely to be especially useful for reconstructing species relationships, even in the presence of gene flow between incipient species, and may in fact be positively misleading.

KEY WORDS: **Dobzhansky–Muller incompatibilities, gene tree discordance, incomplete lineage sorting, reproductive isolation, speciation.**

Impact Summary

The variety of species in nature is kept distinct by the barriers that prevent interbreeding. New species form when genetic changes within different populations prevent them from reproducing. The genetic analysis of reproductive incompatibilities has revealed the identity of genes responsible for reproductive isolation—and thus, speciation—in a number of species. These genes may have a unique pattern of evolution because of their participation in the process of speciation. It has been hypothesized that the evolutionary histories of speciation genes could be especially useful for determining the order in which species diverged. In this article, we formally analyze this hypothesis by combining the prevailing genetic model of speciation with population genetic theory. We find that genetic loci responsible for reproductive isolation do have a unique signal, but in a way that can often be misleading about the order in which species diverged. Our findings contradict existing models and provide a new expectation for the evolutionary history of speciation genes.

Speciation proceeds from the evolution of reproductive isolation between populations. The study of reproductive isolation has advanced our understanding of the genetic basis of speciation for which a common evolutionary model has become established.

Evolution Letters

The Dobzhansky–Muller (DM) model describes how hybrid incompatibilities can arise as the result of epistasis between two or more loci that have diverged between populations (Bateson 1909; Dobzhansky and Dobzhansky 1937; Muller 1942). By having incompatible alleles for these loci arise in separate populations, the DM model allows reproductive isolation to evolve between populations without the appearance of reproductive failure within populations. A growing number of so-called "speciation genes" that isolate species in accordance with the DM model have emerged from the genetic analysis of reproductive isolation in hybrids (e.g., Ting et al. 1998; Barbash et al. 2003; Presgraves et al. 2003; Bomblies and Weigel 2007; Mihola et al. 2009; Phadnis and Orr 2009; Barr and Fishman 2010; Lienard et al. 2016). Combinations of alleles from different species at these genes cause hybrid infertility or inviability.

The identification of speciation genes in multiple model systems has led to a search for the genetic, molecular, and evolutionary commonalities among them (Orr et al. 2004; Wu and Ting 2004; Oliver et al. 2009; Presgraves 2010; Rieseberg and Blackman 2010; Nosil and Schluter 2011; Castillo and Barbash 2017). A major question is whether the genes leading to reproductive isolation differ from other genes in the genome. Various hypotheses have suggested that speciation genes are more likely to be the targets of adaptive evolution (Coyne and Orr 2004), more prone to interact with other genes (Guerrero et al. 2017), or more likely to be involved in genetic conflict (Bomblies and Weigel 2007; Phadnis and Orr 2009; Ågren 2013).

The unique role of speciation genes in establishing species boundaries has also led to arguments asserting that these genes should be especially informative about species relationships (Ting et al. 2000; Rosenberg 2003; Maroja et al. 2009; Zachos 2009; Nosil and Schluter 2011; Cutter 2013). Such a property becomes useful when multiple species are separated by very short times between successive speciation events. In these cases, individual gene trees may have different topologies from one another and from the species tree (Maddison 1997). This phenomenon is not due to low power or sampling error, but represents a real difference in the genealogical history between loci, due to incomplete lineage sorting (ILS) or gene flow. With the high degree of discordance seen in many systems (e.g., Pollard et al. 2006; White et al. 2009; Jarvis et al. 2014; Pease et al. 2016), concordance between the topology of speciation genes and species trees would provide uniquely powerful insight into evolutionary histories. Verbal models have created the impression that speciation loci are biased toward concordance (Ting et al. 2000; Nosil and Schluter 2011; Cutter 2013), but no formal analysis of this idea has been carried out.

Here, we compare the expected genealogical history of loci involved in Dobzhansky-Muller Incompatibilities (DMIs) to the expected history of loci uninvolved in incompatibilities from the same genomes. The appreciation of discordance among gene trees has become acute with whole-genome sequence data, leading to multiple methods that incorporate ILS in the inference of species trees (e.g., Liu et al. 2009; Larget et al. 2010; Drummond et al. 2012; Mirarab and Warnow 2015). However, gene tree discordance has received limited consideration in the context of DMIs. Loci involved in DMIs present additional challenges because of the epistatic nature of incompatibilities, which means that participating loci must act together to produce the incompatible phenotype. In addition, because both alleles involved in an incompatibility cannot segregate in the same population without leading to lower fitness in some individuals, the order in which mutations arise at each locus in a DMI matters.

We find that under a neutral model with ILS, the stochastic processes of mutation and coalescence typically lead to higher rates of species tree discordance at hybrid incompatibility loci. We arrive at this counterintuitive result by examining the probability of ILS at loci participating in a canonical two-locus DMI. Our analysis considers four potential types of gene trees at a hypothetical incompatibility locus (Fig. 1). A key initial insight is recognizing the possibility that incompatible alleles can arise on discordant gene trees and still lead to reproductive isolation between pairs of species. Figure 2 shows how a DMI can arise from two loci, both with discordant gene trees, and isolate one or more species pairs. Because the expected branch lengths for each type of gene tree differ, mutations giving rise to incompatible alleles are not equally likely among the types of gene trees. We consider each combination of topologies for a pair of loci and calculate the probability of a DMI from differences in expected branch lengths. We find that loci participating in DMIs are typically more likely to have discordant gene trees, and that some patterns of isolation between species are more likely when loci are discordant.

Results

PRELIMINARIES

Our genealogical model considers a single pairwise DMI in a three-species complex. DMIs are typically modeled as isolating two taxa, but depending on where an incompatible allele arises on a phylogenetic tree, a DMI can be shared among different species pairs (Moyle and Payseur 2009). The most straightforward way for this to occur is to have an interaction between two derived alleles ("derived–derived" incompatibilities; Orr 1995), where one of the derived alleles is shared between species, having arisen before their divergence (Fig. 2B). Two mutations inherited by the same lineage can also result in shared isolation, with the second derived allele being incompatible with the ancestral allele in other taxa ("derived–ancestral" incompatibilities; Orr 1995; Fig. S1). Generally, a DMI involving only two loci can produce six patterns of isolation among three taxa. The topology and branch lengths for

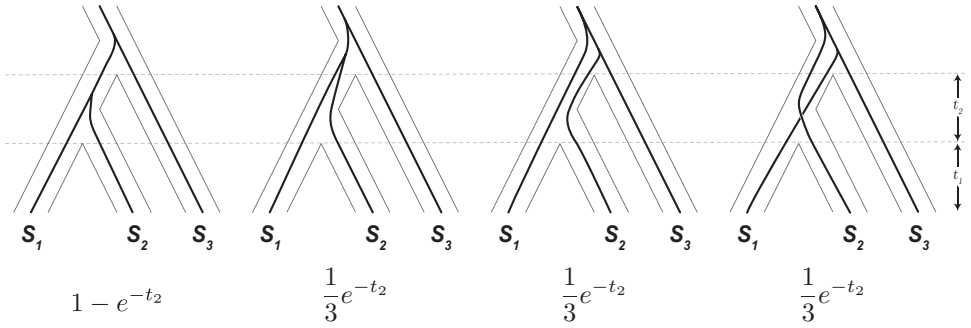

Figure 1. **Four types of gene trees at a DMI locus and their expected frequencies. Two gene trees concordant with the species tree (left), and two that are discordant (right). Only the leftmost gene tree coalesces before the first speciation event. The labeled times, t_1 and t_2, are the time from present to the first speciation event and the time between speciation events, respectively. Below each gene tree is the probability of its occurrence for a random locus in the genome under ILS alone.**

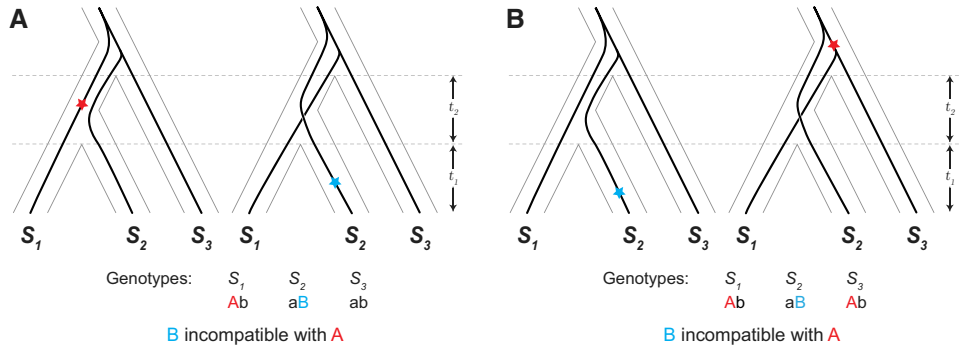

Figure 2. **Incompatibility between one or more species pairs due to alleles from two loci that both have discordant gene trees, (S_2, S_3) S_1 and (S_1, S_3) S_2. Red and blue stars mark the position of the first and second mutations, respectively, that produce incompatible alleles. The ancestral genotype for the two loci is denoted "ab", with the mutations producing derived alleles "A" and "B". (A) Incompatibility between lineages S_1 and S_2. (B) Incompatibility between lineages S_1 and S_2 as well as S_3 and S_2 from shared incompatible alleles.**

the two most recently diverged taxa are interchangeable in many of the subsequent calculations, leaving four unique patterns of reproductive isolation (Fig. 3). As we show below, these patterns of reproductive isolation are more often associated with particular types of gene trees.

