Phylogenomic assessment of the role of hybridization and introgression in trait evolution

Yaxuan Wang, Zhen Cao, Huw A. Ogilvie, Luay Nakhleh
Department of Computer Science, Rice University
6100 Main Street, Houston, TX 77005, USA

September 16, 2020

Abstract

Trait evolution in a set of species—a central theme in evolutionary biology—has long been understood and analyzed with respect to a species tree. However, the field of phylogenomics, which has been propelled by advances in sequencing technologies, has ushered in the era of species/gene tree incongruence and, consequently, a more nuanced understanding of trait evolution. For a trait whose states are incongruent with the branching patterns in the species tree, the same state could have arisen independently in different species (homoplasy) or followed the branching patterns of gene trees, rather than the species tree (hemiplasy). Recent work by Guerrero and Hahn (PNAS 115:12787-12792, 2018) provided a significant step towards teasing apart the roles of homoplasy and hemiplasy in trait evolution by analyzing it with respect to the species tree and the gene trees within its branches.

Another evolutionary process whose extent and significance are better revealed by phylogenomic studies is hybridization between different species. In this work, we present a phylogenomic method for assessing the role of hybridization and introgression in the evolution of bi-allelic traits, including polymorphic ones. We apply the method to simulated evolutionary scenarios to demonstrate the interplay between the parameters of the evolutionary history and the role of introgression in a trait's evolution (which we call *xenoplasy*). Very importantly, we demonstrate, including on a biological data set, that inferring a species tree and using it for trait evolution analysis when hybridization had occurred could provide misleading hypotheses about trait evolution.

Introduction

Evolutionary biology began with the study of traits, and both descriptive and mechanistic explanations of trait evolution are key focuses of macroevolutionary studies today. Most famously, the beaks of Darwin's finches are an example of trait evolution in an adaptive radiation [5, 11, 12, 29].

With the development of next generation-sequencing and scalable computational methods, the use of whole or enriched genomes for phylogenetic inference has turbocharged systematics, synthesizing big genomic data into informative species trees [6, 27]. But the increased focus on species trees is not in competition with studies of trait evolution, rather it has revealed the complex relationship between speciation and trait evolution. Indeed, statistical methods for elucidating interspecific trait evolution without making use of the species tree could produce misleading results [10, 34], leading some to proclaim phylogenetics as the new genetics [31].

Given a hypothesized species tree inferred from available data, "congruent" trait patterns may be parsimoniously explained as having a single origin in some ancestral taxon, and are shared by all descendant taxa. However, many traits are "incongruent" and cannot be explained this way. These may be examples of convergent evolution, where traits have been gained or lost independently in different lineages. This kind of explanation is termed homoplasy, referring to a pattern of similarity which is not the result of common descent [16].

However, incongruent trait patterns can also be produced by discordant gene trees and ancestral character state polymorphism. In such cases, while the trait pattern is incongruence with the species tree, it is congruent with gene trees that differ from the species tree. This explanation when gene tree incongruence is due to incomplete lineage sorting (ILS) [32] is termed hemiplasy [1]. Inference of species trees from genomic data in the presence of ILS had attracted much attention in recent years, resulting in a wide array of species tree inference methods, including [22, 21, 26, 4, 28, 8, 35, 36]. However, the significance of elucidating not only the species tree but also the gene trees within its branches was recently highlighted for its significance in understanding trait evolution [15]. To the best of our knowledge, Guerrero and Hahn [14] devised the first method for assessing the role of hemiplasy in the evolution of a (binary) trait.

Another major source of species/gene tree introgression in eukaryotes is hybridization and consequent introgression [23]. Recently, the multispecies network coalescent was introduced as a model for unifying phylogenomic inference while accounting for both ILS and introgression [41, 42]. Several computational methods for inferring phylogenetic networks based on this model were then developed, many of which are reviewed in [7]. Hybridization and introgression could impact and help explain trait evolution [20], and methods for tracing the evolution of a trait on a phylogenetic network were introduced in [19, 2]. However, these methods do not take a "phylogenomic view" on trait evolution, i.e., they do not account for gene tree incongruence to tease apart homoplasy and hemiplasy.

Hibbins *et al.* [17] recently extended the method of [14] to account for introgression using the phylogenomic view. The focus of this work was still on distinguishing homoplasy from hemiplasy, with the possibility of introgression folded into these two categories. However, in introducing hemiplasy, Avise and Robinson [1] recommended: "Nevertheless, for epistemological clarity we recommend that the term hemiplasy not include these additional (and well appreciated) generators of phylogenetic discordance between gene trees and species trees but instead be confined to discordances that arise from idiosyncratic lineage sorting per se." In this case, the authors were specifically discussing hybridization and introgression as "additional generators of phylogenetic discordance." Following this recommendation, we propose the term "xenoplasy"[1] to explain a trait pattern that is incongruent with the species tree but whose incongruence could be explained by inheritance across species boundaries by means of hybridization and introgression. We illustrate this concept using the scenario of Fig. 1. In this case, taxa B and C diverged from their most recent common ancestor (MRCA) at time T_1, and their MRCA and taxon A diverged from their MRCA at time T_2. Furthermore, hybridization between taxa A and B occurred at time T_r, resulting in B's genome having some material that traces its evolution to the MRCA with C and others that were inherited from A via introgression. The character S is incongruent with the species tree, as A and B share the derived state 1, whereas C has the ancestral state 0. While hemiplasy of this trait is explained with respect to the gene tree drawn in solid lines and whose incongruence is due to ILS, xenoplasy is explained with respect to the gene tree drawn in dashed lines and whose incongruence is due to hybridization. Teasing apart homoplasy, hemiplasy, and xenoplasy has to do with the values of the different divergence and hybridization times, the migration rate, and the character state transition ($0 \rightarrow 1$ and $1 \rightarrow 0$) rates.

It is important to highlight here that in some cases there cannot be clear delineation of homoplasy, hemiplasy, and xenoplasy, as the evolution of trait could simultaneously involved convergence and genes whose evolutionary histories involve both ILS and introgression. In fact, the picture can get even more complex when the effects of gene duplication and loss are involved (maybe necessitating yet another term, e.g., "paraplasy," following the term "paralogy" that is used to describe genes whose ancestor is a duplication event).

Following the hemiplasy risk factor (HRF) of [14], in this work, we introduce the xenoplasy risk factor (XRF) to assess the role of introgression in the evolution of a given binary trait. We show that computing the XRF can be done with the readily available tools of [3] and [43]. An additional benefit of using these tools is accounting for polymorphic trait characters. We evaluated the XRF on simulated data involving ILS, introgression, and incongruent characters. Through this evaluation, we demonstrated the interplay among various factors, including divergence times, hybridization time, migration rate, and character transition rates, and how this interplay impacts the role of introgression in trait evolution. We also conducted a similar study on scenarios involving a polymorphic character. Furthermore, we demonstrated the importance of inferring a species phylogenetic network, instead of a species tree, when hybridization had occurred, in elucidating trait evolution. In particular, we showed how inferring a species tree despite the presence of hybridization and introgression yields misleading results on the roles of hemiplasy and

[1] Following the term "xenology," which was introduced to denote homologous genes that share ancestry through horizontal gene transfer [13].

