Exploring the distribution of phylogenetic networks generated under a birth-death-hybridization process

Joshua A. Justison[1*] and Tracy A. Heath[1]

November 10, 2022

Author Affiliations
[1]Department of Ecology, Evolution, & Organismal Biology, Iowa State University, Ames, Iowa USA
Email
[*]justisnojo@gmail.com

ABSTRACT
Gene-flow processes such as hybridization and introgression play important roles in shaping diversity across the tree of life. Recent studies extending birth-death models have made it possible to investigate patterns of reticulation in a macroevolutionary context. These models allow for different macroevolutionary patterns of gene flow events that can either add, maintain, or remove lineages, thus creating complex diversification scenarios. Further, many reticulate phylogenetic inference methods assume specific reticulation structures or phylogenies belonging to certain network classes. However, the distributions of phylogenetic networks under reticulate birth-death processes are poorly characterized and it is unknown whether they violate common methodological assumptions. We use a combination of analytical and simulation techniques to explore phylogenetic network space under a birth-death-hybridization process. Specifically, we measured the number of lineages through time and role of hybridization in diversification along with the proportion of phylogenetic networks that belong to commonly used network classes (*e.g.,* tree-child, tree-based, or level-1 networks). We find that the growth of phylogenetic networks and class membership is largely affected by the assumptions about macroevolutionary patterns of gene flow. In accordance with previous studies, a lower proportion of networks belonged to these classes based on type and density of reticulate events. However, under a birth-death-hybridization process, these factors form an antagonistic relationship; the type of reticulation events that cause high membership proportions also lead to the highest reticulation density, consequently lowering the overall proportion of phylogenies in some classes. Further, we observed that genetic distance–dependent gene flow and incomplete sampling increase the proportion of class membership, primarily due to having fewer reticulate events. Our results can inform studies if their biological expectations of gene flow are associated with evolutionary histories that satisfy the assumptions of current methodology and aid in finding phylogenetic classes that are relevant for methods development.

Keywords: birth-death-hybridization process, phylogenetic networks, simulation

1 Introduction

Gene flow—the exchange of genetic material between species—has been observed in a number of systems (Taylor and Larson, 2019) and plays an important role in shaping Earth's biodiversity (Stebbins, 1959; Bock, 2010; Mallet et al., 2016). Phylogenetic networks are used to directly infer reticulate histories and are vital in informing our understanding of adaptive radiations (Ottenburghs et al., 2016), mimicry patterns (Edelman et al., 2019), and even the phylodynamics of infectious diseases like COVID-19 (Wang et al., 2022). The field has recently seen an explosion of new methods for estimating phylogenetic networks (Elworth et al., 2019). However, due to the complex topological landscape of phylogenetic networks, robust model-based techniques remain limited in their ability to scale to large datasets (Hejase and Liu, 2016). The birth-death-hybridization process (Morin and Moret, 2006; Woodhams et al., 2016; Zhang et al., 2018) is an extension

of a model commonly used to describe lineage diversification (*i.e.,* the birth-death process, Kendall, 1948; Nee et al., 1994b), and provides a statistical framework for understanding the generation of phylogenetic networks. With this model, we can understand how reticulate processes influence patterns of diversification and affect the distribution of species across the tree of life.

Birth-death processes are used to model the phylogenetic branching process on a macroevolutionary scale, uncovering patterns of speciation and extinction (Kendall, 1948; Raup, 1985; Nee et al., 1994b; Nee, 2006). A rich theoretical framework for this process allows us to describe the distribution of branch lengths (Steel and Mooers, 2010; Mooers et al., 2012), growth dynamics (Nee et al., 1994a; Stadler, 2008), and topological properties (Mooers and Heard, 1997; Lambert and Stadler, 2013) of phylogenetic trees. The birth-death family of models has proven useful for estimating diversification rates (Nee et al., 1994b,a; Höhna et al., 2015) and choosing biologically motivated priors in Bayesian analyses (Rannala and Yang, 1996; Velasco, 2008). However, analyses of divergence times (Pybus and Harvey, 2000) and tree shape (Mooers and Heard, 1997; Heath et al., 2008b; Jones, 2011) have shown that empirical phylogenies differ from those expected under a constant rate birth-death process. To more accurately explain the observed patterns, these statistical models have undergone several extensions to account for a wide range of biological processes and systematic sampling challenges.

The birth-death-hybridization process (formally described below) generates phylogenetic networks under a process that allows lineages to undergo speciation, extinction, and reticulation (hybridization). Although—for eponymous consistency—events are denoted as hybridizations, the process does not make any mechanistic assumptions about gene flow; reticulate events on phylogenetic networks can model the many processes where genetic material is exchanged between lineages. Consequently, phylogenetic network inference has been used to describe histories of introgression (Thawornwattana et al., 2018; Myers et al., 2022), hybridization (Morales-Briones et al., 2018; Dolinay et al., 2021), and lateral gene transfer (Betat et al., 2015). The specific biological interpretation of a given reticulation event is a challenge that depends largely on the system under investigation and is further complicated by how reticulation events are modeled on the phylogenetic network (Hibbins and Hahn, 2022).

Even though phylogenetic network methods are largely agnostic to the underlying cause of reticulation, existing implementations of the birth-death-hybridization process make different assumptions about how reticulation events are represented on the network. We denote each type of reticulation—lineage generative, lineage neutral, and lineage degenerative—based on how the events change the number of lineages on the phylogenetic network (Box 1). The structure and timing of each reticulation type can provide some biological meaning to reticulation events (Box 1; but see Hibbins and Hahn, 2022). Consequently, caution is needed when trying to infer and interpret reticulate patterns; the lineage effects from the types of reticulations can have far reaching implications on species diversification dynamics.

Box 1: Description of Reticulation Types

Patterns of gene flow can take different shapes on phylogenetic networks. We differentiate types of reticulation depending on changes in the number of lineages on the phylogeny after an event (Fig 1).

- **Lineage Generative**: A new lineage is created (denoted as m-type in Janssen and Liu, 2021)
- **Lineage Neutral**: The number of lineages remains constant (n-type)
- **Lineage Degenerative**: The process has one fewer lineages after the event (y-type)

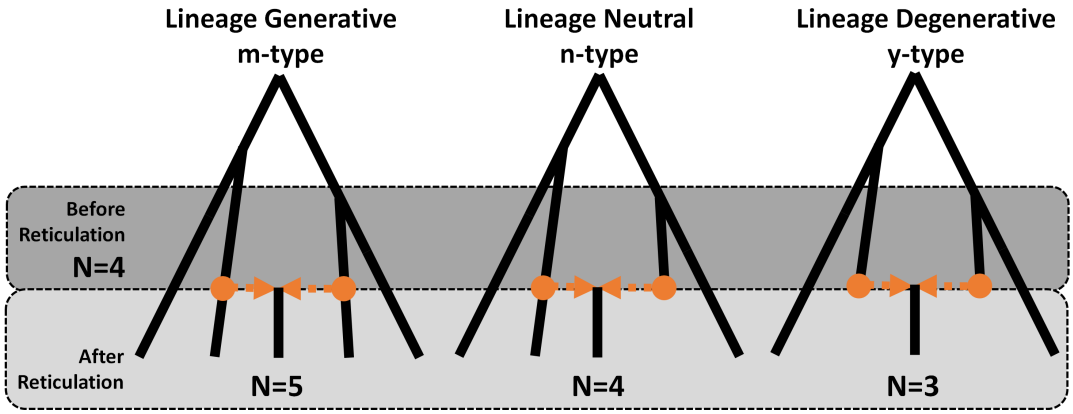

Figure 1: Types of Reticulation. Orange circles denote the parental nodes leading to the reticulate node. The dark grey box delineates the period before the reticulation event when the number of lineages is $N = 4$. The light grey box highlights the period after reticulation. Lineage generative hybridization (m-type) occurs when a reticulation event results in a gain of one lineage ($N = 5$). Lineage neutral hybridization (n-type) results in the a net zero change in the number of lineages ($N = 4$), and degenerative hybridization (y-type) reduces the number of lineages by one ($N = 3$).

The table below outlines possible interpretations for each reticulate pattern. For example, lineage generative events are consistent with hybrid speciation or allopolyploidization, since these events directly create new hybrid lineages. Introgression or lateral gene transfer can be seen as lineage neutral events, as genetic material is exchanged but the number of lineages remains constant. Interestingly, the formation of a new hybrid lineage also has potential to be lineage neutral or even lineage degenerative if genetic swamping occurs (Todesco et al., 2016). Genetic swamping is the loss of parental lineages after hybridization, resulting from either the hybrid having a higher fitness or a breakdown of reproductive barriers, eventually merging the hybrid lineage with its parental lineages. Extinction and incomplete sampling can further complicate the interpretation of reticulation events. Failure to sample the donor lineages of reticulation—so called ghost lineages (Ottenburghs, 2020; Tricou et al., 2021)—may change its structure on the reconstructed phylogenetic network (Fig S1). Without complete sampling, lineage generative and lineage neutral hybridization can appear as a different types of reticulation events.

Reticulate Pattern	**Possible Biological Processes**
Lineage Generative	• Hybrid speciation • Allopolyploidy
Lineage Neutral	• Genetic swamping with one parental lineage • Introgression • Lateral gene transfer • Lineage generative event with one unsampled parental lineage
Lineage Degenerative	• Genetic swamping with both parental lineages • Lineage generative event with both parental lineages unsampled • Lineage neutral event with one unsampled parental lineage

The type of reticulation also affects the topological properties of phylogenetic networks (Janssen and Liu, 2021). Phylogenetic networks can be grouped into a multitude of classes based on their topological

features (see Zhang, 2019; Kong et al., 2022). Understanding the topological properties of phylogenetic networks are exceptionally important because of the scalability challenges associated with their inference (Hejase and Liu, 2016; Elworth et al., 2019). Some network classes have useful identifiabilty results (Willson, 2010b; Pardi and Scornavacca, 2015; Solís-Lemus and Ané, 2016; Francis and Moulton, 2018; Erdős et al., 2019), unique encodings (Van Iersel and Moulton, 2014; Linz and Semple, 2020; Semple and Toft, 2021), or polynomial-time algorithms for reconstruction (Erdős et al., 2019; Murakami et al., 2019; Semple and Toft, 2021). Consequently, some methods will assume *a priori* that the underlying network belongs to a given class. For example, both the algorithms of Bordewich and Tokac (2016) and Wawerka et al. (2022) assume that networks are tree-child while the software packages `SNaQ` (Solís-Lemus and Ané, 2016) and `NANUQ` (Allman et al., 2019) assume that networks are level-1 (see Box 2 for definitions of network classes). Thus, to ensure methodological assumptions are met, it is important to understand the complex topological landscape of phylogenetic networks.

Simulation is an essential technique for characterizing the distribution of phylogenetic networks generated from various processes (Arenas et al., 2008; Janssen and Liu, 2021). Researchers found that increasing the number of hybridizations on a network of fixed size decreased the proportion of networks that belonged to certain classes (*e.g.,* tree-based and tree-child; Janssen and Liu, 2021). Additionally, for branching processes—including a special case of the birth-death-hybridization process—the type of reticulation also affected the topological properties of simulated networks. However, when comparing types of reticulation, a different simulation tool was used for each type, making it difficult to understand the effects of the reticulation type from artifacts of the generators themselves. Lastly, simulating to a fixed number of taxa and hybridizations obfuscates the role of hybridization in lineage diversification. The birth-death-hybridization process offers a complimentary approach for analyzing specific properties of phylogenetic networks. By having all three reticulation types as a part of the model, we can directly assess their effects on topological properties and lineage growth of phylogenetic networks.

In this study, we examine and summarize the distribution of phylogenetic networks generated by the birth-death-hybridization process. Specifically, we used a unified modeling framework to consider different assumptions about the type reticulation events. To understand the impact of hybridization on lineage diversification in phylogenetic networks, we computed and evaluated the expected number of lineages through time. We also simulated phylogenetic networks and characterized the proportion of phylogenies that belong to common network classes. We simulated phylogenies with incomplete lineage sampling and genetic distance-dependent hybridizations to consider how important systematic and biological processes might affect the properties of generated phylogenetic networks.

Our work explores the space of phylogenetic networks generated under the birth-death-hybridization process and assesses how different assumptions about the process of gene flow influence this distribution. Developing an intuition for how hybridization creates and shapes diversity will be valuable for selecting and fitting appropriate macroevolutionary models. Lastly, understanding the distribution of topological properties from a biologically relevant process can give insight into how often network class assumptions are violated in empirical systems, potentially biasing inferences and steering methods development.

2 Methods

2.1 Phylogenetic Network Properties

Phylogenetic networks are similar to phylogenetic trees but have additional reticulate nodes and edges representing genetic material crossing species boundaries. Although these nodes can represent specific biological processes such as lateral-gene transfer or introgression, the network does not distinguish which process occurs. Though, for consistency with the formal descriptions of the birth-death-hybridization process (Morin and Moret, 2006; Woodhams et al., 2016; Zhang et al., 2018), we, at times, refer to reticulation events as hybridization nodes. Like phylogenetic trees, networks can be rooted or unrooted. In this study, we consider rooted phylogenetic networks. Formally, a phylogenetic network is a directed acyclic graph that contains four node types:

- *root*: A node of in-degree 0 and out-degree 1 if the process begins at the root or out-degree 2 if the root represents the most recent common ancestor of the process.

