The effects of host phylogenetic coverage and congruence metric on Monte Carlo-based null models of phylosymbiosis

James G. DuBose[1]

https://doi.org/10.1101/2025.02.07.637028

Abstract

Variation in the structure and composition of host-associated microbial communities often parallels patterns of phylogenetic divergence between hosts. This pattern is generally known as phylosymbiosis and there is great interest in understanding how such patterns emerge from fundamental ecological and evolutionary processes. Understanding of this phenomenon relies initially on quantifying phylosymbiotic signals from across a broad range of host taxa. Quantifying signals of phylosymbiosis is typically achieved by calculating how congruent a host's phylogenetic tree is with a dendrogram that represents patterns of dissimilarity in their associated microbial communities. To statistically assess the degree of congruence, several studies have constructed null models using a Monte Carlo approach to randomly sample trees. Although this approach is becoming more common, it has several features that warrant benchmarking to advise its further use. This approach relies on quantification of congruence between a host's phylogenetic tree its microbial community dendrogram. Therefore, it is important to establish how choice of congruence metric influence null model-based inferences. Furthermore, phylosymbiotic signals may manifest at different scales of host divergence, and it is important to establish the extent of host phylogenetic breadth needed to reliably detect a phylosymbiotic signal. To help improve our study of phylosymbiosis, here I examine how power and type 1 error (false positive) rates associated with this approach varies with choice of congruence metric and host phylogenetic coverage. I generally found that model performance increased rapidly with increasing tree sizes, suggesting lower limits on the host phylogenetic breadth needed to reliably detect phylosymbiotic signals with this approach. Furthermore, I found minor but notable variations in performance between several congruence metrics, which translated into different inferences regarding signal detection. Overall, these findings suggest this approach is effective and highlight several considerations for its implementation.

Keywords: Phylosymbiosis, Monte Carlo, Tree Congruence

[1]Department of Biology, Emory University, 1510 Clifton Rd NE #2006, Atlanta, GA 30322, USA

Correspondence
james.g.dubose@gmail.com

Introduction

The structure and composition of host-associated microbial communities vary significantly across the tree of life. In many host taxa, this variation appears to be predominately explained by environmental factors, such as diet or local habitat (Sullam et al., 2012, Davenport, 2016, Martinson et al., 2017, DuBose et al., 2025). In other cases, this variation mirrors the patterns of evolutionary divergence between hosts, a phenomenon known as *phylosymbiosis* (Mazel et al., 2018, Brooks et al., 2016, Lim and Bordenstein, 2020, Kohl, 2020). Patterns of phylosymbiosis can be generated by the combination of different ecological and evolutionary processes operating with varying degrees of importance (Mazel et al., 2018, Lim and Bordenstein, 2020, Kohl, 2020). In other words, the specific processes that generated any given pattern of phylosymbiosis are likely to differ between groups (Lim and Bordenstein, 2020, Kohl, 2020, Qin et al., 2023). Therefore, understanding the diversity of processes that produce emergent phylosymbiotic patterns requires quantifications of phylosymbiotic signals from across a broad range of taxa (Kohl, 2020).

To this end, there have been several approaches proposed for statistically evaluating if patterns of microbial community dissimilarity significantly parallel patterns of their host's evolutionary relatedness. One popular approach is to transform the host phylogeny into a dissimilarity matrix and use a Mantel test to quantify its correlation with the corresponding microbial community dissimilarity matrix (Mantel, 1967). However, the Mantel test relies on linear associations between corresponding matrix elements, an assumption that may not always be satisfied. An alternative approach that has been more recently employed is to construct a dendrogram representing patterns of microbial community dissimilarity between hosts, and quantify the topological congruence between said dendrogram and the host's phylogenetic tree (Lim and Bordenstein, 2020, Lin et al., 2024). Although this approach relaxes assumptions of linear correspondence between host phylogeny and microbial community dissimilarity, it provides only a single quantification that can be difficult to interpret and use to make statistically informed inferences. In response to this issue, many have used Monte Carlo simulations to generate null models of phylosymbiotic signal (Brooks et al., 2016, Trevelline et al., 2020, Grond et al., 2020). This is accomplished by either by repeatedly sampling random trees (or shuffling tip labels, which yield nearly equivalent results (Mazel et al., 2018)) and quantifying their congruence with the host phylogeny. Using this approach, phylosymbiotic signal can be statistically assessed by calculating the probability that the congruence between the host phylogeny and random trees is at least as great as the congruence between the host phylogeny and the corresponding microbial dendrogram.

The assessment of phylosymbiotic signal using this Monte Carlo-based framework fundamentally relies on quantification of tree congruence, a task that can be approached using a variety of existing metrics. The Robinson-Foulds (RF) distance, which has historically been the most widely used metric, measures the number of splits that are present in one tree but not the other (Robinson and Foulds, 1981). Although widely used, this calculation has been criticized as too conservative because it does not account for similar but not identical splits (Smith, 2022). Therefore, a variety of Robinson-Foulds inspired approaches for addressing this issue have been developed, among the more notable of which are the information-corrected Robinson-Foulds distance (ICRF) Smith, 2020a, Nye similarity (NS) (Nye et al., 2006), and the Jaccard-Robinson-Foulds distance (JRF) (Böcker et al., 2013). The matching split distance (MSD), another popular metric, takes a different perspective on this issue by finding the matching of splits between trees

that minimize dissimilarity (Bogdanowicz and Giaro, 2012). More recently, information-theoretic approaches have been introduced to weight splits based on their information content, which has significantly improved quantification of tree congruence (Smith, 2020a). In addition to the previously described ICRF, this has led to the development of new information-theoretic tree congruence metrics, such as matching split information distance (MSID), mutual clustering information (MCI), and shared phylogenetic information (SPI) (Smith, 2020a).

