# The Diverse Applications of Tree Set Visualization and Exploration

Jeremy M. Brown[1], Genevieve G. Mount[1], Kyle A. Gallivan[2], and James C. Wilgenbusch[3]

[1]Department of Biological Sciences and Museum of Natural Science, Louisiana State University, Baton Rouge, LA, 70803

[2]Department of Mathematics, Florida State University, Tallahassee, FL

[3]Minnesota Supercomputing Institute, University of Minnesota, Minneapolis, MN

## Abstract

All phylogenetic studies are built around sets of trees. Tree sets carry different kinds of information depending on the data and approaches used to generate them, but ultimately the variation they contain and their structure is what drives new phylogenetic insights. In order to better understand the variation in and structure of phylogenetic tree sets, we need tools that are generic, flexible, and exploratory. These tools can serve as natural complements to more formal, statistical investigations and allow us to flag surprising or unexpected observations, better understand the results of model-based studies, as well as build intuition. Here, we describe such a set of tools and provide examples of how they can be applied to relevant questions in phylogenetics, phylogenomics, and species-tree inference. These tools include both visualization techniques and quantitative summaries and are currently implemented in the TreeScaper software package (Huang et al. 2016).

# Introduction to Visualizing and Exploring Tree Sets

All phylogenetic studies are built around sets of trees. Tree sets carry different kinds of information depending on the data and approaches used to generate them, but ultimately the variation they contain and their structure is what drives new phylogenetic insights. However, tree sets also present particular challenges. Trees are complicated objects that are meant to be interpreted visually and each tree, on its own, can carry a huge amount of information. Trees also naturally exist in a high-dimensional space that is challenging to conceptualize.

In some cases, tree sets can be understood and explored in formal statistical frameworks. For instance, the multispecies coalescent model (MSC) describes how stochastic coalescent processes explain variation in the histories of individual genes, and how these histories differ both from each other and from the overarching species tree. Models like the MSC are elegant and incredibly valuable for testing different hypotheses, but their application is specific to those cases where we can reasonably expect the variation in trees to be explained by the process(es) they include. Some tree sets may not have a natural process-based model to describe their variation. For instance, sets of trees meant to represent the uncertainty in a phylogenetic estimate (i.e., bootstrap sets or those drawn from a posterior distribution) have variation that is not described by a biological process. Other variation in trees may result from problems with data quality or poor fit between our observed data and the models we have available (e.g., misalignment, unintended paralogy, or heterogeneous evolutionary processes).

In order to better understand the variation in and structure of tree sets generally, we need tools that are generic, flexible, and exploratory. These tools can serve as natural complements to more formal, statistical investigations and allow us to flag surprising or unexpected observations, better understand the results of model-based studies (e.g., by comparing the trees generated from replicate runs or assuming different kinds of models), as well as build intuition.

Here, we describe a set of tools for understanding variation and structure in tree sets, and provide some examples of how they can be applied to questions related to phylogenetics and species-tree inference. These tools include both visualization techniques and quantitative summaries and are currently implemented in the

TreeScaper software package (Huang et al. 2016). We start with a brief overview of the approaches, then move to examples of how they can be applied.

## Tree Set Visualization

Non-linear dimensionality reduction (NLDR) is a general class of approaches for projecting the position of data that exist in a high-dimensional space onto a lower-dimensional space that can be visualized and explored more easily (typically, in two or three dimensions; Lee and Verleysen 2007). This projection is done in a way that attempts to preserve the pairwise distances among data points as much as possible, to minimize distortions, and to provide insight into the relationships between data points in their original, high-dimensional representation. However, some distortion may be inevitable, since visualization is restricted to no more than three dimensions, but the intrinsic dimensionality of the data (i.e., the number of dimensions required to maintain the original distances between data points) may be larger. TreeScaper (Huang et al. 2016) includes tools to estimate the intrinsic dimensionality of the pairwise distances among a set of phylogenetic trees (Wilgenbusch et al. 2017). When this intrinsic dimensionality is much greater than the number of dimensions used for visualization, the visualization should be interpreted with caution (though the projection may still be useful for other downstream analyses).

NLDR techniques have a long history of development in mathematics and other applied fields, but were first used to visualize phylogenetic trees by Amenta and Klingner (2002), who developed the TreeSetViz module within Mesquite (Maddison and Maddison 2004). Since then, several studies have explored applications of NLDR in phylogenetics (e.g., Hillis et al. 2005, Jombart et al. 2016, Warren et al. 2017). More recently, Wilgenbusch et al. (2017) addressed some outstanding questions about how different NLDR approaches perform in the phylogenetic context. In particular, they compared different stress functions, as well as different algorithms for optimizing projections. While a variety of stress functions and algorithms are available in TreeScaper, we will focus here on examples of how NLDR and complementary techniques can be applied to questions in phylogenetics and species-tree inference, and refer readers to other publications for more technical detail.

Before we apply NLDR, a few important points should be mentioned. First, NLDR projections can be incredibly valuable for the intuitive and visually appealing

summaries they provide, but in most phylogenetic cases the low-dimensional projections will distort the original tree-to-tree distances to varying degrees and should be interpreted with caution. Second, there are several stress functions available in TreeScaper for performing NLDR, which, in our experience, can sometimes meaningfully affect the visualization of tree space. No single function will necessarily always perform best, although Wilgenbusch et al. (2017) suggest that Curvilinear Components Analysis (CCA; Demartines and Herault 1997) best preserved the original tree-to-tree distances in their tests with mitochondrial data. Given these potential differences in results, one approach is to focus on those relationships that remain constant across different functions. Third, NLDR visualizations can be a powerful exploratory technique that suggests further avenues of investigation in more formalized statistical frameworks, and this exploratory role should be included in descriptions of a phylogenomic or species-tree workflow.

## Detecting Structure in Tree Sets

Interpretations of tree space as a network have a deep history in phylogenetics, largely in the context of exploring these networks to find the optimal tree (or set of trees) given some criterion – like a parsimony score, likelihood score, or posterior probability (see Felsenstein 2004 for an excellent overview of approaches for exploring tree space, and Whidden and Matsen 2015 for a recent study on the properties of these searches in Bayesian analyses). In this context, networks are conceptualized with trees as nodes, and edges connecting "neighboring" trees. Whether or not trees are considered neighbors will depend on the type of tree alteration that is used to move through tree space. Note that this type of network is distinct from the networks that are used to describe relationships among species, where nodes represent splitting or fusion events between lineages and branches represent the evolution of different lineages.

Here, we also use networks, but we are not primarily interested in finding the "best" tree. Instead, we are interested in how the tree set itself is structured. For instance, are there regions of tree space where trees in a set are more abundant or dense? Do trees form distinct groups? To investigate these properties, we construct networks with trees as nodes, and we add edges between all pairs of trees in the set (Fig. 1). These edges are then weighted as a function of a pairwise tree distance (several distance options are available in TreeScaper). More specifically, edges are

weighted with the affinity between two trees, which, roughly speaking, is the inverse of distance. Trees that are more similar (have a low pairwise distance) will have a higher pairwise affinity and will be connected by an edge with a higher weight. If trees in a set are clumped in tree space, we expect to find regions of the network where some trees form groups, such that there are large edge weights inside the groups and small edge weights between groups (Fig. 1). Note that while we use these topological networks to detect structure in tree sets, they can be very challenging to visualize for large sets of trees. Therefore, we often display results from these networks by coloring points (trees) in an NLDR plot.