- *tree*: A node of in-degree 1 and out-degree 2. These nodes typically represent speciation events.
- *hybrid*: A node of in-degree 2 and out-degree 1. Hybrid nodes denote reticulation events.
- *leaf*: Nodes of in-degree 1 and out-degree 0 that can represent either extant lineages that survived until the present or they can denote an extinction event.

A directed edge from nodes u to w is denoted as *(u,w)*. We further describe *(u,w)* as a hybrid edge if w is a hybrid node. Hybrid edges additionally have weights (γ and $1-\gamma$) that are used to describe the proportion of genetic material that contributes to the hybrid node. Hybrid nodes are considered stacked if the child of a hybrid node is another hybrid node (*i.e.,* there exists an edge *(x,y)* where both x and y are hybrid nodes). We restrict our analyses to identifying tree-based, fold-unfold stable (FU-stable), tree-child, normal, and level-k networks (Box 2).

Box 2: Description of Network Classes

Phylogenetic network classifications describe the topological properties of a given phylogenetic network. In this study, we focus on five different classifications:

- **Level-k**: Any biconnected component of the network having at most k hybrid nodes (Jansson and Sung, 2006).
- **Tree-based**: A phylogenetic network that can be constructed by starting with a phylogenetic tree and sequentially adding reticulation edges (Francis and Steel, 2015). Non-tree-based networks are identified by locating a zigzag pattern that connects reticulate nodes (Zhang, 2016).
- **Fold-unfold stable (FU-stable)**: A network that can be 'unfolded' into multi-labeled trees (MUL-trees) and then 'refolded' (see Huber and Moulton, 2006) back into itself is FU-stable. Fold-unfold stability occurs when reticulate nodes are not stacked and when no two tree nodes have the same set of child nodes (Huber et al., 2016).
- **Tree-child**: A tree-child network is a phylogeny where all internal nodes have at least one non-reticulate child node (Cardona et al., 2008).
- **Normal**: tree-child networks that contain no directed 'shortcut' edges that directly connects two nodes (u, v) given that there exists another directed path from node u to v (Willson, 2008, 2010b).

The latter four classes have the following nested hierarchy:

$$normal \subset tree\text{-}child \subset FU\text{-}stable \subset tree\text{-}based.$$

Any nested class will also have the topological properties of the all superceding classes (*e.g.,* a tree-child network is also FU-stable and tree-based). Figure 2 gives postive and negative examples of the classifications, but also see Kong et al. (2022) and the webpage "ISIPhyNC (Information System on Inclusions of Phylogenetic Network Classes)"[a] for more information on each classification.

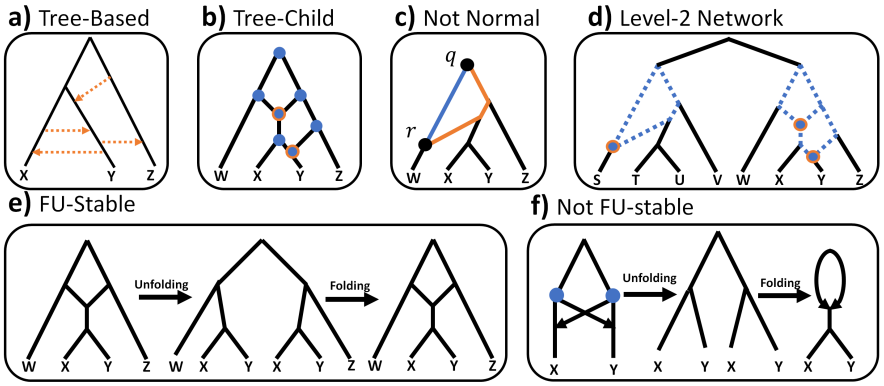

Figure 2: Examples of Network Classes. **a)** A tree-based network, which can be constructed by first starting with a backbone tree (black edges) and adding hybrid edges (orange dotted arrows). **b)** A tree-child network: all internal nodes (blue circles) have at least one child node that is not a hybrid node (circles with an orange outline). **c)** A network that is not normal because of the shortcut edge ((q, r); blue) among the two paths from q to r (blue and orange). **d)** Depicts a level-2 phylogenetic network because one of the two biconnected components (blue dotted lines) contains two hybrid nodes (orange outlined circle). **e)** and **f)** Unfolding a network (left) into a MUL-tree (middle) and then folding back into a network (right). The original network in **e)** is FU-stable because the unfolding and folding operations result in the original network. The network in **f)** is not FU-stable because the two operations generate a different network from the original. This network has two nodes (blue circles) with the same set of child nodes and therefore cannot be FU-stable.

[a] http://phylnet.univ-mlv.fr/isiphync

2.2 Birth-Death-Hybridization Process

We use a birth-death-hybridization process (Morin and Moret, 2006; Woodhams et al., 2016; Zhang et al., 2018) as a generating model for phylogenetic networks. This process has exponentially distributed waiting

times between speciation, extinction, and hybridization events. We consider rates of speciation (λ) and extinction (μ) to be rates on each lineage while the rate of hybridization (ν) is on each lineage pair. That is, when the process has N lineages, the rates of speciation, extinction, and hybridization are $N\lambda$, $N\mu$, and $\binom{N}{2}\nu$, respectively. There are three types of hybridization events: lineage generative hybridization, lineage neutral, and lineage degenerative (Box 1). We use the thinning property of a Poisson process to break the hybridization rate (ν) into separate rates for each type of hybridization: $\nu = \nu_+ + \nu_0 + \nu_-$, where the lineage generative hybridization rate is denoted as ν_+, lineage neutral hybridization as ν_0, and lineage degenerative hybridization as ν_-. We refer to the possible state space—all linear combinations of ν_+, ν_0, and ν_- for a given hybridization rate (ν)—as the hybridization rate simplex, which we represent visually using a ternary plot (see example in Fig 3).

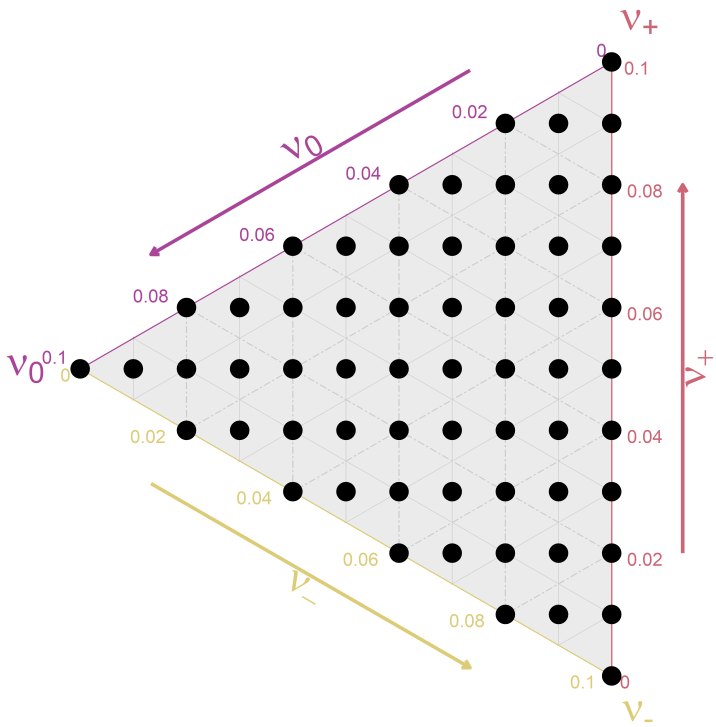

Figure 3: Ternary plot with the state space for the combination of hybridization-type rates for $\nu = 0.1$. Moving towards a corner corresponds to an increase of the rate for one of the hybridization types: the top corner (ν_+) denotes the lineage generative rate, the bottom corner (ν_-) denotes the lineage degenerative rate, and the leftmost corner (ν_0) denotes the lineage neutral rate. Each dot represents one parameter combination used when simulating across the hybrid rate simplex (Table S1). See Section 2.4.1 for details about the simulation procedure.

At a given hybridization event, we denote the genetic contribution of each parental lineage as γ and $1-\gamma$. Each hybridization event has its own associated γ that can be drawn from any distribution on $[0,1]$. We draw the inheritance proportions γ from a $Beta(10,10)$ distribution, creating a symmetric distribution centered on equal inheritance proportion of $\gamma = 0.5$.

In our simulations, hybridization events are more likely to occur if the the two parent lineages are closely related, to accomplish this, we used the approach of Woodhams et al. (2016) for genetic distance–dependent hybridization. Under this model, hybridization events are proposed with rate ν and successfully occur on the phylogeny with a probability that is proportional to the genetic distance between the two hybridizing taxa (d_{ij}):

$$Pr(Hybridization\ Success) \propto f(d_{ij}).$$

We assume a strict molecular clock where the genetic distance between two taxa is taken to be the summation of distinct branch lengths between each lineage and their most recent common ancestor. Moreover, in the case of hybrid taxa, the distance is a weighted summation across paths (see Woodhams et al., 2016). We use a linear decay function to relate the probability of hybridization success to genetic distance:

$$f(d_{ij}) = \begin{cases} 1 - d_{ij}\delta & d_{ij} \leq \frac{1}{\delta} \\ 0 & d_{ij} > \frac{1}{\delta} \end{cases},$$

where δ represents the strength of the genetic distance dependence. A value of $\delta = 0$ makes hybridization events independent of genetic distance while higher values of δ make hybridization primarily occur between more closely related taxa.

2.3 Diversification Dynamics

Measuring the rate of lineage accumulation over time allows us to compare the outcomes of different models. We derived the expected number of lineages through time and extend the rate of diversification to accommodate hybridization. With these concepts, we later explore how different assumptions about hybridization modulate these quantities.

2.3.1 Expected Number of Lineages Through Time

We can represent the change in the expected number of lineages over time with the following differential equation:

$$\frac{dN}{dt} = N\lambda - N\mu + \binom{N}{2}\nu_+ - \binom{N}{2}\nu_-,$$

or equivalently:

$$\frac{dN}{dt} = N^2 s + N(r - s),$$

such that $r = (\lambda - \mu)$ and $s = \frac{1}{2}(\nu_+ - \nu_-)$. Solving this differential equation for a process that originates at time $t = 0$ with N_0 species gives the expected number of extant lineages $N(t)$ at time t:

$$N(t) = \frac{r - s}{\frac{N_0 s + r - s}{N_0} e^{-(r-s)t} - s}.$$

The expected number of lineages does not depend on the lineage-neutral hybridization rate as these events do not affect the overall number of lineages. Instead, there are dependencies on the differences between parameters that generate lineages (λ and ν_+) and those that remove lineages from the process (μ and ν_-).

2.3.2 Diversification Rate

The value of r is often often referred to as the net diversification rate. However, under the birth-death-hybridization process, it cannot have this interpretation since hybridization events also can cause the number of lineages to increase or decrease. Further, since hybridization rates scale quadratically with the number of lineages, net diversification cannot be computed as the difference between all lineage adding and lineage deleting rates, *i.e.,* $(\lambda + \nu_+) - (\mu + \nu_-)$. Instead we can compute the overall diversification for a process with N lineages as:

$$N\lambda - N\mu + \binom{N}{2}\nu_+ - \binom{N}{2}\nu_-$$

or equivalently:

$$Nr + \binom{N}{2} 2s.$$

Thus, we interpret r as the per-lineage diversification rate and $2s$ as the lineage-pair diversification rate. The overall diversification rate is a function of the number of species but allows us to more directly compare the effects of speciation, extinction, and hybridization on lineage accumulation.

2.4 Simulating the Distribution of Phylogenetic Network Classes

In this study, we sought out to assess how different assumptions about hybridization affect the distribution of phylogenetic networks under a birth-death-hybridization process. We simulated phylogenetic networks to evaluate properties and expectations that are difficult to compute analytically under a birth-death-hybridization process.

2.4.1 Simulation Procedure

Phylogenies were simulated using the R package SiPhyNetwork (avaliable on CRAN; Justison et al., 2022). Simulations were conditioned on survival to the present with at least two extant lineages prior to sampling and to have completed in less than 10 seconds on an Dell XPS 15 with an i7 processor. The latter condition is to prevent simulation replicates from spending an infinite amount of time approaching the stopping time without ever reaching it. A product of having hybridization rates increase quadratically with the number of lineages is that some simulation conditions are expected to explode and have infinitely many events in a finite amount of time (Goriely and Hyde, 2000). We expect this behavior for positive lineage-pair diversification rates ($s > 0$; see Section 3.2).

All simulations began with two lineages at $t = 0$, representing the most recent common ancestor at the root, and were allowed to proceed until reaching the stopping time at $t = 1$. We used constant rates of speciation ($\lambda = 4$), extinction ($\mu = 2$), and hybridization ($\nu = 0.1$). For each specific scenario and parameter combination (Tables S2, S3, and S4) we generated 20,000 replicate phylogenetic networks.

2.4.2 Simulating Across the Hybridization-Rate Simplex

To assess how hybridization type affected the properties of simulated phylogenies, we used 61 different linear combinations of rates for the hybrid types ν_+, ν_0, and ν_-. The specific parameter combinations are visualized in Figure 3 and listed in Table S1.

For each combination of hybridization type rates, we used two different sampling strategies of extant-only sampling and complete sampling. First, we considered complete phylogenies with all extant and extinct taxa included in the phylogeny. For these complete phylogenies we had two scenarios where hybridization was either genetic distance–dependent ($\delta = 0.5$) or independent of genetic distance ($\delta = 0$). Next, we sampled extant-only phylogenies by pruning extinct taxa from the phylogeny. For the extant-only phylogenies, we considered both complete extant sampling ($sampling\ fraction = 1$) and incomplete sampling where a constant fraction of extant taxa were sampled ($sampling\ fraction = 0.75$).

