Global Ecology and Biogeography, (Global Ecol. Biogeogr.) () **25,** 368–377

Global Ecology and Biogeography
A Journal of Macroecology

MACROECOLOGICAL METHODS

Contrasting beta diversity among regions: how do classical and multivariate approaches compare?

Joseph R. Bennett[1,2†] and Benjamin Gilbert[3*†]

[1]*Institute of Environmental Science and Department of Biology, Carleton University, Ottawa, ON K1S 5B6, Canada,* [2]*School of Biological Sciences, University of Queensland, Brisbane, Qld 4072, Australia,* [3]*Department of Ecology and Evolutionary Biology, University of Toronto, Toronto, ON M5S 3B2, Canada*

ABSTRACT

Aim Approaches to calculating beta diversity (β) include classical measures based on alpha (α) and gamma (γ) diversity, and multivariate distance-based measures. Species–area relationships cause measurements of γ to vary, making comparisons of classical β among regions contingent on sampling effort. A recent null-modelling approach has attempted to account for variation in γ by calculating the degree to which β deviates from a random expectation. Here, we clarify the mathematical links between classical and multivariate approaches to measuring β, to derive predictions regarding the reliability of classical, null-model and multivariate approaches. Next, we use four ecological datasets and simulated data to test the consistency of these approaches across sampling effort and γ. We focus on an issue that arises when making comparisons among regions, namely that even small changes to the area sampled can differentially increase measured γ in each region, potentially causing artefacts in β that are driven by methodology rather than biology.

Innovation Comparisons among regions using classical and null-model measures change dramatically as sampling effort and γ increase. This change is understood for classical β because of species–area relationships, but not for null-model measures, making comparisons among regions impossible using the null-model approach. Multiple-site dissimilarity shows a similar sensitivity to γ as classical measures. In contrast, pairwise multivariate distances show no systematic effect of sampling effort and γ: increasing the number of sample plots decreases variability but does not alter mean β.

Main conclusions Multivariate pairwise distances are independent of sample size, offering the most robust comparison among regions. The widespread influence of sampling effort and γ indicate that only scale-dependent measures of classical and multiple-site β are comparable, whereas null-model β may not be comparable among regions. However, in cases where γ is well known, multiple-site dissimilarity metrics offer several advantages, and should be strongly considered.

Keywords
Alpha diversity (α), beta diversity (β), classical metrics, gamma diversity (γ), multiple-site dissimilarity, multivariate metrics, null model.

*Correspondence: Benjamin Gilbert, Department of Ecology and Evolutionary Biology, University of Toronto, 25 Harbord Street, Toronto, ON M5S 3B2, Canada.
E-mail: benjamin.gilbert@utoronto.ca
†Authors contributed equally.

INTRODUCTION

Beta diversity (β) is an emergent property of species and the environment; it describes the change in species from one area to another and is considered to be a key signature of dispersal, intra-specific competition, Janzen–Connell effects and the spatial structure of environmental gradients (Harms *et al.*, 2000; Gilbert & Lechowicz, 2004; Legendre *et al.*, 2005). In spatially disjunct patches, such as those found in fragmented habitats, β provides an intuitive measure of both the connectedness and the ecological distinctiveness of patches (e.g. Tscharntke *et al.*, 2002). In contiguous habitat, β among samples can provide estimates of spatial aggregation in species distributions (e.g. Novotny *et al.*, 2007). Because of the ecological relevance of β,

DOI: 10.1111/geb.12413

ecologists are increasingly interested in quantifying how it differs among regions or ecosystems (e.g. Kraft *et al.*, 2011; Myers *et al.*, 2013; Zhang *et al.*, 2013).

Despite the importance of β as an ecological metric, its quantification has inherent challenges. Metrics of β may be influenced by factors including the size of the species pool, the density of individuals in a sample unit, the species abundance distribution and species aggregation in space, making interpretation of differences among regions difficult (Chase & Knight, 2013). As a result, the question of how to test and interpret differences in β among regions remains an important challenge for ecologists (Kraft *et al.*, 2011; Qian *et al.*, 2012, 2013; Tuomisto & Ruokolainen, 2012).