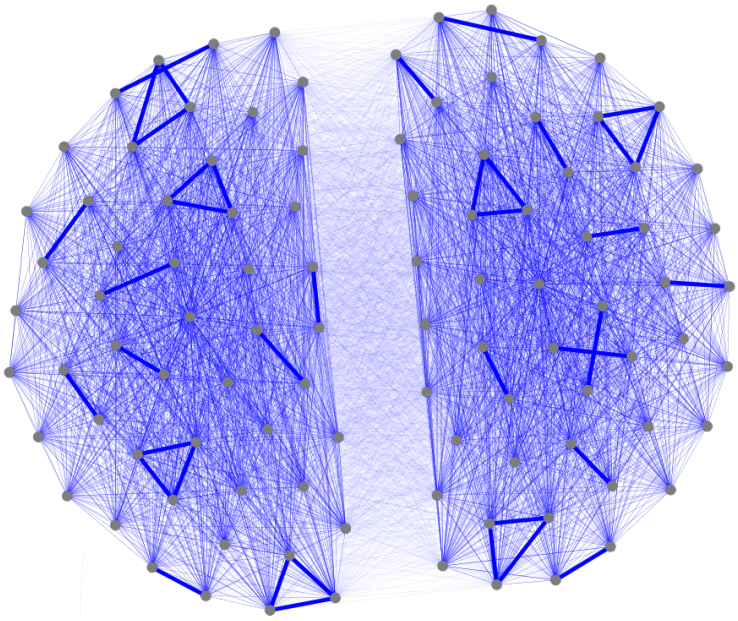

Fig. 1 – An example topological affinity network with 100 25-taxon trees grouped into two distinct communities. Each node is a different tree topology, and the edges between nodes are weighted by the affinity of each pair of trees. Larger affinities (smaller distances) are represented by thicker lines. The arrangement of points was chosen simply to make edges visible, and the placement of nodes is not optimized for two-dimensional representation as it is in an NLDR plot.

In addition to tree-based networks, we also construct bipartition-based networks. In this second type, bipartitions are nodes and the edges connecting pairs of bipartitions indicate how often they are found together in the same trees. More formally, the edges that connect the bipartitions are weighted by their covariances

in presence/absence across trees in a tree set (Fig. 2). Bipartitions that are found together in the same tree more often than expected by chance will have large, positive covariances, and those that are found together less often than expected by chance will have large, negative covariances. If there is very little structure in the tree set, the bipartition network should only contain edges with low weights (both positive and negative; Fig. 2A). However, if the tree set is highly structured (for instance, when combining trees inferred from two genes with strongly conflicting phylogenetic signal), then bipartitions should form strong associations (Fig. 2B). Those bipartitions found in trees inferred with Gene 1 will all have strong, positive covariances with each other, as will those found in trees inferred with Gene 2. However, bipartitions from Gene 1 trees should have strong, negative covariances with bipartitions from Gene 2 trees.

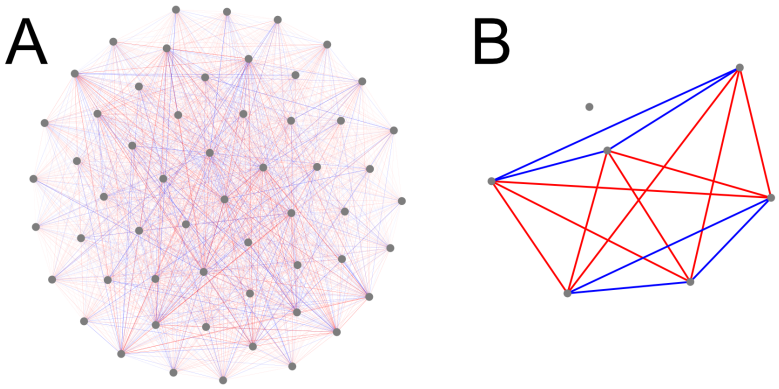


Fig. 2 – Examples of bipartition covariance networks based on sets of seven-taxon trees. Blue branches indicate positive covariances and red branches indicate negative covariances. The line weight indicates the relative magnitude of the covariance. (A) A covariance network based on 1,000 trees sampled from a uniform distribution. (B) A covariance network based on 1,000 trees, where 500 trees have one topology and the other 500 have a distinct topology. The point not connected to any lines indicates a bipartition that is present in every tree in the tree set. The other six bipartitions are found in only one topology or the other.

A major advantage of conceptualizing both trees and bipartitions as parts of a network is the ability to take advantage of the substantial progress mathematicians have made in formalizing the study of networks (Newman 2010). Here, we specifically focus on a class of techniques called community detection methods

(Girvan and Newman 2002, Fortunato 2010). Roughly speaking, the goal of community detection is to find sets nodes that form distinct communities, where nodes in a community are tightly connected by edges with large positive weights and nodes in different communities are weakly connected or are connected by edges with large negative weights. (Note that topological networks will only have positive weights on edges, while bipartition networks will have both positive and negative edge weights.) Community detection methods are able to look for this type of structure in networks without the need to specify the number or size of these groups *a priori*. Many different community detection models can be applied to networks, and TreeScaper implements several different options (Huang et al. 2016, Brown et al. 2020). Here, we focus on use of the Constant Potts model (Traag et al. 2011), because it is better able to detect communities of widely varying sizes than many other models (a property known as being "resolution-limit-free").

Many community detection methods, including Constant Potts, include tuning parameters that can adjust the focus of the model on communities of varying size and number. At one extreme of the tuning parameter values, the model will prefer to put all nodes in one large community. At the other extreme, the model will prefer to place every node in its own community. Since these results are not biologically interesting, we focus on the communities identified by the model with intermediate parameter values. While this gives us a great deal of flexibility, we generally do not know ahead of time which parameter values we should prefer. Instead, we allow the network to tell us. By adjusting the tuning parameter in small increments, we can look for regions of parameter values that all return the same community structure (called plateaus in parameter space). These regions of stability indicate that the detected community structure is a natural property of the network. A given network may display zero, one, or multiple of these intrinsic community structures.

## Applications to Gene Trees, Species Trees, and Phylogenomics

Here, we provide examples of ways in which combinations of tree set visualization and community detection can be applied to questions about gene tree variation, species-tree inference, and phylogenomics. These tools are intended to allow exploration, provide new intuition, and facilitate more detailed investigation. As a result, many of the examples we outline below do not overturn previous understanding, but rather provide a new perspective on it.

