Title: Phylogenetic community structure metrics and null models: a review with new methods and software

Running title: Phylogenetic community structure methods

Word count: 8699 (including references, tables, figure legends and a text box)

Authors: Eliot T. Miller[1*], Damien R. Farine[2,3,4], Christopher H. Trisos[2,5]

[1]Department of Biological Sciences, University of Idaho, Moscow, ID 83844, USA; [2]Edward Grey Institute of Field Ornithology, Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS;[3]Department of Collective Behaviour, Max Planck Institute for Ornithology, 78457 Konstanz, Germany; [4]Department of Biology, University of Konstanz, 78457 Konstanz, Germany; [5] Fynbos Node, South African Environmental Observation Network, Private Bag X7, Rhodes Drive, Cape Town, 7735, South Africa

* Correspondence author. Email: eliotm@uidaho.edu; Phone: (208) 885-6280; Fax: (208) 885-7905; Address: University of Idaho, Department of Biological Sciences, 875 Perimeter MS 3051, Moscow, Idaho, 83844-3051, USA.

Summary

1. Competitive exclusion and habitat filtering are believed to have an important influence on the assembly of ecological communities, but ecologists and evolutionary biologists have not reached a consensus on how to quantify patterns that would reveal the action of these processes. No fewer than 22 phylogenetic community structure metrics and nine null models can be combined, providing 198 approaches to test for such patterns. Choosing statistically appropriate approaches is currently a daunting task.

2. First, we explored the statistical behavior of these metrics and null models, given random community assembly. This provides a baseline against which empirical results can be compared. Second, we developed spatially explicit, agent-based simulations where communities were created according to random, competitive exclusion or habitat filtering assembly rules, and then sampled from these communities to create realistic community data matrices. Third, we quantified the performance of all 198 approaches against each of the three assembly processes.

3. Our first approach showed that many metrics and null models are equivalent, which reduced to ca. 60 the number of truly unique approaches. Moreover, the second component of the analysis, our assessment of type I and II error rates, suggests that an even smaller subset are suitable for testing community assembly patterns. Based on these results, we recommend best practices for accurate detection of phylogenetic community structure.

4. We introduce a flexible new R package, *metricTester*, to facilitate robust analyses of method performance. The package is programmed in parallel to readily accommodate integration of new row-wise matrix calculations (metrics) and matrix-wise

randomizations (null models) to quickly generate expectations and quantify error rates of proposed methods.

Key-words: Phylogenetic community structure, review, phylogenetic metric, null model, habitat filtering, competitive exclusion, phylogenetic clustering, phylogenetic overdispersion, community assembly, metricTester

Introduction

The idea that competition among species increases with relatedness goes back at least to Darwin (1859), who noted that more closely related species tend to be more ecologically similar and should therefore compete more intensely (reviewed in Cavender-Bares *et al.* 2009). Referred to as the competition-relatedness hypothesis (Cahill *et al.* 2008), this competitive exclusion is predicted to result in communities composed of less closely related species (phylogenetic overdispersion) than would be expected if communities were assembled entirely via stochastic processes (Elton 1946; Webb *et al.* 2002; but see Mayfield & Levine 2010), such as speciation and dispersal. In contrast to competitive exclusion, which limits similarity of co-occurring species, habitat filtering is the process whereby only those species possessing similar traits (i.e. those within a specific subset of trait space) are able to survive and reproduce within a given abiotic environment (Harper 1977; Keddy 1992). Thus, to the extent that such traits are evolutionarily conservative, habitat filtering results in local assemblages of species more closely related than expected by chance (phylogenetic clustering; Webb 2000; Cavender-Bares *et al.* 2009). Habitat filtering operates largely independently of individual interactions. In contrast,

competitive exclusion occurs via either direct or indirect agonistic interactions among individuals of different species. Thus, while the patterns thought to be indicative of habitat filtering and competitive exclusion represent opposite ends of a gradient, and the two processes are often studied in concert, they are in fact quite dissimilar. Further, despite these processes having been proposed long ago, few methods existed to test for patterns of relatedness within communities, and those available took a taxonomic rather than a phylogenetic approach (Elton 1946; Vane-Wright *et al.* 1991).

Beginning in the early 1990s, a number of methods were developed to quantify phylogenetic patterns in community structure. The aim of these was to infer the action of community assembly processes. However, misconceptions about the relationships of these metrics to each other and to species richness (reviewed in Box 1) have reduced their impact on our understanding of community assembly. Furthermore, while the metrics introduced by Webb and others (Webb 2000; Webb *et al.* 2002) have been most influential in community ecology, other metrics have also received widespread use, and their performance across different assembly processes has not been comprehensively assessed. Recent reviews (Kraft *et al.* 2007; Kembel 2009; Vamosi *et al.* 2009; Vellend *et al.* 2011) have addressed the performance of some of these metrics, but have evaluated only partially overlapping assortments of metrics, often using different methods. Consequently, results cannot be compared among studies, making the selection of appropriate metrics for empirical research difficult.

Assessing the significance of an observed phylogenetic community structure metric requires comparing the observed data to an expectation in the absence of the process of interest. Such expectations are generally produced by a null model. Since the introduction

of null models, metrics have been linked to null models (Webb 2000), whereas in fact they are independent concepts. A null model requires a reference pool (e.g., a regional species pool, perhaps with abundance or frequency information) which is randomized with certain constraints, the details of which are defined by the null model used. These randomized values are generally used to standardize observed metrics. Thus, the metric for a particular community and phylogeny is fixed, but the significance of that metric varies according to which null model is used (Connor & Simberloff 1979; Diamond & Gilpin 1982; Gotelli 2000). A good null model randomizes those structures in the observed data (e.g., individual co-occurrence patterns) relevant to the null hypothesis, while maintaining structures in the dataset unrelated to the null hypothesis (e.g., species' abundance distributions) (Gotelli & Graves 1996). In practice, null model performance, specifically type I (false positive) and II (false negative) error rates, and redundancy among null models is rarely tested (but see Gotelli 2000).

Here, we compare the performance of 22 phylogenetic community structure metrics (Table 1) and 9 null models (Table 2). We develop spatially explicit, agent-based simulations of community assembly based on habitat filtering, competitive exclusion or the random placement of individuals, and then compare the ability (type I and II error rates) of each metric + null model combination to identify the correct assembly process. We quantify inter-correlations and document cases of equivalency among metrics and null models. We also assess the response of both the metrics and the null models to variation in species richness. We conclude by discussing the implications of our findings for future tests of community assembly processes.

Methods

Null model background

We adopt the following terminology. The community is the spatial extent (i.e. study area) of interest. A research question pertinent at this scale might be, "what assembly processes govern species composition in a rainforest community?" The quadrat is the sampling unit. For instance, 20 1-ha forest plots in the Ecuadorian Amazon would be considered 20 quadrats of this rainforest community. We refer to the quadrat by species data matrix as the community data matrix (CDM).

We test the performance of nine null models (Table 2) designed to randomize patterns in species co-occurrence data. Perhaps the simplest of these is the richness null model, which randomizes species occurrences (or abundances) within quadrats, thereby maintaining species richness (and for abundance data: total abundance and the rank-abundance curve) of each quadrat fixed. In contrast, a frequency null model randomizes occurrences within species in the CDM, which maintains species' occurrence frequencies (or abundances) but not quadrat species richness. For clarity, we refer to this null as the "frequency by quadrat" null, because in our implementation of it, metric values from randomized quadrats are grouped by the quadrat they are associated with, and these quadrat-specific randomized values are compared with those from the corresponding quadrat in the CDM. The species richness of the randomized assemblages resulting from the frequency by quadrat null approximates a normal distribution around the mean species richness in the observed CDM. Thus, this null model is likely to exhibit high type I error rates, particularly at low species richness, as the large variance anticipated of repeated small samples from a larger pool (Efron 1979) is not incorporated in the

expectation, and observed low species richness quadrats tend to be compared to randomized quadrats of the mean species richness. To account for this, Miller *et al.* (2013, Appendix 3 of that paper) developed the "frequency by richness" null model, wherein randomized quadrats are grouped by their species richness values, thereby maintaining both species richness and species' occurrence frequency data structures in the null model. The "independent swap" null model also maintains these same two data structures (Gotelli 2000; Gotelli & Entsminger 2001), but we directly test that null here to confirm that it and the "frequency concatenated by richness" model perform similarly. Similarly, we also test the "trial swap" (Miklós & Podani 2004) and 1s (Hardy 2008) null models, which are functionally equivalent (i.e. they converge on the same expectations) to the independent swap and richness null models, respectively (see Appendix S1 in Supporting Information). We used 10^5 swaps per matrix for these swap algorithms (CHRIS WHAT IS THE CITATION HERE AGAIN? I FORGOT).