The previously described metrics represent popular choices across the various fields that frequently evaluate tree congruence. However, there is reason to suspect their utility for assessing phylosymbiotic signals may vary due to how they emphasize and penalize different features of congruence. For example, the Robinson-Foulds derivatives emphasize overall tree structural similarity more than individual splits, while the information-theoretic approaches place less emphasis on uninformative differences. In other words, Robinson-Foulds derivatives will define trees as similar because most splits are close matches, and information-theoretic approaches will define trees as similar because they share the most informative splits. In practice, it is unclear if these differences translate into different inferences regarding statistical support for phylosymbiotic associations. In addition to uncertainty regarding the effects of congruence metric, how broad of host phylogenetic coverage is needed to detect a phylosymbiotic signal using the previously described Monte Carlo-based approach is also unclear. In other words, it is important to know if the attained breadth of host phylogenetic coverage is sufficient for detecting a phylosymbiotic signal, particuarly given the inherent noise associated with variation in host-associated microbial communities.

To this end, here I examined the variation in performance associated with the aforementioned null model approach when used with different extents of host phylogenetic breadth and congruence metrics. Through simulations under null and alternative hypotheses, I found that null model performance generally increased rapidly with increasing tree size. In this, good power and type 1 error (false positive) rates were typically achieved for trees of between 10-20 leaves. However, I also found several minor variations in performance between congruence metrics, which translated into their frequent disagreement regarding the significance of phylosymbiotic associations. Taken together, these findings provide useful considerations for practitioners by illustrating the conditions in which this null model approach is likely sufficient or insufficient for quantifying patterns of phylosymbiosis.

Methods

Monte Carlo algorithm for generating null models of phylosymbiosis

Let the host phylogeny be T_h and the associated microbial community dendrogram be T_c, each with l leaves. The observed congruence is then $O = C(T_h, T_c)$, where C represents a previously described congruence metric. To generate the null model, n unrooted binary trees with l leaves are randomly sampled with replacement from a uniform tree distribution (all topologies are equally likely), denoted as defined as $\{T_{n_1}, T_{n_2}, \dots, T_{n_i}\}$. The null congruence distribution is then defined as $N = \{C_1, C_2, \dots, C_n\}$, where $C_i = C(T_h, T_{n_i})$. In other words, the null congruence distribution is the congruence vector between the host tree and each of the randomly drawn trees. This distribution can then be used to test the null hypothesis $H_0 : O \leq N$, against the alternative hypothesis $H_A : O > N$. These hypotheses can then be assessed by calculating the proportion of the null distribution that is greater than or equal to the observed congruence (since

this is a one-sided test):

$$P(N \geq O) = \frac{1}{n}\sum_{i=1}^{n} I(C_i \geq O) \tag{1}$$

where I is the indicator function, equal to 1 if the inequality is satisfied and 0 otherwise. Note that for some metrics, greater congruence is associated with a lower value, in which case the direction of the inequality is reversed. A graphical representation of this processes id depicted in Figure 1. For convenience, I hereafter refer to this method as a random tree congruence test. For methodological evaluations described in this text, as well as promoting use of this approach, I have implemented this algorithm in *manticore* (Monte cArlo simulaioN algoriThm for assessIng tree COngRuencE), which is an open-source R library that provides a simple interface for its utilization. Within *manticore*, the *rtopology* function from the *ape* R library (v5.7.1) is used to sample random trees (Paradis and Schliep, 2019) and various functions from the *TreeDist* R library (v2.7.0) are used for quantifying congruence (Smith, 2020b).