## Sensitivity to Models of Sequence Evolution

Species-tree inference can be compromised by systematic errors in gene-tree inference, which can occur when models of sequence evolution do not fit the data well. Many studies have explored how inferred gene trees, or the distribution of uncertainty in these trees, may vary depending on which model of sequence evolution is used in an analysis (e.g., Lemmon and Moriarty 2004, Huelsenbeck and Rannala 2004). However, the comparison of gene trees resulting from analyses that assume different models can be challenging, especially as trees become large. Also, in some cases, we would like to gain intuition for how the tree changes with model assumptions, especially when the best model is not clear.

Tree set visualization and community detection methods can provide intuitive summaries for how results change across analyses, and focus attention on those parts of the tree that strongly conflict across analyses. These methods can also highlight regions of trees that are consistently uncertain in all analyses. To illustrate how these tools can be used in this context, we conducted Bayesian phylogenetic inference using RevBayes (Höhna et al. 2016) for an alignment of *cytb* sequences from primates. We conducted three separate analyses that assumed a JC, HKY, or GTR model (see Yang 2014 for an overview of these and other related models).

We sampled 80 trees from the posterior distributions of each analysis, and then visualized these trees in two dimensions using NLDR based on weighted Robinson-Foulds (wRF) distances (Fig. 3). When trees are colored based on the assumed model, a cursory visual inspection reveals that the results of our analysis are sensitive to the assumed model (Fig. 3A). HKY and GTR produce credible sets of trees that are closer together in tree space, compared to trees from the JC credible set. A few trees from the JC analysis are similar to trees sampled with HKY, but generally not those sampled with GTR. The space occupied by HKY trees tends to sit in between those from JC and GTR, which suggests that an HKY analysis may support some relationships in common with JC and some in common with GTR.

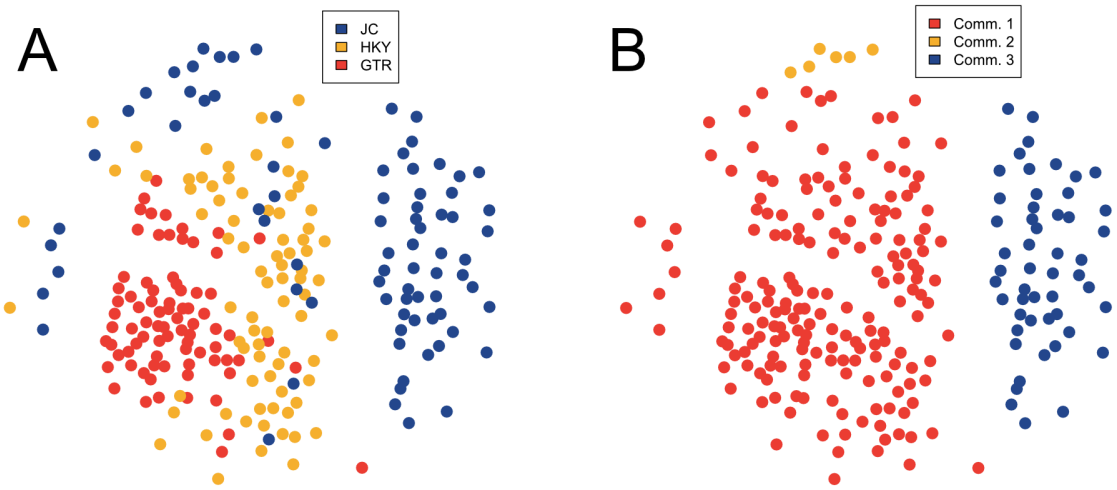


Fig. 3 – Two-dimensional visualizations from non-linear dimensionality reduction (NLDR) using weighted RF (wRF) distances of primate *cytb* trees. (A) Trees are colored by the model assumed in different analyses. (B) Trees are colored based on communities detected from a topological affinity network.

Community detection using topological affinities shows intrinsic structure with three communities. These three communities reinforce the distinction between trees sampled from JC analyses and those from GTR and HKY (Fig. 3B). One community includes all trees from HKY and GTR (and a few from JC), while the other two communities are solely trees from the JC analysis. These communities highlight the wide dispersion and structuring of the JC credible set, as well as the greater similarity between HKY and GTR results.

The results of community detection on bipartition covariance networks identifies the major topological areas of conflict between results from JC and GTR, and reinforces the intermediate position of HKY trees in tree space (Fig. 4). These networks exhibit intrinsic structuring with both two and three communities. When the two communities are mapped onto the network (Fig. 4B), several large, positive covariances are obvious within each community, as are many large, negative covariances that separate the two. Interestingly, the three-community results show how the third community is composed of bipartitions originally included in both of the two, larger communities (Fig. 4C). However, these bipartitions generally had

weak covariances (both positive and negative) connecting them to any other bipartition, and are themselves connected only by weakly positive edges (thin, blue lines). Therefore, this third community tells us less about major topological conflict in our tree set, so we focus on the two, larger communities.

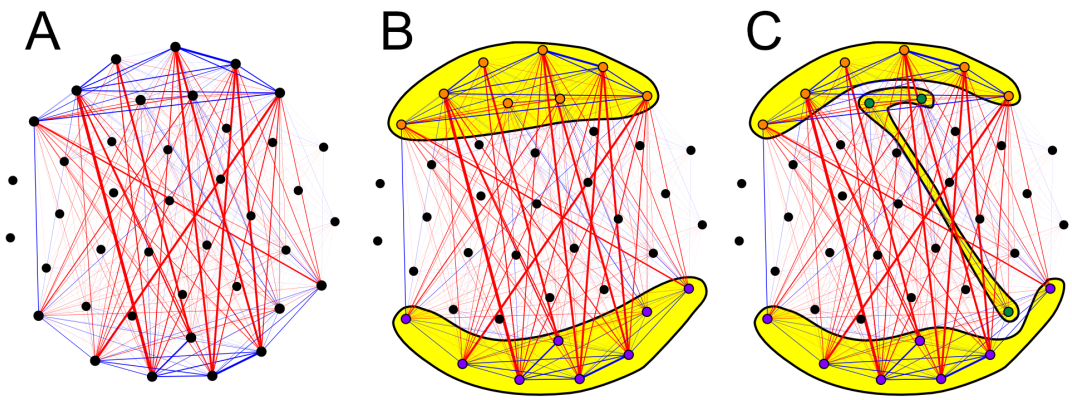


Fig. 4 – Bipartition covariance network based on trees from analyses assuming different models of sequence evolution. (A) The bipartition network with no communities labeled. (B) The bipartition network with two communities labeled. (C) The bipartition network with three communities labeled. Note that nodes in different communities are colored differently.