We additionally simulated phylogenies to isolate the effects of genetic distance dependence and incomplete sampling on the class membership of networks. We used a constant rate for each hybridization type ($\nu_+ = \nu_0 = \nu_- = \frac{1}{30}$) while separately varying the strength of genetic distance dependence and amount of incomplete lineage taxa sampling. For incomplete lineage sampling, we simulated extant only phylogenies and used sampling fractions that ranged from $\rho = 0.25$ to $\rho = 1.0$ For genetic distance dependence, we simulated complete phylogenies with both extinct and extant taxa while varying the strength of distance dependence from $\delta = 0$ to $\delta = 2$.

2.5 Analysis of Simulated Data

We used a combination of summary statistics and network classifications to characterize the distribution of phylogenetic networks under the birth-death-hybridization process. For the replicates of each simulation scenario we recorded the number of total taxa, hybridizations, reticulation density ($\frac{Number\ of\ Hybridizations}{Number\ of\ Tips}$), observed proportion of total hybridization for each type, and the level of the network. Lastly, we assessed whether phylogenies belonged to the tree-based, FU-stable, tree-child, or normal classes.

3 Results

Our aim—through both analytical computation and simulation—was to understand and characterize the properties of phylogenetic networks generated from the birth-death-hybridization process. We also assessed

how our assumptions about gene flow and sampling affected this distribution. We present and outline our key findings below.

3.1 Data Simulation

We generated 20,000 phylogenies that met the criteria defined in Section 2.4.1 for each parameter combination (Tables S3 and S4). Across all completely sampled simulations without genetic-distance dependence, we discarded 339,054 (21.7% of simulation attempts) samples that went extinct or had only one extant taxon. We discarded 2233 (0.14% of simulations) due to the simulation running longer than 10 seconds. A higher proportion of simulations timed out as s increased (Fig 4a depicts how s changes across the hybrid-rate simplex). Increasing the stopping time to 20 seconds for the settings with the highest lineage-pair diversification s did not have a large affect on the number timeouts (1074 timeouts; 0.13% of simulations), indicating that we were likely able to capture the distribution of networks with little bias. Simulations rejected from timing out may have never finished or were outliers in terms of numbers of extant lineages.

3.2 Expected Diversification Dynamics through Time

The shape of lineage accumulation takes different forms depending on the relative rates of lineage-generative and lineage-degenerative hybridization (Fig 4b). Firstly, when $\nu_+ = \nu_-$, $s = 0$ and the equation for the expected number of lineages simplifies to the constant-rate birth-death process. However, when $\nu_+ > \nu_-$, the lineage-pair diversification is positive and $N(t)$ exhibits hyper-exponential growth with a vertical asymptote at:

$$t = \frac{\ln(N_0 s) - \ln((N_0 s + r - s))}{s - r}.$$

Although constant-rate birth-death processes can expect the number of lineages to reach infinity as t approaches infinity, the birth-death-hybridization process can expect an infinite number of lineages in a finite amount of time when $s > 0$. Lastly, for $\nu_+ < \nu_-$ resulting in a negative value for s, the process has logistic growth, approaching a maximum of $\frac{r-s}{s}$ lineages over time.

As the number of lineages increases, hybridization quickly becomes the dominant force in diversification (Fig 4d). The overall diversification rate grows rapidly when s is positive but actually becomes negative with a large number of taxa and when $s < 0$ (Fig 4c). However, when considering a negative value of s and lineages through time, the overall diversification rate approaches zero as the negative effects from hybridization match the per-lineage diversification, resulting in the logistic growth.

3.3 Simulating Across the Hybrid Rate Simplex

Hybridization type affected the diversification dynamics of simulated phylogenies (Fig 5). In line with analytical expectations, simulations with higher lineage-pair diversification rates (s) had an increased number of taxa on average, a result that is consistent across all sampling scenarios (complete, complete with genetic distance–dependent hybridization, extant-only networks, and extant-only networks with incomplete sampling). We also observed the same increasing pattern with respect to s for the number of hybridization events and reticulation density. Interestingly however, extant-only networks had higher reticulation densities than complete networks despite having fewer reticulations (Fig 5; Table S5). Though all extinct lineages are not observed for extant-only sampling, reticulations are only lost on the phylogeny when all the hybrid descendants fail to get sampled. This becomes increasingly difficult for highly intertwined networks with many reticulations. Further, due to incomplete sampling among some parental lineages, extant-only networks had a higher proportion of lineage-degenerative and neutral hybridization than completely sampled networks with the same hybridization rates (Fig S2).

The proportion of networks belonging to each class was differentially affected across the hybrid rate simplex (Fig 6; Table S6). The more deeply contained classes within the nested hierarchy of tree-based, FU-stable, tree-child, and normal classes (see Box 2) have increasingly stringent topological requirements, and, consequently, had lower class proportions (*e.g.,* a smaller proportion of networks were tree-child than tree-based). More deeply nested classes also showed a greater range of proportions across

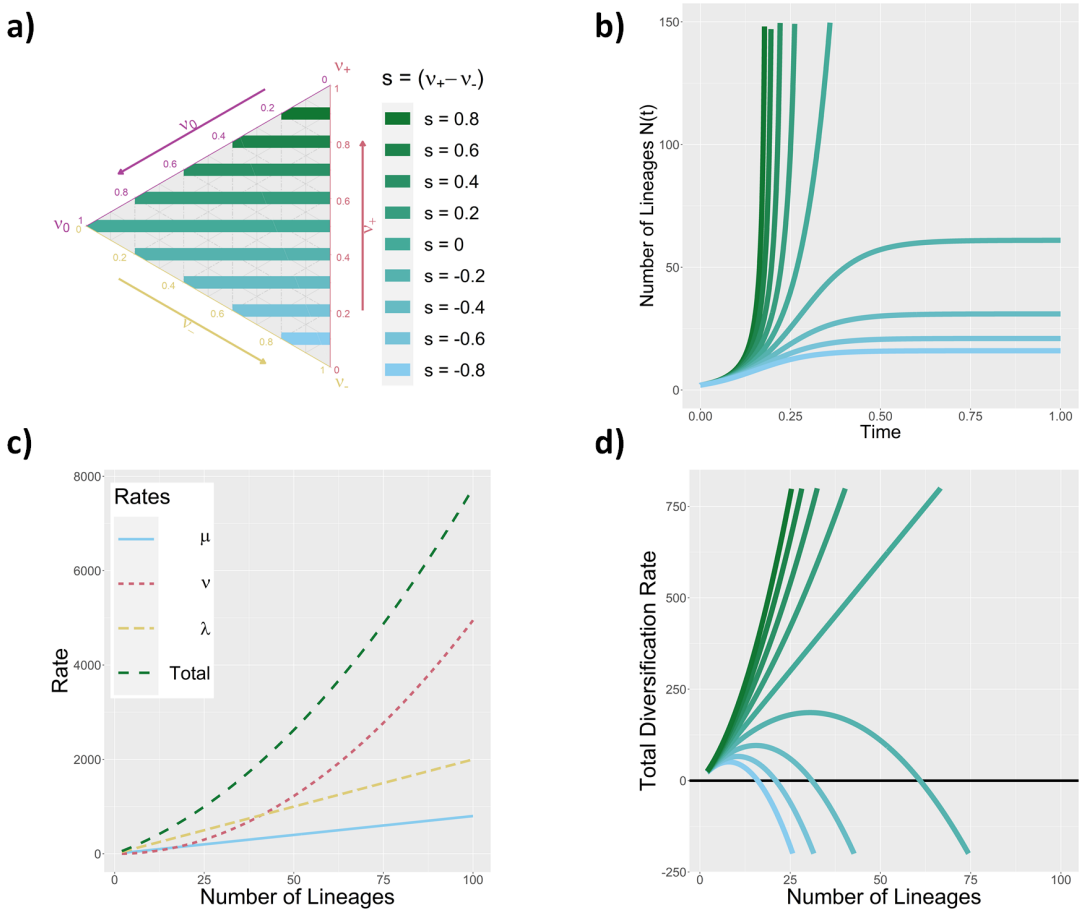

Figure 4: The effect of hybridization type on diversification dynamics. For all panels, we assume a speciation rate of $\lambda = 20$, extinction rate of $\mu = 8$, and a hybridization rate of $\nu = 1$. Higher lineage-pair diversification s corresponds to a darker green while low values of s are denoted with a light blue. **a)** The lineage-pair diversification rate s changes across the hybrid rate simplex. **b)** The expected number of lineages through time with respect to s. **c)** The rate of speciation(λ), extinction(μ), hybridization(ν) events as the number of lineages increase. Panel **d)** shows the overall diversification rate $Nr + \binom{N}{2}2s$ as a function of the number of lineages N.

the hybrid simplex (Table S6). Interestingly, for completely sampled, extant-only and incompletely sampled datasets, the region with high lineage generative rates (*i.e.,* high s) led to the lowest proportion of tree-based and FU-stable networks but resulted in relatively high proportions for tree-child and normal classes, thus indicating tree-based and FU-stable classes could be more sensitive to a high number of hybridization events than other classes. However, for distance–dependent hybridization, the proportion of tree-based and FU-stable networks tended to decrease as s decreased. Furthermore, for all simulation scenarios, tree-child, and normal network proportions decreased in regions with fewer hybridization events but an increased proportion of lineage degenerative hybridizations (*i.e.,* lower s), suggesting hybridization type is and important factor for these classes. When comparing between sampling conditions, class proportions were lower for the incompletely sampled datasets than those that were completely sampled—despite having fewer hybridization events. The complete dataset simulated under genetic distance–dependent hybridization had relatively few reticulation events, resulting in overall higher class proportions than other sampling conditions.

Having only lineage-generative and neutral hybridization always produced networks with certain topological properties (dark diamonds on Fig 6; Table S6). For completely sampled networks, all simulations with a lineage-degenerative rate of zero ($\nu_- = 0$; axis with ν_0) were tree-child. Similarly, at the corner of the hybrid rate simplex with only lineage generative hybridization ($\nu_+ = \nu = 0.1$), all networks were tree-based, FU-stable, tree-child, and normal. Interestingly, the same parameterizations with extant-only or incomplete sampling did not lead to unity with respect to class membership. Unsampled parental lineages caused lineage

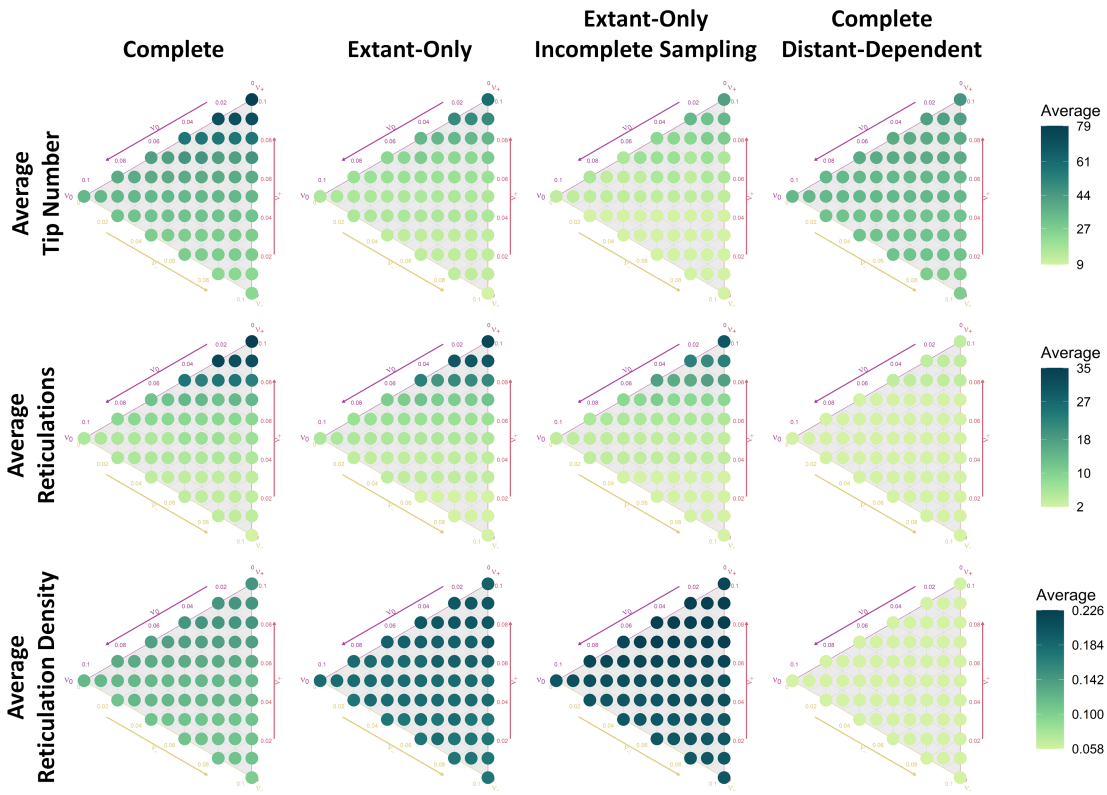

Figure 5: Diversification dynamics of phylogenetic networks simulated across the hybrid rate simplex. Each column shows one of the four simulation conditions used to simulate phylogenetic networks across the hybrid rate simplex (See 2.4.1). The position of each point on the simplex indicates the lineage generative, lineage neutral, and lineage degenerative hybridization rates. Each row summarizes the number of tips, reticulations, and reticulation density ($\frac{reticulations}{tips}$) for each dataset. The color of each point is used to depict the average value of 20,000 replicates.

generative and lineage neutral hybridizations to be observed as another type (*e.g.,* Figs S1 and S2).