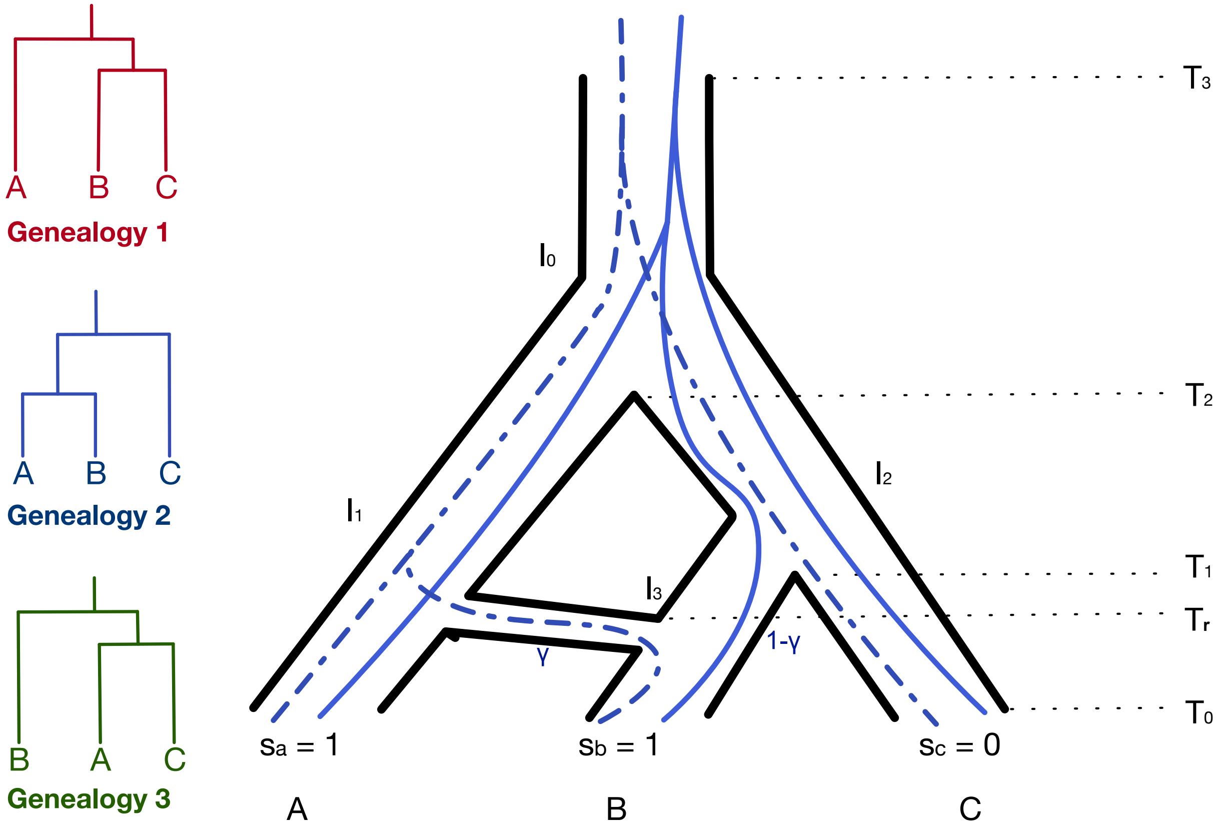

Figure 1: **Phylogenomic view of trait evolution in the presence of incomplete lineage sorting (ILS) and introgression.** Left: The three possible genealogies of three taxa A, B, and C. Right: Phylogenetic network that models an underlying species tree (A,(B,C)) along with a hybridization involving taxa A and C, and whose associate inheritance probability is γ. Two gene genealogies are shown within the branches of the phylogenetic network. The genealogy in solid lines involves ILS but no introgression, whereas the genealogy in dashed lines involves introgression but not ILS. The states S_a, S_b, and S_c of an incongruent binary character are shown at the leaves of the phylogenetic network.

introgression in the evolution of a given trait. Our work provides an additional advance towards bringing together phylogenetic inference and comparative methods in a phylogenomic context where both the species phylogeny and the phylogenies of individual loci are all taken into account.

Results

The Model and the Xenoplasy Risk Factor

Consider the evolutionary history depicted by the phylogenetic network of Fig. 1. If a single individual is sampled from each of the three species A, B, and C, then this network can be viewed as a mixture of two parental trees [44]: The "species" tree (A,(B,C)) and another tree that captures the genomic parts in B of introgressive descent ((A,B),C). The given trait whose character states are 1, 1, and 0 for taxa A, B, and C, respectively, could have evolved down and within the branches of the species tree. In this case, either homoplasy and hemiplasy could explain the trait evolution. To tease these two processes apart, assuming introgression did not play a role, the HRF [14] can be evaluated with respect to the species tree. Furthermore, doing a similar analysis on both parental trees can provide a way for assessing the role of hemiplasy in the presence of introgression, as in [17]. In our case, we are interested in answering a different question: How much does a reticulate evolutionary history involving hybridization and introgression explain the evolution of a trait as opposed to a strictly treelike evolutionary history?

To answer this question, we define the xenoplasy risk factor (XRF) in terms of the posterior odds ratio:

$$XRF(\mathcal{T}, \Psi, \psi, \mathcal{A}) = -\ln \frac{f(\Psi, \psi|\mathcal{A})}{f(\mathcal{T}, \psi|\mathcal{A})}, \tag{1}$$

where $f(.|.)$ is the posterior value, Ψ is a phylogenetic network that includes a species tree $\mathcal{T}$ and the reticulations under investigation, $\mathcal{A}$ is the trait pattern at the leaves of the phylogenies, and ψ is the mutation rate of the trait. In our case, we focus on bi-allelic traits; thus, ψ consists of the forward $(0 \rightarrow 1)$ mutation rate and the backward $(1 \rightarrow 0)$ mutation rate. Here, the phylogenetic network and species tree models consist of the topologies, divergence times, and population mutation rates. In the case of the network, the reticulation edge also includes the hybridization time as well as the inheritance probability [42]. The data consists of the trait pattern given by the states (0 or 1) of the individuals sampled. Polymorphism is allowed where a subset of individuals within the population have state 0 for the trait and the rest of the individuals have state 1. The likelihood of the trait pattern on the species tree integrates over all possible gene histories within the branches of the species tree and can is readily calculated using the method of [3]. Similarly, the likelihood of the trait pattern on the phylogenetic network integrates over all possible gene histories and is calculated using the method of [43]. Both methods work on bi-allelic traits, including polymorphic ones. Furthermore, while the model was illustrated above on three taxa, the methods of [3, 43] allow for any number of taxa and any topology of the phylogenies, including any number of hybridization events. However, in practice, increasing the number of taxa and the number of hybridizations results in a significant increase in the running time. Full details of the model and XRF computation are given in section *Metarials and Methods*.

All the simulation results described and discussed in the next two sections are based on data generated on the phylogenetic network and species tree therein of Fig. 1 while varying their associated evolutionary parameters.

The Interplay Between the Role of Introgression in Trait Evolution and Various Evolutionary Parameters

A phylogenomic view of the evolution of a bi-allelic trait on the phylogenetic network of Fig. 1 involves, in addition to the topologies of the phylogenetic network and species tree, roles for:

- The inheritance probability γ, which measures the proportion of the parental population A in the hybrid population B, and also correlates with the migration rate [42, 37].

- The hybridization time T_r, as it controls the likelihood of inheriting a character state by B from A, as well as the likelihood of such an inherited state becoming fixed in the population.

- The length of the internal species tree branch, $T_2 - T_1$, as it controls the amount of ILS and, consequently, hemiplasy.

- The population mutation rate, $2N_2\mu$, which also controls the amount of ILS and hemiplasy.

- The relative forward and backward character mutation rates ψ, which control the degree of homoplasy.

In this section, the character states are shown at the leaves of the network of Fig. 1. As we varied the values of all five parameters, there are too many plots to visualize the 2-, 3-, 4-, and 5-way interactions among all these parameters. We focus here on two groups of results: XRF as a function of the interplay among the internal branch length, the inheritance probability, and the relative forward/backward character mutation rates (results in Fig. 2), and XRF as a function of the interplay among the hybridization time, population mutation rate, and the relative forward/backward character mutation rates (results in Fig. 3).

As Fig. 2 shows, the role of introgression in the character evolution increases as the internal branch becomes longer and/or the inheritance probability becomes larger. These observations make sense. As the internal branch becomes longer, the amount of ILS and, consequently, hemiplasy decrease, increasing the roles of homoplasy and introgression. Furthermore, as the figure shows, as the forward/backward relative mutation rate increases, the role of introgression decreases, as indicated by decreasing XRF values for the same combination $(T_2 - T_1)$ and γ across the three panels in the figure from left to right. For example, the top right corner has the highest value in the leftmost panel and the lower value in the rightmost panel. This requires a more detailed discussion. The XRF measure,

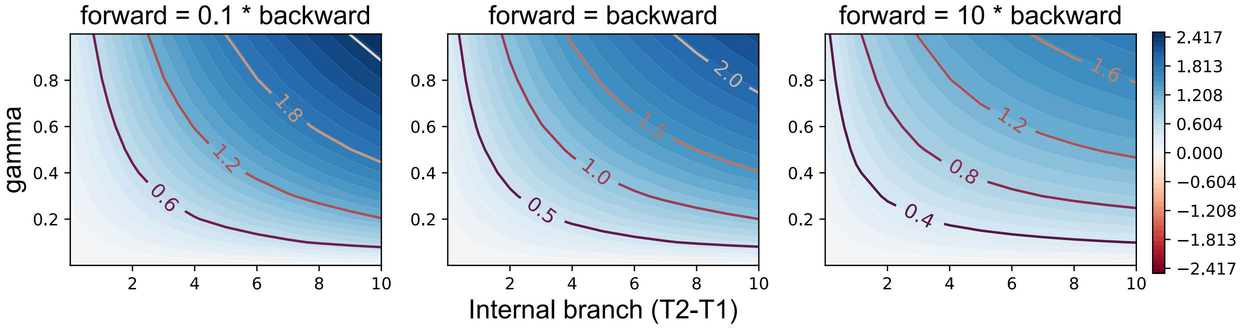

Figure 2: **The interplay among the inheritance probability, internal branch length, and forward/backward character mutation rates, and its impact on the role of introgression in the character's evolution.** Plots of the XRF values as a function of $T_2 - T_1$ and inheritance probability γ, on the x- and y-axis of each panel, respectively, and for three settings of the relative forward and backward character mutation rates. In all panel, $\theta = 0.01$ and $T_r = 0.1$ coalescent units.

as defined above, and implemented does not assume that 0 is the ancestral state and 1 is the derived state. Instead, it sums over both possibilities. Given this fact, let us first focus on the case where backward character mutations are 10 times as likely as forward character mutations (left panel of Fig. 2). As the internal branch becomes longer, the probability of the character states either remaining 0 or reverting back to 0 becomes higher. Therefore, invoking introgression as an explanation for the states 1 at both leaves (A and B) becomes more plausible, resulting in higher XRF values. In the case of the opposite setting where the forward character mutations are 10 times as likely as backward character mutations (right panel of Fig. 2), deriving state 1 at more than one species becomes more plausible through $0 \rightarrow 1$ mutations, especially as the internal branch becomes longer. Therefore, the XRF value is now lower, indicating that the simpler, tree-based hypothesis of homoplasy competes with introgression in explaining the trait pattern.

In the second set of results (Fig. 3), we focused on the interplay between the hybridization time, the population mutation rate, and the relative forward/backward character mutation rate. In this figure, the results are based on a scenario where the internal branch is too long for ILS to occur and, consequently, for hemiplasy to be a factor. Therefore, the two forces underlying trait evolution in this case are homoplasy and xenoplasy.