We allow loci participating in an incompatibility to be discordant with the species tree, defined by the historical timing and order of species' divergence, only through ILS. Specifically, a DMI locus can have one of four potential types of gene trees (Fig. 1). Although there are only three potential topologies for three species, we divide concordant gene trees into those that coalesce in the ancestral population and those that coalesce between the speciation events (i.e., are lineage-sorted). Discordant gene trees must coalesce in the ancestral population of all three species. For a single locus, each of the three ancestrally coalescing gene trees are equally likely, at $\frac{1}{3}e^{-t_2}$, where t_2 is the interspeciation time (Hudson 1983; Nei 1986). This leaves the probability of a concordant, lineage-sorted gene tree at $1 - e^{-t_2}$. The more familiar probability for gene-tree/species-tree concordance, $1 - \frac{2}{3}e^{-t_2}$ (Hudson 1983), includes another $\frac{1}{3}e^{-t_2}$ from concordant trees that coalesce in the ancestral population of all three species. For two independent loci, the joint probability of any particular pair of genealogies is a product of the individual probabilities. However, this is not the case for two loci participating in a DMI.

CALCULATING GENE-TREE/SPECIES-TREE CONCORDANCE AT A DMI LOCUS

For a given pattern of reproductive isolation, certain pairs of gene trees are more likely to give rise to incompatible alleles. Loci participating in a DMI must have experienced mutations on the appropriate branches to form the corresponding pattern of isolation. As an obvious example, a mutation specific to the S_3 lineage, on any of the gene trees shown in Figure 1, cannot participate in an incompatibility that isolates S_1 from S_2. In the standard coalescent model, the probability of a mutation on a given branch is proportional to its length and independent of the coalescent process. Because branch lengths differ among gene trees, the probability of a DMI depends on the types of gene trees at a pair of loci. Conversely, the probability of a specific pair of gene trees at the two loci involved in a DMI (DMI loci) depends on the pattern of

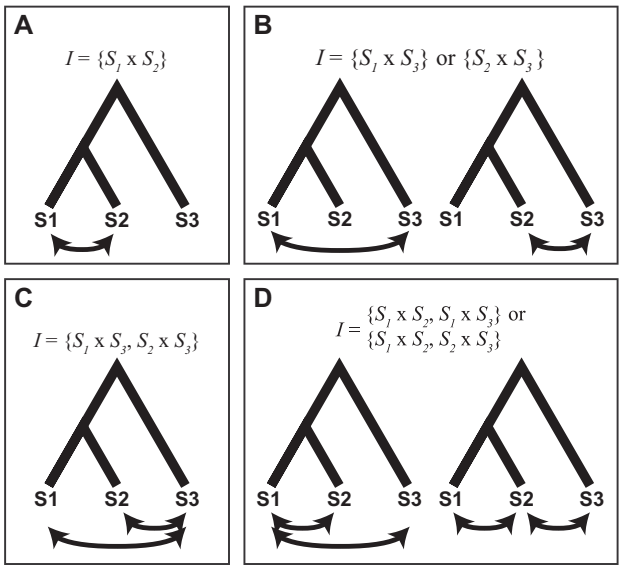

Figure 3. Four patterns of reproductive isolation. A single pairwise DMI can isolate a pair of species, as in panels (A) and (B), or two pairs of species, as in panels (C) and (D). For subsequent calculations, S_1 and S_2 are often interchangeable, leading us to group the two sets of relationships in (B) and (D).

isolation. We can express the relationship between the probability of each pair of gene trees and the probability of an incompatibility through Bayes' theorem.

Let I be the species pair(s) for which a DMI manifests—that is, I specifies the pattern of reproductive isolation between species for a DMI (Fig. 3). The probability that a pair of DMI loci have gene trees of type T_x and T_y, respectively, can be expressed as:

$$P(T_x, T_y|I) = \frac{P\left(I|T_x, T_y\right) P(T_x)\, P\left(T_y\right)}{\sum_{n,m} P(I|T_n, T_m) P(T_n)\, P(T_m)}, \quad (1)$$

where n and m are each indices enumerating the four types of gene trees as ordered in Figure 1 (i.e., T_1 and T_2 are concordant, whereas T_3 and T_4 are discordant), and $P(T_x)P(T_y)$ is the joint probability of T_x and T_y assuming independence as described above.

Assuming incompatibilities are rare between any given pair of loci, the conditional probability $P(I|T_x, T_y)$ can be written as a sum of the probability that two mutations result in an incompatibility, across all branches of T_x and T_y. Let x_α and y_β be indexed branches on trees T_x and T_y, then,

$$P\left(I|T_x, T_y\right) = p \sum_{\alpha,\beta} P(\text{mutation on } x_\alpha)\, P\left(\text{mutation on } y_\beta\right) \times \mathbf{1}(x_\alpha, y_\beta), \quad (2)$$

where $\mathbf{1}(x_\alpha, y_\beta)$ is 1 when two mutations, on branches x_α and y_β, can generate a DMI with isolation pattern I, and 0 otherwise; and p is the probability of an incompatibility forming between untested allelic combinations (Orr 1995; Orr and Turelli 2001). This probability is valid when the mutations are independent, but the order in which mutations occur must be considered for derived-ancestral incompatibilities (see Methods). The probability of at least one mutation on a given branch can be estimated by its length,

$$P(\text{mutation on } x_\alpha) \approx 2N_e\mu \cdot L\,(x_\alpha)\,, \quad (3)$$

where $2N_e\mu$ is the population mutation parameter and $L(x_\alpha)$ denotes the branch length of x_α (see Supporting Information Methods and Hudson 1992). Under an infinite sites model, the expression in equation (3) is equivalent to the probability of observing a derived allele. The following calculations assume such a model, with the probability of a derived allele on branches x_α and y_β calculated from their respective branch lengths.

With the probability of each pair of genealogies for a given pattern of reproductive isolation, we can calculate the probability that the gene tree for a single DMI locus is concordant with the species tree by summing the marginal probabilities for a concordant topology,

$$P(\text{concordance} \mid I) = \frac{1}{2} \sum_{C,m} P(T_C, T_m|I) + \frac{1}{2} \sum_{n,C} P(T_n, T_C|I)\,, \quad (4)$$

where n and m are indices over the four types of gene trees as before, and C includes only the indices for the concordant trees (i.e., T_1 and T_2). The probability of discordance can similarly be calculated from a sum of marginal probabilities.

MUTATIONS THAT ISOLATE SISTER TAXA VIA TWO LOCI WITH DISCORDANT GENE TREES

Determining the limited branch segments on which incompatible alleles for a particular pattern of isolation can arise is central to the calculation of the conditional probability in equation (2). In Figure 4, we illustrate the branch segments on which an incompatible allele isolating the sister taxa, $S_1 \times S_2$, can arise on two discordant gene trees, $(S_2, S_3)\, S_1$ and $(S_1, S_3)\, S_2$. (For the segments on all gene tree pairs, see Appendix 1 in Supporting Information.) We divide segments on the gene trees in Figure 4 by speciation and coalescent events, labeling each segment by its endpoints (e.g., *a–d* describes the segment specific to S_1 from the present to the most recent speciation event). Two mutations on the highlighted segments in each pair of gene trees in Figure 4, one on the left-hand tree and one on the right-hand tree, can produce alleles isolating S_1 and S_2. For example, Figure 4A shows the potential for an $S_1 \times S_2$ incompatibility from a derived allele that arises on the *a–d* segment of the left-hand tree (inherited by S_1) and a derived allele that arises on segments *b–e*, *e–g*, or *g–k* (inherited by S_2). Because we do not allow incompatibilities to arise before

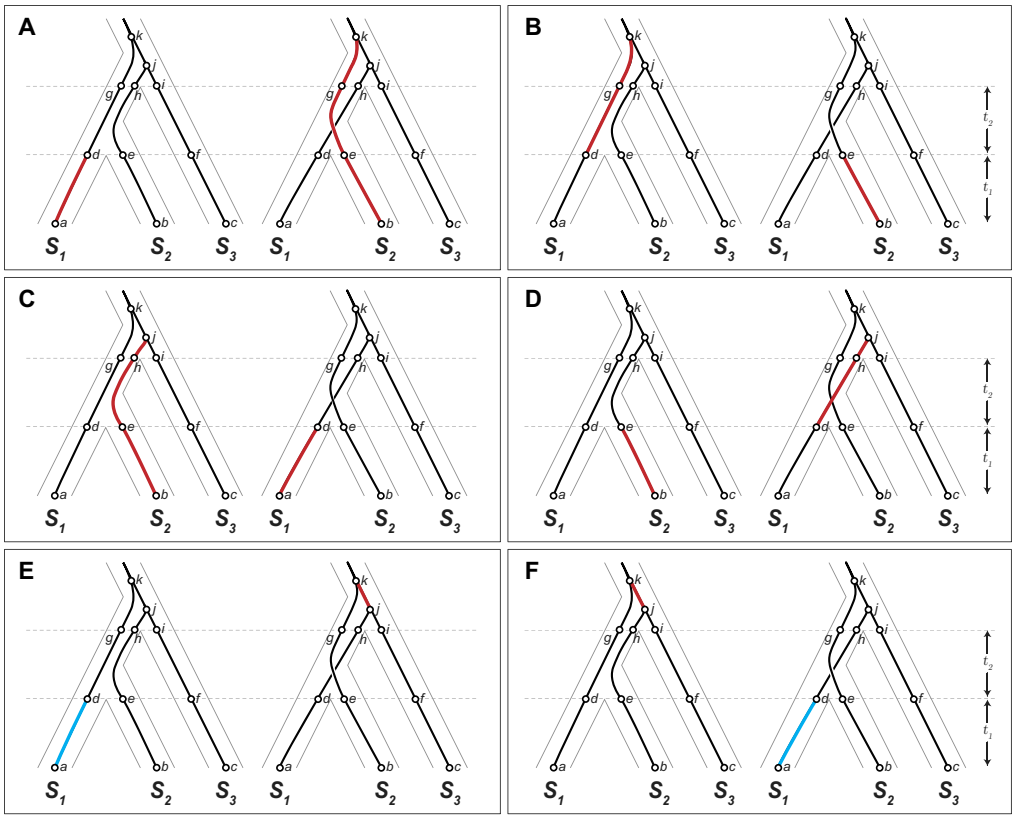

Figure 4. Branch segments on the discordant gene trees, (S_2, S_3) S_1 and (S_1, S_3) S_2, that can give rise to incompatible alleles isolating S_1 and S_2. In (A)–(D), a mutation on the left-hand tree, (S_2, S_3) S_1, and a mutation on the right-hand tree, (S_1, S_3) S_2, can give rise to a derived–derived incompatibility. These panels show all combinations of branch segments on which mutations could give rise to a derived–derived incompatibility on the pair of discordant trees shown. Panels (E) and (F) show all combinations of branch segments on which mutations could give rise to a derived-ancestral incompatibility on this pair of trees. A mutation on branch segment *j–k* (red) leaves S_2 with an ancestral allele that can be incompatible with a derived allele produced by a mutation on branch segment *a–d* (blue) and inherited by S_1 (see main text).