A large portion of reticulate phylogenies were not level-1 (Fig S3). Increasing the lineage-pair diversification rate s led to an increase in the average level of simulated networks and a higher proportion of networks being level-2 or greater. Trends in the level of networks were strongly tied to the number of reticulations (Fig 7). In most cases, the level of the network was the same as number of reticulations. However, there was slightly more variation in network level for networks with few reticulations (Fig 7b).

Increasing the effects of genetic-distance dependence and incomplete sampling decreased the number of reticulations on networks and consequently increased the class proportions of networks (Fig 8). Increasing the effect of incomplete sampling (*i.e.,* decreasing the sampling fraction) only modestly affected the class proportions. However, increasing the effect of distance dependence (*i.e.,* strength of distance dependence δ) more drastically increased class proportions, making nearly all networks tree-based and FU-stable. Higher class proportions can be explained by large decreases in the overall number of hybridization events and an increase in the proportion of phylogenies with no reticulation events, particularly for distance dependent hybridization (Figs S5 and S6). For the incomplete sampling scenario, higher reticulation densities (Fig S5) and changes in the observed hybridization types (Fig S4) likely offset some of the effects of having fewer reticulations, explaining a more limited increase in class proportions. Interestingly, when looking on a per-reticulation basis, the opposite trend is observed for incomplete sampling. Decreasing sampling actually decreased class proportions for a given number of reticulations (Fig 9). When looking on a per-reticulation basis for distance-dependent hybridization, increased dependence strength (δ) is associated with higher proportions of tree-based and FU-stable networks but lower proportions of tree-child and normal networks.

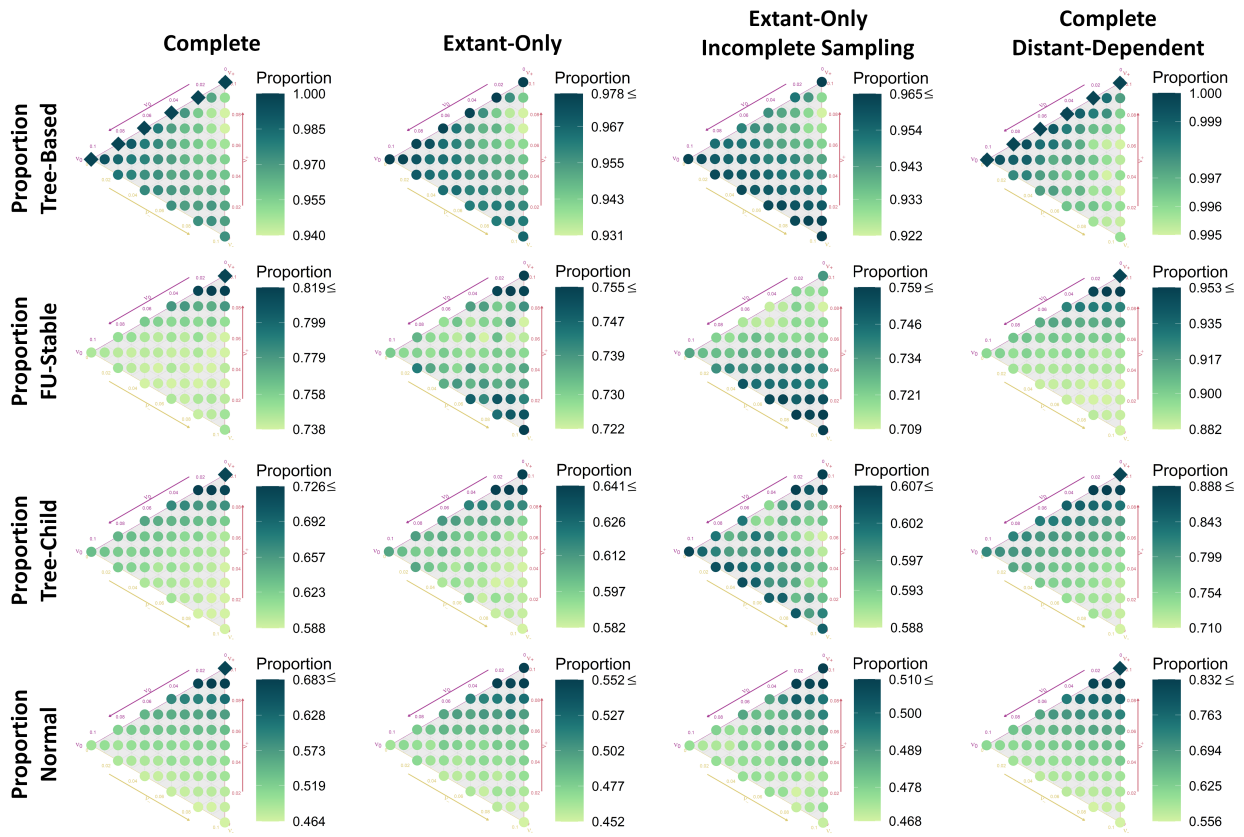

Figure 6: Class membership of simulated phylogenetic networks across the hybrid rate type simplex. Each column shows the simulation conditions used to simulate phylogenetic networks across the hybrid rate simplex (See 2.4.1). The position of each point on the simplex indicates the lineage generative, lineage neutral, and lineage degenerative hybridization rates. Each row summarizes the proportion of phylogenies from a given class. The color of each point is used to depict the proportion of 20,000 replicates that belonged to the class, with a darker color representing a higher proportion. Colors are scaled for each class and dataset. Scales range from the minimum observed proportion to the 95th percentile and all values above the 95th percentile are binned into the same color. Cases where all simulated networks belong to a class (proportion of 1) are indicated with a diamond.

4 Discussion

4.1 Diversification under Birth-Death-Hybridization

With the rate scaling by each lineage-pair, hybridization can eclipse rates of speciation and extinction, resulting in highly reticulate histories. However, in practice, some systems have rampant gene flow (Salzburger, 2018; Edelman et al., 2019; Kozak et al., 2021; Esquerré et al., 2022; Suvorov et al., 2022), while it is much more limited in others (Solís-Lemus and Ané, 2016; Morales-Briones et al., 2018; Zhang et al., 2018; Karimi et al., 2020). The lack of extensive reticulation in some systems could be an artifact of the computational burden associated with estimating complex phylogenetic networks and using genome-scale data (Hejase and Liu, 2016; Elworth et al., 2019). Few inferred reticulation events could also be a biological result due to the build-up of incompatibilities as lineages genetically diverge over time (Orr, 1995; Mallet, 2005; Gourbiere and Mallet, 2010). In fact, we showed that reticulation dependence on genetic distance had great propensity to reduce the number of events on simulated phylogenetic networks. However, there are a number of isolating mechanisms (see Mallet, 2005; Soltis and Soltis, 2009) that could create heterogeneity in the amount gene flow from system to system. Future work towards explicitly incorporating these mechanisms into the birth-death-hybridization processes will be important in assessing their role in the diversification of admixed systems.

Solely changing assumptions about hybridization leads to phylogenies with considerable variation in

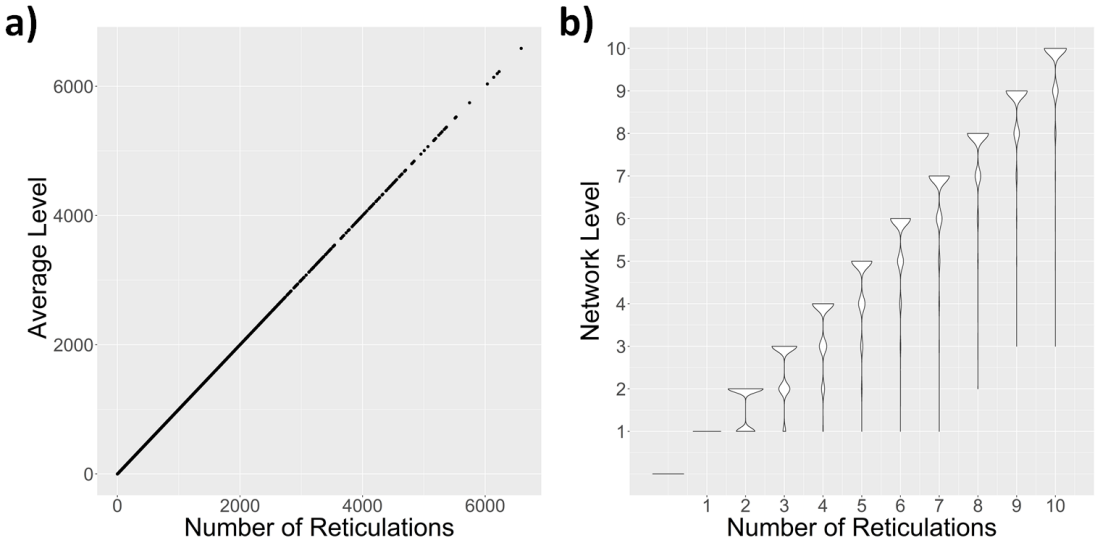

Figure 7: Completely sampled phylogenies summarized across the hybrid rate simplex by the number of reticulations. **a)** A scatterplot depicting a near one to one relationship between the number of reticulations and the average level of simulated phylogenetic networks. **b)** A violin plot showing the distribution of levels for 0 to 10 reticulations. Widths of the violin correspond to the proportion of networks with that level.

size and pattern of diversification. The slowdown of lineage accumulation has been observed in numerous clades across the tree of life (Morlon et al., 2010). Several mechanisms have been proposed to explain slowing diversification (Moen and Morlon, 2014), including density-dependent speciation rates (Rabosky and Lovette, 2008) and protracted speciation (Etienne and Rosindell, 2012). Under a birth-death-hybridization model where $s < 0$, hybridization acts as a diversity-dependent mechanism that reduces the overall degree of diversification, and generates a pattern of slowed growth over time. In fact, estimating the likelihood of the phylogenetic network from a branching process is currently limited to a special case of the birth-death-hybridization process where only lineage-degenerative hybridization occurs ($s < 0$; Zhang et al., 2018). Given that gene flow represents a broad set of processes with heterogeneous effects on diversification, it will be important to be able to estimate hybridization rates for each type of hybridization from the birth-death-hybridization process. Although methods currently do not include extinction or estimate lineage neutral or lineage generative rates, for systems with lineage-generative reticulation, the hyper-exponential growth when $s > 0$ might prove helpful in estimating diversification dynamics. If these systems have ages calibrated by fossil records, it would be sensible to limit the parameter search space to regions of the hybrid-rate simplex that are not expected to explode to infinitely many species over the age of the clade.

4.2 Topological properties of Phylogenetic Networks

Two key results arise from previous simulation studies of topological profiles (Arenas et al., 2008; Janssen and Liu, 2021): (1) as the number of reticulations increases, the proportion of generated networks that belong to certain classes (*e.g.,* tree-child, tree-sibling) decreases and (2) when controlling for the number of reticulations, the type of reticulation affects the proportion of networks belonging to the same given classes. Our results are largely concordant with these conclusions. Though, when considering the birth-death-hybridization process, it creates a complex interaction of these factors: the reticulate events that are the best for phylogenetic some network classes (lineage generative; Janssen and Liu, 2021) also lead to more lineages, and consequently, more reticulations. We found that phylogenetic-network classes responded differently to these factors; tree-based and FU-stable network proportions were lower in the region of the hybrid rate simplex with a high number reticulations, primarily being lineage generative. Tree-child and normal classes had low proportions in the region that had few reticulation events but were mostly lineage degenerative. Thus, it is important to not only consider how each hybridization type affects the diversification and topological properties of systems but also to characterize how each network class individually responds

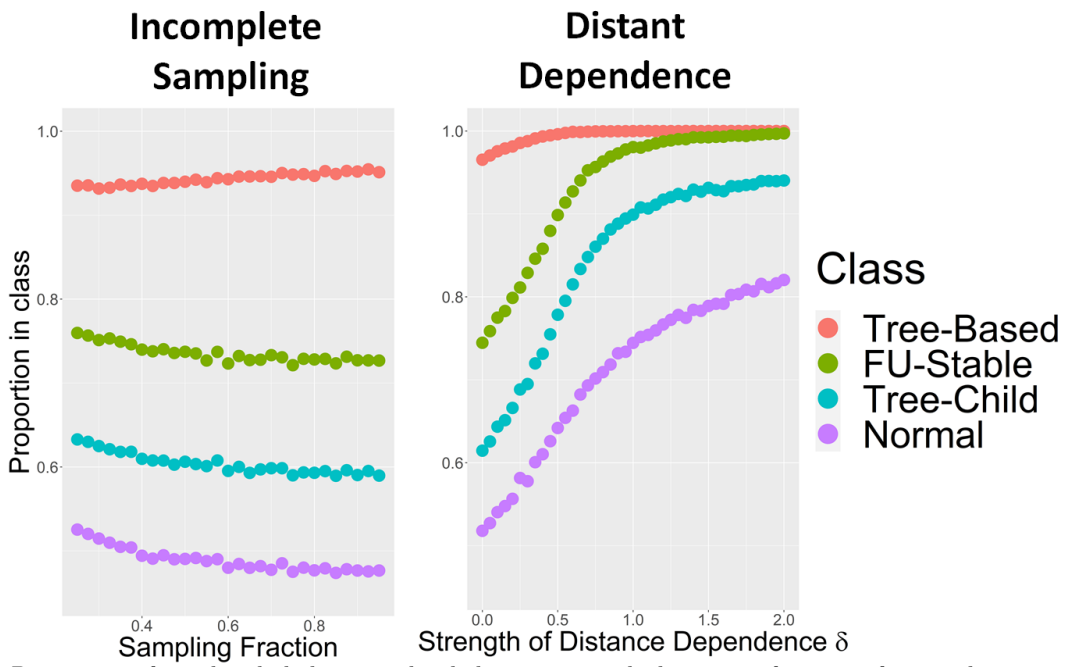

Figure 8: Proportion of simulated phylogenies that belong to network classes as a function of incomplete sampling and genetic distance dependence. Each point represents the class proportions observed from 20,000 simulated phylogenetic networks. The sampling fraction used in the incomplete sampling dataset represent the number of extant tips that are randomly sampled from the network. The value of δ used in the genetic distance-dependant dataset corresponds to the strength of dependence, with higher values of δ indicating that reticulation events primarily only occur between closely related lineages. Extant-only networks were used for the incomplete sampling dataset, while complete phylogenetic networks with extant and extinct species are used for the genetic distance-dependence dataset.

to these factors.