Prior to the development of abundance-weighted metrics, few null models intentionally maintained aspects of abundance distributions. For instance, a species might occur infrequently, but have high abundance when it is present. Hardy (2008) introduced the "2x" and "3x" null models to maintain both species richness and occurrence frequency, as well as either the species or quadrat-level structure of abundance data. The 2x maintains the total abundance and rank-abundance curve of each quadrat, but neither species' abundances nor the set of species-specific abundance distributions. By contrast, the 3x maintains species' abundances and the set of species-specific abundance distributions, but not the abundance distributions of each quadrat. No null model that we know of maintains species richness, species occurrence frequency, species-specific and

quadrat-specific abundance distributions in the same model. We developed (Appendix S1) and tested a model that approximates this behavior, which we call the "regional null". It is meant to simulate a fixed propagule pressure on a local community, where local dynamics have no influence on the regional pool. Instead of using observed species abundance and occurrence frequencies from the community (i.e. study area) of interest, information from a larger, regional pool is used to generate a null expectation; species' colonization probabilities are proportional to regional abundances.

metricTester

We wrote an R software package to run all of the analyses presented in this paper. This package is available from Github, along with associated documentation, and can be directly installed using the *devtools* package (*metricTester*, user name "eliotmiller"). *metricTester* interfaces with functions from the R packages *picante* (Kembel *et al.* 2010), *ape* (Paradis *et al.* 2004), *vegan* (Oksanen *et al.* 2013), *geiger* (Harmon *et al.* 2008), and *spacodiR* (Eastman *et al.* 2011), among others. It also depends on *ecoPD* (Cadotte *et al.* 2010). To simplify conflicts with *picante* we renamed some of the functions in *ecoPD* and rebuilt the package, hosted under the name *ecoPDcorr* in the same Github account. *metricTester* is programmed to run in parallel (on multicore computers) and designed in a manner to easily integrate new metrics, null models and community simulations. Thus, the performance of proposed metrics and null models can be tested against community simulations of the user's choice. Generation of such expectations is not limited to phylogenetic community structure methods, and extends to any row-wise metric calculation with repeated matrix-wise null model randomization.

General behavior of the metrics

To understand the behavior of the 19 focal metrics (Table 1) across variation in species richness we generated a phylogenetic tree that terminated at 50 species using a pure-birth model (birth=0.1), then assembled a CDM that included one "quadrat" at every species richness value between 10 and 40 species (thus, each CDM had 31 rows). We use the term quadrat loosely here, but in keeping with terminology throughout the paper (see *Null model background*). Specifically, we refer to a CDM row with no spatial association. These quadrats were created by randomly sampling from the tips of the phylogeny, and assigning selected species abundances from a log-normal distribution (mean = 3, SD = 1). For each simulated CDM, we calculated the focal metrics for each quadrat, and retained those values. Using the same tree, we repeated the process of filling a new CDM, and using it to calculate and retain all metric values 50,000 times. We then calculated the mean and 95% confidence intervals (CIs) at every sampled species richness value, and plotted these across their respective species richness values.

We performed a Pearson correlation on the retained results to examine intercorrelations among metrics. Because of the large number of simulations, some metrics that appear exactly correlated do in fact differ subtly (Appendix S2). We used these correlations to generate a dendrogram (Fig. 1B) and better visualize relationships among metrics.

General behavior of the null models

We explored the behavior of 9 null models (Table 2) across variation in species richness. We used MPD for this, since null model expectations (CIs) of phylogenetic structure converged relatively quickly (exhibited less stochasticity) for this metric, and MPD is not inherently correlated with species richness (Fig. 1A). Using an abundance-weighted metric did not affect results (not shown). We also explored how expectations changed with increasing numbers of randomizations (Appendix S1). We did this by plotting the expected CIs across the corresponding species richness while increasing the randomization of a given, initial CDM and phylogeny. In sum, this set of analyses identified null models that do or do not converge efficiently on a stable range of expected metric values, and identified functional equivalence among the null models.

Agent-based spatial simulations of community assembly to assess the performance of metric + null combinations

The first two sets of analyses illustrated the underlying behavior of each of the focal metrics and null models. In this third analysis, we assessed the ability of each metric + null model combination to detect a given assembly process. Because of the large number of steps in this analysis, we include a schematic to aid the following explanation. (Appendix S3). Because of the total computing time required to run these tests (>>7 years), we did not systematically examine sensitivity to simulation parameters, but results were very similar across a range of the available parameter space (Appendix S4).

To generate test cases against which to assess each metric + null approach, we created CDMs with three types of spatially explicit community assembly simulations, intended to model the extremes of habitat filtering, competitive exclusion and random assembly.

Each spatial simulation produced a 316 x 316 m (10 ha) community, and 1,009 such communities of each type were generated. We began by generating a phylogeny of 100 species using a pure-birth model (birth = 0.1) and log-normal rank abundance curve, and randomly assigned species abundances from this distribution. We expanded assigned abundances to create a vector of individuals with species identities. In the random assembly spatial simulation, these individuals were then randomly placed within the community.

In habitat filtering simulations, we independently evolved two traits according to a Brownian motion evolutionary process (sigma = 0.1). These traits are meant to mimic two independently evolving environmental preferences, e.g., soil moisture and pH. In our case, we treated these as spatial preferences (i.e. x and y-axis preferences), and scaled the simulated traits to match community bounds. We further smoothed species' spatial preferences, which initially approximated a normal distribution, to a uniform distribution, such that species' preferences were evenly distributed but phylogenetically conserved across the arena. We then placed individuals near their spatial preference, with a controllable degree of variation (exact parameters in Appendix S4). This simulation has the effect of placing related individuals near each other in space.

In competitive exclusion simulations, we first placed individuals using the random assembly process. Following this, each generation, we calculated the mean relatedness of every individual in the simulation to all individuals within 20 m, which we term the "interaction distance". We then identified the 20% of individuals with the highest mean relatedness. For each of these individuals, we identified the individual within their interaction distance to which they were most closely related. We randomly selected one

of the two individuals to remove from the community. At the end of each generation, the same number of individuals as was removed was drawn from the original vector of individuals, and situated randomly in the community. This was repeated for 60 generations for each competitive exclusion simulation. Preliminary analyses indicated that results were similar across different interaction distances and percentages of individuals considered (Appendix S4). All spatial simulations employed 200-400 individuals/ha, which is somewhat less than stem-density in Australian tropical rain forests (Murphy *et al.* 2013) and notably less than those in Ecuador (Valencia *et al.* 2004); results were nearly identical, however, when we performed the same analysis with comparable stem density (Appendix S4).

In all spatial simulations, after a given community was assembled, we randomly placed 20, non-overlapping quadrats of 31.6 x 31.6 m (0.1 ha) within its confines. We recorded the individuals in each quadrat to create a CDM, and calculated observed metrics. To assess significance of these observed metrics, each CDM was randomized 1,000 times according to each null model. Each randomization, we calculated and retained all metrics, which we used to calculate standardized metric scores (SES, e.g., standardized MPD equals NRI, Box 1) and to construct 95% CIs. We did this both concatenating the randomized scores by species richness and by quadrat (Appendix S3). Because results were similar with both approaches, we generally report results for the quadrat concatenation in the main text, with results from the richness method in Appendix S6. As this is the distinguishing feature between the frequency by richness and frequency by quadrat null models, however, we report these separately in the main text.

The regional null is designed to be concatenated by richness, so quadrat method results for this model were discarded.

We used the set of SES (all quadrats and/or all unique species richness) for a given metric, null model and spatial simulation to perform a Wilcoxon signed-rank test. A type I error was defined as a set of SES from the random spatial simulation differing significantly from zero (two-sided test). A type II error was defined as a set of SES from either the filtering or competition simulations not being significantly less or more than zero, respectively (one-sided test). Thus, the overall type I error rate for a given metric + null model approach is the percent of the 1,009 random spatial simulations where the set of SES differed from zero, and the overall type II error rate for a given approach is the mean percent of the filtering and competition simulations where the SES did not differ as expected from zero. The use of confidence intervals to quantify error rates is discussed in Appendix S6.

Results

General behavior of the metrics

We directly evaluated behavior of 19 focal community phylogenetic metrics (Table 1) across variation in species richness. MPD, interspecific AW MPD, PSV and PAE were not correlated with species richness (Fig. 1A). Intraspecific AW MPD, complete AW MPD, PSE, IAC, H_{AED}, H_{ED}, SimpsonsPhy, PD, PD_c, and QE were positively correlated with species richness. MNTD, AW MNTD, PSC, and E_{ED} were negatively correlated with species richness. The intercorrelations (Fig. 1B, Appendix S5) among metrics and post-hoc plotting of absolute metric values against each other revealed that: (1) MPD is

equivalent to PSV; (2) complete AW MPD is equivalent to SimpsonsPhy and QE, and approximately equal to intraspecific AW MPD (Appendix S2) and to PSE; and (3) PSC is equivalent to MNTD. Moreover, MPD, interspecific AW MPD, and intraspecific AW MPD are equivalent to $\Delta+$, $\Delta*$, and Δ, respectively, of Clarke & Warwick (1998) (Box 1, Appendix S2). Based on these intercorrelations (Fig. 1B), we classify the metrics into the following groups: Clade 1 are "total community relatedness" metrics; Clade 2 metrics focus on the relationship between "evolutionary distinctiveness and abundance" (Cadotte *et al.* 2010); Clade 3 are "nearest-relative" metrics; and Clade 4 metrics are closely correlated with species richness, and increase both with the addition of new species, and phylogenetically unique spfecies.