S_1 and S_2 diverge, one mutation must occur on a segment after divergence (i.e., a pair of incompatible alleles cannot arise, for instance, along segment *d–g* on the left-hand tree and *e–g* on the right-hand tree).

Interestingly, loci with discordant gene trees can also produce derived–ancestral incompatibilities between sister taxa (Fig. 4E and F). In Figure 4E, a mutation on segment *j–k* on the right-hand tree produces a derived allele inherited by S_1 and S_3. The second derived allele on segment *a–d* on the left-hand tree is inherited by S_1, but arises in the background of the derived allele from the first mutation, creating the potential for an incompatibility with the ancestral allele on S_2. A derived-ancestral incompatibility between sister taxa is only possible in a model that allows incompatible alleles to arise before divergence.

GENE-TREE/SPECIES-TREE DISCORDANCE IS MORE LIKELY AT LOCI ISOLATING SISTER TAXA

The probability of concordance between the species tree and gene trees at DMI loci depends on the pattern of reproductive isolation considered. Figure 5 shows the probability of concordance for a DMI locus participating in each of the four possible patterns of isolation in a three-species complex. The values depicted represent a ratio of the probability for gene-tree/species-tree concordance at a DMI locus relative to the expected probability of concordance at a random, non-DMI locus: $1 - \frac{2}{3}e^{-t_2}$ (Hudson 1983). The greatest deviations from this background probability of gene-tree/species-tree concordance occur when little time has elapsed since, and between, speciation events; this is true for all four patterns of isolation (Fig. 5). When t_1 and t_2 are short, branch segments on which incompatible alleles can arise vary greatly between the four types of gene trees (Fig. 1); this in turn leads to larger disparities in gene-tree/species-tree discordance among the patterns of isolation (Fig. 5).

Two contrasting patterns emerge as time because divergence grows, depending on whether the locus participates in an incompatibility between sister taxa, $S_1 \times S_2$. For loci participating in an incompatibility isolating sister taxa, the relative probability of gene-tree/species-tree concordance is at a minimum when t_1 and

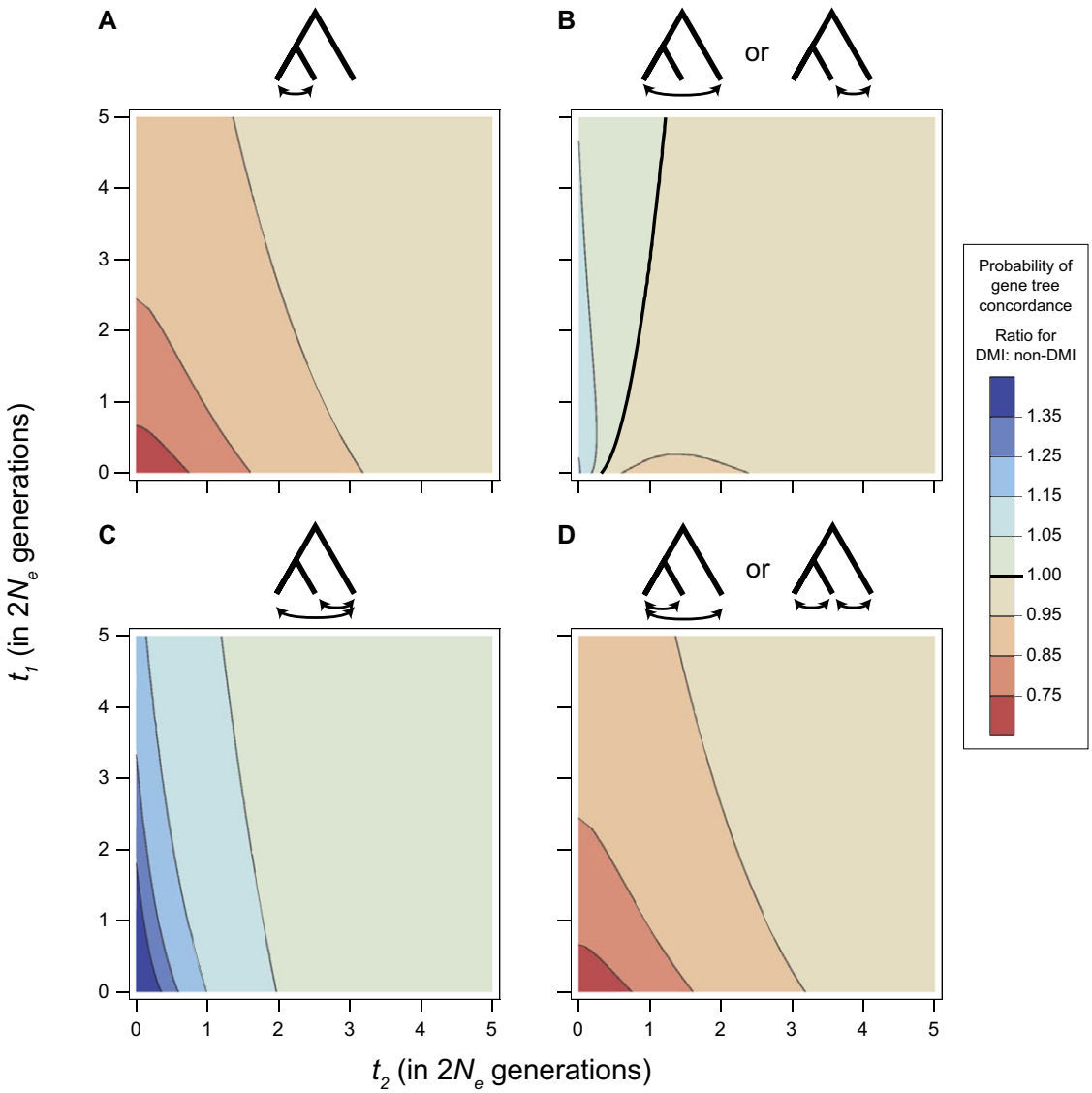

Figure 5. **Relative probability of concordance conditioned on the pattern of reproductive isolation. Contour plots show a ratio of the probability of gene–tree species–tree concordance for a DMI locus relative to a random non-DMI locus. Bold line in (B) and in the legend indicates where the probability of concordance is equal between the two types of loci. All other panels show results that are always either above or below a ratio of 1.**

t_2 are short, and are 33% less likely to be concordant than a random, non-DMI locus as these times approach zero (Figs. 5A and D, and S2a). Loci participating in an incompatibility that does not isolate sister taxa are *more* likely to have gene trees that are concordant with the species tree when times are short, up to 67% more likely to be concordant than a random, non-DMI locus as times approach zero (Figs. 5B and C, and S2b).

The contrast between isolation patterns derives from the restrictions placed on the position of mutations when conditioning on each pattern of reproductive isolation. On concordant trees, alleles involved in sister-taxa incompatibility must arise before (looking backward in time) the coalescence of lineages from the sister species. This coalescence is, on average, deeper on discordant trees, providing more time for the appropriate mutations to arise. Conversely, the deeper coalescence on discordant trees also reduces the shared branch length leading to sister species. The reduced potential for an incompatible allele shared between the sister species on discordant trees increases the chances that a shared incompatibility isolating both $S_1 \times S_3$ and $S_2 \times S_3$ is due to loci with concordant trees.

DMI LOCI ARE ON AVERAGE SLIGHTLY MORE LIKELY TO BE DISCORDANT WITH THE SPECIES TREE

The results above were presented separately for the four different patterns of reproductive isolation among three species. To present the probability of gene-tree/species-tree discordance across all

patterns of isolation, we must take into account the likelihood of each isolation pattern under different histories. For example, pairs of species that have been diverged longer are more likely to harbor incompatibilities, and thus, more likely to be among the species that are isolated. The likelihood that a DMI locus confers a specific pattern of isolation therefore depends on both tip length, t_1, and internal branch length, t_2. As a result, the general, unconditioned probability of gene-tree/species-tree concordance at a DMI locus depends on the likelihood of each isolation pattern.

We calculate the relative probability of each isolation pattern by conditioning on the observation of a DMI. The probability of a particular pattern of isolation, I_0, can be written as

$$P(I_0|\text{ DMI observed}) = \frac{P(I_0)}{\sum_k P(I_k)}\,, \tag{5}$$

where k is an index for the patterns of isolation and $P(I_k)$ is the denominator in equation (1) from the law of total probability,

$$P(I_k) = \sum_{n,m} P(I_k|\, T_n, T_m) P(T_n)\, P(T_m). \tag{6}$$

From this, we compare the relative probability of each isolation pattern in our model, which considers ILS, to a model on a fixed species tree with no ILS. For the model with a fixed species tree, we use the expected number of incompatibilities from Wang et al. (2013) to compute the probability of each isolation pattern. In both models, isolation patterns that include an incompatibility between sister species are most likely when tip lengths, t_1, are long relative to the interspeciation time, t_2 (Figs. S3a and d, and S4a and d). The opposite case, with short tip lengths relative to interspeciation time, favors the isolation pattern where both sister species are incompatible with the third species (Figs. S3c and S4c). For intermediate values of t_1 and t_2, the most likely case is an incompatibility that isolates one of the two more distantly related species pairs, that is, isolating $S_1 \times S_3$ or $S_2 \times S_3$ (Figs. S3b and S4b).