Many network classifications require reticulations to be distant from one another. For example, networks will not be tree-child if reticulations are stacked or networks are not tree-based if they have a specific zig-zagging pattern that connects reticulations (Zhang, 2016)—both only occur if reticulations have few nodes separating them on the phylogenetic network. Reticulation proximity makes it important to consider not only the number of hybridizations when considering topological properties but also their density with respect to the size a phylogenetic network. Both poor taxon sampling and not accounting for extinct lineages with fossil evidence has been shown to reduce the accuracy of phylogenetic inference (Heath et al., 2008a; Nabhan and Sarkar, 2012; Warnock et al., 2020). Incomplete species sampling may also pose a challenge for estimating phylogenetic networks, as the higher reticulation density of poorly sampled networks compared to those that are completely sampled may induce fewer topological properties that can inform phylogenetic estimates.

Both our simulations and those of Janssen and Liu (2021) with genetic distance dependence also highlight the importance of reticulation proximity. Gene flow reduces the observed divergence between species (Slarkin, 1985; Leaché et al., 2014). When considering a fixed number of reticulation events, this creates a positive feedback loop for gene flow events to occur locally on reticulate lineages, thus decreasing the proportion of networks with certain topological properties. However, these same distance-dependent mechanisms also reduce the overall number of gene flow events, potentially mitigating the effects of increased locality. Ultimately, if systems with distance-dependent gene flow (*e.g.,* Chapman and Burke, 2007; Tea et al., 2020; Barley et al., 2022) have an abundance of events, it may pose challenges to methods that make certain topological assumptions of the underlying phylogenetic network.

The proximity of reticulations to one another also explains why only lineage-generative or neutral hybridization (*i.e.,* $nu_{-} = 0$) always produced networks of certain types and why incomplete or extant-only sampling eliminates this pattern (see lemmas 1-4 in Janssen and Liu, 2021). Effectively, lineage-generative and neutral hybridization events create extra tree nodes during the reticulation event. These nodes can put space between reticulation nodes, making it more likely that a network meets certain topological criteria for

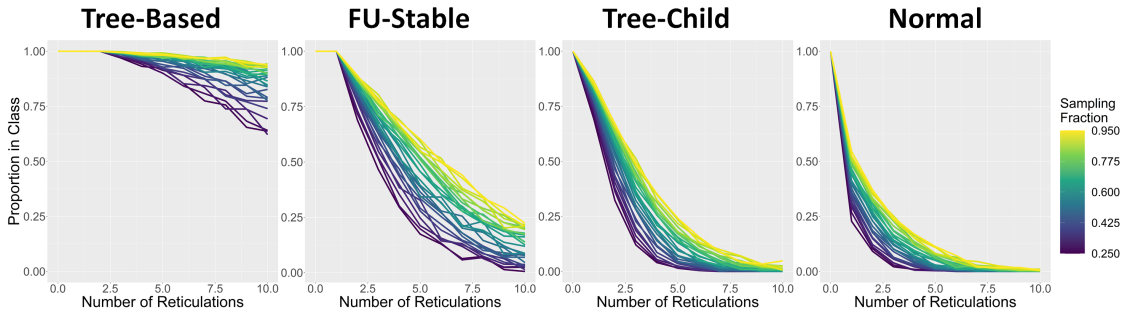

Figure 9: Class membership proportions under incomplete sampling as a function of reticulation number. Replicates are grouped and summarized by the number of reticulations and sampling fraction. The sampling fraction is the proportion of extant lineages sampled from an extant-only phylogenetic network. For each sampling fraction, 20,000 phylogenetic networks were simulated. The proportion in a class represents the proportion of networks with a given number of reticulation events to belong to that class. Points were only drawn when a given distance-dependence strength and number of reticulations had at least 30 replicates.

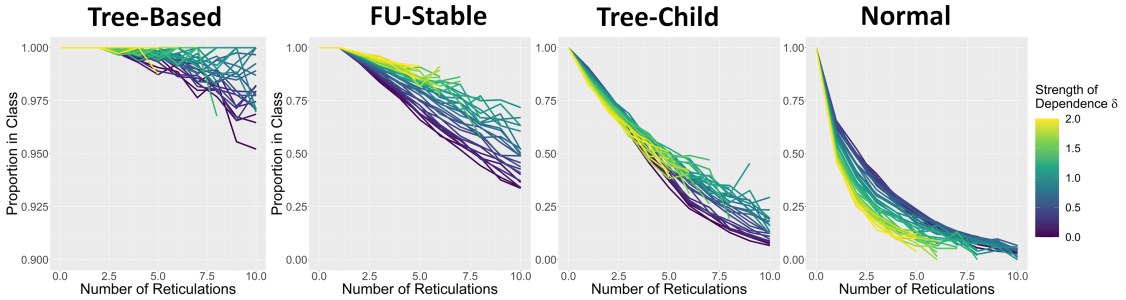

Figure 10: Class membership proportions under genetic-distance dependence and as a function of reticulation number. Phylogenetic networks are grouped by the simulated number of reticulations and strength of distance dependence. For each value of distance dependence (δ), 20,000 phylogenetic networks were simulated. The proportion in a class represents the proportion of networks with a given number of reticulation events to belong to that class.

class assignment.

While not all reticulate histories are tree-based and have an underlying tree as a backbone (Francis and Steel, 2015), those generated from the our parameterizations of birth-death-hybridization process appear largely tree-based. The tree-based class is conceptually satisfying for phylogenetic networks by allowing researchers to think of reticulate evolution as phylogenetic trees with additional arcs. In fact, tree inference on reticulate systems has several uses and advantages over network inference (see Blair and Ané, 2020). Though, inferring the treelike backbone of a network—if such a tree exists—should be done with caution because topology and divergence estimates can be biased (Leaché et al., 2014). Interestingly however, several recent methods attempt to quickly estimate phylogenetic networks by starting with phylogenetic trees and augmenting them into networks (Cao et al., 2019; Molloy et al., 2021).

Some relationships on phylogenetic trees are even impossible to infer (Slowinski, 2001; Lewis et al., 2005), particularly for systems undergoing rapid speciation (Stanley et al., 2011; Suh, 2016). Instead, polytomies can be included in an analysis to reflect the lack of phylogenetic resolution (Lewis et al., 2005; Kemp, 2009). A similar discourse is occurring in the phylogenetic network community with many studies finding only specific topologies identifiable and possible to estimate (Huber et al., 2015; Pardi and Scornavacca, 2015; Solís-Lemus and Ané, 2016; Francis and Moulton, 2018; Gross and Long, 2018; Erdős et al., 2019). Pardi and Scornavacca (2015) proposed an interesting solution to the identifiability problem: infer distinguishable cannonical phylogenetic networks by effectively collapsing some reticulate events into polytomies. There are techniques capable of reconstructing the phylogenetic network if it is of a certain class and returns a reduced form network if not belonging to that class. Normal networks can be reconstructed in polynomial time from their set of displayed trees (Willson, 2010a), and if the underlying network is not normal then the

algorithm will return a reduced form normal network. Not all reticulations may be present in the reduced normal network but these constructions have desirable mathematical properties and summarizes the history into a simpler structure (Francis et al., 2021). Additionally, the folding operation of Huber and Moulton (2006) can reconstruct FU-stable networks from MUL-trees (Huber et al., 2016). If the underlying network is FU-stable then the method returns the true network, otherwise it produces a minimally reticulate network that unfolds into the same MUL-tree. Although estimating the MUL-tree is not a trivial task (Huber et al., 2008), MUL-trees and their implied FU-stable networks have been used to understand polyploid histories (Brysting et al., 2007). These types of approaches represent a shift from the fully resolved phylogenies that biologists have come to expect but they have several appealing qualities: (1) they only return what is distinguishable, (2) if the true network has the right topological properties then it can still be recovered, and (3) canonical forms have a reduced topology search space with computationally efficient construction algorithms. Further, these approaches will still be applicable to the appreciable proportion of networks under the birth-death-hybridization process that do not have ideal topological properties or meet strict assumptions.

Work towards characterizing and understanding the identifiability of level-1 networks (Solís-Lemus and Ané, 2016; Solis-Lemus et al., 2020; Gross et al., 2021; Allman et al., 2022) has led to many reticulate inference methods assuming an underlying level-1 network (Huber et al., 2010; Solís-Lemus and Ané, 2016; Allman et al., 2019; LeMay et al., 2020). These methods have surged in popularity recently, largely due to their computational tractability and—for `SNaQ` (Solís-Lemus and Ané, 2016) and `NANUQ` (Allman et al., 2019)—their ability to account for discordance both from gene flow and incomplete lineage sorting. However, the level-1 assumption was routinely violated for networks generated under the birth-death-hybridization process, particularly for highly reticulate networks. Fortunately, in practice, methods that make this assumption are often used to either infer few reticulate events (*e.g.,* Morales-Briones et al., 2018; Karimi et al., 2020; Myers et al., 2022) or researchers assume that gene flow events are primarily isolated within each major clade (*e.g.,* Esquerré et al., 2022)—with both cases being more likely to have level-1 networks. Further, recent identifiability results for level-2 networks (Van Iersel et al., 2009; van Iersel et al., 2020, 2022) may prove useful in extending the level-1 limitation of some methods. Nonetheless, it will be important assess how robust these methods are to model violations and characterize possible biases when estimating networks that are not level-1.

5 Conclusions

In this study we have shown that the specific assumptions about reticulation and sampling can have important effects on distribution of generated phylogenetic networks from the birth-death-hybridization process. There are many mechanisms that leave a specific reticulate pattern (see table in Box 1); both the biological process and the sampling framework affect how reticulate patterns (lineage generative, degenerative, neutral) manifest on phylogenetic networks. Methods that can directly investigate which type of reticulation occurs on phylogenetic networks (Hibbins and Hahn, 2019; Flouri et al., 2020) will be vital as we infer patterns of gene flow. Further, system-specific knowledge—factors like life history (Wolf et al., 2001; Rumpho et al., 2008; Wendel et al., 2009; Schulte et al., 2012; Montanari et al., 2016; Meier et al., 2019), geographic distributions (López-Caamal et al., 2014; Ottenburghs et al., 2017; Dolinay et al., 2021), and genomic architecture (Wendel and Cronn, 2002; Ottenburghs et al., 2017; Edelman et al., 2019)—should be incorporated to help inform how to model and interpret each type of reticulation. The specific reticulation types affect diversification and topology of phylogenetic networks, potentially violating assumptions of some methods or biasing diversification rate estimates if not taken into account. Overall, our work highlights importance of thinking carefully and deliberately about how we model gene flow with the birth-death-hybridization process.

6 Additional

6.1 Acknowledgements

We would like to thank members of the Heath lab for helpful comments and feedback on the manuscript.

6.2 Data Availability Statement

All scripts used to simulate data, perform analyses, and create figures are available at: https://github.com/jjustison/BDH_simulation. All simulated data is avaliable on Zenodo at: https://doi.org/10.5281/zenodo.7271696

6.3 References

References

Allman, E. S., H. Baños, and J. A. Rhodes. 2022. Identifiability of species network topologies from genomic sequences using the logdet distance. Journal of Mathematical Biology 84:1–38.

Allman, E. S., H. Baños, and J. A. Rhodes. 2019. NANUQ: a method for inferring species networks from gene trees under the coalescent model. Algorithms for Molecular Biology 14:1–25.

Arenas, M., G. Valiente, and D. Posada. 2008. Characterization of reticulate networks based on the coalescent with recombination. Molecular Biology and Evolution 25:2517–2520.

Barley, A. J., A. Nieto-Montes de Oca, N. L. Manríquez-Morán, and R. C. Thomson. 2022. The evolutionary network of whiptail lizards reveals predictable outcomes of hybridization. Science 377:773–777.

Betat, H., T. Mede, S. Tretbar, L. Steiner, P. F. Stadler, M. Mörl, and S. J. Prohaska. 2015. The ancestor of modern Holozoa acquired the CCA-adding enzyme from Alphaproteobacteria by horizontal gene transfer. Nucleic Acids Research 43:6739–6746.