As the figure shows, the role of introgression increases as T_2 decreases, since the state 1 in taxon B is inherited from taxon A more recently and has less time to mutate back to state 0. The impact of θ in general could be more complex. A larger value of θ could mean a larger population size or a larger mutation rate (or both). Let us assume the mutation rate is fixed and that a larger value of θ stems from a larger population size. This means both more mutations and more ILS. In other words, we are now looking at a situation of all three forces of homoplasy, hemiplasy, and xenoplasy simultaneously at play. However, as we stated above, given that the internal branch is too long, an increase in θ from 0.001 to 0.01 is not sufficient here to cause much ILS. This is why we observe the lowest XRF

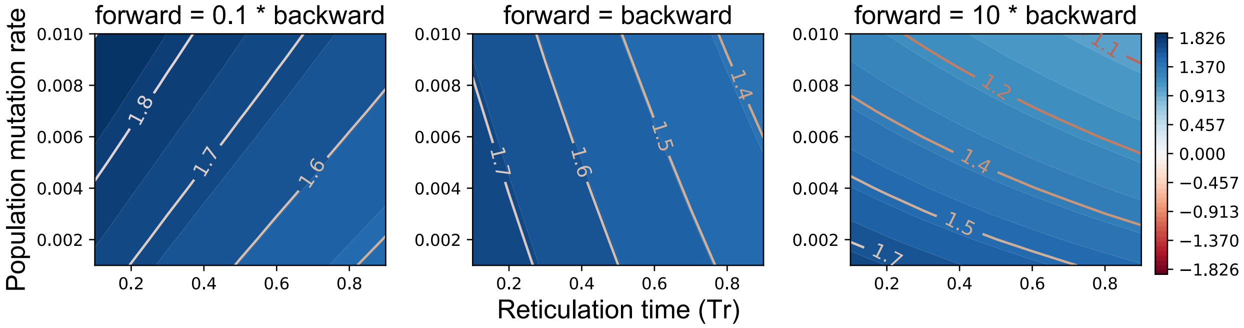

Figure 3: **The interplay among the hybridization time, population mutation rate, and forward/backward character mutation rates, and its impact on the role of introgression in the character's evolution.** Plots of the XRF values as a function of T_r and θ (population mutation rate), on the x- and y-axis of each panel, respectively, and for three settings of the relative forward and backward character mutation rates. In all panels $T_2 - T_1 = 10$ coalescent units and $\gamma = 0.5$.

values in the case with the higher character forward mutation rate, as homoplasy's role increases.

As the relative rate of forward character mutations to backward character mutations increases, the relation between T_r and θ with respect to impacting the XRF value changes from positive to negative. For smaller forward rates, increase in both T_r and θ maintains the same XRF value. For larger forward rates, an increase in T_r and decrease in θ maintain the same XRF value. More specifically, very recent hybridization and a high population mutation rate increase the role of introgression when forward mutations are less likely than backward mutations. The reason is that as T_r increases, the role of xenoplasy would decrease; however, as θ increases, the number of back mutations also increases, thus lowering the role of homoplasy. In other words, the combined effect of homoplasy and xenoplasy remain the same. On the other hand, very recent hybridization and a lower population mutation rate increase the role of xenoplasy when the forward mutation rate is higher. This is because lower θ means a smaller role of homoplasy, as we would have fewer $0 \rightarrow 1$ transitions.

Introgression and Polymorphic Traits

The elimination of polymorphism may greatly decrease the precision of relative analyses [39]. The XRF as defined above applies to polymorphic characters where some individuals within a species have state 0 and others have state 1. Furthermore, the methods of [3, 43] allow for polymorphic bi-allelic characters.

As above, we studied how various evolutionary parameters interact with introgression in explaining a trait pattern, but in this case, since we are interested in polymorphic characters, we added one more parameter to the model, which is the frequencies of the two states in taxon B (we assume taxa A and C to be monomorphic). A quick inspection of the results in Fig. 4 and Fig. 5 shows that the XFR values can now, under certain conditions, be much higher than the values we observed in the figures above. This indicates that introgression can potentially play a larger role in trait polymorphism. Once again, in these results, the internal branch is too long for ILS and homoplasy to be discernible factors.

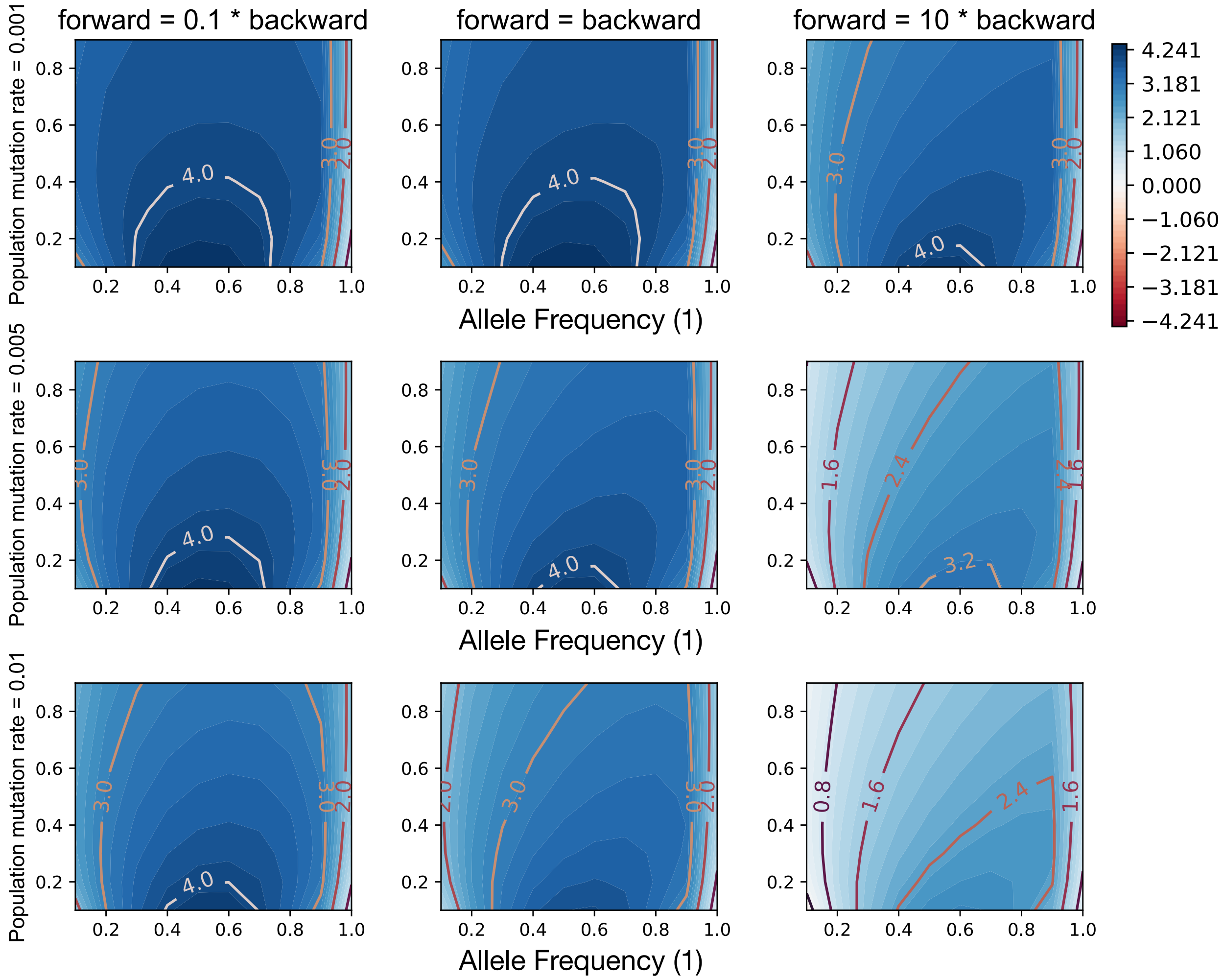

Figure 4: **XRF values in the presence of trait polymorphism.** The x- and y-axis in each panel correspond to the frequency of character state 1 in taxon B and the hybridization time T_r. Columns correspond to three different relative forward/backward character mutation rates and rows correspond to three different population mutation rates. In all panels $T_2 - T_1 = 10$ coalescent units and $\gamma = 0.5$.

In terms of individual evolutionary parameters, Fig. 4 shows that changes in the role of introgression as a function of the population mutation rate become more pronounced when the backward mutations occur at a much lower rate than the forward mutations.

Furthermore, as the forward character mutation rate increases, the role of introgression is larger for larger frequencies of allele 1 as observed by a right-shift of the darkest regions in the panels from left to right. We also observe that introgression plays the largest role in the case of the lowest forward mutation rate and lowest population mutation rate, and plays the smallest role in the case of the highest forward mutation and highest population mutation rate. The explanation for this is that in the former case homoplasy is less likely to determine the trait pattern due to lower $0 \rightarrow 1$ transitions.