General behavior of the null models

The CIs from the richness null model matched statistical expectations (Fig. 2), with more variance observed at smaller subsamples of the regional species pool (i.e. a confidence funnel; Clarke & Warwick 1998). The 1s and richness null models were equivalent (i.e. converged on similar expectations, Fig. S1.1). We found (Fig. S1.4) that the frequency by richness null was equivalent to the independent swap null. Similarly, the trial swap null seemed to converge slowly (i.e. after $>10^6$ randomizations) on the same expectations as these two nulls (Fig. S1.2). The CIs of the frequency by quadrat null model did not form a confidence funnel. Instead, the value beyond which an observed metric needed to deviate to be considered significant was approximately the same for all quadrats, irrespective of underlying species richness of the quadrat (Fig. 2). We also found that the expectations when concatenated by richness from the 2x and 3x null

models were equivalent, but did not form a confidence funnel (Fig. 2, Fig. S1.5). Finally, expectations for the independent swap varied depending upon relationships between occurrence frequency and phylogenetic uniqueness. For instance, if phylogenetically unique species occurred more frequently in the input CDM, then CIs were shifted upwards from those obtained without incorporating occurrence frequency (Fig. S1.6).

Performance of metric + null approaches

We ran 1,009 complete tests (all spatial simulations, null models and metrics). On an 8-core computer, each complete test takes approximately 8 hours. Thus, to finish the simulations we ran analyses on four different computing clusters and also employed spot instances from Amazon Web Services (http://aws.amazon.com).

There was a great deal of variation in performance of different metric + null approaches. Across all metrics for both competitive exclusion and habitat filtering assembly simulations, the frequency by quadrat null showed high rates of type II error, particularly for metrics that were correlated with species richness (e.g., MNTD, PD). The 2x and 3x nulls showed high type II error rates, particularly for the detection of habitat filtering and for metrics tailored to be sensitive to differences in abundance distributions (Cadotte *et al.* 2010). The independent swap, trial swap and frequency by richness null models performed reasonably well in habitat filtering simulations when used with some metrics (e.g., PD and MPD, Fig. 3), but poorly in competitive exclusion simulations with most metrics (Fig. 4). Finally, the richness, 1s and regional nulls performed well with most metrics in both the habitat filtering and competitive exclusion simulations, but the richness and 1s exhibited high type I error rates.

Given a community assembled according to habitat filtering, PD, PD_c, MNTD and AW MNTD outperformed other metrics, though Clade 1 metrics also performed well (Fig. 3). Given a community assembled according to competitive exclusion, PD and PD_c were also well-suited to detecting overdispersion (Fig. 4), though here they were outperformed by Clade 1 metrics. Clade 3 metrics exhibited elevated type II error rates. If we take overall metric performance as the mean of the type I error rates across all null models for the random simulations, and the type II error rates across all null models for the habitat filtering and competitive exclusion simulations, then Clade 1 metric performed best overall, followed closely by PD and PD_c, and then by Clade 3 metrics (Fig. 5). Some metrics (E_{AED}, PAE, IAC, H_{AED}) exhibited type I error rates on par with those of the more successful metrics (i.e. 10-11%), but also failed more often than they succeeded to detect simulated community assembly processes.

Discussion

The unification of phylogenetic community structure methods with age-old questions of community assembly has revolutionized the fields of ecology and evolution. Since Webb's seminal papers (Webb 2000; Webb *et al.* 2002), there has been an explosion of interest in these matters, including a wide variety of "improvements" upon existing measures (Box 1). Many of these, however, have never been adequately tested, and others are equivalent, as we show here (Fig. 1B). Our objective was to assess a wide range of available methods in order to identify those with demonstrable utility, and to identify those that measure unique aspects of phylogenetic community structure.

Which metrics are best? The results of our study suggest that the answer depends in part on which community assembly process are of interest, and which null models are used. However, some clear and general answers did emerge. Across most null models and all community assembly simulations, PD (Faith 1992) consistently performed well (Fig. 5), showing low type I error rates and more power than most other metrics; it was particularly good at detecting the effects of habitat filtering (Fig. 3). Clade 1 ("total relatedness") metrics (Fig. 1) also performed well, particularly at detecting effects of competitive exclusion (Fig. 4). Like Kembel (2009), and unlike Kraft *et al.* (2007), we found that Clade 3 ("nearest-relative") metrics were not as powerful as Clade 1 metrics at detecting competitive exclusion, though we did not directly probe changes in community size as did Kraft *et al*. Instead, we found that Clade 3 metrics were slightly better-suited to detecting habitat filtering.

In contrast to previous speculation on the subject (Miller *et al.* 2013), and to our results from a previous version of this manuscript, abundance-weighted forms of Clade 1 metrics and MNTD showed slightly higher type I error rates than non-abundance-weighted forms. We presumed that because the non-abundance-weighted metrics can be strongly influenced by the presence or absence of a single individual, randomly assembled communities would more frequently exhibit type I errors. We encourage additional exploration of the circumstances under which abundance-weighted versions of these metrics throw type I errors. We emphasize, however, that these differences in error rates among the Clade 1 metrics were small, and they performed comparatively well.

The metrics introduced by Cadotte *et al.* (2010) generally showed poor performance, particularly PAE and H_{AED} (PD_c is an exception, but see Box 1. Also, E_{ED} performed best

of the Clade 3 metrics). As suggested by Cadotte *et al.* (2010), the metrics do indeed measure unique aspects of phylogenetic community structure (Fig. 1B). These aspects, however, do not seem to be related to traditionally recognized community assembly processes. When used with the regional null, IAC did show strong power to detect non-random patterns; this node-based metric does not incorporate branch length information. H_{ED} was closely correlated with PD (r = 0.94), but did not perform as well as that metric.

Which null models are best? Again, our results suggest that the answer depends in part on the choice of metric and the community assembly process of interest. In general, we do not recommend use of a frequency by quadrat null. The CIs for this null model account for neither the increased variance in expectations at smaller subsamples of the regional species pool (Clarke & Warwick 1998), nor the correlation of many metrics with species richness (Fig. 1). Under certain parameters (e.g., low quadrat species richness as compared with the regional pool), this is expected to result in high rates of type I errors. Early versions of this manuscript did indeed find this to be true, particularly for metrics that were correlated with species richness. While our results do not as strongly indicate these pitfalls, we suggest this null should be used with prudence.

The 2x and 3x null models showed mixed performance. While they exhibited fairly low type I error rates (Hardy 2008), they also exhibited limited power to detect expected phylogenetic signals. When these nulls are concatenated by richness, they exhibit further elevated type II error rates (Appendix S6). We suspect that the extreme constraints imposed on the matrix randomizations by these nulls results in biased exploration of reasonable phylogenetic space; when randomizations from different quadrats of the same species richness are combined, it results in wide CIs. Regardless of the reason for this

lack of power, the instability across species richness shown by the CIs for the 2x and 3x null models (Fig. 2) means that the expectations for a given metric can change dramatically based on whether N or N+1 species are present in an observed community. Nevertheless, these null models are intended to be concatenated by quadrat, and when used in this manner, they performed better than all but the regional null.

The regional null (Appendix S1) was designed to simulate propagule pressure/dispersal probability on a local community (study area) of interest, such that deviations from these dispersal pressures (e.g., the product of environmental filters) can be readily detected, and local community dynamics (e.g., competition) do not obfuscate expectations. For instance, given strong competitive exclusion, local communities may show widespread phylogenetic overdispersion, where certain species are generally excluded. When these observed occurrence frequencies are taken as regional occurrence frequencies and randomized accordingly (as in the independent swap), it becomes difficult to detect phylogenetic overdispersion, since the randomized CDMs will tend to contain distantly related species. The regional null avoids this issue by using expectations from a larger, fixed pool as the standard against which to compare observations from the study area. However, it is difficult to quantify dispersal pressure on a community of interest, and this null model may not be practical for many researchers. Future studies should investigate what information might be used to construct these expectations (e.g., range sizes), and whether this null can be of widespread utility.