Blair, C. and C. Ané. 2020. Phylogenetic trees and networks can serve as powerful and complementary approaches for analysis of genomic data. Systematic Biology 69:593–601.

Bock, R. 2010. The give-and-take of DNA: horizontal gene transfer in plants. Trends in Plant Science 15:11–22.

Bordewich, M. and N. Tokac. 2016. An algorithm for reconstructing ultrametric tree-child networks from inter-taxa distances. Discrete Applied Mathematics 213:47–59.

Brysting, A. K., B. Oxelman, K. T. Huber, V. Moulton, and C. Brochmann. 2007. Untangling complex histories of genome mergings in high polyploids. Systematic Biology 56:467–476.

Cao, Z., J. Zhu, and L. Nakhleh. 2019. Empirical Performance of Tree-Based Inference of Phylogenetic Networks. Pages 21:1–21:13 *in* 19th International Workshop on Algorithms in Bioinformatics (WABI 2019) (K. T. Huber and D. Gusfield, eds.) vol. 143 of *Leibniz International Proceedings in Informatics (LIPIcs)* Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany.

Cardona, G., F. Rosselló, and G. Valiente. 2008. Comparison of tree-child phylogenetic networks. IEEE/ACM Transactions on Computational Biology and Bioinformatics 6:552–569.

Chapman, M. A. and J. M. Burke. 2007. Genetic divergence and hybrid speciation. Evolution 61:1773–1780.

Dolinay, M., T. Nečas, B. M. Zimkus, A. Schmitz, E. B. Fokam, E. M. Lemmon, A. R. Lemmon, and V. Gvoždík. 2021. Gene flow in phylogenomics: Sequence capture resolves species limits and biogeography of Afromontane forest endemic frogs from the Cameroon Highlands. Molecular Phylogenetics and Evolution 163:107258.

Edelman, N. B., P. B. Frandsen, M. Miyagi, B. Clavijo, J. Davey, R. B. Dikow, G. García-Accinelli, S. M. Van Belleghem, N. Patterson, D. E. Neafsey, et al. 2019. Genomic architecture and introgression shape a butterfly radiation. Science 366:594–599.

Elworth, R. A., H. A. Ogilvie, J. Zhu, and L. Nakhleh. 2019. Advances in computational methods for phylogenetic networks in the presence of hybridization. Pages 317–360 *in* Bioinformatics and Phylogenetics. Springer.

Erdős, P. L., C. Semple, and M. Steel. 2019. A class of phylogenetic networks reconstructable from ancestral profiles. Mathematical Biosciences 313:33–40.

Esquerré, D., J. S. Keogh, D. Demangel, M. Morando, L. J. Avila, J. W. Sites Jr, F. Ferri-Yáñez, and A. D. Leaché. 2022. Rapid radiation and rampant reticulation: Phylogenomics of South American Liolaemus lizards. Systematic Biology 71:286–300.

Etienne, R. S. and J. Rosindell. 2012. Prolonging the past counteracts the pull of the present: protracted speciation can explain observed slowdowns in diversification. Systematic Biology 61:204.

Flouri, T., X. Jiao, B. Rannala, and Z. Yang. 2020. A Bayesian implementation of the multispecies coalescent model with introgression for phylogenomic analysis. Molecular Biology and Evolution 37:1211–1223.

Francis, A., D. H. Huson, and M. Steel. 2021. Normalising phylogenetic networks. Molecular Phylogenetics and Evolution 163:107215.

Francis, A. and V. Moulton. 2018. Identifiability of tree-child phylogenetic networks under a probabilistic recombination-mutation model of evolution. Journal of Theoretical Biology 446:160–167.

Francis, A. R. and M. Steel. 2015. Which phylogenetic networks are merely trees with additional arcs? Systematic Biology 64:768–777.

Goriely, A. and C. Hyde. 2000. Necessary and sufficient conditions for finite time singularities in ordinary differential equations. Journal of Differential Equations 161:422–448.

Gourbiere, S. and J. Mallet. 2010. Are species real? The shape of the species boundary with exponential failure, reinforcement, and the "missing snowball". Evolution 64:1–24.

Gross, E. and C. Long. 2018. Distinguishing phylogenetic networks. SIAM Journal on Applied Algebra and Geometry 2:72–93.

Gross, E., L. van Iersel, R. Janssen, M. Jones, C. Long, and Y. Murakami. 2021. Distinguishing level-1 phylogenetic networks on the basis of data generated by markov processes. Journal of Mathematical Biology 83:1–24.

Heath, T. A., S. M. Hedtke, and D. M. Hillis. 2008a. Taxon sampling and the accuracy of phylogenetic analyses. Journal of Systematics and Evolution 46:239–257.

Heath, T. A., D. J. Zwickl, J. Kim, and D. M. Hillis. 2008b. Taxon sampling affects inferences of macroevolutionary processes from phylogenetic trees. Systematic Biology 57:160–166.

Hejase, H. A. and K. J. Liu. 2016. A scalability study of phylogenetic network inference methods using empirical datasets and simulations involving a single reticulation. BMC Bioinformatics 17:1–12.

Hibbins, M. S. and M. W. Hahn. 2019. The timing and direction of introgression under the multispecies network coalescent. Genetics 211:1059–1073.

Hibbins, M. S. and M. W. Hahn. 2022. Phylogenomic approaches to detecting and characterizing introgression. Genetics 220:iyab173.

Huber, K. T., M. Lott, V. Moulton, and A. Spillner. 2008. The complexity of deriving multi-labeled trees from bipartitions. Journal of Computational Biology 15:639–651.

Huber, K. T. and V. Moulton. 2006. Phylogenetic networks from multi-labelled trees. Journal of Mathematical Biology 52:613–632.

Huber, K. T., V. Moulton, M. Steel, and T. Wu. 2016. Folding and unfolding phylogenetic trees and networks. Journal of Mathematical Biology 73:1761–1780.

Huber, K. T., L. Van Iersel, S. Kelk, and R. Suchecki. 2010. A practical algorithm for reconstructing level-1 phylogenetic networks. IEEE/ACM Transactions on Computational Biology and Bioinformatics 8:635–649.

Huber, K. T., L. Van Iersel, V. Moulton, and T. Wu. 2015. How much information is needed to infer reticulate evolutionary histories? Systematic Biology 64:102–111.

Höhna, S., M. R. May, and B. R. Moore. 2015. TESS: an R package for efficiently simulating phylogenetic trees and performing Bayesian inference of lineage diversification rates. Bioinformatics 32:789–791.

Janssen, R. and P. Liu. 2021. Comparing the topology of phylogenetic network generators. Journal of Bioinformatics and Computational Biology 19:2140012.

Jansson, J. and W.-K. Sung. 2006. Inferring a level-1 phylogenetic network from a dense set of rooted triplets. Theoretical Computer Science 363:60–68.

Jones, G. R. 2011. Tree models for macroevolution and phylogenetic analysis. Systematic Biology 60:735–746.

Justison, J. A., C. Solis-Lemus, and T. A. Heath. 2022. SiPhyNetwork: An R package for simulating phylogenetic networks. bioRxiv 10.1101/2022.10.26.513953.

Karimi, N., C. E. Grover, J. P. Gallagher, J. F. Wendel, C. Ané, and D. A. Baum. 2020. Reticulate evolution helps explain apparent homoplasy in floral biology and pollination in baobabs (adansonia; bombacoideae; malvaceae). Systematic Biology 69:462–478.

Kemp, T. S. 2009. Phylogenetic interrelationships and pattern of evolution of the therapsids: testing for polytomy. Palaeontologica Africana 44:1–12.

Kendall, D. G. 1948. On the generalized "birth-and-death" process. The Annals of Mathematical Statistics 19:1–15.

Kong, S., J. C. Pons, L. Kubatko, and K. Wicke. 2022. Classes of explicit phylogenetic networks and their biological and mathematical significance. Journal of Mathematical Biology 84:1–44.

Kozak, K. M., M. Joron, W. O. McMillan, and C. D. Jiggins. 2021. Rampant genome-wide admixture across the *Heliconius* radiation. Genome Biology and Evolution 13:evab099.

Lambert, A. and T. Stadler. 2013. Birth–death models and coalescent point processes: The shape and probability of reconstructed phylogenies. Theoretical Population Biology 90:113–128.

Leaché, A. D., R. B. Harris, B. Rannala, and Z. Yang. 2014. The influence of gene flow on species tree estimation: a simulation study. Systematic Biology 63:17–30.

LeMay, M., R. Libeskind-Hadas, and Y.-C. Wu. 2020. A polynomial-time algorithm for minimizing the deep coalescence cost for level-1 species networks. bioRxiv .

Lewis, P. O., M. T. Holder, and K. E. Holsinger. 2005. Polytomies and Bayesian phylogenetic inference. Systematic Biology 54:241–253.

Linz, S. and C. Semple. 2020. Caterpillars on three and four leaves are sufficient to reconstruct binary normal networks. Journal of Mathematical Biology 81:961–980.

López-Caamal, A., Z. Cano-Santana, J. Jiménez-Ramírez, R. Ramírez-Rodríguez, and E. Tovar-Sánchez. 2014. Is the insular endemic *Psidium socorrense* (Myrtaceae) at risk of extinction through hybridization? Plant Systematics and Evolution 300:1959–1972.

Mallet, J. 2005. Hybridization as an invasion of the genome. Trends in Ecology & Evolution 20:229–237.

Mallet, J., N. Besansky, and M. W. Hahn. 2016. How reticulated are species? BioEssays 38:140–149.

Meier, J. I., R. B. Stelkens, D. A. Joyce, S. Mwaiko, N. Phiri, U. K. Schliewen, O. M. Selz, C. E. Wagner, C. Katongo, and O. Seehausen. 2019. The coincidence of ecological opportunity with hybridization explains rapid adaptive radiation in Lake Mweru cichlid fishes. Nature Communications 10:1–11.

Moen, D. and H. Morlon. 2014. Why does diversification slow down? Trends in Ecology & Evolution 29:190–197.

Molloy, E. K., A. Durvasula, and S. Sankararaman. 2021. Advancing admixture graph estimation via maximum likelihood network orientation. Bioinformatics 37:i142–i150.

Montanari, S. R., J.-P. A. Hobbs, M. S. Pratchett, and L. Van Herwerden. 2016. The importance of ecological and behavioural data in studies of hybridisation among marine fishes. Reviews in Fish Biology and Fisheries 26:181–198.

Mooers, A., O. Gascuel, T. Stadler, H. Li, and M. Steel. 2012. Branch lengths on birth–death trees and the expected loss of phylogenetic diversity. Systematic Biology 61:195–203.

Mooers, A. O. and S. B. Heard. 1997. Inferring evolutionary process from phylogenetic tree shape. The Quarterly Review of Biology 72:31–54.

Morales-Briones, D. F., A. Liston, and D. C. Tank. 2018. Phylogenomic analyses reveal a deep history of hybridization and polyploidy in the Neotropical genus *Lachemilla* (Rosaceae). New Phytologist 218:1668–1684.

Morin, M. M. and B. M. E. Moret. 2006. NetGen: generating phylogenetic networks with diploid hybrids. Bioinformatics 22:1921–1923.

Morlon, H., M. D. Potts, and J. B. Plotkin. 2010. Inferring the dynamics of diversification: a coalescent approach. PLoS Biology 8:e1000493.

Murakami, Y., L. van Iersel, R. Janssen, M. Jones, and V. Moulton. 2019. Reconstructing tree-child networks from reticulate-edge-deleted subnetworks. Bulletin of Mathematical Biology 81:3823–3863.

Myers, E. A., D. G. Mulcahy, B. Falk, K. Johnson, M. Carbi, and K. de Queiroz. 2022. Interspecific gene flow and mitochondrial genome capture during the radiation of Jamaican *Anolis* lizards (Squamata; Iguanidae). Systematic Biology 71:501–511.

Nabhan, A. R. and I. N. Sarkar. 2012. The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy. Briefings in Bioinformatics 13:122–134.

Nee, S. 2006. Birth-death models in macroevolution. Annual Review of Ecology, Evolution, and Systematics 37:1–17.

Nee, S., E. C. Holmes, R. M. May, and P. H. Harvey. 1994a. Extinction rates can be estimated from molecular phylogenies. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 344:77–82.

Nee, S., R. M. May, and P. H. Harvey. 1994b. The reconstructed evolutionary process. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 344:305–311.

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

Ottenburghs, J. 2020. Ghost introgression: spooky gene flow in the distant past. Bioessays 42:2000012.

Ottenburghs, J., R. H. Kraus, P. van Hooft, S. E. van Wieren, R. C. Ydenberg, and H. H. Prins. 2017. Avian introgression in the genomic era. Avian Research 8:1–11.

Ottenburghs, J., P. van Hooft, S. E. van Wieren, R. C. Ydenberg, and H. H. Prins. 2016. Birds in a bush: Toward an avian phylogenetic network. The Auk: Ornithological Advances 133:577–582.

Pardi, F. and C. Scornavacca. 2015. Reconstructible phylogenetic networks: do not distinguish the indistinguishable. PLoS Computational Biology 11:e1004135.

Pybus, O. G. and P. H. Harvey. 2000. Testing macro–evolutionary models using incomplete molecular phylogenies. Proceedings of the Royal Society of London. Series B: Biological Sciences 267:2267–2272.