The nature of the conflict between tree sets from different analyses is clarified by mapping bipartitions from different communities to the maximum *a posteriori* (MAP) trees from each analysis (Fig. 5). The first community (in orange) from the two-community analysis contains bipartitions that map primarily to the GTR MAP tree. These bipartitions either strongly conflict between the GTR and JC/HKY analyses or have intermediate posterior probabilities across all three analyses. In both cases (strong conflict between analyses or consistent uncertainty across them), there are identifiable sets of trees with alternative resolutions for these clades. The second community (in purple) largely contains those bipartitions that conflict with the GTR tree and are primarily found in the JC tree. One exception is the single purple bipartition that appears in the GTR MAP tree (Fig. 5). This bipartition is found on the far left of the purple community in Fig. 4B and is connected by positive covariances to nodes in both the orange and purple communities, indicating that it does not strongly conflict with other bipartitions. The community detection methods that we currently use for analysis of these networks do not allow nodes to be assigned to more than one community or occupy intermediate positions. The use of community

detection methods that allow communities to overlap (i.e., allow for a node to belong to more than one community) may better accommodate bipartitions with weak, positive covariances to other bipartitions in multiple communities.

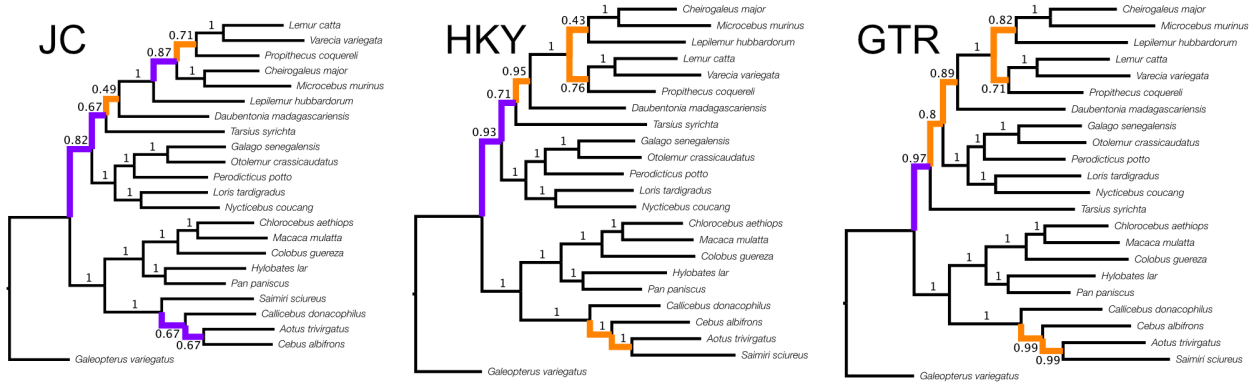


Fig. 5 – Maximum *a posteriori* trees from analyses of primate *cytb* sequences, assuming different models of sequence evolution. Branches are labeled with posterior probabilities and colored branches correspond to those bipartitions found in the two communities in Fig. 4B.

In this example, the perspective offered by visualization and community detection methods reinforces the conflict that is also possible to see by manual inspection of the posterior probabilities on the MAP trees from different analyses. However, visualization and community detection tools give a richer view of the nature and strength of the conflict. In particular, they help us to recognize how the trees preferred by HKY are essentially intermediate between JC and GTR. For analyses with more taxa, manual inspection of summary trees (MAP or consensus) can be incredibly tedious and conflict can be difficult to summarize. The tools that we highlight here can be applied at much larger scales to automatically focus attention on relevant parts of trees where conflict exists (rather than where uncertainty is consistent across analyses) and give insights into the conflict's structure. They may also give the conflict a sense of direction, by allowing us to see if each analysis occupies a completely distinct region of tree space, or if we move in consistent directions as model assumptions change.

We focused on a relatively restricted set of models here for the purpose of illustrating clearly how visualization and community detection tools work and how they can be helpful in exploring model sensitivity, but the power of these

approaches will become most apparent in analyses with larger trees and when comparing larger sets of models.

## Joint Versus Independent Inference of Gene Trees

By modeling biological processes that cause gene trees to vary, we can jointly estimate topologies and branch lengths for all gene trees, while also inferring the overarching species tree. The most common of these models is the multispecies coalescent model (MSC). The MSC describes the variation across gene trees that we would expect to occur due to stochasticity in the coalescent process. While allowing gene trees to vary, the MSC still constrains their variation in important ways. For instance, gene trees should be more similar along those branches of the species tree that had small population sizes and relatively infrequent speciation events. Conversely, gene trees should vary more when ancestral population sizes were large and speciation events were rapid. In any case, sequences from all the genes in a joint analysis provide information that influences the inferred species tree, which in turn influences inference of each gene tree.

While we know that joint inference can alter each gene tree, we might like to have some sense for how strong this effect is and whether it acts in the same way across different genes. For instance, is the space of gene trees sampled during joint inference a subset of the space that is sampled when gene trees are inferred independently? Does joint inference change the precision of gene tree estimates or does it shift the overall distribution of sampled gene trees more substantially? Can we identify distinct regions in tree space occupied by gene trees resulting from joint versus independent inference? NLDR visualizations and community detection can help shed some light on answers to these questions, which we illustrate here with Bayesian analyses of 10 genes sampled from 23 primate species.

By comparing gene trees from independent inference to those from joint inference under the MSC, we can see several interesting patterns (Fig. 6). First, in this case the tree topologies sampled during joint inference generally do represent a subset of those sampled during independent inference, and they do not seem to occupy a completely distinct region of tree space (Fig. 6, top row). This result should be comforting and suggests that the topological signal in each gene is roughly concordant with the topologies implied by the species tree. However, we can also see some interesting differences across genes. While the topologies sampled during

joint inference generally occupy a subset of the space sampled during independent inference, the distribution of topologies changes (particularly for gene 1), indicating shifts in the posterior probability of different trees between the two analyses. Also, the tree space sampled by gene 1 includes a greater number of unique topologies, compared to genes 2 and 3, for both joint and independent inference. Both of these results suggest that gene 1 contains less topological information than genes 2 and 3. (Note that the overall size of the NLDR visualization is scaled to be equal for each plot, so the sizes of the plots themselves do not give us information about the expanse of tree space that they depict.)