The introduction of ILS substantially increases the proportion of incompatibilities that isolate more than one species pair (i.e., isolation patterns in Fig. 3C and D). On a fixed species tree with no ILS, no more than one-third of incompatibilities ever isolate multiple species pairs, but in a model with ILS, more than half isolate multiple species pairs when t_2 is short relative to t_1 (Fig. S4). This difference arises from the additional branch length that is specific to one lineage on discordant topologies. Coalescence on discordant topologies can only occur in the ancestral population of all three species, substantially increasing lineage-specific branch lengths. When mutations that produce incompatible alleles are inherited by the same lineage, a derived-ancestral incompatibility forms with other lineages (Fig. S1). Discordant topologies increase the chances for these shared incompatibilities, especially when t_2 is short relative to t_1.

Putting together the probability of each isolation pattern with its probability of concordance, the general, unconditioned probability of gene-tree/species-tree concordance can be calculated by the sum,

$$P(\text{concordance} \mid \text{DMI observed}) = \sum_k P(\text{concordance} \mid I_k) \times P(I_k \mid \text{DMI observed}). \tag{7}$$

Figure 6A shows the general probability of concordance between the species tree and gene trees from a DMI locus, across all patterns of isolation. When tip lengths, t_1, are short, gene-tree/species-tree concordance is slightly more likely for DMI loci, up to 27% as both times approach 0. However, for most combinations of times, t_1 and t_2, DMI loci are slightly less likely to be concordant than a random, non-DMI loci. Overall, gene trees for a locus participating in a DMI have a probability of concordance very slightly below the background value of $1 - \frac{2}{3}e^{-t_2}$.

ALLOWING INCOMPATIBILITIES TO ARISE IN THE SAME POPULATION

Our model of DMI loci featuring discordant gene trees has, thus far, explicitly prevented incompatibilities from arising in the same population. This prohibition assumes that selection against incompatibilities is strong enough to prevent the persistence of incompatible alleles in a population. Evidence for the variability of reproductive isolation within populations suggests that the strength of selection may be insufficient to prevent polymorphic incompatibilities from existing (e.g., Corbett-Detig et al. 2013). To address this possibility in our model, we relax this prohibition, allowing incompatible alleles to arise and segregate in ancestral populations as long as extant lineages do not individually carry the incompatible genotype.

Because our model considers the genealogical history of a DMI locus, incompatibilities that arise in the same population can be incorporated with relative ease. The restriction against these incompatibilities has been enforced by requiring at least one derived allele in an incompatibility to arise after divergence between species pairs. This restriction can be lifted by allowing derived alleles to arise up to (backward in time) the point of coalescence. For derived–ancestral incompatibilities, we continue to enforce the restrictions from mutation order. That is, a derived allele participating in a derived-ancestral incompatibility may only arise after a mutation has already produced the compatible allele (see Supporting Information Methods). The probability of concordance when considering incompatibilities that arise within populations before divergence can then be calculated by using a relaxed indicator function in equation (2). (The indicator function of each gene tree pair is available in Appendix 2 in Supporting Information Materials).

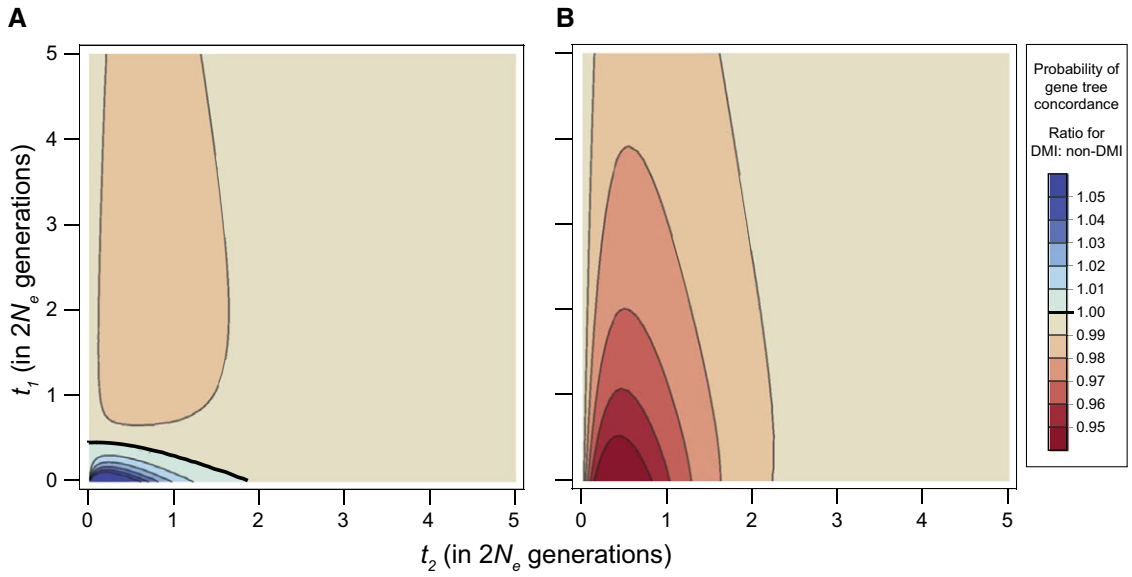

Figure 6. **Relative probability of concordance for a DMI locus. (A) Probability of gene–tree/species–tree concordance for a DMI locus in a model restricting incompatibilities from arising in the same population. Concordance is slightly more likely when times are short. Bolded line shows the contour where concordance is equal to the canonical expectation from coalescent theory. (B) Probability of concordance for a DMI locus in a model allowing incompatibilities to arise in the same population. Concordance is always less likely.**

Overall, allowing the unrestricted emergence of incompatibilities within populations reduces the probability that DMI loci will have gene trees that are concordant with the species tree (Fig. 5B). When tip lengths, t_1, are short, the probability of concordance can be reduced by up to 7% relative to the background expectation. In contrast to the model that restricts DMIs from emerging before divergence, concordance is always less likely than the expectation for non-DMI loci.

A model that allows incompatibilities to arise within populations also increases the probability of DMIs that isolate sister species (Fig. S5a and d). This results from the additional opportunities for isolating mutations on inner branch segments. Unlike incompatibilities isolating more distantly related species pairs, incompatible alleles on inner branch segments were consistently restricted among sister species from producing incompatibilities within populations. This model also causes the pattern of reproductive isolation to have an even greater impact on the probability of gene-tree/species-tree concordance. Qualitatively, the patterns of concordance conditioned on each isolation pattern are similar (see Fig. S6). However, loci participating in incompatibilities between sister taxa, S_1 x S_2, are much less likely to have gene trees that are concordant with the species tree, up to 66% less likely than non-DMI loci as t_1 and t_2 approach 0. Meanwhile, a locus participating in an incompatibility shared between species pairs $S_1 \times S_3$ and $S_2 \times S_3$ is up to 133% more likely than a non-DMI locus to have a gene tree that is concordant. As before, the differences in branch length and topology between the types of gene trees are greatest when t_1 and t_2 are short, but these differences are amplified when incompatibilities can arise in the same population before divergence.

INCOMPATIBLE ALLELES ARE LIKELY TO HAVE ARISEN IN ANCESTRAL POPULATIONS

In the previous section, we allowed pairs of incompatible alleles to arise in the same population before divergence. Because of selection against incompatibilities within populations, such a history for DMI loci should be less common. However, an incompatible allele in a DMI pair could have arisen in an ancestral population without ever having caused an incompatibility within populations. Such an allele would have arisen in the ancestral population and then fixed in one lineage before becoming incompatible with a new mutation after divergence (e.g., the pair of DMI loci in Fig. 4D). As taxa spend more time diverged, incompatibilities between them become more likely to be the result of interactions between new mutations that arise postdivergence. In contrast to this scenario, many formulations of the DM model only allow incompatibilities to form from new mutations after divergence (Orr 1995; Orr and Turelli 2001; Fierst and Hansen 2010; Livingstone et al. 2012; Wang et al. 2013; Fraisse et al. 2014).

To examine the extent to which an incompatible allele is likely to have arisen prior to speciation, we consider its probability in our DM model with ILS. Mutations that occur before divergence give rise to incompatible alleles (but not incompatibilities) in ancestral populations. Branch segments on gene trees positioned before divergence bear such mutations. We can

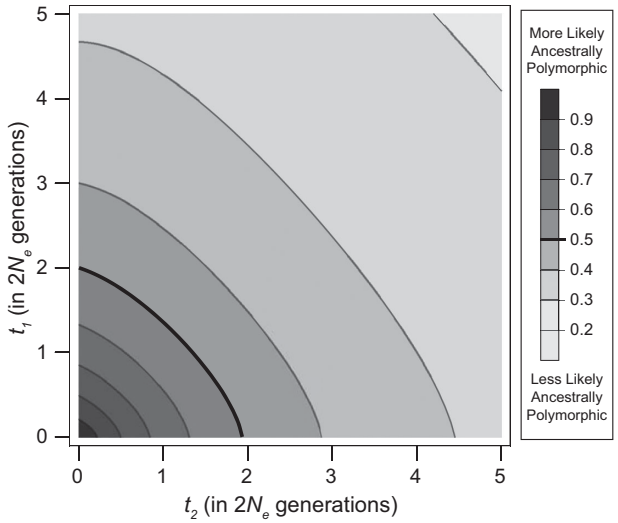

Figure 7. Probability that an incompatibility involves an allele that arose prior to speciation. Bold line shows the contour where a pairwise DMI is equally likely to form from at least one ancestrally arising mutation as from alleles appearing completely after divergence.

calculate the probability that an incompatibility involves ancestrally arising alleles by adjusting equation (2) to count only mutations on predivergence branch segments,

$$P(I_{\mathrm{ANC}}|T_x,\ T_y) = \ p \sum_{\alpha,\beta} P(\text{mutation on } x_\alpha)\, P(\text{mutation on } y_\beta) \times \mathbf{1}(x_\alpha, y_\beta)\mathbf{1}_{\mathbf{ANC}}\left(x_\alpha, y_\beta\right), \quad (8)$$

where $\mathbf{1}_{\mathbf{ANC}}(x_\alpha, y_\beta)$ is 1 when either branch segment is positioned before divergence and 0 otherwise (see Supporting Information Methods). The unconditional probability that an incompatibility involves one ancestrally arising allele can then be calculated following equation (5). Figures 7 and S7 show the proportion of incompatibilities in a three-species complex involving one incompatible allele that arose in an ancestral population. As branch lengths increase, incompatibilities are more likely to form solely from alleles that arose after divergence.