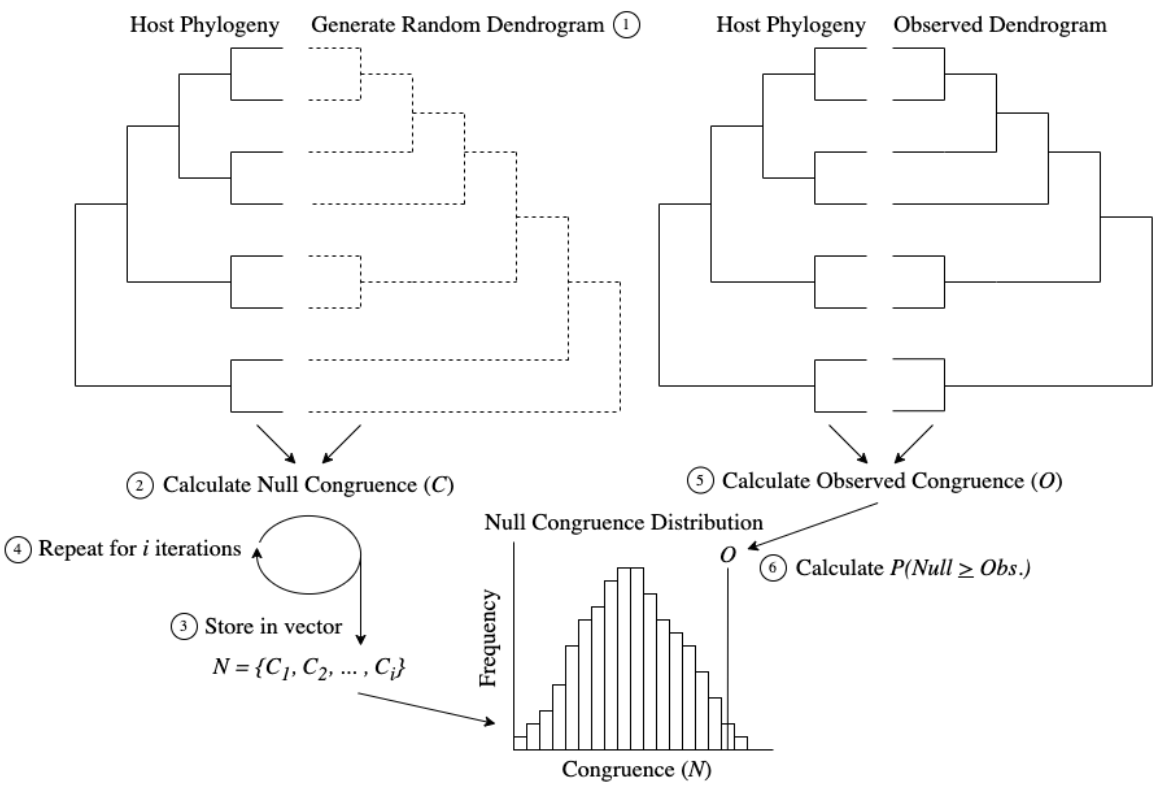

Figure 1 – A graphical depiction of the Monte Carlo algorithm for generating null congruence distributions. The processes starts by (1) sampling a random dendrogram and (2) calculating how congruent it is to the host phylogeny. This congruence value (denoted C) is stored in a vector, and (4) steps 1-3 are repeated for i iterations. The resulting vector then serves as the null distribution that will be used for hypothesis testing. (5) Calculate the observed congruence between the host phylogeny and the corresponding dendrogram (denoted O). Then (6) use the previously described null distribution to calculate the probability that randomly generated trees that showed greater than or equal to congruence to the host phylogeny relative to the observed congruence.

Simulations under H_0 and H_A

To evaluate the type 1 error (false positive) rate associated with each metric when used in random tree congruence testing, I first simulated 1000 pairs of independently randomly sampled host phylogenies and microbial community dendrograms (under H_0) for trees with 5, 10, 20, 40, 60, 80, and 100 leaves. For each pair, I conducted a random tree congruence test using each congruence metric with 1000 iterations and recorded $P(Null \geq Observed)$ values.

To evaluate the power (true positive rates) associated with each metric when used in random tree congruence testing, I simulated variation in microbial community dissimilarities using a subtree pruning and regrafting (SPR) approach to mutate host phylogenies (under H_A). Specifically, this process first generates a host phylogeny. To generate associated microbial community dendrograms, the host phylogeny is then subject to a number of subtree prunings and regraftings, which I varied to represent portions of the total leaves that ranged from 0 to 0.5, by 0.1. This process recapitulates what could be expected if factors that do not necessarily track host's phylogeny (e.g., diet, local environment) also play a role in shaping their associated microbial communities. In other words, more subtree pruning and regrafting introduces more noise in the phylosymbiotic association. For each previously described tree size and congruence metric, I conducted 50 replicate simulations, and calculated the proportion of significant tests as a function of the number of subtree prunes and regrafts. I used the *SPR* function from the *phangorn* R library (v.2.11.1) for all subtree pruning and regrafting (Schliep, 2011).

Evaluating true and false positive rates

To evaluate the power and false positive rates associated with each metric and aforementioned tree size, I used the previously described simulations to construct receiver operating characteristic (ROC) curves. ROC curves describe how well a statistical test distinguishes between the null and alternative hypotheses based on the $P(Null \geq Observed)$ values attained from both simulations. To describe performance more quantitatively, I calculated the area under each ROC curve (AUC). Here, values closer to 1 indicate that the distribution of $P(Null \geq Observed)$ attained from the H_A and H_0 simulations are more distinguishable, while values closer to 0.5 indicate they are less distinguishable. I performed all ROC curve construction and AUC calculations using the *SciPy* Python library (v.1.13.1) (Virtanen et al., 2020).

It is important to note that ROC curve construction utilized all H_A simulations, which span a gradient of noise levels. This provides a more general picture of how this test can be expected to perform in practice, but it does not provide intuition about how power may vary with increased noise, or the interaction of this function with tree size. Therefore, I also constructed curves that examine power, which is the proportion of significant results at $\alpha = 0.05$ under H_A simulations, as a function of the extent of subtree pruning and regrafting. To facilitate comparison across tree sizes, I divided the number of subtree prunes and regrafts by tree size. Similarly, I also constructed curves that examine type 1 error (false positive) rates as a function of tree size. Specifically, I calculated the proportion of simulations under H_0 that were greater than or equal to $\alpha = 0.05$, where the expected type 1 error rate for a well calibrated test should be approximately 0.05 for said α.