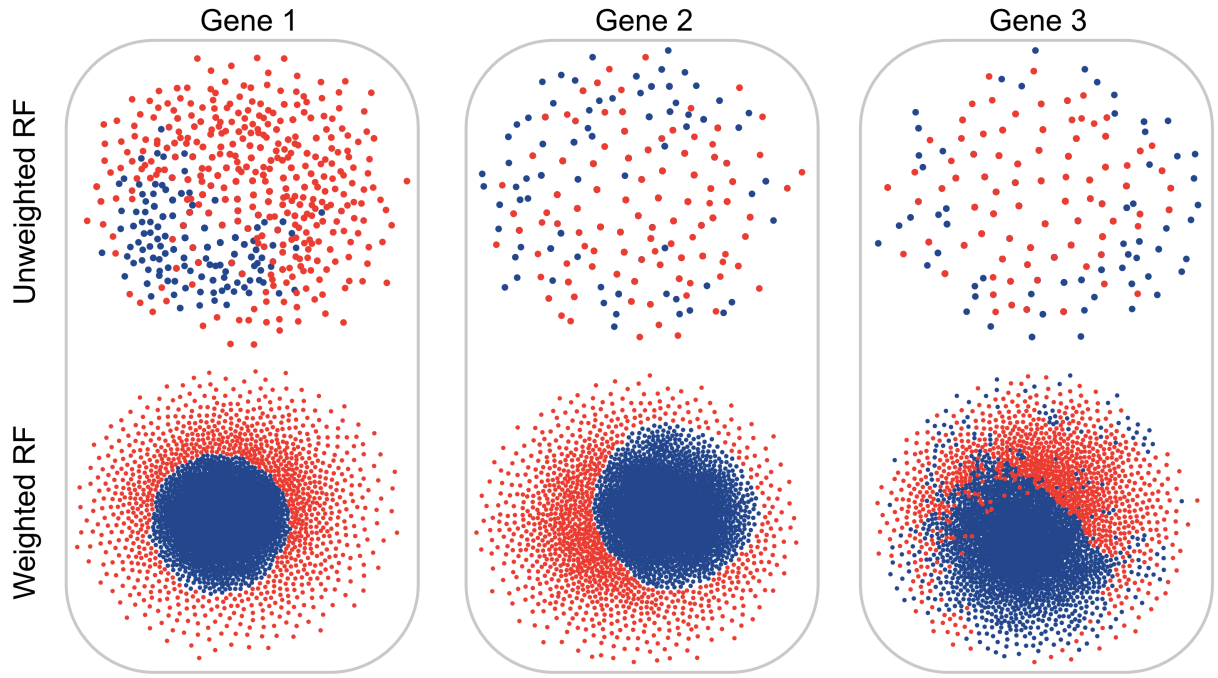


Fig. 6 – NLDR visualization of trees sampled during independent gene-tree inference (red) or joint inference under the multispecies coalescent model (MSC; blue) for three genes using two distances (weighted and unweighted RF).

Comparing projections of tree space generated using distances that do or do not take branch lengths into account also provides important information. The top row of Fig. 6 shows projected tree space based on unweighted RF distances for the three genes, where only the topology contributes to the distance. However, the bottom row shows projections based on weighted RF distances, where branch lengths are used, and these projections show much clearer distinctions between the trees sampled from joint and independent inference. The boundaries between the

two are very clear for genes 1 and 2, although the two tree sets still overlap to some degree for gene 3. The increased distinction of the boundary in these plots suggests that the joint inference is having a substantial impact on estimated divergence times, possible more than on the topology.

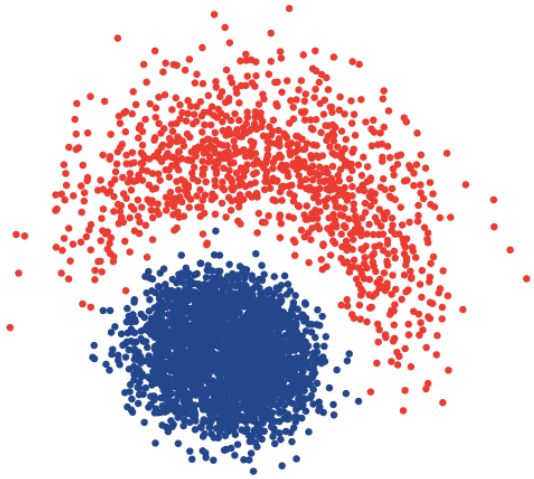

Fig. 7 – Alternate NLDR projection for Gene 1 from Fig. 6 using weighted RF distances. Kruskal-1 stress was used for this projection, while CCA stress was used in Fig. 6.

A word of caution is also warranted here about the potential overinterpretation of NLDR visualizations. As discussed above, these visualizations frequently include some distortion of the true tree-to-tree distances, and these distortions can be resolved in different ways, depending on the chosen NLDR method. While CCA stress (Demartines and Herault 1997) was optimized to generate all of the projections in Fig. 6, Kruskal-1 (Kruskal 1964) stress was optimized to generate an alternative projection for Gene 1 with weighted RF distances (Fig. 7). The CCA projection in Fig. 6 seems to suggest that the joint gene trees for Gene 1 are nested inside the space of trees sampled from the independent analysis, but the Kruskal-1 projection suggests that these spaces are more distinct. In both cases, the space occupied by trees from the independent analysis remains larger than the space occupied by those from the joint analysis, and the distinction between the independent and joint tree sets is pronounced. A conservative practice with NLDR is to try several projections and see if patterns of interest persist. In cases where one is interested in whether sets of trees are distinct in tree space, community detection with a topological affinity network has an advantage over NLDR, since it uses the

original tree-to-tree distances with no distortions from projection. Community detection on a topological affinity network can cleanly delineate the independent and joint trees as distinct communities for gene 1 (results not shown).

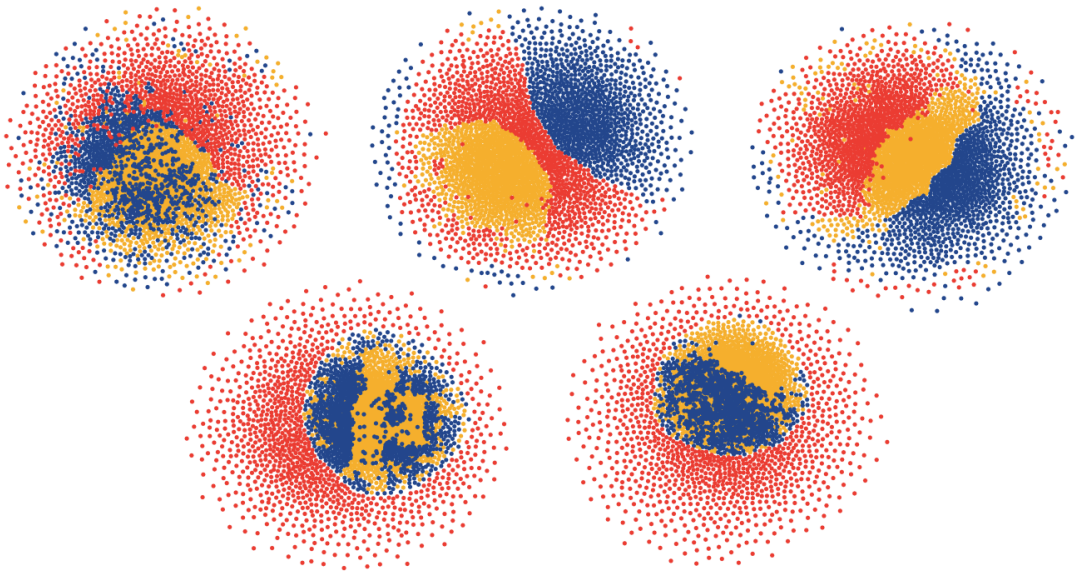

Fig. 8 – NLDR projections of gene-tree space for five different genes, based on three different analyses – independent gene tree analysis (red), replicate one of a joint MSC analysis (blue), and replicate two of a joint MSC analysis (yellow).