The pattern of isolation at a DMI has a substantial influence on whether ancestrally arising alleles participate in the incompatibility (Fig. S8). When considering incompatibilities between sister taxa, $S_1 \times S_2$, the proportion of incompatibilities that have at least one ancestrally arising allele depends only on tip length, t_1, and effective population size, N_e. This is because the branch length on which incompatible alleles can arise in the ancestral population depends only on the expected time to coalescence of two lineages from the point of their divergence, which equals $2N_e$. The probability of an ancestral allele in a DMI for this case is proportional to the product of the two predivergence segments unique to each lineage and the postdivergence segment of its counterpart (see Fig. S9); this is equal to $2 \cdot 2N_e \cdot t_1$. Meanwhile, the probability that a DMI involves only mutations arising postspeciation is proportional to the product of the postdivergence lengths, ${t_1}^2$. For other patterns of isolation, the probability that ancestral alleles participate in the DMI decreases with interspeciation time, t_2.

Overall, for examined speciation times, incompatibilities are likely to involve at least one ancestrally arising allele. When incompatibility loci are allowed to arise in the same population, the probability that incompatibilities result from ancestrally arising alleles increases, but the qualitative patterns are not substantially different from those described above (Fig. S10).

Discussion

Our results show that DMI loci are slightly more likely to have gene trees that are discordant with the species tree because discordant trees offer more opportunities for the formation of incompatible alleles. This finding follows inevitably from the fact that the mutational target size for a DMI at a particular locus depends on its gene tree, coupled with the constraint that incompatible alleles in a pairwise DMI can only arise on certain pairs of branch segments. A DMI locus is more likely to have a discordant genealogy if these branch segments are longer on discordant trees, and vice versa. The relevant branch segments differ for different patterns of isolation, such that alleles isolating sister taxa are more likely to form on discordant trees, whereas alleles isolating both sister taxa from a third taxon are more likely to form on concordant trees. Our results are in opposition to previous verbal models and suggest that gene tree concordance is unlikely to be useful for identifying DMI loci across the genome. Given the slight excess of discordance expected at DMI loci, gene tree discordance is also unlikely to be useful for this task.

The results presented here assume a neutral model with only ILS acting, leading to several important limitations. Among these are two important considerations that have sometimes been used to argue in favor of greater concordance at speciation loci: postdivergence gene flow and positive selection on incompatibility alleles. The effects of these two phenomena on concordance at DMI loci depend critically on the particulars of the scenario (e.g., the species involved in postdivergence gene flow, the timing of selection, and the species isolated by the incompatibility) and therefore require some discussion. Overall, we argue that the probability of concordance at incompatibility loci is not consistently increased by gene flow or selection during the process of speciation.

At loci conferring reproductive isolation, gene flow can be reduced by the lower fitness of hybrids that inherit the incompatible genotype. This implies that while substantial gene flow between two species may complicate the genealogical history for most of the genome, DMI loci should retain their original history.

There is no guarantee that the original history of a DMI locus is concordant with the species tree (i.e., ILS can still occur at this locus), and they therefore may not be any more likely to be concordant than other nonintrogressed loci. Nevertheless, compared to introgressed loci, the relative probability of concordance at DMI loci can be either increased, decreased, or unaffected by postdivergence gene flow.

To see why all of these outcomes are possible, consider two key details concerning patterns of gene flow. First, gene flow must have occurred between the same species isolated by the DMI to have any effect on gene-tree/species-tree concordance. General patterns of gene flow between taxa that do not express the incompatibility will not affect the relative probability of concordance. Second, the particular pair of lineages exchanging genes will determine whether DMI loci will be more or less concordant than non-DMI loci. When gene flow occurs between nonsister taxa (e.g., taxa S_1 and S_3 in Fig. 1), loci with introgressed histories will be more likely to have discordant topologies. Therefore, DMI loci will be relatively more likely to be concordant, though the direction of introgression can have a large effect on the magnitude of this increase (cf. Hibbins and Hahn 2018). Alternatively, when gene flow occurs between sister taxa (taxa S_1 and S_2 in Fig. 1), loci with introgressed histories will actually be more likely to have *concordant* topologies. This occurs because gene flow effectively lengthens the internal branch along which introgressed loci can coalesce, increasing their chances for lineage sorting. The original history, retained by DMI loci, becomes less likely to be concordant with the species tree relative to the history of introgressed loci. Finally, note that theory suggests DMI loci that do not provide a selective advantage in the lineage on which they arose are unlikely to persist in the presence of gene flow (Gavrilets 1997; Bank et al. 2012). In other words, not all DMI loci will be resistant to introgression.

Loci involved in DMIs often bear the signature of positive selection (Coyne and Orr 2004; Orr et al. 2006). Although positive selection on a DMI locus can only increase the probability that its gene tree is concordant with the species tree—because selection reduces N_e, and consequently, the time to coalescence (Kaplan et al. 1989)—the scenarios in which this can occur are limited. To increase the relative probability of concordance, selection on an incompatibility locus must occur in the ancestral population of the sister lineages (S_1 and S_2 in Fig. 1). Lineage-specific selection, acting on a DMI locus in only one species, can have no effect on the process of lineage sorting that determines concordance, nor can selection in the common ancestor of all three species. When selection on a DMI locus does occur between speciation events, its effects on the relative probability of gene-tree/species-tree concordance are greatest when selection is strong and divergence times are short. Figure S11 shows the effects of selection on the probability of concordance compared to a random unselected locus, demonstrating a larger effect with stronger selection. Note also that scenarios involving selection on DMIs that arise in the ancestral population of species S_1 and S_2 can only lead to specific patterns of isolation (in this case, the patterns in Fig. 3B or C). Positive selection in the ancestor of two sister taxa is not expected for DMI loci that isolate them from one another (the patterns in Fig. 3A and D).

Ultimately, the magnitude by which DMI loci are more likely to be concordant depends on how often they are targets of positive selection relative to non-DMI loci. The results shown in Figure S11 assume that non-DMIs are under no selection, exaggerating the effects of positive selection on concordance. In fact, a higher chance of concordance from positive selection is not unique to loci that confer reproductive isolation; any locus that has experienced linked selection in the ancestral population is more likely to be concordant with the species tree (Slatkin and Pollack 2006; Stukenbrock et al. 2011; Dutheil et al. 2015). An examination of loci likely to be affected by linked selection is a more direct way of gaining insight into species relationships, regardless of whether they are DMI loci (e.g., Scally et al. 2012; Pease and Hahn 2013; Munch et al. 2016).

As in the traditional DM model, we assumed that incompatibilities have an equal probability of forming between untested allelic combinations, which arise at independent, unlinked loci. An important simplification for our model is the consideration of only a single history for an incompatibility locus in each lineage; that is, we consider only a haploid history for incompatibility loci. Among diploids (and systems with higher ploidy), this simplification is equivalent to assuming that incompatible alleles have fixed in their extant lineages without having passed through the incompatible genotype. Although incompatibility loci can be polymorphic in both extant and ancestral populations (Cutter 2012), tracking this polymorphism in extant populations would require the consideration of multiple lineages in each species. This would necessitate a model of dominance and fitness for each genotypic combination of incompatibility loci. Such a model is outside the scope of our focus here on ILS and the stochastic accumulation of incompatible alleles, but may be interesting for future work. Similarly, we restricted our calculations to three species to produce a tractable model—the number of possible genealogies and isolation patterns grows more than exponentially with the number of species. Although we expect that the main results would remain the same when extended to more species, unique patterns of isolation among multiple species could harbor novel phylogenetic signals.

In examining the possible histories for incompatible alleles, our results highlight how likely they are to have arisen prior to speciation (i.e., in ancestral populations) when they do not affect fitness in conspecific backgrounds. Our results suggest that pairwise DMIs are likely to involve at least one incompatible

allele that arose ancestrally until three to four N_e generations after species divergence. That incompatible alleles are more likely to be from postspeciation mutations as species diverge may not be surprising from a population genetics perspective, but this result may help to clarify an argument on the relative importance of derived versus ancestral alleles in incompatibilities (Cutter 2012). Derived alleles are expected to play a larger role in the formation of incompatibilities because each pairwise incompatibility must involve at least one derived allele (Orr 1995). The distinction between derived and ancestral alleles should, however, not be confused with the genealogical history of participating loci. Mutations that yield incompatible derived alleles can arise both before and after populations diverge; that is, "ancestral" alleles are not derived alleles that arose in ancestral populations. Similarly, a derived–ancestral incompatibility can be the product of two mutations along the same lineage after divergence. Whether an incompatible allele arose in the ancestral population of two or more species depends on the timing of the mutation rather than its state relative to a common ancestor. Thus, the percentage of incompatibilities that involve alleles that arose in ancestral populations decreases with time, but the percentage of ancestral alleles participating in incompatibilities (from derived–ancestral interactions) remains the same. Because the number of potential interactions remains the same, ILS does not change the prediction that incompatibilities should accumulate faster than linearly with divergence time (Orr 1995, Orr and Turelli 2001). This prediction of "snowballing" incompatibilities was made with respect to any two species embedded in a larger tree, a comparison that is not affected by ILS.