Results

Determining sufficient simulation effort

Due to their random nature, the number of simulations to conduct is an important feature to consider when making inferences based on Monte Carlo-based null models. Previous studies have handled this issue by performing an excessively large number of simulations. However, this approach may be too computationally demanding for some studies, particularly those that might wish to conduct scans across many sets of trees (e.g., Pollock et al., 2018). Fortunately, because Monte Carlo simulation is binomial sampling, known properties of the binomial distribution can be used to assess whether simulation efforts were sufficient. Let $p = P(Null \geq Observed)$ represent the true p-value one would obtain from exact probability calculations or infinite Monte Carlo simulations. Using finite sampling, p is estimated by conducting i replicate simulations and calculating the proportion $\hat{p}$ of simulations where the null statistic is greater than or equal to the observed. Because this is equivalent to binomial sampling, the standard error of $\hat{p}$ is:

$$\text{SE}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{i}} \tag{2}$$

This allows one to assess uncertainty in the estimated p-value without conducting excessive simulations. For example, if ε is the desired standard error threshold, the number of simulations conducted is sufficient if

$$i \geq \frac{p(1-p)}{\varepsilon^2} \tag{3}$$

Furthermore, although less commonly applied, a 95% confidence interval can be estimated as

$$\hat{p} \pm 1.96 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{i}} \tag{4}$$

Practitioners can use this framework to determine the number of simulations necessary to make confident inferences. For example, if lower bound $<$ upper bound $< \alpha$, one can confidently reject the null. Likewise, if lower bound $< \alpha <$ upper bound, the simulation is too uncertain and more simulations should be conducted. One can also use this logic to specify ε and therefore solve for the number of simulations needed. Intuitively, the margin of error should be less than the distance between $\hat{p}$ and α. This can be used to redefine Equation 3 in terms of α:

$$i \geq \frac{\hat{p}(1-\hat{p}) \cdot (1.96)^2}{(\hat{p}-\alpha)^2} \tag{5}$$

Test performance rapidly increases with tree size and shows minor variation by congruence metric

To assess how well random tree congruence testing discriminates between null and alternative simulations based on the $P(Null \geq Observed)$ attained from each, I constructed ROC curves that show the true positive rate as function of the false positive rate. Here, curves that rapidly approach a true positive rate of 1 indicate that the test is correctly discriminating true positives from false positives. Visual inspection of these curves across congruence metrics generally show this pattern is typically achieved for trees with 20 leaves or greater (Figure 2). Conversely, the

test was less able to discriminate true from false positives for trees with 10 leaves, and even less able for trees with 5 leaves. The aforementioned pattern was general across congruence metrics, with the exception of ICRF. Here, use of ICRF resulted in decreased discriminatory power from across tree sizes (Figure 2).

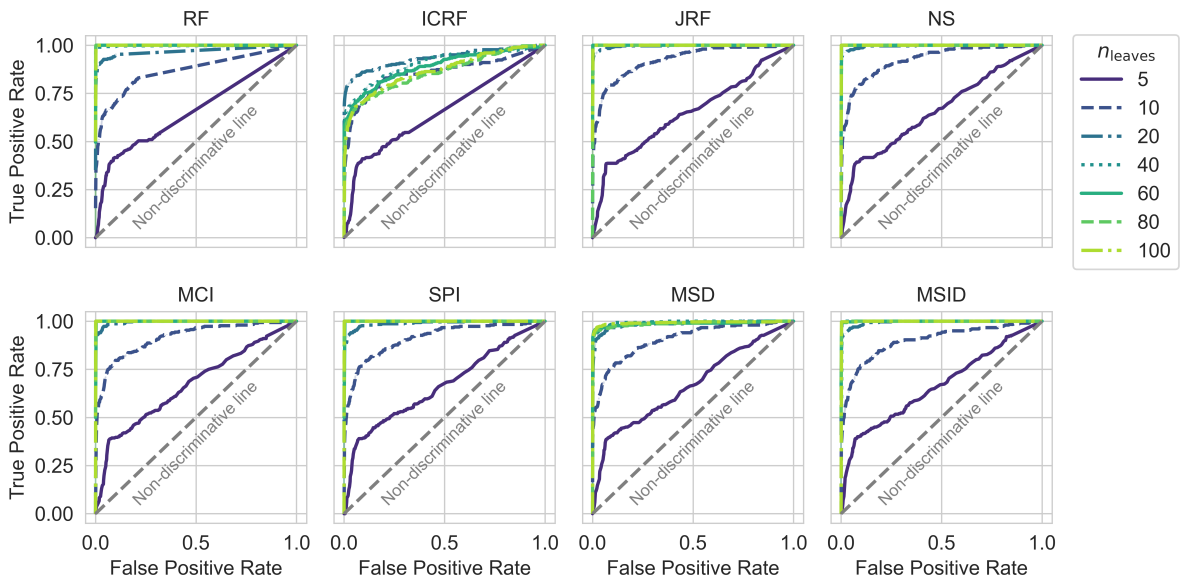

Figure 2 – Receiver operating characteristic (ROC) curves showing the random tree congruence test's ability to discriminate between true and false positives. Each panel corresponds to a different congruence metric and shows the true positive rate (y-axis) as a function of false positive rate (x-axis). Different colored and patterned lines represent the ROC curve for different tree sizes. The gray y=x line in each panel represents the non-discriminative line, which is the expected function for a test that is completely unable to discriminate between true and false positives.

To more quantitatively examine the previously described trends, I plotted the area under each ROC curve as a function of tree size. This showed that for all congruence metrics besides ICRF, values of AUC ≈ 1 were achieved for trees with 20 leaves and greater (Figure 3). Likewise, the test showed significantly reduced discriminative ability for trees with 10 leaves, which was further reduced to near the non-discriminative line for trees with 5 leaves (Figure 3). Consistent with my visual account of the ROC curves, use of ICRF resulted in decreased discriminative ability across tree sizes. Although there was a significant increase in area under ROC curve from trees with 5 leaves to 20 leaves, use of this metric never resulted in AUC ≈ 1 for any tree size (Figure 3).