Fig. 5 shows that higher frequencies of allele 1 and higher inheritance probabilities increase the role of introgression in general, especially as the hybridization is more recent. Regions in red colors in the figure correspond to a situation where introgression can be ruled out as a factor in the evolution of the trait pattern. The extreme case occurs when the 1 allele frequency is 10%, the forward mutation rate is 10 times that of the backward mutation rate, the hybridization is recent, and the inheritance probability is high. In this scenario, having an inheritance probability close to 1.0 and a very low 1 allele frequency are contradictory, since we would expect that at such a high inheritance probability almost all individuals in B to have state 1, in particular given that the hybridization is recent and inherited character state has not had enough time to revert back to state 0. Observe that this pattern becomes less likely as the frequency of allele 1 increases.

Misleading Results When Using an Inferred Species Tree Despite Introgression

We now turn our attention to a different aspect of trait evolution in the presence of hybridization. It was shown in [37, 7] that when the evolutionary history of a set of species is reticulate, inferring a species tree could result in a tree with much shorter branches. In such cases, the role of hemiplasy could be erroneously estimated to be larger simply because of the estimated short branches. This could in turn give the false impression that introgression did not play a role in the trait's evolutionary history. In other words, inferring a species tree despite the presence of hybridization could lead to misleading results not only about the evolutionary history of the species but also about the evolution of traits.

We illustrate this phenomenon on both a biological data set and a simulated data set. For the biological data set, we analyzed six species from the plant genus *Jaltomata*. In [25], the authors inferred a species tree of the *Jaltomata* species and hypothesized that trait patterns arose mostly by means of homoplasy. The study of [40] indicated that the evolutionary history of these species was reticulate, yet no phylogenetic network was inferred. We inferred a species phylogeny of this group in two different ways: We inferred a tree, ignoring the possibility of introgression, using the method of [3], and inferred a network using the method of [43]. The tree and network are shown in Fig. 6.

We first evaluated the HRF values of the species tree and the major tree inside the network of Fig. 6. We observed that one branch in the major tree is shorter than its

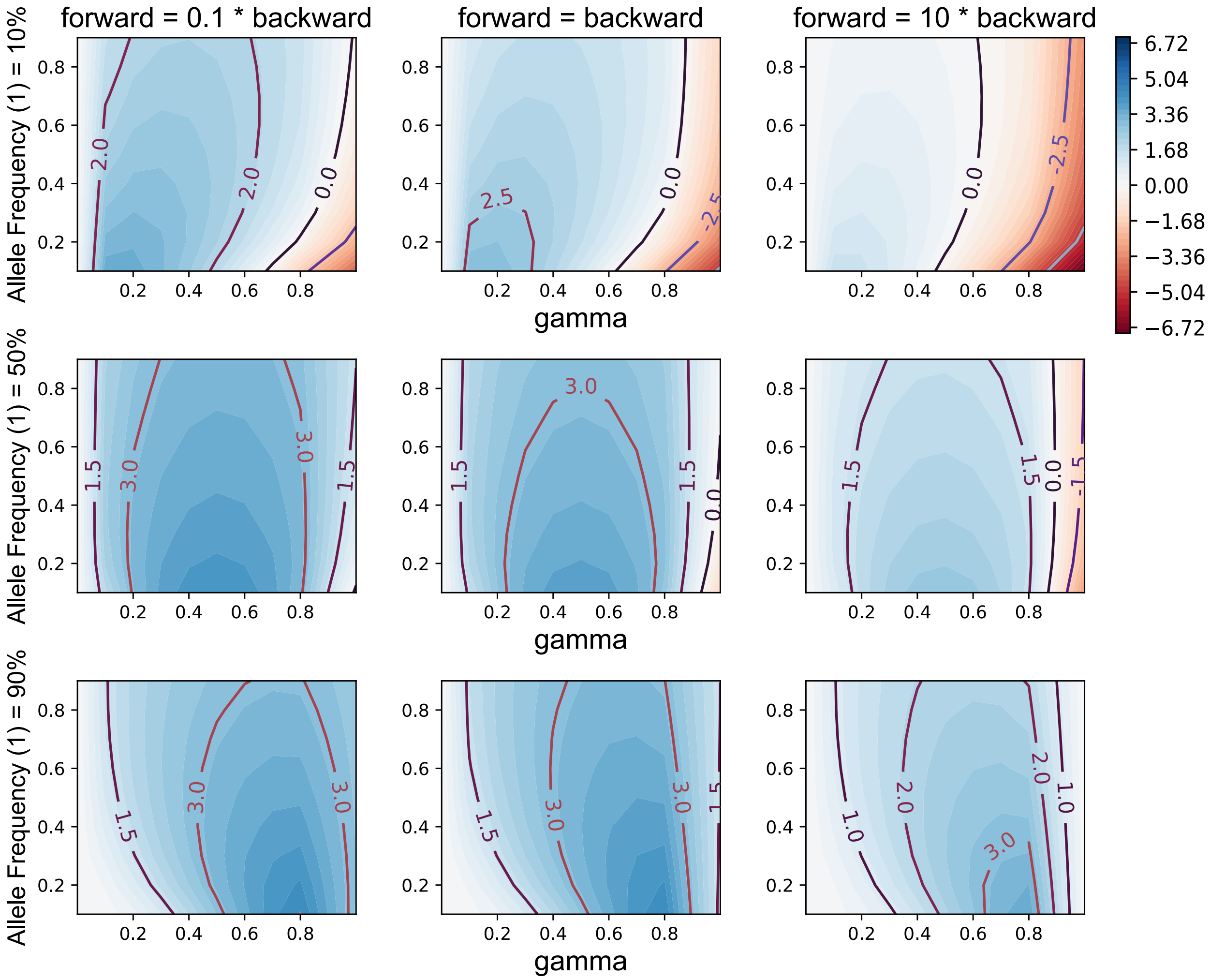

Figure 5: **XRF values in the presence of trait polymorphism.** The x- and y-axis in each panel correspond to the inheritance probability γ and hybridization time T_r, respectively. Columns correspond to three different relative forward/backward character mutation rates, and rows correspond to three different frequencies of all 1 in taxon B. In all panels $T_2 - T_1 = 10$ coalescent units and $\theta = 0.01$.

counterpart in the species tree, resulting in two orders of magnitude decrease in the HRF value. All other branches were longer in the major tree than their counterparts in the species tree, resulting in changes in the HRF values as large as nine and 21 orders of magnitude; Fig. 7 and Fig. 8.

Furthermore, we computed the XRF of the network of Fig. 6 and the network formed by adding a reticulation to the species tree of Fig. 6 to make it identical in topology to the network in the same figure. Here as well we found that, based on the XRF values of three trait patterns, as shown in Table 1, introgression played a larger role in the evolution of incongruent traits when using the phylogenetic network of Fig. 6 than when using the

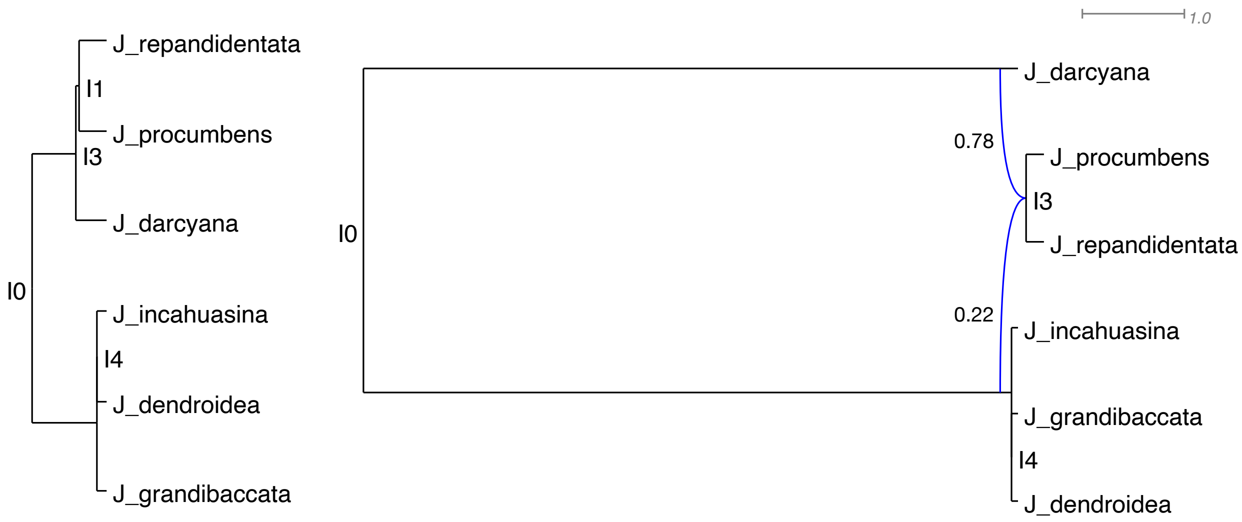

Figure 6: **Inferred species tree and network of the *Jaltomata* data set.** (a) Species tree inferred by [3]. (b) Species network inferred by [43]. The major tree inside the network is obtained by removing the red reticulation edge with inheritance probability 0.22.