The results presented here should be applicable to a number of empirical systems that have been the focus of speciation research. All that is required are short internal branches on a species tree and the ability to carry out crosses between multiple pairs of species. Some examples include species of wild tomato in the genus *Solanum* (Moyle and Nakazato 2010; Pease et al. 2016) and subspecies of mouse within *Mus musculus* (White et al. 2009; Wang et al. 2015). One of the most interesting systems to which these predictions can be applied are the three species in the *Drosophila simulans* clade, which were the focus of one of the seminal studies purporting to demonstrate the unique phylogenetic signal at speciation loci (Ting et al. 2000). However, this study produced what is now thought to be a discordant gene tree from a DMI locus. The tree of the *D. simulans* clade constructed from the hybrid incompatibility locus *OdsH* in Ting et al. (2000) places *D. simulans* as most closely related to *D. mauritiana*. More recently, whole-genome sequence data (Garrigan et al. 2012) found the best-fitting maximum-likelihood tree to be one that groups *D. simulans* with *D. sechellia* to the exclusion of *D. mauritiana*. This inference is supported by an excess of gene trees concordant with this topology in regions of low recombination, where ILS should have the least effect (Pease and Hahn 2013).

The identification of loci participating in hybrid incompatibilities between multiple closely related species pairs has spurred many new analyses, including phylogenetic comparisons among species (Cattani and Presgraves 2009; Scarpino et al. 2013; Sherman et al. 2014; Wang et al. 2015). These comparative approaches can help to elucidate the timing and progression of reproductive isolation (Moyle and Payseur 2009). But the analysis of incompatibilities between multiple species pairs also introduces genealogical ambiguity at incompatibility loci. When incompatible alleles in a DMI arise on gene trees with different topologies, identifying which branch of the species tree they arose on becomes challenging. In fact, incompatible alleles arising on discordant trees may be mapped onto the wrong branches of the species tree by standard methods. Although a derived allele in a pairwise DMI will always be inherited by at least one of the lineages isolated by the incompatibility, the other allele can originate on branches that are shared with uninvolved lineages. For example, in the absence of ILS an incompatibility isolating only the sister species among three taxa (Fig. 3A) would be interpreted as the result of two mutations, each uniquely inherited by one of the sister species. However, in the presence of ILS such an incompatibility could involve a mutation inherited by one of the sister species and a third taxon (as in Fig. 4C and D). Discordance between gene trees is most pronounced when little time separates speciation events, and this is often the case for model systems of speciation, where the ability to perform interspecific crosses is a useful tool for genetic investigation (e.g., True et al. 1996; Slotman et al. 2004; Sweigart et al. 2006; Moyle and Nakazato 2008; Matute and Coyne 2010; White et al. 2011). Under these circumstances, inferences about the origin of incompatibilities from comparative mapping need to be mindful of the genealogical history at DMI loci.

Methods

CALCULATING BRANCH LENGTHS

To calculate the probability of gene-tree/species-tree concordance at a DMI locus, we find the probability of concordance conditioned on a particular pattern of isolation, I. This requires calculating the branch lengths from each pair of gene trees, and the indicator function representing the opportunity to produce an incompatibility (eq. 2). The indicator function can be represented as a matrix, $\mathbf{1}_{T_x,T_y}$, with rows and columns corresponding to the branches of the respective gene trees T_x and T_y. With four types of gene trees, there are 10 unique matrices (see Appendix 1 in Supporting Information Materials).

Branch segments on each of the four types of gene trees are labeled in Figure 8. The probability of a mutation on the branch segments of each tree, T, can be assembled into the vector B_T. We

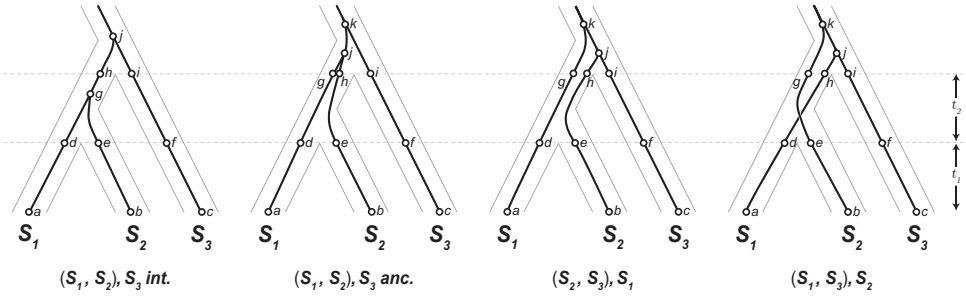

Figure 8. **Four types of gene trees with branch segments labeled. Concordant trees are divided into those that coalesce in the time between species divergences (*int*er-) and those that coalescence in the *anc*estral population.**

order the lengths of each segment (using the segment names as labels for their expected lengths for convenience, that is, $ij = L(i - j)$) in the vectors as follows:

$$\begin{aligned} B_{(S1,S2)S3\ int.} &= [ad, dg, be, eg, cf, fi, ij, gh, hj], \\ B_{(S1,S2)S3\ anc.} &= [ad, dg, gj, be, eh, hj, cf, fi, ik, jk], \\ B_{(S2,S3)S1} &= [ad, dg, gk, be, eh, hj, cf, fi, ij, jk], \\ B_{(S1,S3)S2} &= [ad, dh, hj, be, eg, gk, cf, fi, ij, jk]. \end{aligned} \quad (9)$$

The length of segments whose ends are not coalescent events are simply t_1 or t_2. Segments that coalesce in the ancestral population of all three species have expected lengths (in $2N_e$ coalescent units) of 1/3, 1, and 4/3, corresponding to the time to coalescence from three-to-two, two-to-one, and three-to-one lineages, respectively.

Several segments on the (S_1, S_2) S_3 *int.* tree have expected values that are conditioned on the first coalescence occurring by t_2. This condition distinguishes this type of tree from the ancestrally coalescing tree with the same topology. Let v be a random variable from 0 to t_2, representing the time from the most recent population divergence to the first coalescence event in the (S_1, S_2) S_3 *int*. tree. The probability distribution function of v is given by the exponential distribution function for the coalescence of two lineages, e^{-t}, divided by the probability of coalescence by t_2 (see Mendes and Hahn 2017). Thus,

$$f_v(t) = \frac{e^{-t}}{1 - e^{-t_2}}; \quad 0 \le t \le t_2. \quad (10)$$

Let $q(t_2)$ be the expected value of v for a given value of t_2, then the expected time from the first population divergence to the first coalescence in the (S_1, S_2) S_3 *int*. tree is

$$q(t_2) = \int_0^{t_2} \frac{te^{-t}}{1 - e^{-t_2}} \mathrm{d}t = 1 - \frac{t_2}{e^{t_2} - 1}. \quad (11)$$

Filling in the segment lengths from equation (9), the values for B_T become

$$\begin{aligned} B_{(S1,S2)S3\ int.} &= [t_1, q(t_2), t_1, q(t_2), t_1, t_2, 1, t_2 - q(t_2), 1], \\ B_{(S1,S2)S3\ anc.} &= \left[t_1, t_2, \frac{1}{3}, t_1, t_2, \frac{1}{3}, t_1, t_2, \frac{4}{3}, 1\right], \\ B_{(S2,S3)S1} &= \left[t_1, t_2, \frac{4}{3}, t_1, t_2, \frac{1}{3}, t_1, t_2, \frac{1}{3}, 1\right], \\ B_{(S1,S3)S2} &= \left[t_1, t_2, \frac{1}{3}, t_1, t_2, \frac{4}{3}, t_1, t_2, \frac{1}{3}, 1\right]. \end{aligned} \quad (12)$$

With the terms in equation (2) written as vectors and matrices, $P(I|T_x, T_y)$ can be written conveniently as the matrix product:

$$P(I|T_x, T_y) = pB_{T_x}\mathbf{1}_{T_x,T_y}B_{T_y}. \quad (13)$$

For four gene trees, there are 16 combinations of T_x, T_y. We form a 4×4 matrix, $\boldsymbol{D}(I)$, for each isolation pattern I, whose entries are the above matrix product (eq. 13) for each gene tree pair. The rows and columns in $\boldsymbol{D}(I)$ are ordered so that they are consistent with the order of trees in Figure 8.

To arrive at the probability of each gene tree pair, Bayes' theorem is applied (eq. 1). The numerator from equation (1) can be calculated from each element of the $\boldsymbol{D}(I)$ product matrix, multiplied by the unconditional probability of each gene tree pair. The probability for each type of gene tree (Fig. 1) can be assembled into a vector of tree probabilities, T_{prob}, as

$$T_{\text{prob}} = \left[1 - e^{-t_2}, \frac{1}{3}e^{-t_2}, \frac{1}{3}e^{-t_2}, \frac{1}{3}e^{-t_2}\right]. \quad (14)$$

The numerator in equation (1) for each gene tree pair can then be represented as the elementwise product of $\boldsymbol{D}(I)$ with the outer product of T_{prob} with itself. We call this 4×4 matrix $\boldsymbol{P}(I)$,

$$\boldsymbol{P}(I) = \left(T_{\text{prob}} \otimes T_{\text{prob}}\right) \circ \boldsymbol{D}(I). \quad (15)$$

The denominator for equation (1) is the sum of all elements in the $\boldsymbol{P}(I)$ matrix.

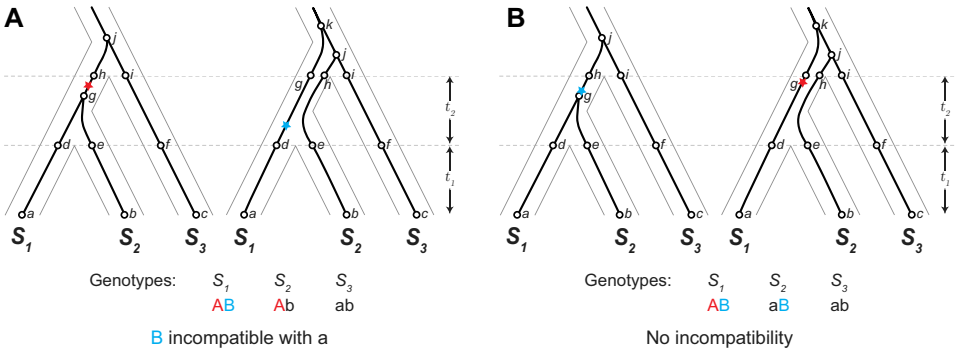

Figure 9. **Mutation order determines the potential for a derived-ancestral incompatibility. Red and blue stars mark the position of the first and second mutations producing incompatible alleles. The ancestral genotype for the two loci is denoted "ab", with the mutations producing derived alleles "A" and "B". (A) Derived–ancestral incompatibility from the derived allele in the S_1 lineage with the ancestral allele in the S_3 lineage. (B) No incompatibility forms because the derived–ancestral genotype persists in the S_2 lineage.**

Returning to equation (4), the probability of concordance for a single DMI locus is given by the sum of all elements in the first two rows and columns of $\boldsymbol{P}(I)$, divided by twice the sum of all elements. Note that the parameters p and $2N_e\mu$ do not appear in the probability of concordance as they are cancelled by the denominator in equation (1).