We emphasize that null model choice cannot be driven entirely by statistical properties. There may be sound biological reasons for why a given null should be employed (Gotelli & Graves 1996), even if its statistical performance is not on par with

others. For example, there could be instances where not every species in the pool could reasonably disperse to every site, and a constrained null model might need to be developed. However, such reasoning should not come at the expense of statistical common sense. For instance, if a phylogenetically unique species occurs only infrequently in observed communities, then a null such as the independent swap that maintains species' occurrence frequencies should be used; failure to do so would result in a loss of power to detect phylogenetic overdispersion. Conversely, if a CDM is not thought to be representative of a regional species pool (e.g., biased sampling across study areas), then the independent swap will only confuse interpretation of results.

What approach do we suggest? The richness null may offer the simplest results to interpret by making the clearest assumptions (any species can occur anywhere); more constrained null models raise questions of sampling artifacts and the efficiency of swap algorithms. We emphasize that little should be made of the deviation of any single community beyond null model expectations; the high type I error rates of most approaches (particularly when using the richness null) casts doubt on the interpretation of single community tests. When multiple communities are available, these can be arranged along an environmental gradient to test hypotheses. Here, the slope of the overall relationship is of interest, rather than the significance of any given community (Miller *et al.* 2013). Hypothesis testing in this manner minimizes the necessity of a null model and, if the metrics in question are not correlated with species richness (e.g., PSV), also the need to standardize the metrics. Raw metric values, which often have intrinsic meaning, can then be used instead of standardized scores. For instance, the MPD of a community, given a time-calibrated phylogeny, is equal to the mean evolutionary time separating co-

occurring taxa. Some metrics, however, are correlated with species richness, and should be standardized if the researcher is interested in phylogenetic community structure (as opposed to, e.g., phylogenetic diversity itself). In short, researchers need to consider what they are measuring with their metric(s) of choice, whether they need to standardize those metrics, and why or why not they might procure significant results.

By making the assumption that the traits responsible for community assembly covary with phylogeny, this study maintains the sometimes questionable dogma that habitat filtering leads to phylogenetic clustering, and that competitive exclusion leads to phylogenetic overdispersion (Webb *et al.* 2002; Mayfield & Levine 2010). If trait data are available, we encourage researchers who use these methods to fit explicit models of evolution to traits pertinent to the assembly processes in question (Butler & King 2004), and to also investigate patterns of community structure in functional traits. Whilst we did not test approaches that account for variation among quadrats in species co-occurrence probabilities (e.g., Cavender-Bares *et al.* 2004; Hardy & Senterre 2007), *metricTester* could be adapted to investigate these metrics. There is also an expansive assortment of existing (and yet to be created), hypothetically useful null models whose behavior and performance remains to be tested (e.g., Ulrich & Gotelli 2010). Ultimately, advanced approaches (Ives & Helmus 2011) could prove more powerful and gain wider use than current phylogenetic community structure metrics, but the existing arsenal remains well suited to addressing a wide variety of questions.

Acknowledgements

We thank Vincenzo Ellis, Amy Zanne and the Ricklefs lab for input and feedback, and the Oslo Bioportal, the University of Missouri Lewis Cluster, and the Domino Data Lab for providing computing resources. DRF was funded by a BBSRC (BB/ L006081/1) grant to Ben Sheldon.

Data accessibility

metricTester is available from GitHub (https://github.com/eliotmiller/metricTester), and can be directly installed into an active R session using the *devtools* package. It requires the package *ecoPDcorr*, which can also be directly installed with *devtools* (https://github.com/eliotmiller/ecoPDcorr).

References

Allen, B., Kon, M. & Bar-Yam, Y. (2009). A new phylogenetic diversity measure generalizing the Shannon index and its application to phyllostomid bats. *The American Naturalist*, **174**, 236–243. Retrieved December 2, 2013,

Butler, M.A. & King, A.A. (2004). Phylogenetic comparative analysis: a modeling approach for adaptive evolution. *The American Naturalist*, **164**, 683–695.

Cadotte, M.W., Jonathan Davies, T., Regetz, J., Kembel, S.W., Cleland, E. & Oakley, T.H. (2010). Phylogenetic diversity metrics for ecological communities: integrating species richness, abundance and evolutionary history. *Ecology Letters*, **13**, 96–105. Retrieved December 2, 2013,

Cahill, J.F.J., Kembel, S.W., Lamb, E.G. & Keddy, P.A. (2008). Does phylogenetic relatedness influence the strength of competition among vascular plants? *Perspectives in Plant Ecology, Evolution and Systematics*, **10**, 41–50. Retrieved May 1, 2014,

Cavender-Bares, J., Ackerly, D.D., Baum, D.A. & Bazzaz, F.A. (2004). Phylogenetic overdispersion in Floridian oak communities. *American Naturalist*, **163**, 823–843. Retrieved February 12, 2010,

Cavender-Bares, J., Kozak, K.H., Fine, P.V.A. & Kembel, S.W. (2009). The merging of community ecology and phylogenetic biology. *Ecology Letters*, **12**, 693–715. Retrieved February 12, 2010,

Clarke, K.R. & Warwick, R.M. (1998). A taxonomic distinctness index and its statistical properties. *Journal of Applied Ecology*, **35**, 523–531.

Clarke, K.R. & Warwick, R.M. (1999). The taxonomic distinctness measure of biodiversity: Weighting of step lengths between hierarchical levels. *Marine Ecology Progress Series*, **184**, 21–29.

Connor, E.F. & Simberloff, D. (1979). The assembly of species communities: chance or competition? *Ecology*, **60**, 1132–1140.

Darwin, C. (1859). *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. John Murray, London.

Diamond, J.M. & Gilpin, M.E. (1982). Examination of the 'Null' Model of Connor and Simberloff for Species Co-Occurrences on Islands. *Oecologia*, **52**, 64–74.

Eastman, J.M., Paine, C.E.T. & Hardy, O.J. (2011). spacodiR: structuring of phylogenetic diversity in ecological communities. *Bioinformatics*, **27**, 2437–2438. Retrieved November 30, 2012,

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, **7**, 1–26.

Elton, C. (1946). Competition and the structure of ecological communities. *Journal of Animal Ecology*, **15**, 54–68. Retrieved March 13, 2014,

Faith, D.P. (1992). Conservation evaluation and phylogenetic diversity. *Biological Conservation*, **61**, 1–10. Retrieved October 22, 2012,

Faith, D.P. (2007). The role of the phylogenetic diversity measure, PD, in bio-informatics: getting the definition right. *Evolutionary Bioinformatics Online*, **2**, 277–283. Retrieved February 10, 2014,

Giehl, E.L.H. & Jarenkow, J.A. (2012). Niche conservatism and the differences in species richness at the transition of tropical and subtropical climates in South America. *Ecography*, **35**, 933–943. Retrieved December 2, 2013,

Gotelli, N.J. (2000). Null model analysis of species co-occurrence patterns. *Ecology*, **81**, 2606–2621. Retrieved October 23, 2012,

Gotelli, N.J. & Entsminger, G.L. (2001). Swap and fill algorithms in null model analysis: rethinking the knight's tour. *Oecologia*, **129**, 281–291. Retrieved December 2, 2013,

1 Gotelli, N.J. & Graves, G.R. (1996). *Null models in ecology*. Smithsonian Institution
2 Press, Washington.

3 Hardy, O.J. (2008). Testing the spatial phylogenetic structure of local communities:
4 statistical performances of different null models and test statistics on a locally
5 neutral community. *Journal of Ecology*, **96**, 914–926. Retrieved October 23,
6 2012,

7 Hardy, O.J. & Senterre, B. (2007). Characterizing the phylogenetic structure of
8 communities by an additive partitioning of phylogenetic diversity. *Journal of*
9 *Ecology*, **95**, 493–506.

10 Harmon, L.J., Weir, J.T., Brock, C.D., Glor, R.E. & Challenger, W. (2008). GEIGER:
11 investigating evolutionary radiations. *Bioinformatics*, **24**, 129–131. Retrieved
12 May 1, 2013,

13 Harper, J.L. (1977). *Population biology of plants.* Academic Press, London.

14 Helmus, M.R., Bland, T.J., Williams, C.K. & Ives, A.R. (2007). Phylogenetic measures
15 of biodiversity. *The American Naturalist*, **169**, E68–E83.

16 Ives, A.R. & Helmus, M.R. (2011). Generalized linear mixed models for phylogenetic
17 analyses of community structure. *Ecological Monographs*, **81**, 511–525.
18 Retrieved February 2, 2014,

19 Keddy, P.A. (1992). Assembly and response rules: two goals for predictive community
20 ecology. *Journal of Vegetation Science*, **3**, 157–164.

21 Kembel, S.W. (2009). Disentangling niche and neutral influences on community
22 assembly: assessing the performance of community phylogenetic structure tests.
23 *Ecology Letters*, **12**, 949–960.

24 Kembel, S.W., Cowan, P.D., Helmus, M.R., Cornwell, W.K., Morlon, H., Ackerly, D.D.,
25 Blomberg, S.P. & Webb, C.O. (2010). Picante: R tools for integrating phylogenies
26 and ecology. *Bioinformatics*, **26**, 1463–1464.