One significant practical challenge of using hierarchical models like the MSC in a Bayesian statistical framework is making sure that the MCMC analysis adequately samples tree space for each individual gene, as well as the species tree. Here, NLDR and community detection can also be helpful in checking for topological convergence. If replicate runs are sampling tree space in the same way, there should not be visually obvious differences in the distributions of trees from those runs and community detection methods on topological affinity networks should not be able to delineate the trees from different runs as separate communities. As an example of how these tools can help with assessing convergence, we separately colored trees from replicate MSC analyses (Fig. 8), which were intentionally stopped before convergence was reached. For several genes, the replicate MSC analyses (in blue and yellow) sampled the same general part of tree space, but clearly did not yet mix well across this space. In other cases, the two replicates are sampling completely distinct regions of tree space, perhaps because they have encountered

local optima. Such diagnostics can be helpful in rapidly identifying loci with poor topological mixing.

## Understanding Variation Across Genomes

Modern phylogenomic studies now often involve the inference of phylogenies from every gene (or genomic region) across a set of completely sequenced genomes – a remarkable amount of information! This genome-wide perspective can shed light on the interplay between diverse biological processes that cause gene histories to differ from species histories. *Heliconius* butterflies are one well-studied group where multiple biological processes and evolutionary forces (ILS, horizontal gene flow, and selection) have combined to influence the true phylogenetic histories underlying different sections of genomes (i.e., the gene trees).

Edelman et al. (2019) recently explored genome-wide patterns in gene-tree variation for *Heliconius* and related groups. After filtering for quality and combining their data with other available genomes, they analyzed genome-wide alignments from 25 taxa in order to reconstruct the relationships among diverse *Heliconius* lineages. In doing so, they found strong evidence for the influence of gene flow in explaining the historical relationships among species. Gene flow and recombination will cause different sections of the genome to support different bifurcating phylogenetic trees, and Edelman et al. (2019) were able to map the distribution of support for different trees provided by 50-kb windows spanning the entire genome. They focused especially on the *erato-sara* clade, for which they had sampled six species, since many sections of the genome support a relatively small number of topologies (two topologies, predominantly). To demonstrate how tree set visualization and community detection approaches could complement other analyses of these data, we focused on trees of the *erato-sara* clade inferred from 50-kb sliding windows across chromosomes 20 and 21.

Two- and three- dimensional visualizations of tree space based on NLDR projection with wRF distances show clear structuring for the trees sampled from these two chromosomes (Figure 9). Labeling trees by chromosome allows us to clearly see that the two chromosomes do not share the same distribution of phylogenetic histories (Figure 9A). Regions of chromosome 20 exhibit more variation in the trees they support, with trees from chromosome 21 being concentrated in a projected space roughly half the size of chromosome 20. This difference between

chromosomes 20 and 21 (or, more broadly, between chromosome 21 and all others) was also noted by Edelman et al. (2019) using tallies of different topologies across windows.

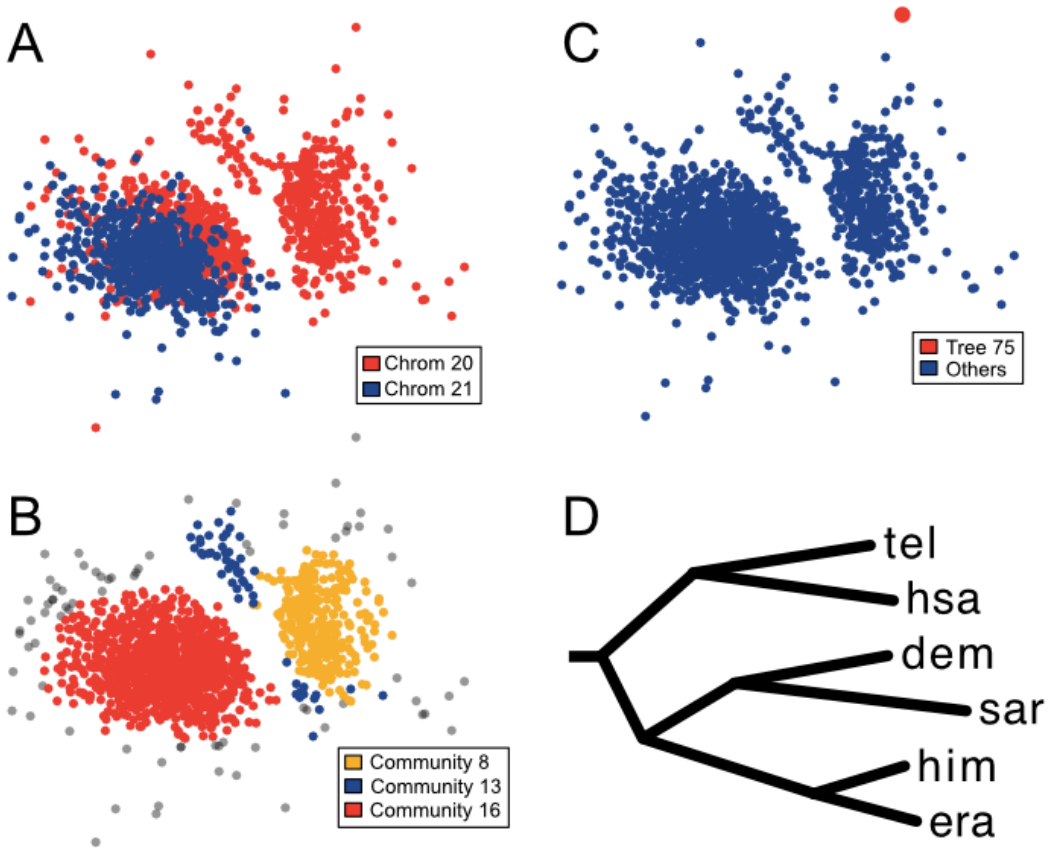


Figure 9 – NLDR visualizations of trees from chromosomes 20 and 21 of the *erato-sara* clade. (A) Trees are colored by chromosome. (B) Trees are colored based on the results of community detection on a topology network with 24 communities. The three largest communities are shown in colors, while trees from all other communities are shown in gray. (C) Community detection on a topology network with two communities, where one community only contains a single tree (tree 75 from chromosome 20). (D) Phylogram of tree 75, with the outgroup removed.

Community detection analyses based on topological networks give a more organized perspective on the variation across trees (Fig. 9B,C). Community detection (using the Constant Potts model) first showed some natural structuring with two communities. One community contained only a single tree (inferred from the 75$^{th}$ 50-

kb window on chromosome 20), while the other community contained all other trees (Fig. 9C). Further examination of tree 75 (Fig. 9D) shows that it has a topology distinct from the eight most common topologies across these genomes (Fig. 9D and 10D). As we increased the value of the Constant Potts tuning parameter, other similar trees were added to the community with tree 75 (results not shown). These trees are uncommon in the genome and come from regions that may have been subject to rare hybridization events or influenced by unique evolutionary processes.