Rabosky, D. L. and I. J. Lovette. 2008. Explosive evolutionary radiations: decreasing speciation or increasing extinction through time? Evolution 62:1866–1875.

Rannala, B. and Z. Yang. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. Journal of Molecular Evolution 43:304–311.

Raup, D. M. 1985. Mathematical models of cladogenesis. Paleobiology 11:42–52.

Rumpho, M. E., J. M. Worful, J. Lee, K. Kannan, M. S. Tyler, D. Bhattacharya, A. Moustafa, and J. R. Manhart. 2008. Horizontal gene transfer of the algal nuclear gene psbo to the photosynthetic sea slug *Elysia chlorotica*. Proceedings of the National Academy of Sciences 105:17867–17871.

Salzburger, W. 2018. Understanding explosive diversification through cichlid fish genomics. Nature Reviews Genetics 19:705–717.

Schulte, U., M. Veith, and A. Hochkirch. 2012. Rapid genetic assimilation of native wall lizard populations (*Podarcis muralis*) through extensive hybridization with introduced lineages. Molecular Ecology 21:4313–4326.

Semple, C. and G. Toft. 2021. Trinets encode orchard phylogenetic networks. Journal of Mathematical Biology 83:1–20.

Slarkin, M. 1985. Gene flow in natural populations. Annual Review of Ecology and Systematics 16:393–430.

Slowinski, J. B. 2001. Molecular polytomies. Molecular Phylogenetics and Evolution 19:114–120.

Solís-Lemus, C. and C. Ané. 2016. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. PLoS Genetics 12:e1005896.

Solis-Lemus, C., A. Coen, and C. Ane. 2020. On the identifiability of phylogenetic networks under a pseudolikelihood model. arXiv preprint arXiv:2010.01758 .

Soltis, P. S. and D. E. Soltis. 2009. The role of hybridization in plant speciation. Annual Review of Plant Biology 60:561–588.

Stadler, T. 2008. Lineages-through-time plots of neutral models for speciation. Mathematical Biosciences 216:163–171.

Stanley, E. L., A. M. Bauer, T. R. Jackman, W. R. Branch, and P. L. F. N. Mouton. 2011. Between a rock and a hard polytomy: rapid radiation in the rupicolous girdled lizards (Squamata: Cordylidae). Molecular Phylogenetics and Evolution 58:53–70.

Stebbins, G. L. 1959. The role of hybridization in evolution. Proceedings of the American Philosophical Society 103:231–251.

Steel, M. and A. Mooers. 2010. The expected length of pendant and interior edges of a Yule tree. Applied Mathematics Letters 23:1315–1319.

Suh, A. 2016. The phylogenomic forest of bird trees contains a hard polytomy at the root of Neoaves. Zoologica Scripta 45:50–62.

Suvorov, A., B. Y. Kim, J. Wang, E. E. Armstrong, D. Peede, E. R. D'agostino, D. K. Price, P. J. Waddell, M. Lang, V. Courtier-Orgogozo, et al. 2022. Widespread introgression across a phylogeny of 155 *Drosophila* genomes. Current Biology 32:111–123.

Taylor, S. A. and E. L. Larson. 2019. Insights from genomes into the evolutionary importance and prevalence of hybridization in nature. Nature Ecology & Evolution 3:170–177.

Tea, Y.-K., J.-P. A. Hobbs, F. Vitelli, J. D. DiBattista, S. Y. Ho, and N. Lo. 2020. Angels in disguise: sympatric hybridization in the marine angelfishes is widespread and occurs between deeply divergent lineages. Proceedings of the Royal Society B 287:20201459.

Thawornwattana, Y., D. Dalquen, and Z. Yang. 2018. Coalescent analysis of phylogenomic data confidently resolves the species relationships in the *Anopheles gambiae* species complex. Molecular Biology and Evolution 35:2512–2527.

Todesco, M., M. A. Pascual, G. L. Owens, K. L. Ostevik, B. T. Moyers, S. Hübner, S. M. Heredia, M. A. Hahn, C. Caseys, and D. G. Bock. 2016. Hybridization and extinction. Evolutionary Applications 9:892–908.

Tricou, T., E. Tannier, and D. M. de Vienne. 2021. Ghost lineages highly influence the interpretation of introgression tests. Systematic Biology 71:1147–1158.

Van Iersel, L., J. Keijsper, S. Kelk, L. Stougie, F. Hagen, and T. Boekhout. 2009. Constructing level-2 phylogenetic networks from triplets. IEEE/ACM Transactions on Computational Biology and Bioinformatics 6:667–681.

van Iersel, L., S. Kole, V. Moulton, and L. Nipius. 2022. An algorithm for reconstructing level-2 phylogenetic networks from trinets. Information Processing Letters 178:106300.

Van Iersel, L. and V. Moulton. 2014. Trinets encode tree-child and level-2 phylogenetic networks. Journal of Mathematical Biology 68:1707–1729.

van Iersel, L., V. Moulton, and Y. Murakami. 2020. Reconstructibility of unrooted level-k phylogenetic networks from distances. Advances in Applied Mathematics 120:102075.

Velasco, J. D. 2008. The prior probabilities of phylogenetic trees. Biology & Philosophy 23:455–473.

Wang, Y., Y. Zhao, and Q. Pan. 2022. Advances, challenges and opportunities of phylogenetic and social network analysis using COVID-19 data. Briefings in Bioinformatics 23:bbab406.

Warnock, R. C., T. A. Heath, and T. Stadler. 2020. Assessing the impact of incomplete species sampling on estimates of speciation and extinction rates. Paleobiology 46:137–157.

Wawerka, M., D. Dąbkowski, N. Rutecka, A. Mykowiecka, and P. Górecki. 2022. Embedding gene trees into phylogenetic networks by conflict resolution algorithms. Algorithms for Molecular Biology 17:1–23.

Wendel, J. F., C. Brubaker, I. Alvarez, R. Cronn, and J. M. Stewart. 2009. Evolution and natural history of the cotton genus. Pages 3–22 *in* Genetics and Genomics of Cotton. Springer.

Wendel, J. F. and R. C. Cronn. 2002. Polyploidy and the evolutionary history of cotton. Advances in Agronomy 87:139–186.

Willson, S. 2010a. Regular networks can be uniquely constructed from their trees. IEEE/ACM Transactions on Computational Biology and Bioinformatics 8:785–796.

Willson, S. J. 2008. Reconstruction of certain phylogenetic networks from the genomes at their leaves. Journal of Theoretical Biology 252:338–349.

Willson, S. J. 2010b. Properties of normal phylogenetic networks. Bulletin of Mathematical Biology 72:340–358.

Wolf, D. E., N. Takebayashi, and L. H. Rieseberg. 2001. Predicting the risk of extinction through hybridization. Conservation Biology 15:1039–1053.

Woodhams, M. D., P. J. Lockhart, and B. R. Holland. 2016. Simulating and summarizing sources of gene tree incongruence. Genome Biology and Evolution 8:1299–1315.

Zhang, C., H. A. Ogilvie, A. J. Drummond, and T. Stadler. 2018. Bayesian inference of species networks from multilocus sequence data. Molecular Biology and Evolution 35:504–517.

Zhang, L. 2016. On tree-based phylogenetic networks. Journal of Computational Biology 23:553–565.

Zhang, L. 2019. Clusters, trees, and phylogenetic network classes. Pages 277–315 *in* Bioinformatics and Phylogenetics. Springer.

7 Supplementary Material

Table S1: Hybridization type proportions. The overall hybridization rate (ν) can be split into three separate rates: lineage generative (ν_+), lineage neutral (ν_0), and lineage degenerative (ν_-). The values in the table represent the proportion of $\nu = 0.1$ for each hybridization type for simulations across the hybridization rate simplex.

	ν_+	ν_-	ν_0		ν_+	ν_-	ν_0		ν_+	ν_-	ν_0
1	0.00	1.00	0.00	22	0.15	0.35	0.50	42	0.35	0.15	0.50
2	0.10	0.90	0.00	23	0.10	0.30	0.60	43	0.30	0.10	0.60
3	0.05	0.85	0.10	24	0.05	0.25	0.70	44	0.25	0.05	0.70
4	0.00	0.80	0.20	25	0.00	0.20	0.80	45	0.20	0.00	0.80
5	0.20	0.80	0.00	26	0.50	0.50	0.00	46	0.70	0.30	0.00
6	0.15	0.75	0.10	27	0.45	0.45	0.10	47	0.65	0.25	0.10
7	0.10	0.70	0.20	28	0.40	0.40	0.20	48	0.60	0.20	0.20
8	0.05	0.65	0.30	29	0.35	0.35	0.30	49	0.55	0.15	0.30
9	0.00	0.60	0.40	30	0.30	0.30	0.40	50	0.50	0.10	0.40
10	0.30	0.70	0.00	31	0.25	0.25	0.50	51	0.45	0.05	0.50
11	0.25	0.65	0.10	32	0.20	0.20	0.60	52	0.40	0.00	0.60
12	0.20	0.60	0.20	33	0.15	0.15	0.70	53	0.80	0.20	0.00
13	0.15	0.55	0.30	34	0.10	0.10	0.80	54	0.75	0.15	0.10
14	0.10	0.50	0.40	35	0.05	0.05	0.90	55	0.70	0.10	0.20
15	0.05	0.45	0.50	36	0.00	0.00	1.00	56	0.65	0.05	0.30
16	0.00	0.40	0.60	37	0.60	0.40	0.00	57	0.60	0.00	0.40
17	0.40	0.60	0.00	38	0.55	0.35	0.10	58	0.90	0.10	0.00
18	0.35	0.55	0.10	39	0.50	0.30	0.20	59	0.85	0.05	0.10
19	0.30	0.50	0.20	40	0.45	0.25	0.30	60	0.80	0.00	0.20
20	0.25	0.45	0.30	41	0.40	0.20	0.40	61	1.00	0.00	0.00
21	0.20	0.40	0.40								

Table S2: Parameters common to all simulations

Diversification Parameters	$\lambda = 4$ $\mu = 2$ $\nu = 0.1$
Simulation Settings	$age = 1$ 20,000 replicates
Rejection Conditions	Extinct Phylogeny Only one extant taxa Simulation $>$ 10 seconds

Table S3: Simulations across the hybrid-rate simplex. Each combination of hybridization rates is given in Table S1 and visually depicted in Figure 3.

Dataset	Parameters
Complete	**Sampling Settings** Extinct species kept $Sampling\ fraction = 1$ **Genetic Distance** $\delta = 0$
Distance Dependent	**Sampling Settings** Extinct species kept $Sampling\ fraction = 1$ **Genetic Distance** $\delta = 0.5$
Extant-only	**Sampling Settings** Extinct species pruned $Sampling\ fraction = 1$ **Genetic Distance** $\delta = 0$
Incomplete Sampling	**Sampling Settings** Extinct species pruned $Sampling\ fraction = 0.75$ **Genetic Distance** $\delta = 0$

Table S4: Simulations using constant hybridization rates for each reticulation type ($\nu_+ = \nu_- = \nu_0 = \frac{1}{30}$).

Dataset	Parameters
Distance Dependent	**Sampling Settings** Extinct species kept $Sampling\ fraction = 1$ **Genetic Distance** δ varies from 0 to 2.0
Incomplete Sampling	**Sampling Settings** Extinct species pruned $Sampling\ fraction$ from 0.20 to 1.0 **Genetic Distance** $\delta = 0$

Table S5: Summary statistics across the hybrid simplex. Summary statistics were computed for each dataset by first taking the average value or proportion within each of the 61 hybrid rate proportions and then summarizing between each hybrid rate proportion (See 3 for hybrid rate proportions).

	Tips				**Reticulations**				**Reticulation Density**				**Proportion with Zero Reticulations**			
Dataset	mean	var	min	max	mean	var	min	max	mean	var	min	max	mean	var	min	max
Complete	38.14	141.68	24.89	78.65	10.20	61.14	3.57	34.80	0.13	1.26e-4	0.11	0.15	0.27	0.00	0.26	0.28
Extant-Only	22.25	97.42	12.39	58.05	8.79	55.40	2.72	32.62	0.18	5.52e-05	0.17	0.20	0.32	0.00	0.31	0.34
Incomplete Sampling	16.51	48.26	9.41	43.82	7.36	35.50	2.34	29.13	0.21	8.84e-5	0.19	0.23	0.35	0.00	0.34	0.37
Distance Dependence	34.77	13.89	28.37	44.79	3.11	0.39	2.13	4.86	0.06	5.89e-7	0.06	0.06	0.35	0.00	0.34	0.36

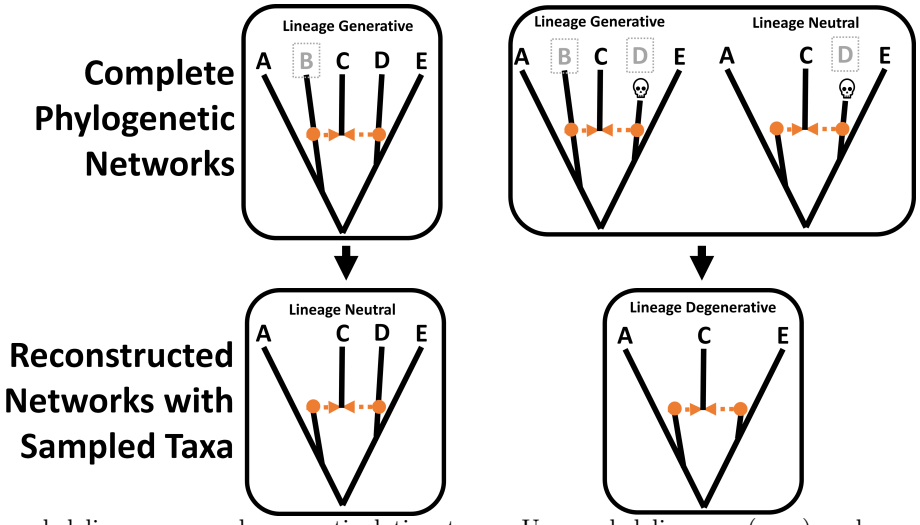

Figure S1: Unsampled lineages can change reticulation type. Unsampled lineages (grey) and are pruned in the reconstructed phylogenies. Extinct species are depicted with a skull.