27 Kraft, N.J.B., Cornwell, W.K., Webb, C.O. & Ackerly, D.D. (2007). Trait evolution,
28 community assembly, and the phylogenetic structure of ecological communities.
29 *The American Naturalist*, **170**, 271–283. Retrieved October 23, 2012,

30 Mayfield, M.M. & Levine, J.M. (2010). Opposing effects of competitive exclusion on the
31 phylogenetic structure of communities. *Ecology Letters*, **13**, 1085–1093.
32 Retrieved October 23, 2012,

33 Miklós, I. & Podani, J. (2004). Randomization of presence-absence matrices: comments
34 and new algorithms. *Ecology*, **85**, 86–92. Retrieved December 2, 2013,

Miller, E.T., Zanne, A.E. & Ricklefs, R.E. (2013). Niche conservatism constrains Australian honeyeater assemblages in stressful environments. *Ecology Letters*, **16**, 1186–1194.

Murphy, H.T., Bradford, M.G., Dalongeville, A., Ford, A.J. & Metcalfe, D.J. (2013). No evidence for long-term increases in biomass and stem density in the tropical rain forests of Australia. *Journal of Ecology*, **101**, 1589–1597. Retrieved January 29, 2014,

Nipperess, D.A. & Matsen, F.A. (2013). The mean and variance of phylogenetic diversity under rarefaction. *Methods in Ecology and Evolution*. Retrieved December 2, 2013, from http://onlinelibrary.wiley.com/doi/10.1111/2041-210X.12042/full

Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H. & Wagner, H. (2013). *vegan: Community Ecology Package*. Retrieved from http://CRAN.R-project.org/package=vegan

Paradis, E., Claude, J. & Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**, 289–290.

Rao, C.R. (1982). Diversity and dissimilarity coefficients: a unified approach. *Theoretical Population Biology*, **21**, 24–43. Retrieved December 2, 2013,

Simpson, E.H. (1949). Measurement of diversity. *Nature*, **163**, 688. Retrieved April 1, 2014,

Tsirogiannis, C. & Sandel, B. (2013). Computing the skewness of the phylogenetic mean pairwise distance in linear time. *Algorithms in Bioinformatics*, **8126**, 170–184. Retrieved December 2, 2013,

Ulrich, W. & Fattorini, S. (2013). Longitudinal gradients in the phylogenetic community structure of European Tenebrionidae (Coleoptera) do not coincide with the major routes of postglacial colonization. *Ecography*. Retrieved December 2, 2013, from http://onlinelibrary.wiley.com/doi/10.1111/j.1600-0587.2013.00188.x/full

Ulrich, W. & Gotelli, N.J. (2010). Null model analysis of species associations using abundance data. *Ecology*, **91**, 3384–3397. Retrieved May 1, 2014,

Valencia, R., Foster, R.B., Villa, G., Condit, R., Svenning, J.-C., Hernández, C., Romoleroux, K., Losos, E., Magård, E. & Balslev, H. (2004). Tree species distributions and local habitat variation in the Amazon: large forest plot in eastern Ecuador. *Journal of Ecology*, **92**, 214–229. Retrieved March 17, 2015,

Vamosi, S.M., Heard, S.B., Vamosi, J.C. & Webb, C.O. (2009). Emerging patterns in the comparative analysis of phylogenetic community structure. *Molecular Ecology*, **18**, 572–592.

Vane-Wright, R.I., Humphries, C.J. & Williams, P.H. (1991). What to protect?—Systematics and the agony of choice. *Biological Conservation*, **55**, 235–254. Retrieved October 22, 2012,

Vellend, M., Cornwell, W.K., Magnuson-Ford, K. & Mooers, A.O. (2011). Measuring phylogenetic biodiversity. *Biological diversity: frontiers in measurement and assessment*, pp. 193–206. Oxford University Press, Oxford.

Villalobos, F., Rangel, T.F. & Diniz-Filho, J.A.F. (2013). Phylogenetic fields of species: cross-species patterns of phylogenetic structure and geographical coexistence. *Proceedings of the Royal Society B: Biological Sciences*, **280**. Retrieved December 2, 2013, from http://rspb.royalsocietypublishing.org/content/280/1756/20122570.short

Warwick, R.M. & Clarke, K.R. (1995). New "biodiversity' measures reveal a decrease in taxonomic distinctness with increasing stress. *Marine Ecology Progress Series*, **129**, 301–305.

Warwick, R.M. & Clarke, K.R. (1998). Taxonomic distinctness and environmental assessment. *Journal of Applied Ecology*, **35**, 532–543.

Webb, C.O. (2000). Exploring the phylogenetic structure of ecological communities: an example for rain forest trees. *The American Naturalist*, **156**, 145–155.

Webb, C.O., Ackerly, D.D. & Kembel, S.W. (2008). Phylocom: software for the analysis of phylogenetic community structure and trait evolution. *Bioinformatics*, **24**, 2098–2100.

Webb, C.O., Ackerly, D.D., McPeek, M.A. & Donoghue, M.J. (2002). Phylogenies and community ecology. *Annual Review of Ecology and Systematics*, **33**, 475–505. Retrieved June 30, 2010,

Table 1. The 22 phylogenetic community structure metrics reviewed in this paper. We paraphrase (or sometimes directly quote) the original description of the metric. While some metrics we discuss are in fact equivalent, these original descriptions often emphasized their uniqueness. IAC is a node-based metric. We multiplied it by -1 such that decreases in its value corresponded with increased clustering.

Metric	Abbreviation	Description	Citation
Quadratic entropy	QE	Within community diversity based on species dissimilarity.	(Rao 1982)
Phylogenetic diversity	PD	Sum of total branch lengths for a set of species, and length to root if set does not span it.	(Faith 1992)
Non-abundance-weighted mean pairwise phylogenetic distance	MPD	Mean of all pairwise branch lengths for a set of species.	(Webb 2000; Webb *et al.* 2002)
Non-abundance-weighted mean nearest taxon distance	MNTD	Mean of the branch lengths separating each species from its closest relative in the set of species.	(Webb 2000; Webb *et al.* 2002)
Taxonomic diversity*	Δ	Average phylogenetic distance between any two individuals from a set.	(Clarke & Warwick 1998)
Taxonomic distinctness*	Δ^*	Average phylogenetic distance between any two heterospecific individuals.	(Clarke & Warwick 1998)
Presence-absence case of taxonomic diversity*	$\Delta+$	Average phylogenetic distance between any two species from a set.	(Clarke & Warwick 1998)
Phylogenetic species variability	PSV	Measures how phylogenetic relatedness decreases the variance of a hypothetical Brownian motion trait shared by all species in the community.	(Helmus *et al.* 2007)

Table 1 *continued*

Metric	Abbreviation	Description	Citation
Phylogenetic species clustering	PSC	Modified form of PSV incorporating maximum off-diagonal element matrix of community phylogenetic correlation structure.	(Helmus *et al.* 2007)
Phylogenetic species evenness	PSE	Modified form of PSV incorporating species abundance.	(Helmus *et al.* 2007)
Phylogenetic form of Simpson's index	SimpsonsPhy	Extension of Simpson diversity index that incorporates phylogenetic information.	(Simpson 1949; Hardy & Senterre 2007)
Abundance-weighted MNTD	AW MNTD	Abundance-weighted form of MNTD.	(Webb *et al.* 2008)
Phylogenetic diversity without regard to a larger regional pool	PD_c	Sum of total branch lengths for a set of species, not including length to root.	(Faith 2007; Cadotte *et al.* 2010)
Phylogenetic abundance evenness	PAE	"Phylogenetic evenness of abundance distribution scaled by branch length."	(Cadotte *et al.* 2010)
Imbalance of abundance	IAC	IAC. "Relative per-node imbalance in individual distribution."	(Cadotte *et al.* 2010)
Community evolutionary distinctiveness	H_{ED}	"Entropic measure of diversity of evolutionary distinctiveness among species."	(Cadotte *et al.* 2010)
Equitability evolutionary distinctiveness	E_{ED}	"Equitability of H_{ED}."	(Cadotte *et al.* 2010)
Community abundance-weighted evolutionary distinctiveness	H_{AED}	"Entropic measure of diversity of evolutionary distinctiveness among individuals."	(Cadotte *et al.* 2010)

Table 1 *continued*

Metric	Abbreviation	Description	Citation
Equitability abundance-weighted evolutionary distinctiveness	E_{AED}	"Equitability of H_{AED}."	(Cadotte *et al.* 2010)
Complete abundance-weighted MPD	complete AW MPD	An abundance-weighted form of MPD. Average phylogenetic distance between two individuals from a set, possibly between the same individual.	(Webb *et al.* 2008, Appendix S2 of this paper)
Intraspecific abundance-weighted MPD	intra AW MPD	An abundance-weighted form of MPD. Average phylogenetic distance between any two individuals from a set.	(Appendix S2 of this paper)
Interspecific abundance-weighted MPD	inter AW MPD	An abundance-weighted form of MPD. Average phylogenetic distance between two heterospecific individuals.	(Miller *et al.* 2013, Appendix S2 of this paper)

1 * Denotes three metrics not directly assessed here due to equivalency with other metrics

2 (see Appendix S2), leaving 19 focal metrics in this paper.