MUTATION ORDER AND DERIVED–ANCESTRAL INCOMPATIBILITIES

Thus far, the probability of each mutation in a DMI pair has been treated as an independent event. That is, the joint probability is calculated as a product of the mutation probabilities on each branch (eq. 2). Although this assumption is valid for derived–derived incompatibilities, greater care must be taken for derived–ancestral incompatibilities due to the order in which mutations must occur. Consider a potential derived–ancestral incompatibility between a mutation on segment *g*–*h* of the (S_1, S_2) S_3 *int.* tree and a mutation on segment *d*–*g* of the (S_2, S_3) S_1 tree, as in Figure 9.

A mutation on segment *g*–*h*, followed by a mutation on *d*–*g* can result in a derived–ancestral incompatibility between S_1 and S_3. The converse, a mutation that occurs first on segment *d*–*g* followed by a mutation on *g*–*h*, cannot result in a derived–ancestral incompatibility because the ancestral allele persists in the S_2 lineage. As the incompatible allelic combination already exists in S_2, this combination cannot be the cause of an incompatibility between S_1 and S_3.

This asymmetry from mutation order occurs because the derived allele in a derived–ancestral incompatibility must arise from the second mutation. If the derived allele arose from the first mutation, it would immediately produce the incompatible genotype. In the previous example, when the second mutation arises on segment *d*–*g*, the derived allele is inherited by the S_1 lineage, whereas when the second mutation arises on segment *g*–*h*, both the S_1 and S_2 lineage inherit the derived allele. Between branch segments that have contemporaneous endpoints, this asymmetry does not exist and the order of mutations does not matter. For example, the first mutation on segment *a*–*d* of either tree in Figure 9 leads to an ancestral allele that can be incompatible with a second mutation at segment *a*–*d* on the other tree.

Differences in mutation order produce this asymmetry only when the inheritance of the derived allele is made ambiguous by the timing of a coalescent event. This is only possible when the mutations are from segments that overlap temporally. Most pairs of segments where mutation order is ambiguous occur before any population divergence has occurred (Fig. 8). When incompatibilities are restricted to arising only after populations diverge, these segments do not cause any incompatibilities. The only gene tree pairs where mutation order must be considered in this model are those that involve the (S_1, S_2) S_3 *int.* tree. Here, the coalescence in the ancestral population of S_1 and S_2 can change the identity of the derived allele in an $S_1 \times S_3$ incompatibility and an $S_2 \times S_3$ incompatibility.

Returning to our earlier example, we calculate the joint probability of two mutations leading to a derived–ancestral incompatibility from segments *g*–*h* on an (S_1, S_2) S_3 *int.* tree and *d*–*g* on an (S_2, S_3) S_1 tree. Such an incompatibility requires the first mutation to be on segment *g*–*h*, restricting the timing of the second mutation on *d*–*g*. The probability of the first mutation is unrestricted and is equal to the branch length of *g*–*h*, which has an expected value of $t_2 - q(t_2)$. Let τ be the time from the S_1, S_2 divergence to the first mutation on segment *g*–*h*. Assuming that a single mutation occurs on segment *g*–*h*, the probability of the first mutation should be uniformly distributed along the length of segment *g*–*h*. Then,

$$\begin{aligned} \mathrm{E}[\tau] &= q(t_2) + \frac{1}{2}(t_2 - q(t_2)) \\ &= \frac{1}{2}(t_2 + q(t_2)). \end{aligned} \tag{16}$$

Mutations that occur before the first mutation on *g–h* do not cause an incompatibility, thus the expected value of τ is also the expected length of the subsegment of *d–g* from which a second mutation may arise. The probability of an incompatibility from these two segments can then be calculated from the product of $E[\tau]$ and $t_2 - q(t_2)$.

This probability applies for each of the gene tree pairs involving (S_1, S_2) S_3 *int.* and any other gene tree for producing a derived–ancestral incompatibility between S_1 and S_3. Similarly, the same reasoning can be applied for interactions between mutations on *g–h* and *e–h* for incompatibilities between S_2 and S_3. When both gene trees are (S_1, S_2) S_3 *int.*, the calculation is more involved due to the coalescence at the end of branch segments from both trees (see Supporting Information Methods).

ACKNOWLEDGMENTS

The authors would like to thank R. Guerrero and L. Moyle for their stimulating discussions and feedback in the development of this manuscript.

AUTHOR CONTRIBUTIONS

RJW and MWH conceived and designed the study. RJW performed calculations. RJW and MWH drafted and revised the manuscript.

LITERATURE CITED

Ågren, J. A. 2013. Selfish genes and plant speciation. Evol. Biol. 40:439–449.

Bank, C., R. Bürger, and J. Hermisson. 2012. The limits to parapatric speciation: Dobzhansky–Muller incompatibilities in a continent–island model. Genetics 191:845–863.

Barbash, D. A., D. F. Siino, A. M. Tarone, and J. Roote. 2003. A rapidly evolving MYB-related protein causes species isolation in *Drosophila*. Proc. Natl. Acad. Sci. 100:5302–5307.

Barr, C. M., and L. Fishman. 2010. The nuclear component of a cytonuclear hybrid incompatibility in *Mimulus* maps to a cluster of pentatricopeptide repeat genes. Genetics 184:455–465.

Bateson, W. 1909. Heredity and variation in modern lights. Pp. 85–101 *in* A. C. Seward, ed. Darwin and Modern Science. Cambridge Univ. Press, Cambridge, England.

Bomblies, K., and D. Weigel. 2007. Hybrid necrosis: autoimmunity as a potential gene-flow barrier in plant species. Nat. Rev. Genet. 8:382–393.

Castillo, D. M., and D. A. Barbash. 2017. Moving speciation genetics forward: modern techniques build on foundational studies in *Drosophila*. Genetics 207:825–842.

Cattani, M. V., and D. C. Presgraves. 2009. Genetics and lineage-specific evolution of a lethal hybrid incompatibility between *Drosophila mauritiana* and its sibling species. Genetics 181:1545–1555.

Corbett-Detig, R. B., J. Zhou, A. G. Clark, D. L. Hartl, and J. F. Ayroles. 2013. Genetic incompatibilities are widespread within species. Nature 504:135–137.

Coyne, J. A., and H. A. Orr. 2004. Speciation. Sinauer Associates, Inc., Sunderland, MA.

Cutter, A. D. 2012. The polymorphic prelude to Bateson–Dobzhansky–Muller incompatibilities. Trends Ecol. Evol. 27:209–218.

Cutter, A. D. 2013. Integrating phylogenetics, phylogeography and population genetics through genomes and evolutionary theory. Mol. Phylogenet. Evol. 69:1172–1185.

Dobzhansky, T. 1937. Genetics and the origin of species. Columbia Univ. Press, New York, NY.

Drummond, A. J., M. A. Suchard, D. Xie, and A. Rambaut. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. 29:1969–1973.

Dutheil, J. Y., K. Munch, K. Nam, T. Mailund, and M. H. Schierup. 2015. Strong selective sweeps on the X chromosome in the human-chimpanzee ancestor explain its low divergence. PLOS Genetics 11:e1005451.

Fierst, J. L., and T. F. Hansen. 2010. Genetic architecture and postzygotic reproductive isolation: evolution of Bateson–Dobzhansky–Muller incompatibilities in a polygenic model. Evolution 64:675–693.

Fraïsse, C., J. A. D. Elderfield, and J. J. Welch. 2014. The genetics of speciation: are complex incompatibilities easier to evolve? J. Evol. Biol. 27:688–699.

Garrigan, D., S. B. Kingan, A. J. Geneva, P. Andolfatto, A. G. Clark, K. R. Thornton et al. 2012. Genome sequencing reveals complex speciation in the *Drosophila simulans* clade. Genome Res. 22:1499–1511.

Gavrilets, S. 1997. Hybrid zones with Dobzhansky-type epistatic selection. Evolution 51:1027–1035.

Guerrero, R. F., C. D. Muir, S. Josway, and L. C. Moyle. 2017. Pervasive antagonistic interactions among hybrid incompatibility loci. PLOS Genetics 13:e1006817.

Hibbins, M. S., and M. W. Hahn. 2018. Population genetic tests for the direction and relative timing of introgression. bioRxiv. https://doi.org/10.1101/328575.

Hudson, R. R. 1983. Testing the constant-rate neutral allele model with protein sequence data. Evolution 37:203–217.

Hudson, R. R. 1992. Gene trees, species trees and the segregation of ancestral alleles. Genetics 131:509–513.

Jarvis, E. D., S. Mirarab, A. J. Aberer, B. Li, P. Houde, C. Li et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346:1320–1331.

Kaplan, N. L., R. R. Hudson, and C. H. Langley. 1989. The "hitchhiking effect" revisited. Genetics 123:887–899.

Larget, B. R., S. K. Kotha, C. N. Dewey, and C. Ané. 2010. BUCKy: gene tree/species tree reconciliation with Bayesian concordance analysis. Bioinformatics 26:2910–2911.

Liénard, M. A., L. O. Araripe, and D. L. Hartl. 2016. Neighboring genes for DNA-binding proteins rescue male sterility in *Drosophila* hybrids. Proc. Natl. Acad. Sci. 113:E4200–E4207.