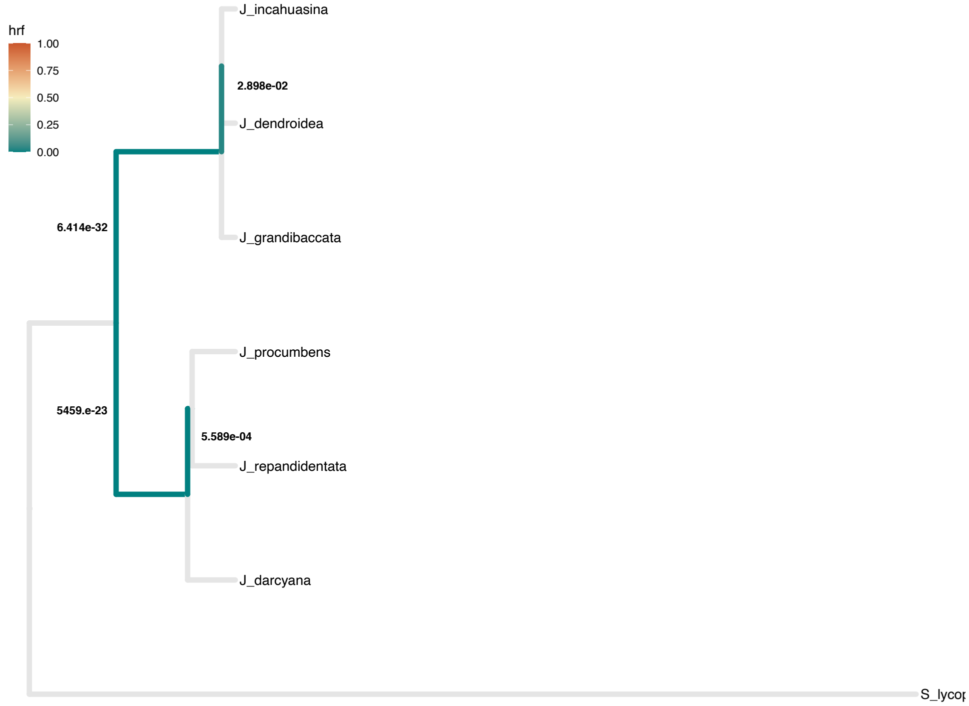

Figure 7: **HRF values of the *Jaltomata* species tree.** HRF values of the internal branches of the species tree (the tree shown in Fig. 6).

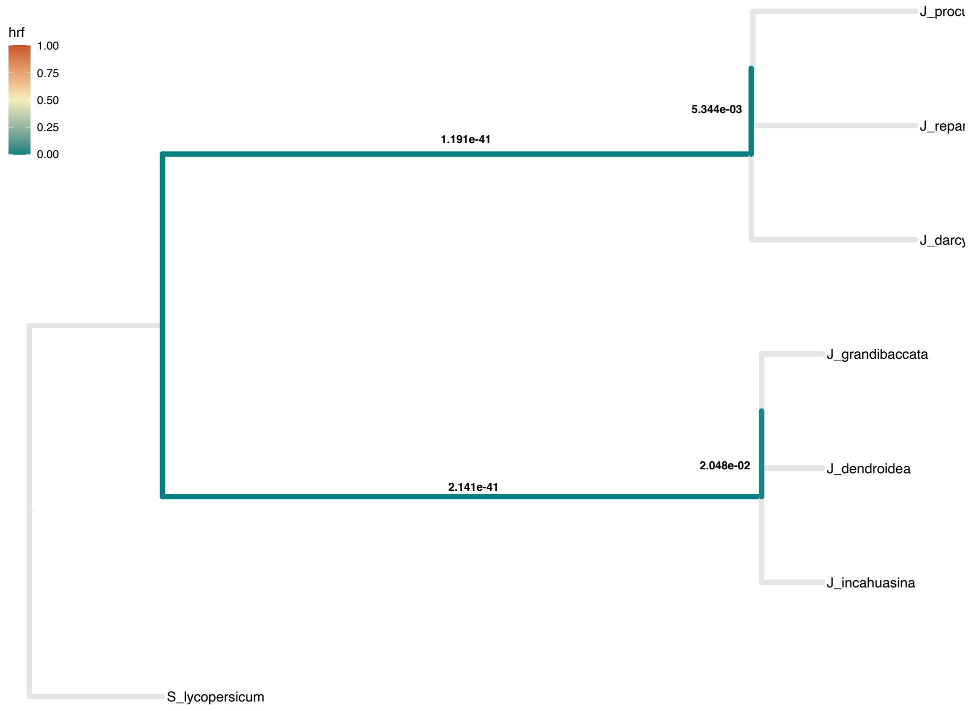

Figure 8: **HRF values of the *Jaltomata* major tree inside the phylogenetic network.** HRF values of the internal branches of the major tree inside the phylogenetic network in Fig. 6.

Table 1: **Three trait patterns of the Jaltomata species.** Congruence in this case is with respect to the species tree (inside the network).

Species	Congruent	Partial congruent	Incongruent
J. repandidentata	0	0	1
J. procumbens	0	1	1
J. darcyana	0	0	0
J. dendroidea	1	1	1
J. incahuasina	1	1	1
J. grandibaccata	1	1	1

network obtained from the species tree of Fig. 6; results in Fig. 9 and Fig. 10.

To further confirm these results, we repeated the same analysis but on a simulated data set where the true evolutionary history is known and involves two hybridizations (see Methods). Using molecular sequence data, we inferred a phylogenetic network and a species tree (see Methods). We then computed the HRF values of the species tree and the major tree inside the network, and found the values were much smaller in the latter case due to longer internal branches (which were the true values); Fig. 11 and Fig. 12.

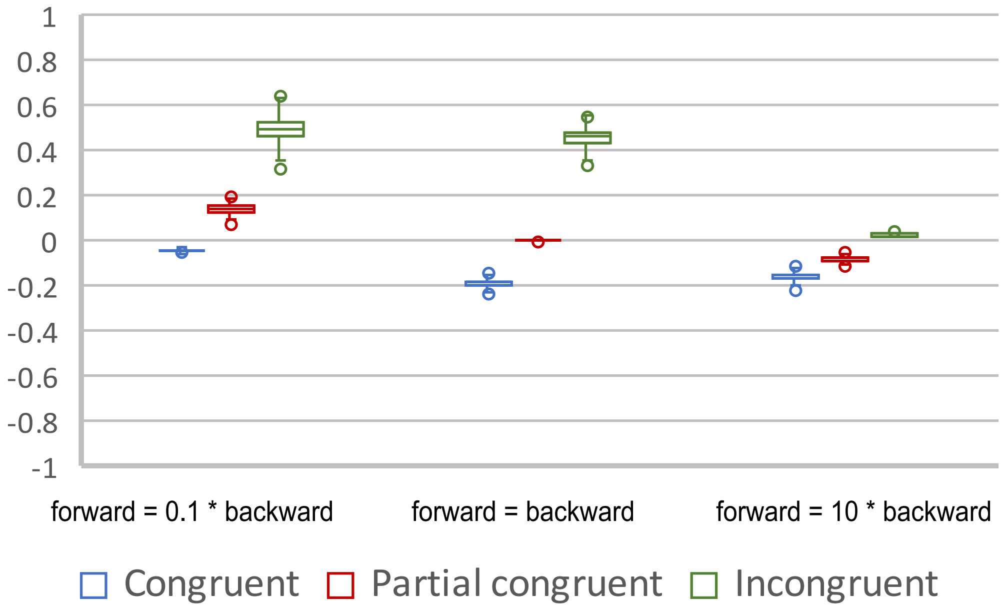

Figure 9: **XRF values of the three trait patterns of Table 1.** The evolutionary histories on which the XRF values were calculated based on the network and major tree inside it from Fig. 6. The x axis shows three different settings for the character mutation rates (forward: $0 \rightarrow 1$ mutation; backward: $1 \rightarrow 0$ mutation). The y axis shows the XRF values. Each box plot summarizes 3,000 XRF values obtained from the 3,000 networks sampled by MCMC_BiMarkers from the posterior distribution.

When calculating the XRF values, we found that since the inferred network was identical in topology and almost identical in the values of its evolutionary parameters, the posteriors of the inferred network (given trait patterns) were higher than those of the major tree inside the inferred network which, in turn, were higher than those of the inferred species tree; Fig. 13 and Fig. 14.

Discussion

The extent of hybridization and introgression continues to be revealed in an increasingly larger number of clades in the Eukaryotic branch of the Tree of Life [24]. Inferring a reticulate evolutionary history of a set of species is often complicated by the co-occurrence of ILS. Recently, much progress has been made on inferring phylogenetic networks of species in the presence of ILS [7]. These developments allow us to take a phylogenomic view of trait evolution that extends beyond homoplasy and hemiplasy, both of which, by their original definitions, tree-based. In this paper, we introduced the concept of xenoplasy to capture the inheritance of morphological character states via hybridization and introgression. We demonstrated how various evolutionary parameters impact the role