Declines in power associated with increasing noise are generally recovered in larger tree sizes but vary by congruence metric

The previously described analyses considered the test's general ability to discriminate between true and false positives, but did not consider how true positive detection varies with increasing noise. To evaluate this property, I plotted power as a function of noise level. Here, power is the proportion of significant results under H_A simulations at $\alpha = 0.05$, and noise level is the number of subtree prunes and regraphs divided by the number of leaves (to facilitate cross comparison between tree sizes). Examining said curves showed several general patterns. First,

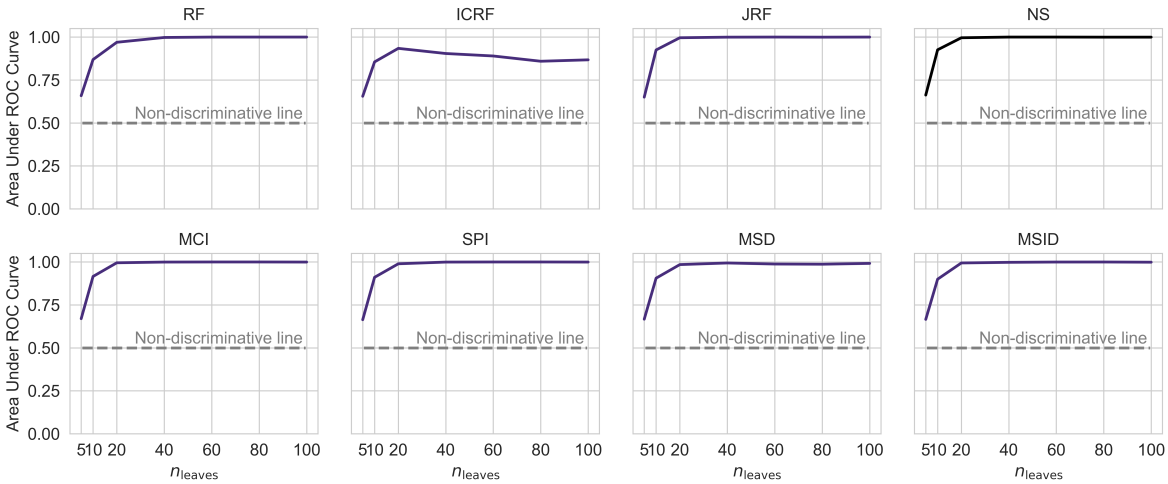

Figure 3 – Ability to discriminate between true and false positives rapidly increases with tree size. Each panel corresponds to a different congruence metric and shows the area under the ROC curve (y-axis) as a function of tree size (x-axis). Here, values closer to 1 indicate greater ability to discriminate between true and false positives, while values closer to 0.5 represent total inability to discriminate between true and false positives (denoted as the gray dashed line in each panel).

consistent with the previously described analyses, virtually no trees with only 5 leaves were detected as significant (Figure 4). However, the test did show some power to detect true positives for trees with 10 leaves at lower levels of noise. The noise level at which this curve drops below a power of 0.8 (a typical standard) was generally between 0.2 and 0.3. Use of RF and ICRF reduced this noise level to between 0.1 and 0.2 (Figure 4), indicating less power.

Increasing tree size to 20 leaves recovered a significant degree of power across congruence metrics. Notably, use of JRF, NS, MCI, SPI, and MSID resulted in acceptable degrees of power (≥ 0.8) across the full span of noise tested, although there was a decline in power at higher levels (Figure 4). Alternatively, use of MSD did confer greater power across the noise spectrum, but power levels did drop below acceptable at high levels of noise (between 0.4 and 0.5) (Figure 4). Use of RF and ICRF still resulted in significantly reduced power, which dropped below acceptable levels at noise levels of between 0.3 and 0.4 (Figure 4). Finally, further increases to tree size ($n_{\text{leaves}} \geq 40$) generally recovered acceptable power across the noise spectrum (note that although use of MSD still showed declines with higher levels of noise, power did not drop below 0.8). The power associated with use of ICRF proved to be an exception to this pattern, which generally dropped below the acceptable level of 0.8 at noise levels of ≈ 0.2 for even larger tree sizes (Figure 4).

Variation in Type 1 error rate by congruence metric and tree size

Similar to the analyses presented in previous section, it is useful to decompose ROC curves to examine if and how type 1 error (false positive) rates vary across congruence metrics and tree sizes. Therefore, I visualized the proportion of simulations under H_0 that were significant ($\alpha = 0.05$). A well calibrated test should show that said rate is approximately equal to 0.05, which is indeed the general case across congruence metrics and tree sizes (Figure 5). However, there are two notable deviations from this trend. First, the test did not reliably identify significant signals of congruence for trees with 5 leaves, regardless of congruence metric (Figure 5). This is