Three large communities emerge from the topological network as the tuning parameter is increased further (Fig. 9B). While no specific number of communities is strongly preferred, the structure of these large communities remains relatively stable. Visualizing these communities (Fig. 9B) shows both how they occupy central positions in tree space with high tree density, and also how the use of networks based on original tree-to-tree distances is different than clustering in low-dimensional space. For instance, trees assigned to community 13 (Fig. 9B) are not nearest neighbors in the NLDR projection. Majority-rule consensus trees constructed from these three communities correspond to three of the most common topologies across the genome (community 8 = topology 1, community 13 = topology 5, and community 16 = topology 2; Fig. 10D), although the communities include some trees that are not identical to these topologies. Community 16 is the largest, which makes sense given the prevalence of topology 2 on chromosome 21. The fact that none of these three communities corresponds directly to tree 3 (the third most common topology in the genome and second most common topology on these two chromosomes) suggests that trees with topology 3 may have relatively short internal branches, and be close to other topologies in weighted RF tree space.

Bipartition networks show substantial structure and are dominated by two edges with large, negative weights (Fig. 10A). Natural community structure exists with two communities, although the exact composition of these communities varies depending on how stringently low- and high-frequency bipartitions are filtered (Fig. 10B,C). Less stringent filtering (removing only bipartitions present at frequencies above 0.99 or below 0.01) results in one community that includes several weakly connected bipartitions (Fig. 10B). Slightly more stringent filtering (removing bipartitions present at frequencies above 0.95 or below 0.05) results in two communities with two bipartitions each (Fig. 10C). Such filtering can be helpful, because bipartitions present at either very low or very high frequencies often have weak covariances and are difficult to assign to any community with confidence. With

either stringency, however, the two communities are separated by two large, negative edges.

The two large, negative edges correspond to the conflict between sets of bipartitions found in topology 1 and topology 3 (Fig. 10D). Topology 2, however, is composed of one bipartition from each community. This result highlights an important aspect of interpreting communities in bipartition networks. Bipartition communities may not precisely correspond to the most frequent topologies. For instance, the bipartition community in the bottom right of Fig. 10C contains bipartitions found in topology 1, while the bipartition community in the upper left contains bipartitions that are found in topology 3 (Fig. 10D). However, no bipartition community corresponds to topology 2, even though it is the most frequent topology across windows from chromosomes 20 and 21. At first, this result may seem troubling, but in fact it is informative. By mapping the bipartitions in different communities onto the most common topologies, the community detection results highlight parts of these topologies that conflict most strongly. The lack of large, positive covariances also tells us that the bipartitions in a community do not really "prefer" each other specifically. If only topologies 1 and 3 were present in the tree set, not topology 2, we would see strong, positive covariances within each of these communities. Further, the fact that the bipartitions in these communities map to these topologies reinforces that they explain most of the conflict. (One bipartition also maps to topology 4, but this seems to be only because tree 4 shares that bipartition with topology 1). While these results from bipartition networks confirm what we could learn by manually examining the frequency of different topologies, manual inspection will be much more difficult for larger trees with more unique topologies.

In aggregate, these tools allow us to visualize the variation in phylogenetic signal across the genome in a way that is agnostic about the underlying evolutionary process, rapidly identify "outlier" regions that may warrant particular attention, and summarize the phylogenetic relationships that conflict most strongly across the genome.

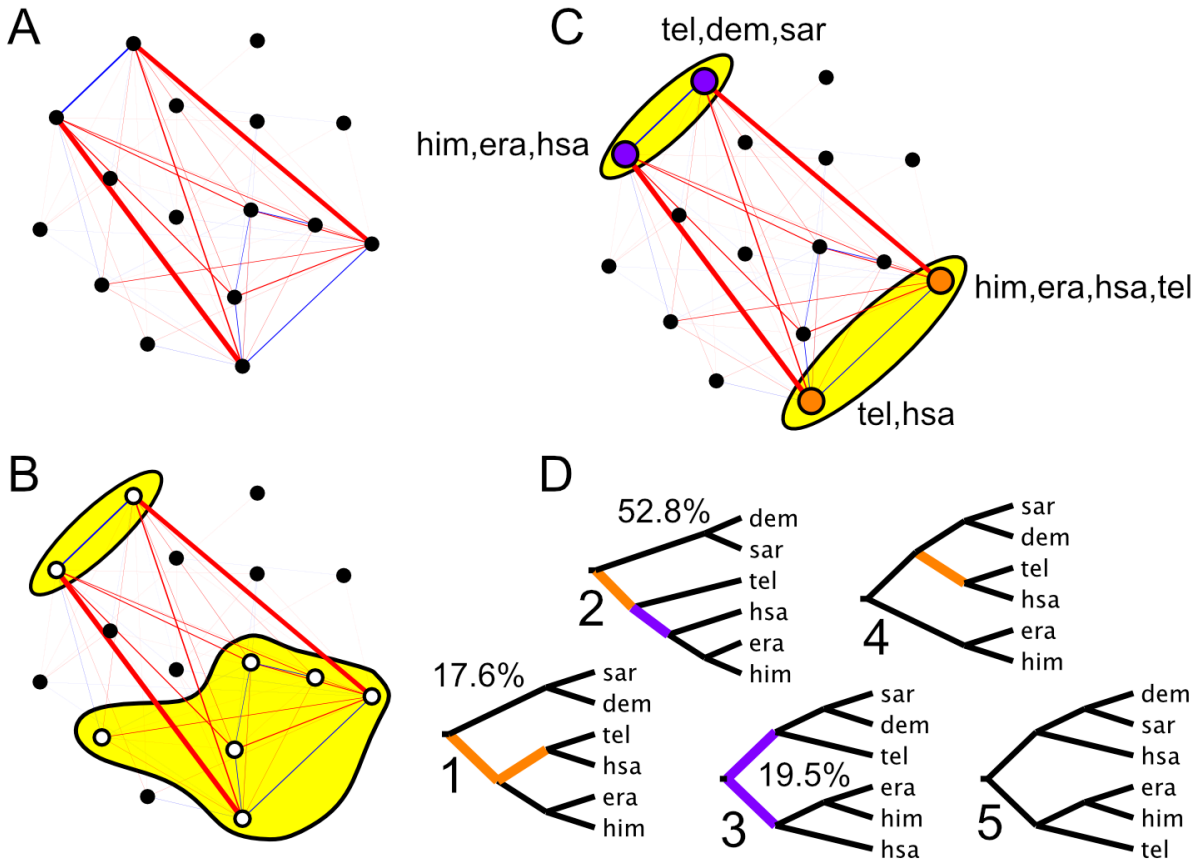


Figure 10 – Bipartition covariance networks of the *erato-sara* clade, based on trees inferred from 50-kb windows along chromosomes 20 and 21. (A) The covariance network with nodes (black circles) representing different bipartitions. Thicker red lines represent strong negative covariances between bipartitions, while thicker blue lines represent strong positive covariances. (B) The covariance network with two communities (in yellow shading) detected by community detection using bipartitions with frequencies $\geq 0.01$ and $\leq 0.99$. (C) The covariance network with two communities detected by community detection using bipartitions with frequencies $\geq 0.05$ and $\leq 0.95$. Bipartitions in these two communities are labeled with the taxa on one side of the induced split. (D) The five most common topologies observed in whole-genome analyses of the *erato-sara* clade, as identified by Edelman et al. (2019). Note that the root of these trees is defined by the position of an outgroup that has been pruned. The frequencies of the first three topologies across chromosomes 20 and 21 are shown.