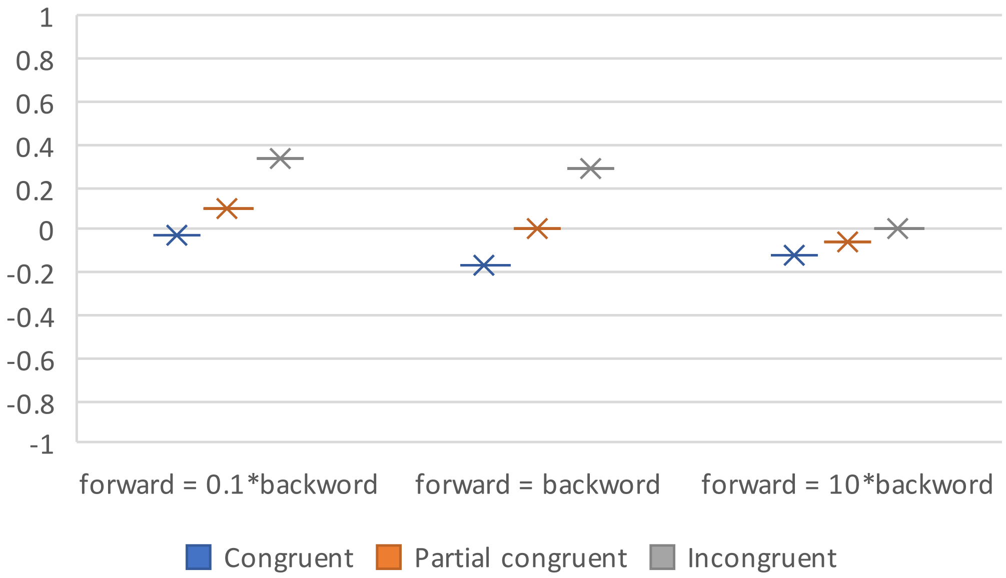

Figure 10: **XRF values of the three trait patterns of Table 1.** The evolutionary histories on which the XRF values were calculated based on the tree of Fig. 6 and the network obtained by adding a reticulation edge to it. The x axis shows three different settings for the character mutation rates (forward: $0 \to 1$ mutation; backward: $1 \to 0$ mutation). The y axis shows the XRF values.

that introgression could play in the evolution of a given trait, including polymorphic traits. We also demonstrated how inferring a species tree when the evolutionary history involved hybridization, thus effectively ignoring hybridization, yields misleading hypothesis about the forces behind trait evolution (in addition to producing an incorrect estimate of the evolutionary history of the species).

The XRF as defined above assumes the trait is bi-allelic and does not assume a species ancestral state. Allowing for a specific ancestral requires a simple modification to the methods of [3, 43], but moving beyond two states could require more substantial changes to the algorithms underlying these two methods. Furthermore, as we stated above, these two methods are in theory applicable to species trees and networks with any number of species and hybridizations and any number of individuals per species. However, the running time of both methods increases substantially on larger data sets, and the increase is much more significant in the case of networks than trees.

Finally, both methods of [3, 43] are based on the multispecies coalescent and its network extension to handle hybridization. This fact means that the XRF neither accounts for selection nor for other causes of incongruence such as gene duplication and loss. We believe the framework we presented here is general enough to be extended to handle such cases, though some extensions would require significant algorithmic developments.

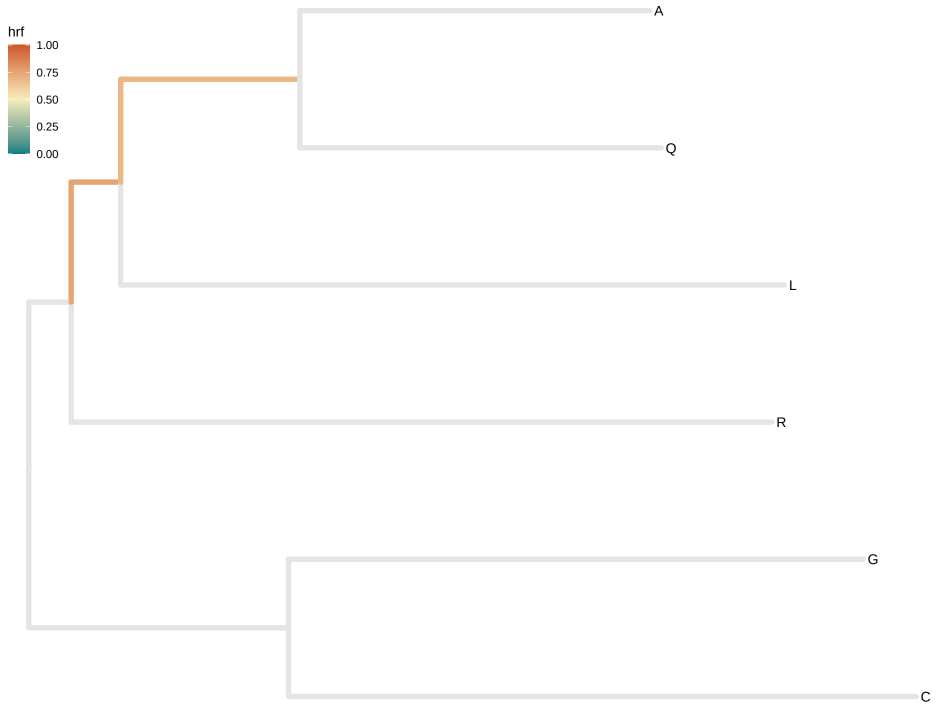

Figure 11: The HRF of the species tree estimated by starBEAST2.

Methods

The Xenoplasy Risk Factor

Consider that a bi-allelic trait evolving along the branches of species tree or species network Ψ with population sizes and divergence times Θ. The trait is given by **A** which specifies for each species whether the species has state 0 or state 1. In the case of a polymorphic trait, for each species, **A** specifies the fraction of individuals within the species with state 0. Furthermore, let ψ consist of two parameters: the forward character mutation rate (the character mutating from state 0 to state 1) and the backward character mutation rate (the character mutating from state 1 to state 0).

The posterior probability of the species phylogeny and associated parameters given **A** is given by

$$f\left(\Psi,\Theta|\mathcal{A}\right) = \frac{1}{f\left(\mathcal{A}\right)} f\left(\Psi,\Theta\right) f\left(\mathcal{A}|\Psi,\Theta\right) \propto f\left(\Psi,\Theta\right) f\left(\mathcal{A}|\Psi,\Theta\right), \tag{2}$$

where $f\left(\Psi,\Theta\right)$ is the prior on the species phylogeny and associated parameters and $f\left(\mathcal{A}|\Psi,\Theta\right)$ is the likelihood.

In the phylogenomic view of trait evolution, the evolutionary history of **A** is intertwined with that of the gene genealogies evolving inside the species phylogeny. To calcu-

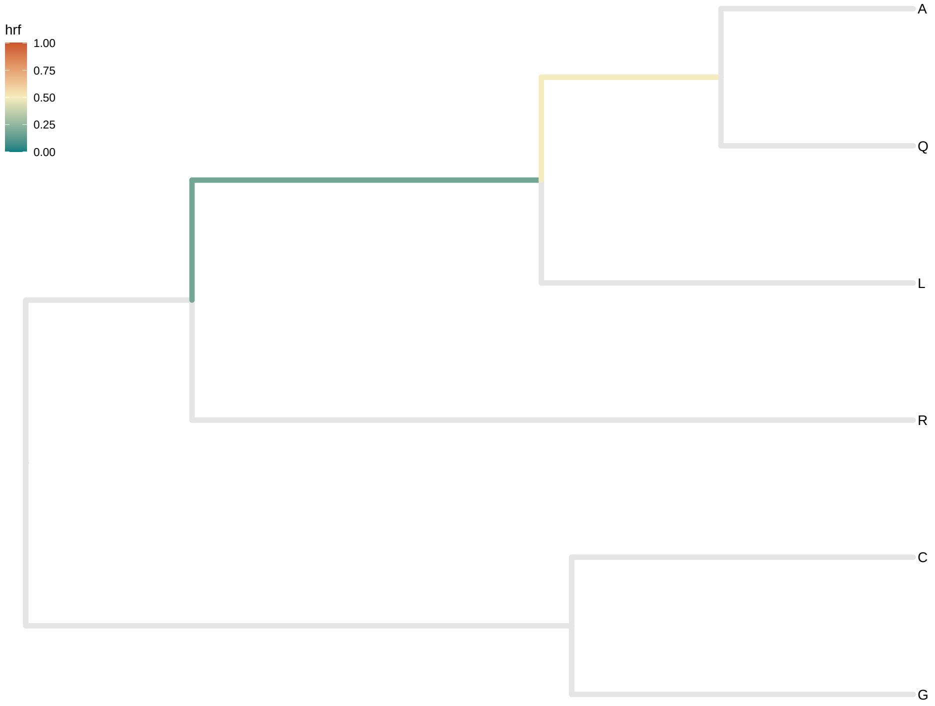

Figure 12: The HRF of the major tree of the species network estimated by MCMC_SEQ.