Table S6: Class proportions across the hybrid simplex. Here, each rate combination is summarized and taken used as a point in analysis.

	Tree-based			**FU-Stable**			**Tree-Child**			**Normal**		
Dataset	mean	min	max	mean	min	max	mean	min	max	mean	min	max
Complete	0.971	0.940	1.000	0.757	0.738	1.000	0.631	0.589	1.000	0.536	0.464	1.000
Extant-Only	0.959	0.931	0.996	0.737	0.722	0.810	0.601	0.582	0.702	0.486	0.452	0.613
Incomplete Sampling	0.951	0.922	0.967	0.733	0.709	0.773	0.598	0.588	0.624	0.483	0.468	0.535
Distance Dependence	0.997	0.995	1.000	0.903	0.882	1.000	0.789	0.710	1.000	0.658	0.556	1.000

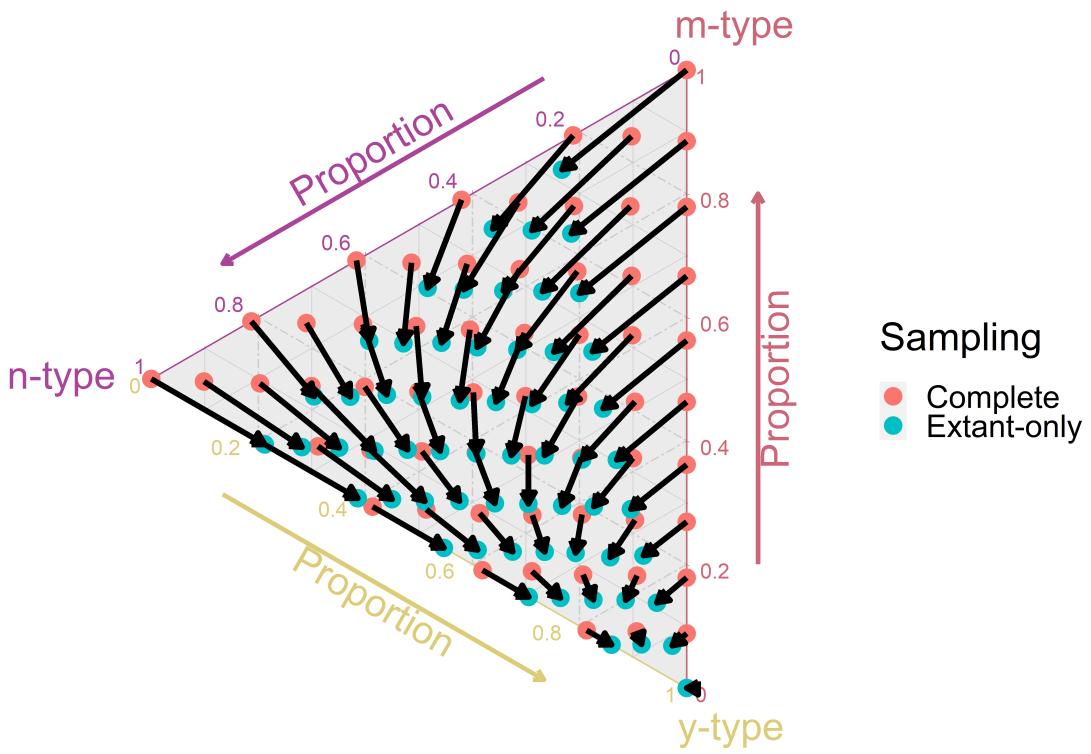

Figure S2: Observed hybridization types on simulated phylogenetic networks. Axes represent the observed proportion of total hybridizations for each hybridization type: lineage generative (m-type), lineage neutral (n-type), and lineage degenerative (y-type). Each circle represents the average proportion 20,000 replicates of a given hybrid rate combination. Red circles denote completely sampled phylogenies while blue circles denote extant-only sampling. Circles connected with an arrow denote the exact same simulation conditions, with the exception of sampling.

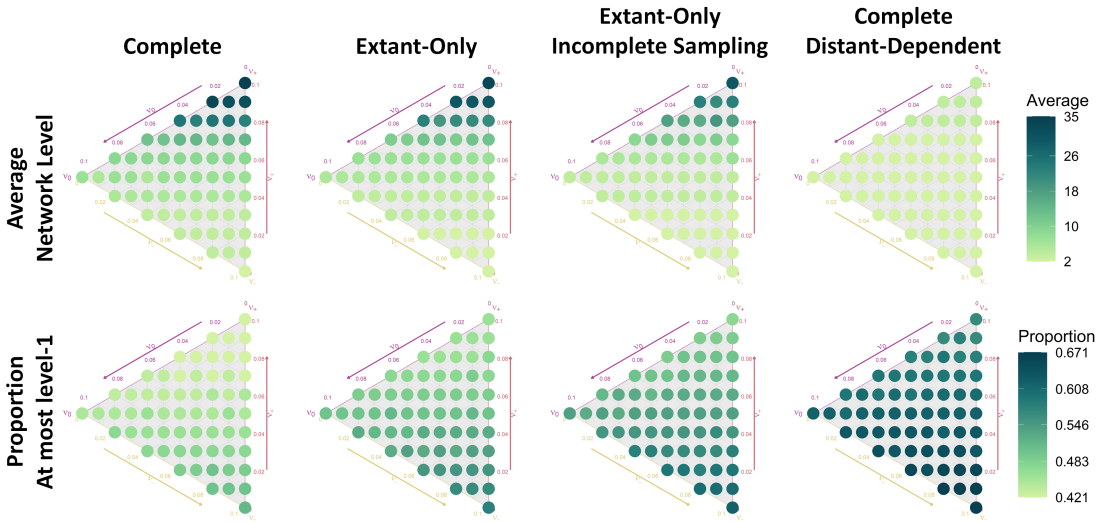

Figure S3: Phylogenetic level across the hybrid rate simplex. Each column denotes the simulation conditions used to simulate phylogenetic networks across the hybrid rate simplex (See 2.4.1). Each row summarizes the average network level and proportion of phylogenetic networks that were level-1 or lower. The color of each point is used to depict the average value observed from 20,000, with darker colors representing a higher values.

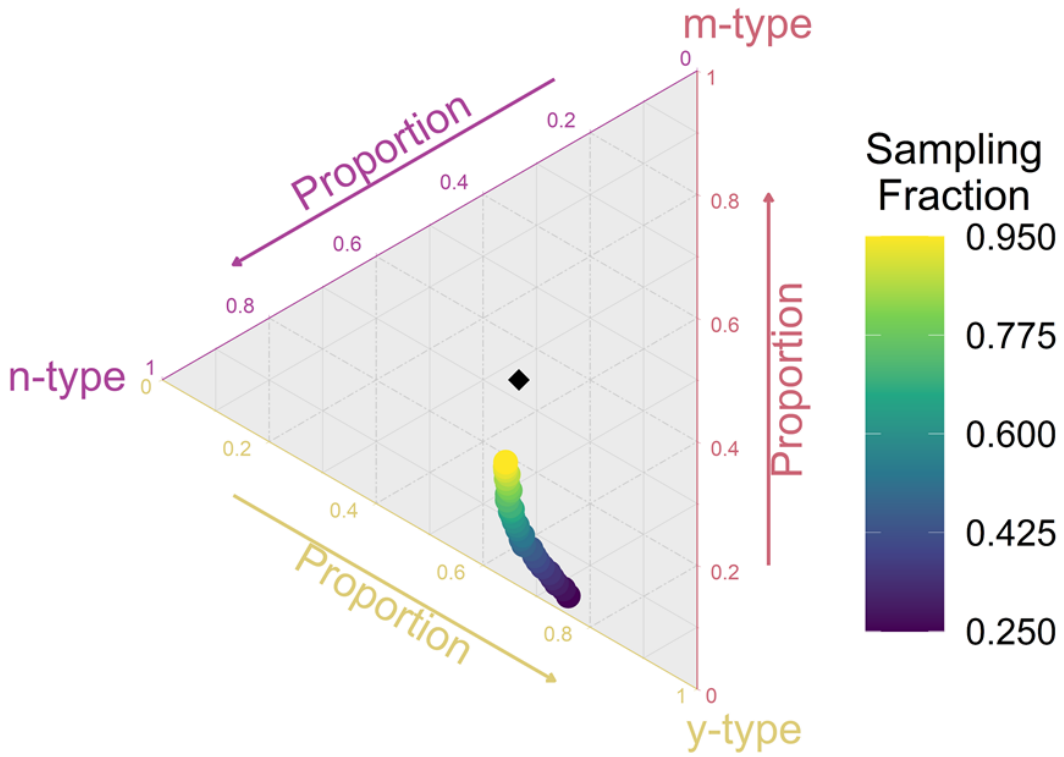

Figure S4: Observed hybridization types on simulated phylogenetic networks. Axes represent the observed proportion of total hybridizations for each hybridization type: lineage generative (m-type), lineage neutral (n-type), and lineage degenerative (y-type). Each circle represents the average proportion 20,000 replicates of a given hybrid rate combination. The black diamond represents the expected proportion with complete sampling.

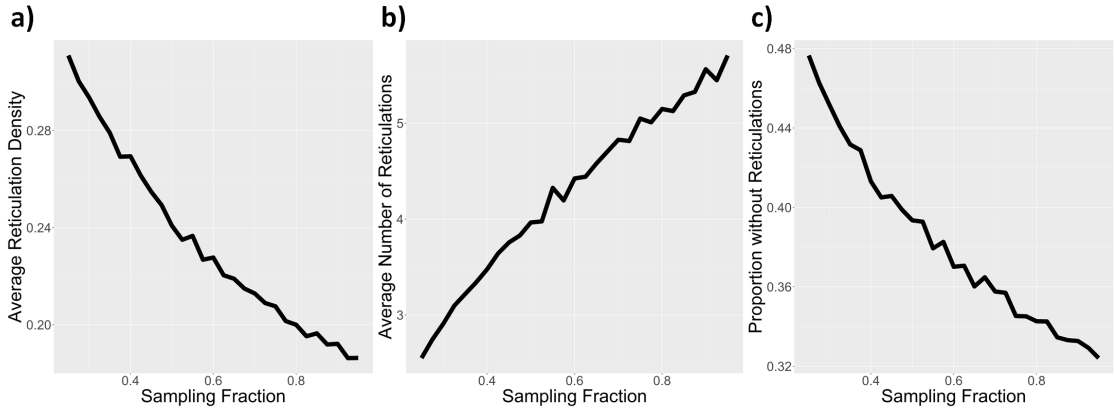

Figure S5: Reticulation summary statistics as a function of incomplete sampling. Values represent observed the averages taken from 20,000 simulated phylogenetic networks for each sampling fraction value. The sampling fraction represents the number of extant tips that are randomly sampled from the network. Extant-only networks were used for the incomplete sampling dataset while complete phylogenetic networks with extant and extinct species are used for the genetic distance-dependent dataset

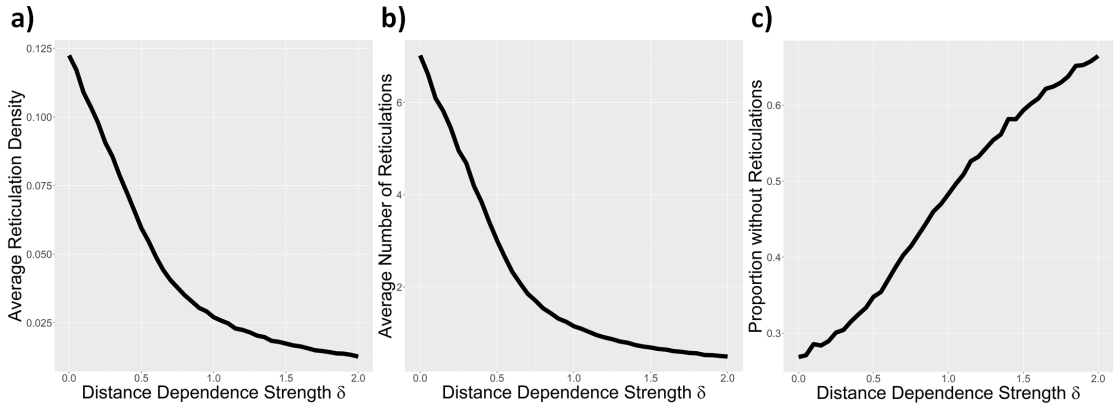

Figure S6: Reticulation summary statistics as a function of genetic distance-dependence. Values represent the observed averages taken from 20,000 simulated phylogenetic networks for each genetic distance-dependence strength (δ). The value of *delta* corresponds to the strength of the genetic distance dependence, with higher values of δ meaning reticulation events primarily only occur between closely related lineages.