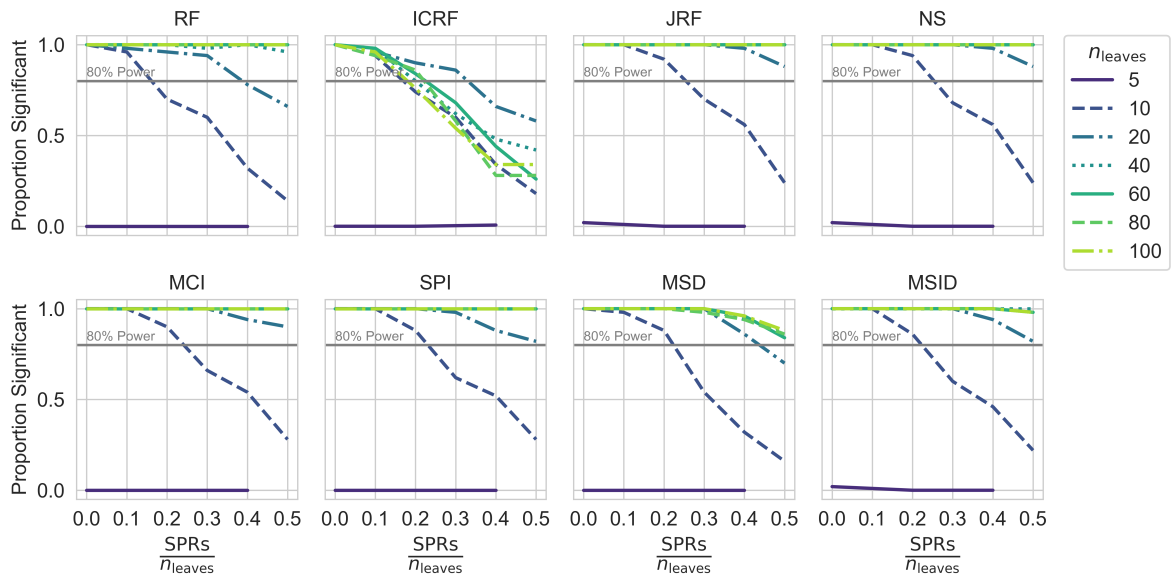

Figure 4 – Variation in the relationship between power and noise level by tree size and congruence metric. Each panel corresponds to a different congruence metric and shows the proportion of significant results under H_A simulations (power) (y-axis) as a function the number of subtree prunes and regrafts divided by the number of leaves (noise level) (x-axis). Each panel also has a gray line that depicts a power level of 0.8, which is a standard benchmark for concluding sufficient power.

unsurprising given the shapes of the ROC curves. Second, use of RF conferred significantly lower type 1 error rates than expected across tree sizes, indicating its conservative nature (Figure 5).

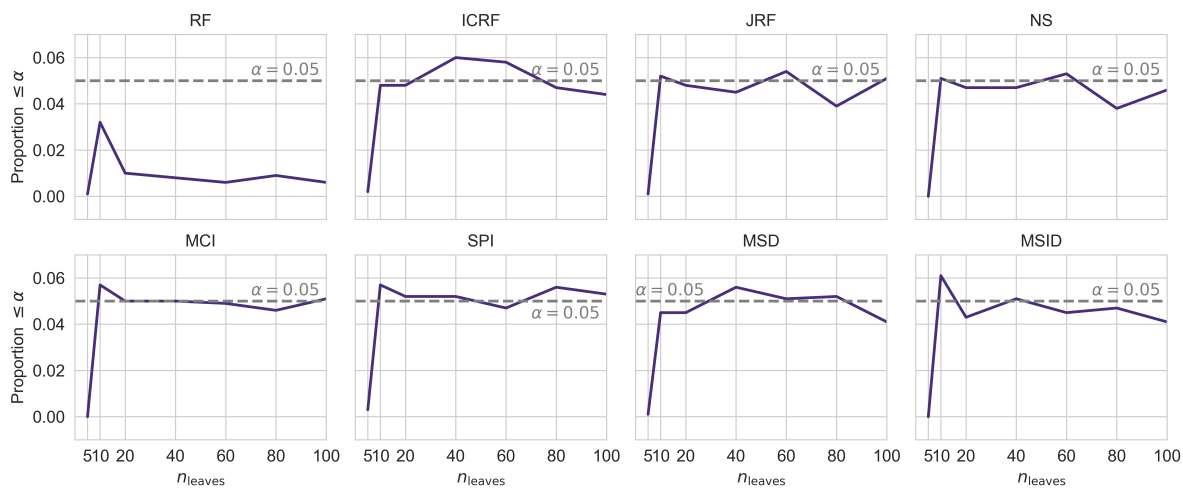

Figure 5 – Variation in type 1 error rates associated with different tree sizes and congruence metrics. Each panel corresponds to a different congruence metric and shows the proportion of simulations under H_0 that were significant at $\alpha = 0.05$ (y-axis) as a function of tree size (x-axis). The gray in each panel depicts the expected type 1 error rate for the corresponding α.

Use of different congruence metrics generally lead to the same inference regarding the detection of a phylosymbiotic signal

Now that the power and type 1 error associated with different congruence metrics have been described, it is important to ask whether or not these differences translate into meaningfully different inferences regarding whether a phylosymbiotic relationship is statistically supported. To answer this question, I used the previously described H_A and H_0 simulations to calculate the proportion of times each pair of metrics agreed on whether a phylosymbiotic relationship was statistically supported (at $\alpha = 0.05$). Overall, this showed that for both H_A (Figure 6A) and H_0 (Figure 6B) simulations different metrics tended to result in the same statistical inference. Specifically, most metrics agreed between $93 - 100\%$ of the time. Unsurprisingly, the notable exception to this was ICRF, which showed agreement approximately 80% of the time with all other metrics (Figure 6A). However, higher levels of agreement were achieved in H_0 simulations (Figure 6B), which is consistent with the previous differences in power and type 1 error rates associated with ICRF.