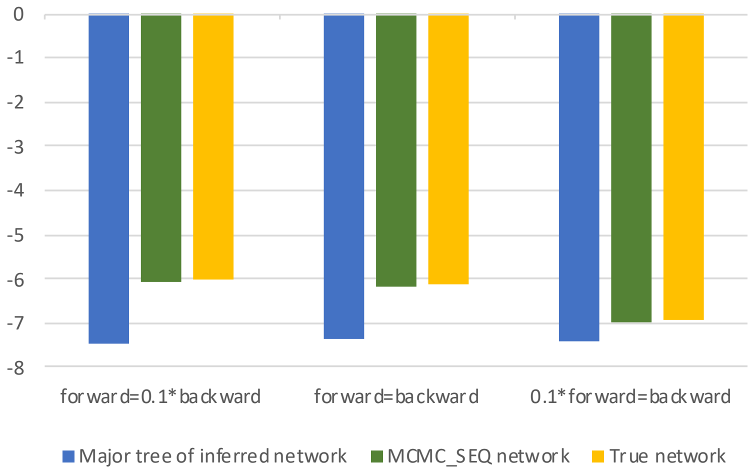

Figure 13: The natural logarithm of the posterior probability of species phylogenies given the trait pattern that A and C got derived trait pattern.

late the likelihood of a given trait pattern, we need to integrate over all possible genealo-

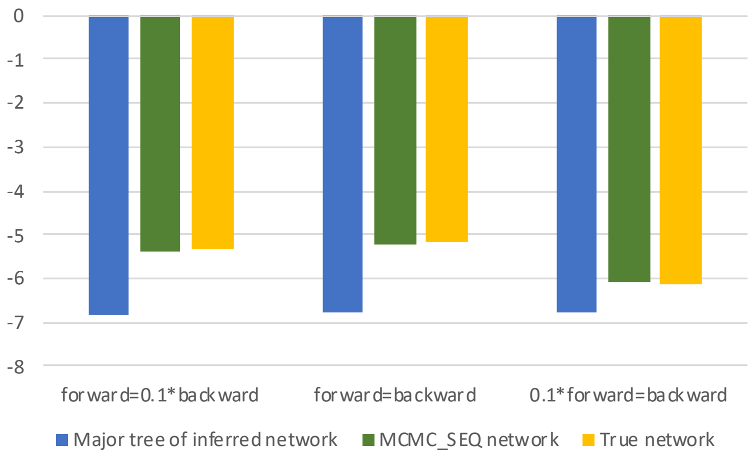

Figure 14: The natural logarithm of the posterior probability of species phylogenies given the trait pattern that Q and R got derived trait pattern.

gies G that can derive **A** and substitution model parameters ψ:

$$f\left(\mathcal{A}|\Psi,\Theta\right)=\int_{\psi}\int_{G} f\left(\mathcal{A}|G,\psi\right) f\left(G|\Psi,\Theta\right) f\left(\psi\right) dGd\psi. \tag{3}$$

Here, $f\left(\mathcal{A}|G,\psi\right)$ is the likelihood of a gene genealogy given the trait pattern, $f\left(G|\Psi,\Theta\right)$ is the likelihood function under the multispecies coalescent (or multispecies network coalescent), and $f\left(\psi\right)$ is the prior on the same parameters.

The methods of [3, 43] calculate $f\left(\Psi,\Theta|\mathcal{A}\right)$ according to Eq. (2) when Ψ is a species tree and when Ψ is a species phylogenetic network. Both methods are implemented in PhyloNet [33, 38].

Finally, the XRF is calculated as the negative natural log of the posterior odds ratio, as given by (1).

Parameter Settings of the 3-taxon Species Network Simulation Study

The Newick string for the phylogenetic network of Fig. 1 is

$$((A, I3\#H1)I1, ((B)I3\#H1, C)I2)I0;$$

We varied the ILS level by varying the internal branch length $(T_2 - T_1)$. The default value of each branch is one coalescent unit while we varied $(T_2 - T_1)$ from 0.001 to 10 to represent a range from very high to very low levels of ILS. Two factors controlled the introgression: the inheritance probability γ and the hybridization time T_r. The inheritance probability γ was varied between 0.0 and 1.0. The hybridization time T_r was varied between 0.0 and 1.0 coalescent units. We varied the population mutation rate θ between 0.001 and 0.01. For the character mutation rate, we used three settings: forward = 0.1×backward, forward = backward and forward = 10×backward. For the polymorphic trait, we varied the frequency of allele '1' in taxon B from 0 to 1.

The Jaltomata Data Set

We obtained the bi-allelic marker data from the original data of [40] The original dataset contained the sequence alignments of 6,431 orthologous genes of the six selected species. To derive conditionally independent bi-allelic markers of the original dataset, we randomly selected one site from each gene and obtained 6,409 valid bi-allelic markers in total.

We inferred the species tree and phylogenetic network of the Jaltomata species with MCMC_BiMarkers [43] as implemented in PhyloNet [33, 38] with chain length 5×10^6, burn-in 2×10^6, and sample frequencies 1000, using the following command:

```
MCMC_BiMarkers -taxa (JA0701, JA0456, JA0694, JA0010, JA0719, JA0816)
-cl 5000000 -bl 2000000 -sf 1000 -mr 1
```

Setting the `-mr` value to 0 amounts to running the method of [3] to infer the species tree (as the number of reticulations is set to 0). The MAP (Maximum a posteriori) estimates of the species tree and phylogenetic network are shown in Fig. 1. The *effective sample size* (ESS) of the parameter values of the MCMC chains were higher than 2321 for the species tree and higher than 1583 for the species network.

Simulated Data Set for Showing the Effect of Inferring Species Tree Despite Introgression

We simulated sequence alignments on 128 loci from the phylogenetic network shown in Fig. 15, whose topology was discovered as the phylogeny of Anopheline mosquitoes by [9] and was analyzed with simulated data in [37].

Generating gene trees We generated 128 gene trees with ms [18] given the species network in Fig. 15. The command is as follows.

```
ms 6 128 -T -I 6 1 1 1 1 1 1 -es 0.25 5 0.3 -es 0.25 3 0.8
 -ej 0.5 7 3 -ej 0.5 8 2 -ej 0.75 6 5 -ej 1.0 3 4 -ej 1.0 2 1
 -ej 2.0 5 4 -ej 2.5 4 1
```

Generating multi-locus sequence alignments At each locus, we generated the sequence alignment given the gene tree with seq-gen [30]. We set the length of sequences to be 500 bps, and utilized GTR model with base frequencies 0.2112,0.2888,0.2896,0.2104 (A,C,G,T) and transition probabilities 0.2173,0.9798,0.2575,0.1038,1.0,0.207. We set the population mutation rate $\theta = 0.036$, so the scale $-s$ is 0.018. The command is as follows.

```
seq-gen -mGTR -s0.018 -f0.2112,0.2888,0.2896,0.2104
-r0.2173,0.9798,0.2575,0.1038,1.0,0.207  -l500
```

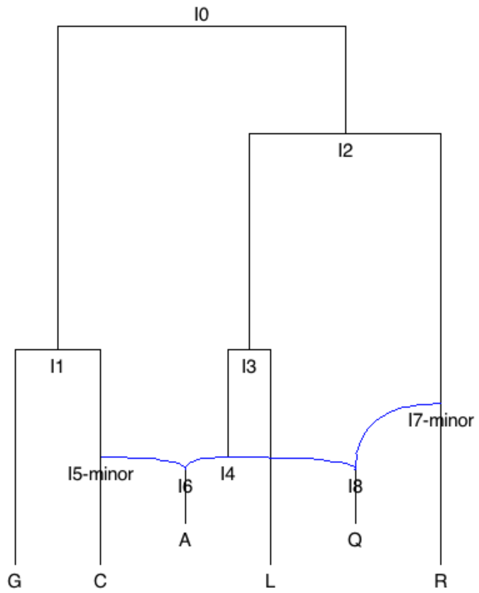

Figure 15: The true species network used in the simulation study.

Species Tree and Network Inference Similar to [37], we analyzed the phylogenetic network with MCMC_SEQ [37] in PhyloNet [33, 38] and phylogenetic tree with StarBEAST2 [28].

We inferred the network with MCMC_SEQ under GTR model with chain length 5×10^7, burn-in 1×10^7 and sample frequencies 5000. We fixed the population mutation rate $\theta = 0.036$ and GTR parameters to be true parameters.

```
MCMC_SEQ -cl 60000000 -bl 10000000 -sf 5000 -pl 8
-tm <A:A_0;C:C_0;G:G_0;L:L_0;Q:Q_0;R:R_0> -fixps 0.036
-gtr (0.2112,0.2888,0.2896,0.2104,0.2173,0.9798,0.2575,0.1038,1,0.2070);
```

The phylogenetic network inferred by MCMC_SEQ is shown in Fig. 16.

We ran StarBEAST2 to infer a species tree on the same dataset with chain length 10^8 and sample frequency $50,000$ under GTR model with empirical base frequencies and transition probabilities used for generating data. Population sizes were sampled for the individual branches (i.e., a fixed population size across all branches was *not* assumed). The species tree inferred by StarBEAST2 is shown in Fig. 17.

Acknowledgments

This work was funded by National Science Foundation Grants DBI 2030604, CCF 1514177, CCF 1800723, and DMS 1547433 (to L.N.).

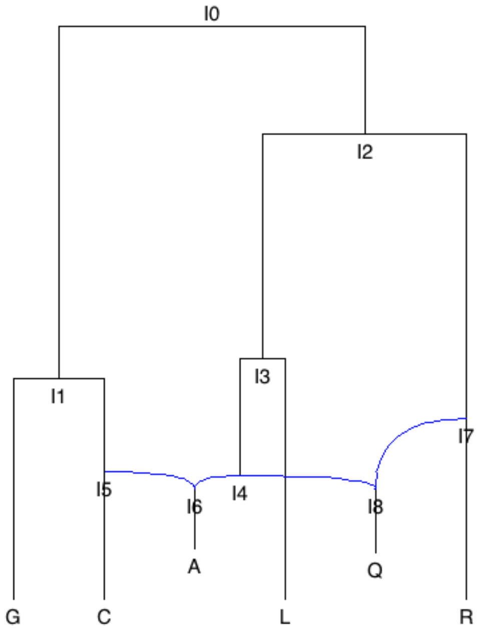

Figure 16: The species network estimated by MCMC_SEQ.