Table 2. The nine null models reviewed in this paper. A community data matrix (CDM) where quadrats (i.e. sites or samples) are rows and species are columns is used as the input. The citation lists either the simulation name from Gotelli (2000), or gives a more recent citation where necessary.

		Constraints (data features left unchanged after randomization)				
Null model	Description	Quadrat species richness	Quadrat rank-abundance curve (and quadrat total abundance)	Species occurrence frequency	Species-specific abundance distribution (and species total abundance)	Citation
Richness	Randomizes species' occurrences (or abundances) among species, independently within each quadrat.	X	X			SIM3
1s*	Randomizes species' occurrences (or abundances) among the tips of a phylogeny. With respect to the CDM, this shuffles entire columns among species.	X	X		†	Hardy (2008)
Frequency by quadrat	Often simply called a "frequency" null. Shuffles species' occurrences (or abundances) independently within each species.			X	X	SIM2
Frequency by richness	The same randomization as above null, but then groups randomized quadrats by their species richness. Observed values compared only to values from randomized quadrats of corresponding species richness.	X‡		X	X	Miller et al. 2013; Appendix S1
Independent swap	Transposes randomly chosen submatrices of the form (0,1)(1,) or (1,0)(0,1) in the CDM. When CDM contains abundance data, treats non-zero elements as 1.	X		X	§	SIM9

		Constraints (data features left unchanged after randomization)				
Null model	Description	Quadrat species richness	Quadrat rank-abundance curve (and quadrat total abundance)	Species occurrence frequency	Species-specific abundance distribution (and species total abundance)	Citation
Trial swap*	Same as independent swap, but guarantees equidistribution of results (evenly distributed randomized results).	X		X	§	Miklós & Podani (2004)
2x	Modified form of independent swap for abundance data. Transposes randomly chosen submatrices, switching elements of the submatrices within quadrats.	X	X	X		Hardy (2008)
3x	As for 2x, but switches elements within species.	X		X	X	Hardy (2008)
Regional	Described in detail in Appendix S1 of this paper.	X‡	Strictly maintains total abundance, approximately maintains rank-abundance curve	Approx.	Approx.	Appendix S1

* These were not included in tests of metric + null model performance due to equivalency with the richness and independent swap (Appendix S1), leaving seven focal models in this paper.

† Because columns are moved as a unit, each randomized CDM contains the same set of species-specific abundance distributions as the original CDM, though these abundance distributions are disassociated from their original species (i.e. the set of columns is the same, but each column is now associated with a different species).

1 ‡ The randomized matrices do not always contain quadrats with species richness values the same as those of the original CDM, but by
2 concatenating results later by randomized quadrat species richness, observed quadrats are compared to random quadrats of the same
3 species richness.
4 § Intended for use with presence/absence data, thus the fact that the *picante* (and *metricTester*) implementations also maintain column
5 sums (and not just the sum of non-zero elements), and therefore also maintain species-specific abundance distributions is an
6 unintentional consequence of the way these null models are coded.
7

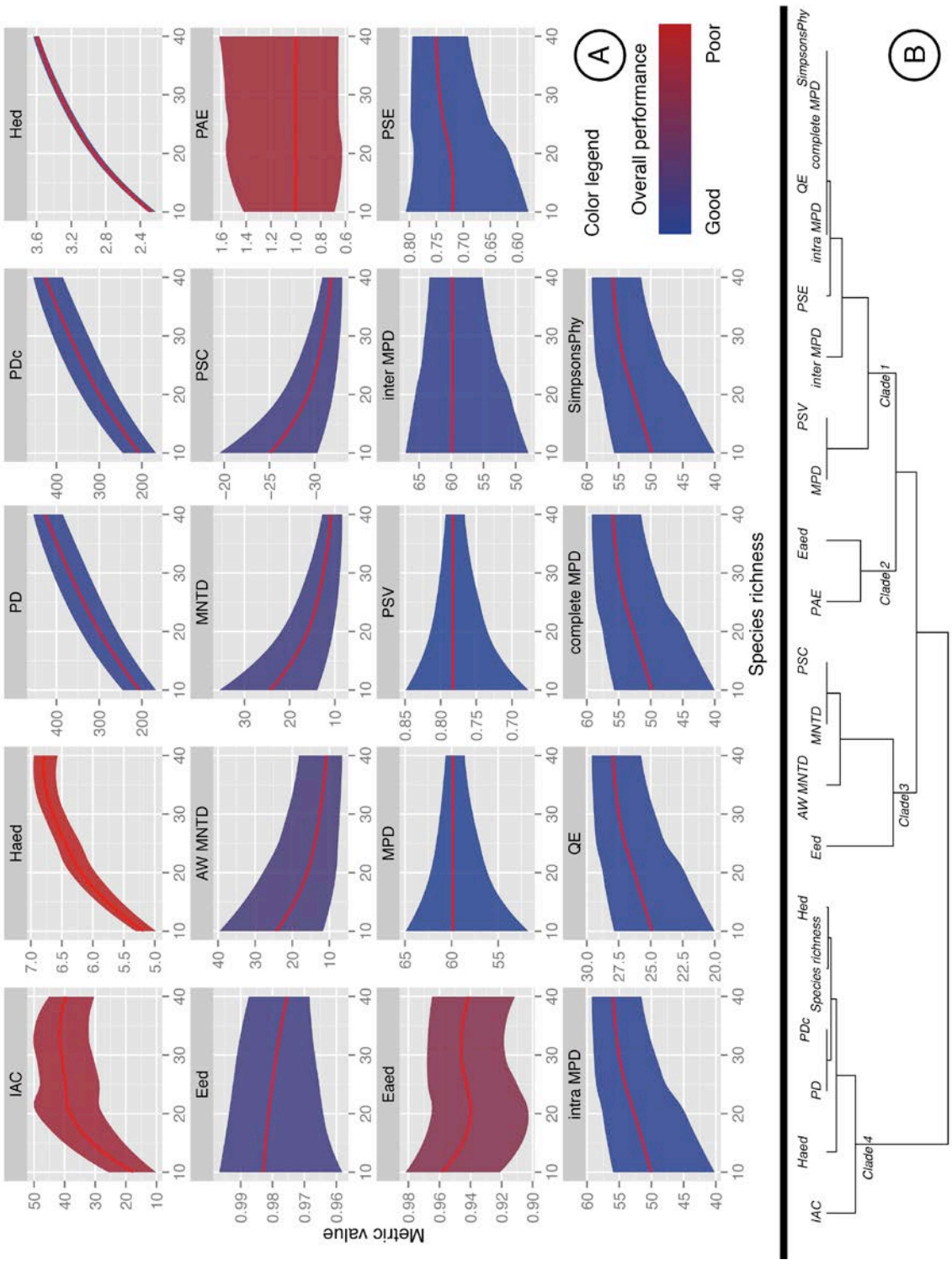
A
B
Color legend
Overall performance
Good
Poor
Species richness
Metric value
IAC
Haed
PD
PDc
Species richness
Hed
Eed
AW MNTD
MNTD
PSC
PAE
Eaed
MPD
PSV
inter MPD
PSE
intra MPD
QE
complete MPD
SimpsonsPhy
Clade1
Clade2
Clade3
Clade4

Figure 1.

Figure 1. (A) Behavior of 19 focal phylogenetic community structure metrics (Table 1) across variation in species richness. Panels are color-coded from blue (good) to red (poor) according to the mean of type I and II errors across all simulated assembly processes. **(B)** Dendrogram of intercorrelations among the phylogenetic community structure metrics (and species richness itself). Closely correlated metrics are annotated along branches. Clade 1 metrics focus on "total community relatedness"; Clade 2 metrics on the relationship between "evolutionary distinctiveness and abundance"; Clade 3 on "nearest-relative" measures of community relatedness; and Clade 4 metrics are particularly closely correlated with species richness.