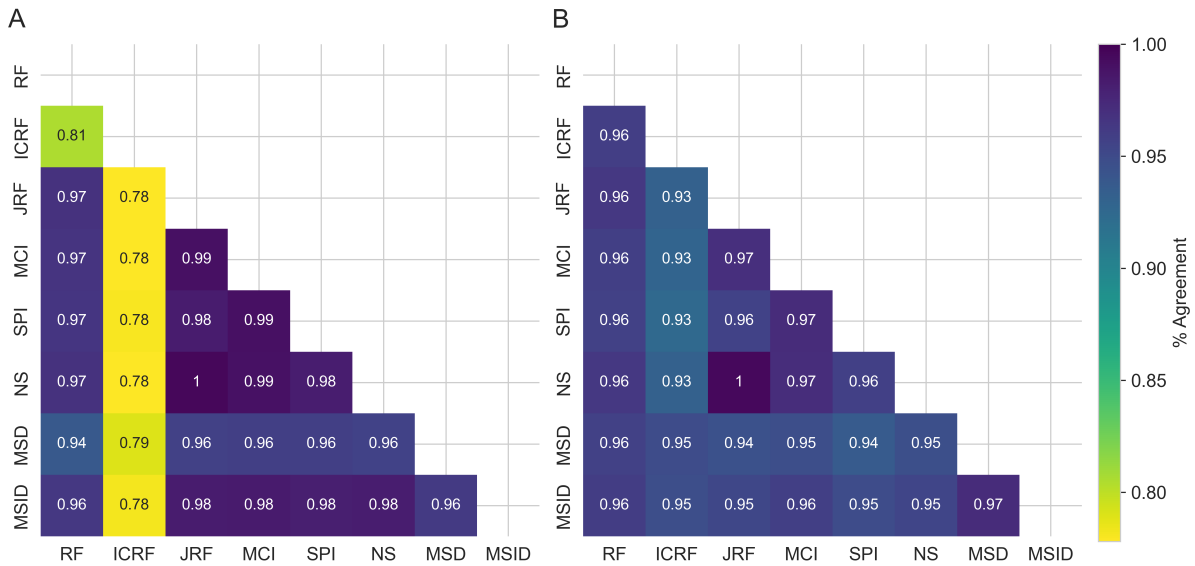

Figure 6 – Use of different congruence metrics tend to result in consistent inferences regarding the statistical support of phylosymbiotic relationships. Panels show heatmaps that depicts the proportion of simulations where the corresponding metrics agreed on the statistical support for a phylosymbiotic relationship (at $\alpha = 0.05$) across A) H_A simulations and B) H_0 simulations. The number in each cell indicates the proportion of agreements for the corresponding metrics.

Discussion

Understanding patterns of phylosymbiosis relies on our ability to detect and quantify said patterns. Here, I examined the suitability and nuances of a previously employed Monte Carlo-based null model approach for accomplishing this task (Brooks et al., 2016, Mazel et al., 2018, Trevelline et al., 2020, Grond et al., 2020). Through null and alternative model simulations, I found that this approach generally showed adequate power and type 1 error (false discovery) rates. Unsurprisingly, trees with only 5 leaves conferred virtually no power and has substantially reduced type 1 error rates. This is intuitive, as the tree space from with the null model is built only consists

of 15 unique trees for unrooted binary trees with 5 leaves. This issue was largely alleviated by increasing host phylogenetic breadth to greater than 10 leaves, which resulted resulted in good power and type 1 error (Figures 2, 3 and 5). Likewise, the extent of decline in power associated with increasing noise was largely alleviated by in trees with more than 10 leaves, though some metrics still showed decent power for trees with 10 leaves at lower levels of noise (Figure 4). I did however see an exception to these generalizations when using ICRF, which showed significant declines in power associated with greater noise across all tree sizes examined, as well as less power overall. This reduction in power translated into significantly different inferences regarding statistical support of phylosymbiotic signals relative to other metrics (Figure 6).

The performance associated with RF was somewhat surprising, as it has been criticized as too conservative (Smith, 2020a). This conservatism did show up in H_0 simulations, as use of RF conferred type 1 error rates well below expectations across all tree sizes. Despite this conservatism, use of RF still conferred degrees of power comparable to other metrics for detecting true (simulations under H_A) phylosymbiotic signals. Therefore, with the exception of ICRF, use of different congruence metrics were overall unlikely to lend different statistical inferences (Figure 6). However, there was some variation in the rate at which power declined with increasing noise for smaller trees (10-20 leaves). Therefore, it may be useful for practitioners to consult the results section describing these nuances, particularly if they find somewhat different results when using different metrics.

More generally, the majority of tree distributions and simulation approaches have been developed and evaluated in the context of phylogenetic trees and have unknown applicability to trees that describe other relationships, which in this case are microbial community dendrograms. Therefore, sampling from a uniform tree distribution via Monte Carlo simulation does not require explicitly modeling process that may not be applicable. While this allows for greater flexibility, this sampling approach is limited in that it does not directly model ecological/evolutionary processes, which would allow for evaluation of more specific hypotheses. Therefore, random tree congruence testing is best used for initial identification of a phylosymbiotic signal, and more processes-based models and experimentation should be used for further descriptions and validation (Brooks et al., 2016).

Fundings

The author declares that they have received no specific funding for this study.

Conflict of interest disclosure

The author declares no financial conflicts of interest in relation to the content of the article.

Data, script, code, and supplementary information availability

All code written for implementing the algorithm described here has been packed as an R library called *manticore*, which is available on GitHub at `https://github.com/gabe-dubose/manticore` and archived with Zenodo at `https://doi.org/10.5281/zenodo.14828441` (DuBose, 2025). All scripts written for the benchmarks and evaluations presented in this manuscript are present in the same GitHub repository but under the "evaluation" directory (`https://github.com/gabe-dubose/manticore/tree/main/evaluation`).

References

Böcker, Canzar, Klau (2013). Algorithms in Bioinformatics. In: *The generalized Robinson-Foulds metric. In: Algorithms in Bioinformatics*. Vol. LNBI 8126. Springer Berlin Heidelberg, pp. 156–169.