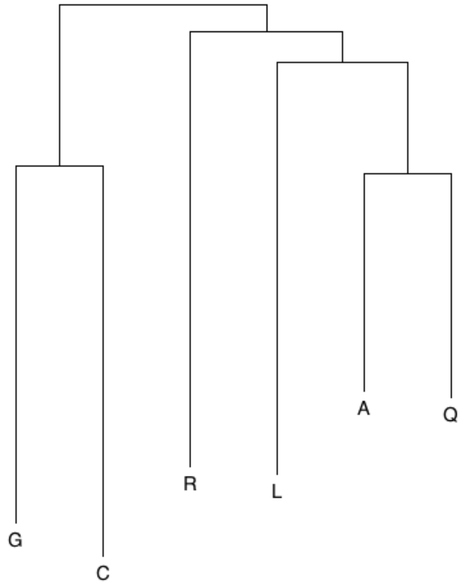

Figure 17: The species tree estimated by starBEAST2.

References

[1] John C. Avise and Terence J. Robinson. Hemiplasy: A New Term in the Lexicon of Phylogenetics. *Systematic Biology*, 57(3):503–507, 06 2008.

[2] Paul Bastide, Claudia Solís-Lemus, Ricardo Kriebel, K William Sparks, and Cécile Ané. Phylogenetic comparative methods on phylogenetic networks with reticulations. *Systematic biology*, 67(5):800–820, 2018.

[3] David Bryant, Remco Bouckaert, Joseph Felsenstein, Noah A. Rosenberg, and Arindam RoyChoudhury. Inferring Species Trees Directly from Biallelic Genetic Markers: Bypassing Gene Trees in a Full Coalescent Analysis. *Molecular Biology and Evolution*, 29(8):1917–1932, 03 2012.

[4] Julia Chifman and Laura Kubatko. Quartet inference from snp data under the coalescent model. *Bioinformatics*, 30(23):3317–3324, 2014.

[5] C. Darwin. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favored Races in the Struggle for Life*. Modern Library, New York, 1859.

[6] Scott V Edwards. Is a new and general theory of molecular systematics emerging? *Evolution: International Journal of Organic Evolution*, 63(1):1–19, 2009.

[7] RA Leo Elworth, Huw A Ogilvie, Jiafan Zhu, and Luay Nakhleh. Advances in computational methods for phylogenetic networks in the presence of hybridization. In *Bioinformatics and Phylogenetics*, pages 317–360. Springer, 2019.

[8] Tomáš Flouri, Xiyun Jiao, Bruce Rannala, and Ziheng Yang. Species tree inference with bpp using genomic sequences and the multispecies coalescent. *Molecular Biology and Evolution*, 35(10):2585–2593, 2018.

[9] Michael C Fontaine, James B Pease, Aaron Steele, Robert M Waterhouse, Daniel E Neafsey, Igor V Sharakhov, Xiaofang Jiang, Andrew B Hall, Flaminia Catteruccia, Evdoxia Kakani, et al. Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science*, 347(6217), 2015.

[10] László Zsolt Garamszegi. *Modern phylogenetic comparative methods and their application in evolutionary biology: concepts and practice*. Springer, 2014.

[11] Peter R Grant. Speciation and the adaptive radiation of darwin's finches: the complex diversity of darwin's finches may provide a key to the mystery of how intraspecific variation is transformed into interspecific variation. *American Scientist*, 69(6):653–663, 1981.

[12] Peter R Grant and B Rosemary Grant. Adaptive radiation of darwin's finches: Recent data help explain how this famous group of galapagos birds evolved, although gaps in our understanding remain. *American Scientist*, 90(2):130–139, 2002.

[13] Gary S Gray and Walter M Fitch. Evolution of antibiotic resistance genes: the DNA sequence of a kanamycin resistance gene from Staphylococcus aureus. *Molecular Biology and Evolution*, 1(1):57–66, 1983.

[14] Rafael F. Guerrero and Matthew W. Hahn. Quantifying the risk of hemiplasy in phylogenetic inference. *Proceedings of the National Academy of Sciences*, 115(50):12787–12792, 2018.

[15] Matthew W Hahn and Luay Nakhleh. Irrational exuberance for resolved species trees. *Evolution*, 70(1):7–17, 2016.

[16] Brian K Hall. Descent with modification: the unity underlying homology and homoplasy as seen through an analysis of development and evolution. *Biological Reviews*, 78(3):409–433, 2003.

[17] Mark S Hibbins, Matthew JS Gibson, and Matthew W Hahn. Determining the probability of hemiplasy in the presence of incomplete lineage sorting and introgression. *bioRxiv*, 2020.

[18] Richard R Hudson. Generating samples under a wright–fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338, 2002.

[19] Dwueng-Chwuan Jhwueng and Brian C O'Meara. Trait evolution on phylogenetic networks. *bioRxiv*, page 023986, 2015.

[20] Nisa Karimi, Corrinne E Grover, Joseph P Gallagher, Jonathan F Wendel, Cécile Ané, and David A Baum. Reticulate evolution helps explain apparent homoplasy in floral biology and pollination in baobabs (adansonia; bombacoideae; malvaceae). *Systematic Biology*, 69(3):462–478, 2020.

[21] Liang Liu and Lili Yu. Estimating species trees from unrooted gene trees. *Systematic Biology*, 60(5):661–667, 2011.

[22] Liang Liu, Lili Yu, and Scott V Edwards. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evolutionary Biology*, 10(1):302, 2010.

[23] Wayne P. Maddison. Gene Trees in Species Trees. *Systematic Biology*, 46(3):523–536, 09 1997.

[24] James Mallet, Nora Besansky, and Matthew W Hahn. How reticulated are species? *BioEssays*, 38(2):140–149, 2016.

[25] Ryan J Miller, Thomas Mione, Hanh-La Phan, and Richard G Olmstead. Color by numbers: Nuclear gene phylogeny of jaltomata (solanaceae), sister genus to solanum, supports three clades differing in fruit color. *Systematic Botany*, 36(1):153–162, 2011.

[26] Siavash Mirarab, Rezwana Reaz, Md S Bayzid, Théo Zimmermann, M Shel Swenson, and Tandy Warnow. Astral: genome-scale coalescent-based species tree estimation. *Bioinformatics*, 30(17):i541–i548, 2014.

[27] Luay Nakhleh. Computational approaches to species phylogeny inference and gene tree reconciliation. *Trends in ecology & evolution*, 28(12):719–728, 2013.

[28] Huw A Ogilvie, Remco R Bouckaert, and Alexei J Drummond. Starbeast2 brings faster species tree inference and accurate estimates of substitution rates. *Molecular Biology and Evolution*, 34(8):2101–2114, 2017.

[29] K Petren, PR Grant, BR Grant, and LF Keller. Comparative landscape genetics and the adaptive radiation of darwin's finches: the role of peripheral isolation. *Molecular Ecology*, 14(10):2943–2957, 2005.

[30] Andrew Rambaut and Nicholas C Grass. Seq-gen: an application for the monte carlo simulation of dna sequence evolution along phylogenetic trees. *Bioinformatics*, 13(3):235–238, 1997.

[31] Stacey D Smith, Matthew W Pennell, Casey W Dunn, and Scott V Edwards. Phylogenetics is the new genetics (for most of biodiversity). *Trends in Ecology & Evolution*, 2020.

[32] Fumio Tajima. Evolutionary relationship of dna sequences in finite populations. *Genetics*, 105(2):437–460, 1983.

[33] Cuong Than, Derek Ruths, and Luay Nakhleh. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics*, 9(1):322, 2008.

[34] Josef C Uyeda, Rosana Zenil-Ferguson, and Matthew W Pennell. Rethinking phylogenetic comparative methods. *Systematic Biology*, 67(6):1091–1109, 2018.

[35] Yaxuan Wang and Luay K Nakhleh. Towards an accurate and efficient heuristic for species/gene tree co-estimation. *Bioinformatics*, 34 17:i697–i705, 2018.

[36] Yaxuan Wang, Huw A Ogilvie, and Luay Nakhleh. Practical Speedup of Bayesian Inference of Species Phylogenies by Restricting the Space of Gene Trees. *Molecular Biology and Evolution*, 37(6):1809–1818, 02 2020.

[37] Dingqiao Wen and Luay Nakhleh. Coestimating reticulate phylogenies and gene trees from multilocus sequence data. *Systematic Biology*, 67(3):439–457, 2017.

[38] Dingqiao Wen, Yun Yu, Jiafan Zhu, and Luay Nakhleh. Inferring phylogenetic networks using PhyloNet. *Systematic Biology*, 67(4):735–740, 2018.

[39] John J. Wiens. Polymorphism in systematics and comparative biology. *Annual Review of Ecology and Systematics*, 30(1):327–362, 1999.

[40] Meng Wu, Jamie L. Kostyun, Matthew W. Hahn, and Leonie C. Moyle. Dissecting the basis of novel trait evolution in a radiation with widespread phylogenetic discordance. *Molecular Ecology*, 27(16):3301–3316, 2018.

[41] Yun Yu, James H Degnan, and Luay Nakhleh. The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. *PLoS Genet*, 8(4):e1002660, 2012.

[42] Yun Yu, Jianrong Dong, Kevin J Liu, and Luay Nakhleh. Maximum likelihood inference of reticulate evolutionary histories. *Proceedings of the National Academy of Sciences*, 111(46):16448–16453, 2014.

[43] Jiafan Zhu, Dingqiao Wen, Yun Yu, Heidi M Meudt, and Luay Nakhleh. Bayesian inference of phylogenetic networks from bi-allelic genetic markers. *PLoS computational biology*, 14(1):e1005932, 2018.

[44] Jiafan Zhu, Yun Yu, and Luay Nakhleh. In the light of deep coalescence: revisiting trees within networks. *BMC bioinformatics*, 17(14):415, 2016.