There are two general approaches to measuring β (Anderson *et al.*, 2011): 'classical metrics', calculated using alpha diversity (α) and gamma diversity (γ), and multivariate measures of species turnover based on statistical summaries of dissimilarity metrics (Legendre *et al.*, 2005; Baselga, 2010; Legendre & De Cáceres, 2013). Although these approaches measure different phenomena, both fit Whittaker's original concept of measuring variation in species composition along environmental or spatial gradients (Whittaker, 1960, 1972; Anderson *et al.*, 2011; Legendre & De Cáceres, 2013).

Recognizing the dependence of classical β on γ, recent studies have proposed and implemented a null-model approach to correct for γ dependence and allow comparison of classical β among regions (Kraft *et al.*, 2011; Mori *et al.*, 2013; Stegen *et al.*, 2013). A separate branch of research has proposed using multivariate distance measures of β for inter-region comparisons (Legendre & De Cáceres, 2013). To our knowledge, no study has systematically tested the consistency of classical and multivariate β across variations in sampling effort and γ, nor the relative effectiveness of these approaches for comparing β among regions.

The development of distinct approaches to quantifying β raises important questions that can clarify how this field should progress. First, how are the different measures of β related to each other, and do these relationships suggest a single best method of testing β? Second, how well do the different measures perform when comparing β among regions? Although this second question is difficult to address, as there is no single perfect measure of β, we can nonetheless use sampling strategies on real data to explore the effect of known differences among regions, such as differences in γ measured from samples.

Here, we examine the reliability of currently used classical and multivariate approaches for analysing β within and among regions. We first explore the challenges of comparing β among regions using classical and multivariate approaches, examining the relationships between the approaches. We then use these relationships to clarify how classical and multivariate measures are linked. Next, we use four data sets to test the validity of β metrics attained using each approach. We show that classical metrics produce consistent biases, unlike their pairwise multivariate counterparts. These results allow us to recommend general methods for measuring and interpreting variation in β among regions.

MATERIALS AND METHODS

Beta diversity metrics: classical β

Classical metrics partition γ into α and β components, either multiplicatively or additively (Whittaker, 1972; Lande, 1996). A fundamental challenge of using these metrics for comparing β among regions stems from this dependence of β on both α and γ. For example, in the multiplicative β partitioning model used by Kraft *et al.* (2011) and throughout this paper, β is measured as $\beta = 1 - (\bar{\alpha}/\gamma)$. Here, $\bar{\alpha}$ is the mean number of species per sample plot and γ is the total number of species among plots, which is a sample-based estimate of the true regional species pool that we refer to as γ_{true}. Comparisons of β among regions that use this metric will be biased if either $\bar{\alpha}$ or γ are not representative of the region. This concern is not important for $\bar{\alpha}$ if the sampling technique (e.g. plot size and survey method) is the same among regions – randomly selected plots will produce $\bar{\alpha}$ with a variance that can be estimated, whether this is based on raw species values from complete censuses or richness estimates (Colwell & Coddington, 1994).

A more difficult problem lies in measuring γ, and controlling for variation in γ when comparing β among regions. Determining the true regional species pool (γ_{true}) can be difficult even for well-documented areas (Lessard *et al.*, 2012), and in many areas the vast majority of species are too poorly studied to characterize species pools (Kraft *et al.*, 2011). Sample-based measurements of γ may be dependent on sampling effort (defined here as the number of equivalent sample units used) and species–area relationships (SARs). When SARs in different regions have different scaling exponents, measured γ based on the total number of species among all plots will vary depending on sampling effort, making it difficult to compare classical β between regions (Gotelli & Colwell, 2001; Chao & Jost, 2012). A similar complication occurs when quantifying β over time. Species pools can change over time due to colonizations and extinctions, making it difficult to determine the species pool that has produced an observed assemblage. In addition, the interplay of local (α-level) and regional (γ-level) colonization–extinction dynamics further complicates temporal comparisons of classical β among regions (Olden & Rooney, 2006).