## Prospects for Future Development and Application

The combination of NLDR visualization and community detection methods show promise for a wide variety of applications in phylogenomics and species-tree inference, and there is considerable room for future development. Since these tools are inherently exploratory, their most fruitful application may be in combination with software and databases that allow users to collect new data or test new hypotheses inspired by their initial explorations. The results from these analyses also need not be static. In the future, researchers could interactively explore integrated results. For example, one could slide the tuning parameter for community detection up and down and watch a network or NLDR plot to see how the community structure changes. If researchers are investigating genome-wide variation in gene trees, they could click on nodes in a topological network or points in an NLDR plot and immediately see summaries of the tree and its corresponding genomic region. They could also click on sets of bipartitions in a covariance network and have them highlighted in the corresponding topologies.

Mathematically, opportunities remain to explore approaches for community detection that incorporate properties found to be important for topological and bipartition networks in the context of phylogenetic studies (such as the ability to assign a particular node to more than one community simultaneously). Characterizing both network types for a range of biological data sets may also suggest new, tailored community detection models. Another opportunity could be to integrate additional information about the data that generated particular subsets of trees (e.g., characteristics of genes like size, function, or variability) and use it to label nodes in topological networks. These labels could then be incorporated into the process of community detection.

Biologically, the application of these approaches (particularly community detection) in phylogenetics is new and it will take time to understand what properties to expect of networks in different situations, and how different community detection models may behave depending on the nature of the tree set (Mount et al. 2020). Bipartition covariance networks, in particular, seem to carry a lot of information, but the best way to characterize and interpret the structure of these networks is an active focus of research.

## Acknowledgements

The development of tools applied here has been supported by the National Science Foundation through grants DBI-1262571 and DBI-1934156 to JMB, DBI-1262476 to JCW and KAG, DBI-1934182 to JCW, and DBI-1934157 to KAG.

# References


N Amenta and J Klingner. 2002. Case study: Visualizing sets of evolutionary trees. Proceedings of the IEEE Symposium on Information Visualization (INFOVIS 2002). pp. 71-74.

JM Brown, W Huang, M Marchand, G Zhou, GG Mount, D Morris, J Ash, KA Gallivan, and JC Wilgenbusch. 2020. Using networks to identify structure in phylogenetic tree sets. *In review*.

NB Edelman, PB Frandsen, M Miyagi, B Clavijo, J Davey, RB Dikow, G García-Accinelli, SM Van Belleghem, N Patterson, DE Neafsey, R Challis, S Kumar, GRP Moreira, C Salazar, M Chouteau, BA Counterman, R Papa, M Blaxter, RD Reed, KK Dasmahapatra, M Kronforst, M Joron, CD Jiggins, WO McMillan, F Di Palma, AJ Blumberg, J Wakeley, D Jaffe, and J Mallet. 2019. Genomic architecture and introgression shape a butterfly radiation. *Science* 366:594-599.

J Felsenstein. 2004. *Inferring Phylogenies*. Sinauer Associates, Inc: Sunderland, Massachusetts.

S Fortunato. 2010. Community detection in graphs. *Physics Reports*. 486:75-174.

M Girvan and MEJ Newman. 2002. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA.* 99:7821-7826.

DM Hillis, TA Heath, and K St. John. 2005. Analysis and visualization of tree space. *Syst. Biol.* 54:471-482.

S Höhna, MJ Landis, TA Heath, B Boussau, N Lartillot, BR Moore, JP Huelsenbeck, and F Ronquist. 2016. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Syst. Biol.* 65:726-736.

W Huang, G Zhou, M Marchand, JR Ash, D Morris, P Van Dooren, JM Brown, KA Gallivan, and JC Wilgenbusch. 2016. TreeScaper: Visualizing and extracting phylogenetic signal from sets of trees. *Mol. Biol. Evol.* 33:3314-3316.

JP Huelsenbeck and B Rannala. 2004. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. *Syst. Biol.* 53:904-913.

T Jombart, M. Kendall, J Almagro-Garcia, and C Colijn. 2017. Treespace: Statistical exploration of landscapes of phylogenetic trees. *Mol. Ecol. Res.* 17:1385-1392.

JB Kruskal. 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*. 29:1-27.

J Lee and M Verleysen. 2007. *Nonlinear Dimensionality Reduction*. New York:Springer Science+Business Media, Inc.

AR Lemmon and EC Moriarty. 2004. The importance of proper model assumption in Bayesian phylogenetics. *Syst. Biol.* 53:265-277.

GG Mount, J Ash, W Huang, M Marchand, KA Gallivan, JC Wilgenbusch, and JM Brown. 2020. The performance of network-based community detection for identifying structure in phylogenetic tree sets. *In review*.

MEJ Newman. 2010. *Networks: An Introduction*. Oxford University Press: Oxford.

V Traag, P Van Dooren, and Y Nesterov. 2011. Narrow scope for resolution-limit-free community detection. *Phys. Rev. E.* 84:016114.

DL Warren, AJ Geneva, and R Lanfear. 2017. RWTY (R We There Yet): An R package for examining convergence of Bayesian phylogenetic analyses. *Mol. Biol. Evol.* 34:1016-1020.

C Whidden, FA Matsen IV. 2015. Quantifying MCMC exploration of phylogenetic tree space. *Syst. Biol*. 64:472-491.

JC Wilgenbusch, W Huang, and KA Gallivan. 2017. Visualizing phylogenetic tree landscapes. *BMC Bioinformatics*. 18:85.

Z Yang. 2014. *Molecular Evolution: A Statistical Approach*. Oxford University Press:Oxford.

# Appendix

## Installing TreeScaper

Executable versions of TreeScaper with a graphical user interface (GUI) are available from https://github.com/TreeScaper/TreeScaper/releases for Mac and Linux operating systems, along with a manual and tutorial files. The TreeScaper code (written in C++) is available from https://github.com/TreeScaper/TreeScaper. The TreeScaper manual contains instructions for compiling versions of TreeScaper that can be run from the command line.

## CloudForest

The National Science Foundation (NSF) recently funded a new collaborative project to expand the interoperability of TreeScaper with other phylogenetic tools (DBI-1934156 to JMB, DBI-1934157 to KAG, and DBI-1934182 to JCW). The resulting cyberinfrastructure framework is called CloudForest and will be able to run on a range of computing platforms varying in size from a desktop computer, to a university-maintained high-performance computing cluster, to commercial cloud computing resources. An initial version of CloudForest intended for use with a desktop computer is planned for release in 2021. Updates will be available at https://github.com/TreeScaper/.