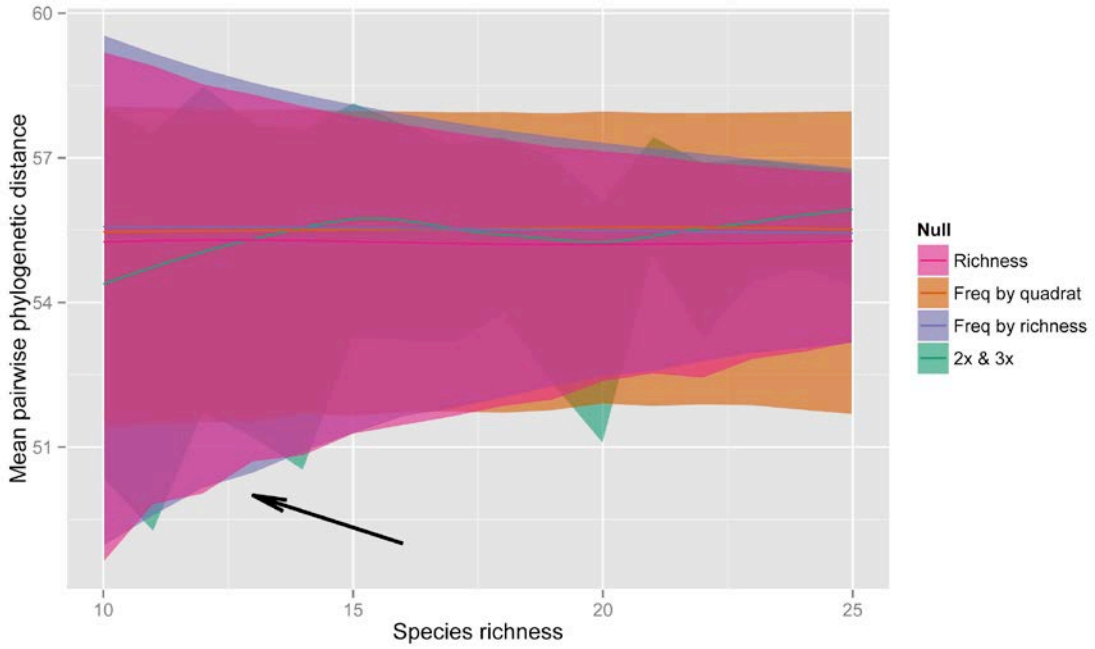

Figure 2. Confidence intervals (95%) for the richness, both forms of the frequency, 2x and 3x null models (Table 2) across variation in species richness. Expectations shown here are the result of 10^5 randomizations. Because the 2x and 3x nulls follow identical distributions (Fig. S1.5), only a single layer is included in this figure. The arrow indicates a region of particular concern for type I error when using the frequency by quadrat null. Other null model behavior (including the independent swap, trial swap, and regional models) is summarized in Appendix S1.

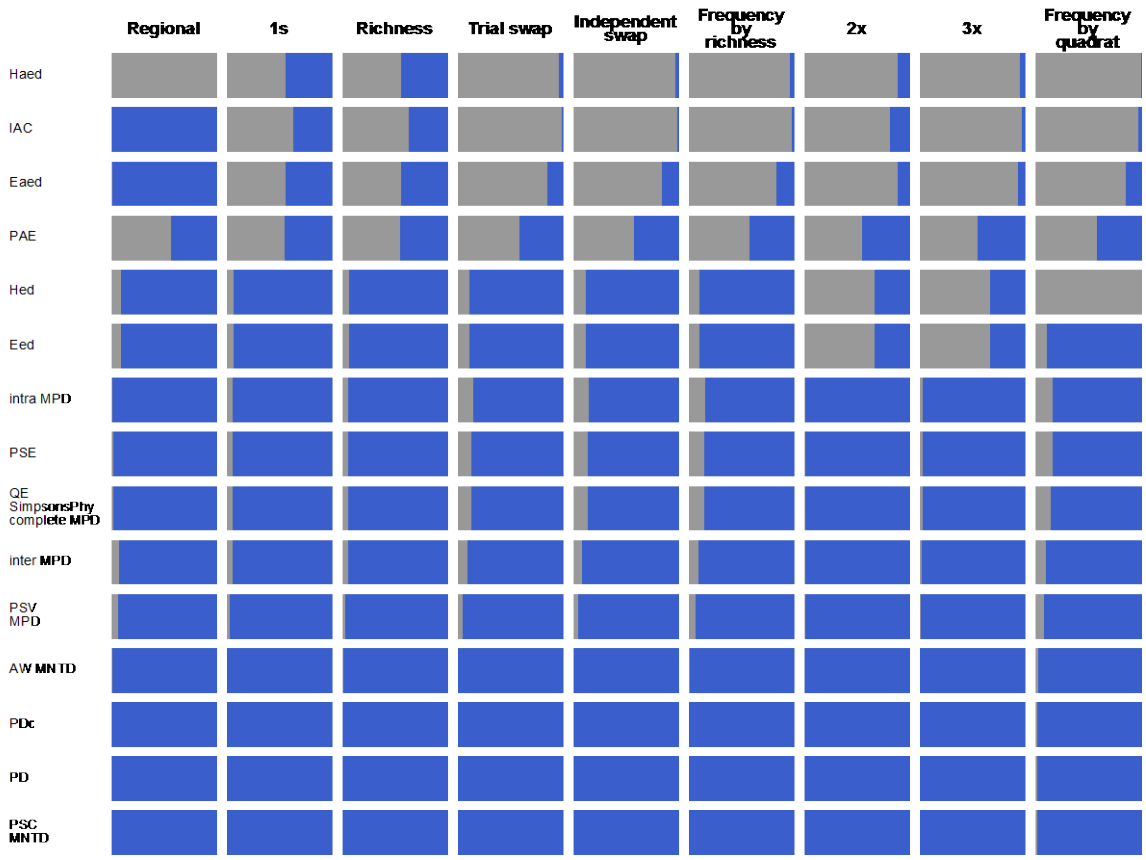

Figure 3. Performance of metric + null model approaches at detecting phylogenetic clustering given habitat filtering. Blue bars summarize the proportion of the total 1,009 simulations where the mean of the standardized effect sizes was significantly less than zero (one-way Wilcoxon signed-rank test). Gray bars summarize the proportion where the mean did not differ from zero (type II errors). Metrics and nulls are arranged in order from habitat filtering-specific best-performing to worst, with the best approaches in the bottom left corner.

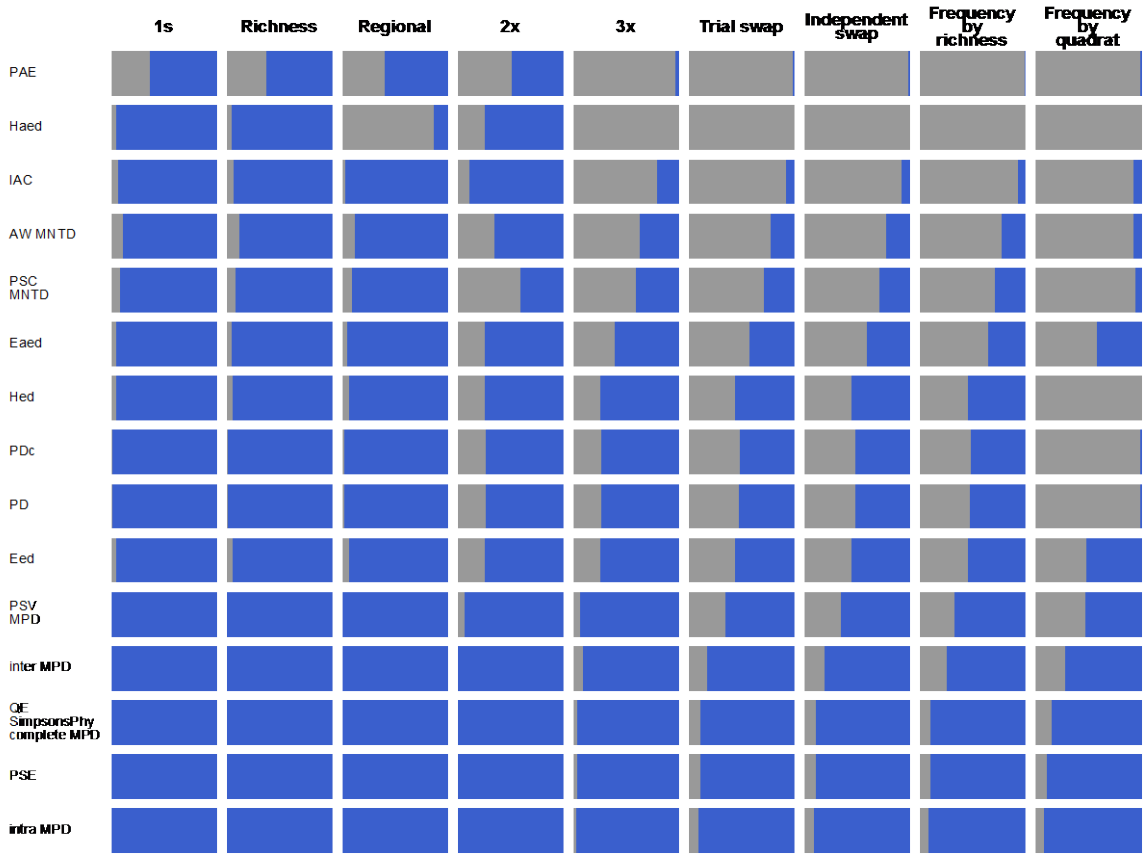

Figure 4. Performance of metric + null model approaches at detecting phylogenetic overdispersion given competitive exclusion. Blue bars summarize the proportion of the total 1,009 simulations where the mean of the standardized effect sizes was significantly greater than zero (one-way Wilcoxon signed-rank test). Gray bars summarize the proportion where the mean did not differ from zero (type II errors). Metrics and nulls are arranged in order from competitive exclusion-specific best-performing to worst, with the best approaches in the bottom left corner.