Bogdanowicz D, Giaro K (2012). Matching Split Distance for Unrooted Binary Phylogenetic Trees. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **9**, 150–160. https://doi.org/10.1109/TCBB.2011.48.

Brooks AW, Kohl KD, Brucker RM, Van Opstal EJ, Bordenstein SR (2016). Phylosymbiosis: Relationships and Functional Effects of Microbial Communities across Host Evolutionary History. *PLOS Biology* **14**, e2000225. https://doi.org/10.1371/journal.pbio.2000225.

Davenport ER (2016). Elucidating the role of the host genome in shaping microbiome composition. *Gut Microbes* **7**, 178–184. https://doi.org/10.1080/19490976.2016.1155022.

DuBose JG, Crook TB, Matzkin LM, Haselkorn TS (2025). The relative importance of host phylogeny and dietary convergence in shaping the bacterial communities hosted by several Sonoran Desert *Drosophila* species. *Journal of Evolutionary Biology* **38**, 180–189. https://doi.org/10.1093/jeb/voae143.

DuBose JG (2025). Data, scripts and supplementary material for "The effects of host phylogenetic coverage and congruence metric on Monte Carlo based null models of phylosymbiosis". https://doi.org/10.5281/zenodo.14828441.

Grond K, Bell KC, Demboski JR, Santos M, Sullivan JM, Hird SM (2020). No evidence for phylosymbiosis in western chipmunk species. *FEMS Microbiology Ecology* **96**, fiz182. https://doi.org/10.1093/femsec/fiz182.

Kohl KD (2020). Ecological and evolutionary mechanisms underlying patterns of phylosymbiosis in host-associated microbial communities. *Philosophical Transactions of the Royal Society B: Biological Sciences* **375**, 20190251. https://doi.org/10.1098/rstb.2019.0251.

Lim SJ, Bordenstein SR (2020). An introduction to phylosymbiosis. *Proceedings of the Royal Society B: Biological Sciences* **287**, 20192900. https://doi.org/10.1098/rspb.2019.2900.

Lin LQ, Tembrock LR, Wang L (2024). Prevalence and underlying mechanisms of phylosymbiosis in land plants. *Journal of Plant Ecology* **17**, rtae051. https://doi.org/10.1093/jpe/rtae051.

Mantel N (1967). The detection of disease clustering and generalized regression approach. *Cancer Research* **27**.

Martinson VG, Douglas AE, Jaenike J (2017). Community structure of the gut microbiota in sympatric species of wild *Drosophila*. *Ecology Letters* **20**, 629–639. https://doi.org/10.1111/ele.12761.

Mazel F, Davis KM, Loudon A, Kwong WK, Groussin M, Parfrey LW (2018). Is Host Filtering the Main Driver of Phylosymbiosis across the Tree of Life? *mSystems* **3**, e00097–18. https://doi.org/10.1128/mSystems.00097-18.

Nye TM, Liò P, Gilks WR (2006). A novel algorithm and web-based tool for comparing two alternative phylogenetic trees. *Bioinformatics* **22**, 117–119. https://doi.org/10.1093/bioinformatics/bti720.

Paradis E, Schliep K (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528. https://doi.org/10.1093/bioinformatics/bty633.

Pollock FJ, McMinds R, Smith S, Bourne DG, Willis BL, Medina M, Thurber RV, Zaneveld JR (2018). Coral-associated bacteria demonstrate phylosymbiosis and cophylogeny. *Nature Communications* **9**, 4921. https://doi.org/10.1038/s41467-018-07275-x.

Qin M, Jiang L, Qiao G, Chen J (2023). Phylosymbiosis: The Eco-Evolutionary Pattern of Insect–Symbiont Interactions. *International Journal of Molecular Sciences* **24**, 15836. https://doi.org/10.3390/ijms242115836. (Visited on 07/05/2025).

Robinson D, Foulds L (1981). Comparison of phylogenetic trees. *Mathematical Biosciences* **53**, 131–147. https://doi.org/10.1016/0025-5564(81)90043-2.

Schliep KP (2011). phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592–593.

Smith MR (2020a). Information theoretic generalized Robinson–Foulds metrics for comparing phylogenetic trees. *Bioinformatics* **36**. https://doi.org/10.1093/bioinformatics/btaa614.

Smith MR (2022). Robust Analysis of Phylogenetic Tree Space. *Systematic Biology* **71**, 1255–1270. https://doi.org/10.1093/sysbio/syab100.

Smith MR (2020b). *TreeDist: Distances between Phylogenetic Trees. R package version 2.6.2*. Publication Title: Comprehensive R Archive Network. https://doi.org/10.5281/zenodo.3528124.

Sullam KE, Essinger SD, Lozupone CA, O'Connor MP, Rosen GL, Knight R, Kilham SS, Russell JA (2012). Environmental and ecological factors that shape the gut bacterial communities of fish: a meta-analysis. *Molecular Ecology* **21**, 3363–3378. https://doi.org/10.1111/j.1365-294X.2012.05552.x.

Trevelline BK, Sosa J, Hartup BK, Kohl KD (2020). A bird's-eye view of phylosymbiosis: weak signatures of phylosymbiosis among all 15 species of cranes. *Proceedings of the Royal Society B: Biological Sciences* **287**, 20192988. https://doi.org/10.1098/rspb.2019.2988.

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, et al. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* **17**, 261–272. https://doi.org/10.1038/s41592-019-0686-2.