Liu, L., L. Yu, L. Kubatko, D. K. Pearl, and S. V. Edwards. 2009. Coalescent methods for estimating phylogenetic trees. Mol. Phylogenet. Evol. 53:320–328.

Livingstone, K., P. Olofsson, G. Cochran, A. Dagilis, K. MacPherson, and K. A. Seitz. 2012. A stochastic model for the development of Bateson–Dobzhansky–Muller incompatibilities that incorporates protein interaction networks. Math. Biosci. 238:49–53.

Maddison, W. P. 1997. Gene trees in species trees. Syst. Biol. 46:523–536.

Maroja, L. S., J. A. Andrés, and R. G. Harrison. 2009. Genealogical discordance and patterns of introgression and selection across a cricket hybrid zone. Evolution 63:2999–3015.

Matute, D. R., and J. A. Coyne. 2010. Intrinsic reproductive isolation between two sister species of *Drosophila*. Evolution 64:903–920.

Mendes, F. K., and M. W. Hahn. 2017. Why concatenation fails near the anomaly zone. Syst. Biol. 67:158–169.

Mihola, O., Z. Trachtulec, C. Vlcek, J. C. Schimenti, and J. Forejt. 2009. A mouse speciation gene encodes a meiotic histone H3 methyltransferase. Science 323:373–375.

Mirarab, S., and T. Warnow. 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics 31:i44–i52.

Moyle, L. C., and T. Nakazato. 2008. Complex epistasis for Dobzhansky–Muller hybrid incompatibility in *Solanum*. Genetics 181:347–351.

Moyle, L. C., and B. A. Payseur. 2009. Reproductive isolation grows on trees. Trends Ecol. Evol. 24:591–598.

Moyle, L. C., and T. Nakazato. 2010. Hybrid incompatibility "snowballs" between Solanum species. Science, 329:1521–1523.

Muller, H. J. 1942. Isolating mechanisms, evolution and temperature. Pp. 71–125 *in* T. Dobzhansky, ed. Biological Symposia, Vol. 6. Jaques Cattell Press, Lancaster, PA.

Munch, K., K. Nam, M. H. Schierup, and T. Mailund. 2016. Selective sweeps across twenty millions years of primate evolution. Mol. Biol. Evol. 33:3065–3074.

Nei, M. 1986. Stochastic errors in DNA evolution and molecular phylogeny. Pp. 133–147 *in* H. Gershowitz, D. L. Rucknagel, and R. E. Tashian, eds. Evolutionary Perspectives and the New Genetics. Alan R. Liss, New York, NY.

Nosil, P., and D. Schluter. 2011. The genes underlying the process of speciation. Trends Ecol. Evol. 26:160–167.

Oliver, P. L., L. Goodstadt, J. J. Bayes, Z. Birtle, K. C. Roach, N. Phadnis, and C. P. Ponting. 2009. Accelerated evolution of the Prdm9 speciation gene across diverse metazoan taxa. PLoS genetics, 5:e1000753.

Orr, H. A. 1995. The population genetics of speciation: the evolution of hybrid incompatibilities. Genetics 139:1805–1813.

Orr, H. A., and M. Turelli. 2001. The evolution of postzygotic isolation: accumulating Dobzhansky–Muller incompatibilities. Evolution 55:1085–1094.

Orr, H. A., J. P. Masly, and D. C. Presgraves. 2004. Speciation genes. Curr. Opin. Genet. Dev. 14:675–679.

Orr, H. A., J. P. Masly, and N. Phadnis. 2006. Speciation in *Drosophila*: from phenotypes to molecules. J. Hered. 98:103–110.

Pease, J. B., and M. W. Hahn. 2013. More accurate phylogenies inferred from low-recombination regions in the presence of incomplete lineage sorting. Evolution 67:2376–2384.

Pease, J. B., D. C. Haak, M. W. Hahn, and L. C. Moyle. 2016. Phylogenomics reveals three sources of adaptive variation during a rapid radiation. PLoS Biol. 14:e1002379.

Phadnis, N., and H. A. Orr. 2009. A single gene causes both male sterility and segregation distortion in *Drosophila* hybrids. Science 323:376–379.

Pollard, D. A., V. N. Iyer, A. M. Moses, and M. B. Eisen. 2006. Widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting. PLOS Genet. 2:e173.

Presgraves, D. C. 2010. Speciation genetics: search for the missing snowball. Curr. Biol. 20:R1073–R1074.

Presgraves, D. C., L. Balagopalan, S. M. Abmayr, and H. A. Orr. 2003. Adaptive evolution drives divergence of a hybrid inviability gene between two species of *Drosophila*. Nature 423:715–719.

Rieseberg, L. H., and B. K. Blackman. 2010. Speciation genes in plants. Ann. Bot. 106:439–455.

Rosenberg, N. A. 2003. The shapes of neutral gene genealogies in two species: probabilities of monophyly, paraphyly, and polyphyly in a coalescent model. Evolution 57:1465–1477.

Scally, A., J. Y. Dutheil, L. W. Hillier, G. E. Jordan, I. Goodhead, J. Herrero et al. 2012. Insights into hominid evolution from the gorilla genome sequence. Nature 483:169–175.

Scarpino, S. V., P. J. Hunt, F. J. Garcia-De-Leon, T. E. Juenger, M. Schartl, and M. Kirkpatrick. 2013. Evolution of a genetic incompatibility in the genus *Xiphophorus*. Mol. Biol. Evol. 30:2302–2310.

Sherman, N. A., A. Victorine, R. J. Wang, and L. C. Moyle. 2014. Interspecific tests of allelism reveal the evolutionary timing and pattern of accumulation of reproductive isolation mutations. PLOS Genet. 10:e1004623.

Slatkin, M., and J. L. Pollack. 2006. The concordance of gene trees and species trees at two linked loci. Genetics 172:1979–1984.

Slotman, M., A. Della Torre, and J. R. Powell. 2004. The genetics of inviability and male sterility in hybrids between *Anopheles gambiae* and *An. arabiensis*. Genetics 167:275–287.

Stukenbrock, E. H., T. Bataillon, J. Y. Dutheil, T. T. Hansen, R. Li, M. Zala, et al. 2011. The making of a new pathogen: insights from comparative population genomics of the domesticated wheat pathogen *Mycosphaerella graminicola* and its wild sister species. Genome Res. 21:2157–2166.

Sweigart, A. L., L. Fishman, and J. H. Willis. 2006. A simple genetic incompatibility causes hybrid male sterility in *Mimulus*. Genetics 172:2465–2479.

Ting, C.-T., S.-C. Tsaur, M.-L. Wu, and C.-I. Wu. 1998. A rapidly evolving homeobox at the site of a hybrid sterility gene. Science 282:1501–1504.

Ting, C.-T., S.-C. Tsaur, and C.-I. Wu. 2000. The phylogeny of closely related species as revealed by the genealogy of a speciation gene, *Odysseus*. Proc. Natl. Acad. Sci. 97:5313–5316.

True, J. R., B. S. Weir, and C. C. Laurie. 1996. A genome-wide survey of hybrid incompatibility factors by the introgression of marked segments of *Drosophila mauritiana* chromosomes into *Drosophila simulans*. Genetics 142:819–837.

Wang, R. J., C. Ané, and B. A. Payseur. 2013. The evolution of hybrid incompatibilities along a phylogeny. Evolution 67:2905–2922.

Wang, R. J., M. A. White, and B. A. Payseur. 2015. The pace of hybrid incompatibility evolution in house mice. Genetics 201:229–242.

White, M. A., C. Ané, C. N. Dewey, B. R. Larget, and B. A. Payseur. 2009. Fine-scale phylogenetic discordance across the house mouse genome. PLOS Genet. 5:e1000729.

White, M. A., B. Steffy, T. Wiltshire, and B. A. Payseur. 2011. Genetic dissection of a key reproductive barrier between nascent species of house mice. Genetics 189:289–304.

Wu, C.-I., and C.-T. Ting. 2004. Genes and speciation. Nat. Rev. Genet. 5:114–122.

Zachos, F. E. 2009. Gene trees and species trees – mutual influences and interdependences of population genetics and systematics. J. Zool. Syst. Evol. Res. 47:209–218.

Associate Editor: S. Wright

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Figure S1. Derived–ancestral incompatibility shared between $S_1 \times S_2$ and $S_1 \times S_3$ due to a shared ancestral allele.
Figure S2. Relative probability of concordance for DMI loci with two different patterns of isolation.
Figure S3. Relative probability of concordance conditioned on the pattern of reproductive isolation (polymorphic incompatibilities allowed).
Figure S4. Probability for different patterns of reproductive isolation; model of DMI with loci on a fixed species tree.

Figure S5. Probability for different patterns of reproductive isolation; model of DMI with loci subject to ILS.
Figure S6. Probability that an incompatibility involves an allele that arose prior to speciation.
Figure S7. Probability for different patterns of reproductive isolation; model of DMI with potentially polymorphic loci in addition to being subject to ILS.
Figure S8. Probability that an incompatibility involves an ancestrally arising locus conditioned on different patterns of reproductive isolation.
Figure S9. The probability of an ancestrally arising allele in a DMI isolating sister taxa.
Figure S10. Probability that an incompatibility involves an ancestrally arising locus; model of DMI with potentially polymorphic loci in addition to being subject to ILS.
Figure S11. Effects of selection on the probability of concordance at a DMI locus.
Figure S12. Calculating constraints on derived–ancestral incompatibilities due to mutation order on a pair of (*S1*, *S2*) *S3* int. trees.
Table S1. Probabilities for incompatibility-participating mutations that depend on mutation order note that patterns A2 and A5 correspond to calculations from equations (16) (main text) and (S4), respectively.
Supplementary Material: Appendix 1. Incompatibility matrices.
Supplementary Material: Appendix 2. Incompatibility matrices with incompatibilities allowed to arise in the same population.