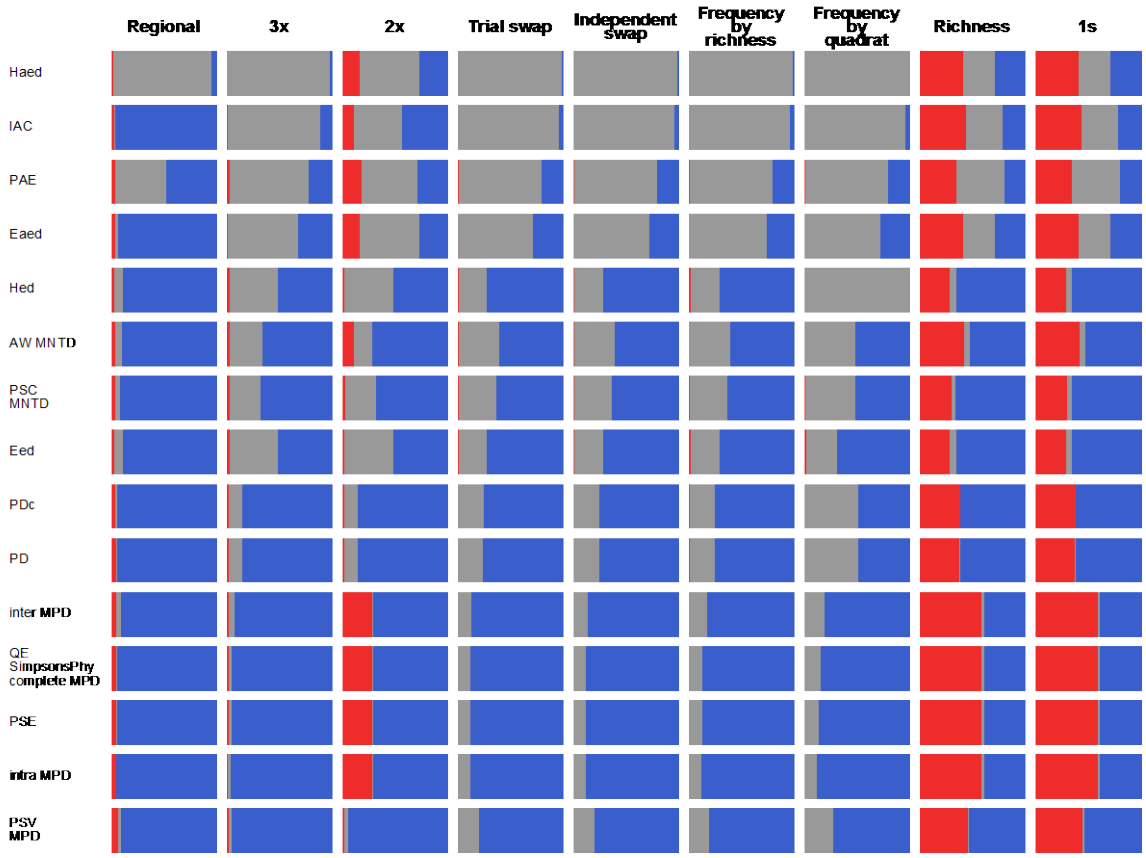

Figure 5. Overall performance of metric + null model approaches. Red bars (type I errors) summarize the proportion of the total 1,009 random community assembly simulations where the mean of the standardized effect sizes differed significantly from zero (two-way Wilcoxon signed-rank test). Gray bars summarize the mean type II error rates from Figs. 3 and 4. Blue bars provide an indication of the success of each approach, and are defined as the proportion of runs that did not generate type I and II errors. Metrics and nulls are arranged in order from overall best-performing to worst, with the best approaches in the bottom left corner.

Box 1: Abbreviated history of phylogenetic community structure metrics.

Faith (1992) introduced PD, a metric that quantifies the unique evolutionary history represented by co-occurring taxa. It was intended (and is often used) as a conservation tool. While PD built upon previous work by Vane-Wright *et al.* (1991) and others, it was the first to explicitly incorporate phylogeny. Since PD is the sum of all branch lengths connecting the species in a community (Table 1), the assumption that it increases with additional species, and is therefore correlated with species richness, was implicit (exact solution provided by Nipperess & Matsen 2013).

Subsequently, Clarke and Warwick introduced metrics (Δ, Δ+, Δ*) focused on the average branch length among a group of taxa or individuals, again linking their methodology to conservation decisions (Warwick & Clarke 1995, 1998; Clarke & Warwick 1998, 1999). Their pioneering papers explored some statistical properties of the metrics, including the fact that mean expected Δ+ is not correlated with species richness, but the width of its confidence intervals decreases with species richness (creating a "confidence funnel"). Yet, the conservation-specific scope of their papers limited their impact on community ecology. In fact, we were unaware of these metrics until after we had run our initial analyses.

Webb (2000) introduced two new metrics--MPD and MNTD--and the standardized forms of these, NRI (net relatedness index) and NTI (nearest taxon index). Initially, MPD was slightly different than Clarke and Warwick's metrics, only incorporating nodal distances, but by Webb *et al.* (2002) the definition had expanded to incorporate branch length, and was therefore equivalent to Δ+ (Appendix S2). Yet, by linking community assembly processes with these phylogenetic patterns, it was MPD and MNTD that

revolutionized the field of community ecology. Moreover, despite the equivalency of MPD and Δ+, Webb stated that both MPD and MNTD are correlated with species richness when only MNTD is (Fig. 1A), and devised standardization procedures to "correct" for this. This misperception occasionally persists to the present (e.g., Ulrich & Fattorini 2013), despite empirical solutions to the contrary (Tsirogiannis & Sandel 2013).

Helmus *et al.* (2007) introduced PSE, the "first" metric to incorporate abundance information. While this is not entirely true (Rao 1982; Warwick & Clarke 1995; Hardy & Senterre 2007), their focus on community assembly linked their approach with venerable evolutionary questions. Helmus *et al.* (2007) also introduced two other metrics intended to be similar but superior to NRI and NTI--PSV and PSC. The noted advantage to these is the lack of need for a reference species pool, and therefore the ability of these metrics to transcend the particulars of the phylogeny and community data matrix at hand, and allow raw metric values to be directly compared. However, these should therefore have been compared with MPD and MNTD, respectively. Had this been done, it would have been noted that PSV and PSC are directly proportional to MPD and MNTD, respectively, a still all but unknown fact (though see Vellend *et al.* 2011). Instead, PSV and PSC were compared with NRI and NTI. As a further complication, the PSC function in *picante* (Kembel *et al.* 2010) returns the inverse of PSC (M. Helmus, pers. comm.). This has confounded subsequent papers (e.g. Giehl & Jarenkow 2012; Villalobos *et al.* 2013). Some authors have incorrectly claimed that PSC is not inherently correlated with species richness.

Cadotte *et al.* (2010) introduced metrics focused on phylogenetic abundance distributions. We review seven of those here: PD_c (this was actually discussed earlier,

Faith (2007)), PAE, IAC, ED, H_{ED}, E_{ED}, H_{AED}, and E_{AED} (see Table 1). Cadotte *et al.* (2010) showed their metrics ranked communities differently than each other and than metrics like PSV and MNTD, but offered no discussion of the metrics' statistical properties, nor has any subsequent paper. The metrics are available in *ecoPD* (http://r-forge.r-project.org/projects/ecopd/).

We discuss six additional metrics in this paper: QE (Rao 1982), SimpsonsPhy (Hardy & Senterre 2007), abundance-weighted (AW) MNTD, and three variants of AW MPD (Table 1, Appendix S2). Both complete AW MPD and AW MNTD were introduced in *Phylocom* (Webb *et al.* 2008) and *picante* without accompanying publication, and their statistical properties and relationship to other metrics remains essentially unknown. Interspecific AW MPD was introduced in (Miller *et al.* 2013), and intraspecific AW MPD is "first" described in the current paper (Appendix 2), though as we subsequently discovered, it is equivalent to Δ (Clarke & Warwick 1998). Similarly, after exploring the behavior of QE and SimspsonsPhy and finding them equivalent, we realized this was already known (Hardy & Senterre 2007; Allen *et al.* 2009).