One approach to improve the effectiveness of comparisons of classical β between regions is based on sampling effort: either both regions may be sufficiently sampled to reach a reasonable estimate of the regional species pool (γ_{true}) or sample sizes may be rarefied to correct for bias. If a region is sampled until the addition of species is low enough to be trivial, β can be calculated with a measured variance attributable to variation of α among plots. Because sampling until γ_{true} is known is often impossible, various rarefaction techniques have been proposed to standardize sampling between regions (Gotelli & Colwell, 2001; Colwell *et al.*, 2012). Recently, Chao & Jost (2012) proposed rarefaction by sample coverage, because rarefaction to equal numbers of individuals underestimates the richness of more diverse communities. However, one disadvantage in this approach is uncertainty in estimates of sample coverage. We do not further explore the rarefaction approach, and instead focus on approaches that do not require estimates of sample coverage.

The second method to allow comparison of classical β among regions is a statistical correction to account for variations in γ from one region to another. Kraft *et al.* (2011) developed a procedure to correct for variation in γ across regions using a null model designed to allow unbiased comparison of β among regions, which we refer to as β deviation (β_{dev}). β_{dev} is adapted from an earlier method that tests when the distribution of species among sample plots differs from a random assembly model (Crist *et al.*, 2003). The β_{dev} approach has since been used by others (e.g. Mori *et al.*, 2013; Stegen *et al.*, 2013), and has been adapted to further test the underlying determinants of β (Myers *et al.*, 2013) as well as phylogenetic β (Stegen *et al.*, 2012). The method for calculating β_{dev} involves generating a null distribution (β_{null}) with a measured standard deviation ($\beta_{SD.null}$) by randomly shuffling species among sites while holding species abundances and plot densities constant (Kraft *et al.*, 2011). Unlike standard randomization tests that compare observed β against null distributions (Crist *et al.*, 2003), β_{dev} represents an effect size, calculated as:

$$\beta_{dev} = \frac{\beta_{observed} - \bar{\beta}_{null}}{\beta_{SD.null}} \tag{1}$$

which can equally be written as:

$$\beta_{dev} = \frac{\bar{\alpha}_{randomized} - \bar{\alpha}}{SD(\bar{\alpha}_{randomized})}. \tag{2}$$

It should be noted that β_{dev} differs from classical β in that it measures the difference in β from a random sample, relative to the variation in the randomized data (equation 2). The overall validity of β_{dev} in being able to account for systematic changes in γ has not been fully tested – initial tests of the performance of the null model across variations in γ have only been performed on randomly generated data that had an expected β_{dev} of zero (Kraft *et al.*, 2011), and may therefore behave differently from ecological data.

Beta diversity metrics: multivariate β

Conceptual differences between classical and multivariate approaches are reviewed by Tuomisto (2010) and Anderson *et al.* (2011). Both approaches have been used extensively to measure β (Anderson *et al.*, 2011; Legendre & De Cáceres, 2013). Many multivariate measures of β are distinguished from classical metrics in that they use distance or resemblance matrices to quantify compositional differences among plots or sites (e.g. Baselga, 2010; Legendre & De Cáceres, 2013) rather than directly calculating variation using synoptic measures of γ and $\bar{\alpha}$ (Anderson *et al.*, 2011). We restrict our consideration of multivariate measures to asymmetric presence–absence metrics such as the Sørensen distance in order to compare to classical β (Legendre & Legendre, 2012). In Appendix S2 we show the mathematical relationship between two pairwise distances, Sørensen and Jaccard, and classical β (equations S1–S7 in Appendix S2). Due to the similarities in the pairwise distances, we focus only on the Sørensen distance in the main article and present tests of the Jaccard index in Appendix S2.

Although the classical and multivariate approaches can be mathematically linked, using the pairwise Sørensen distance to determine variation in diversity among regions differs from classical β in three important ways. First, so long as sites are randomly selected and the sampling technique (i.e. plot size and survey technique) is consistent among samples, we would not expect a change in pairwise Sørensen distance with the number of plots sampled because increasing the number of plots increases neither $\bar{\alpha}$ nor the total number of species in any two plots (γ_2). Thus, unlike classical β, we would not expect the pairwise Sørensen distance to change with measured γ, and the type of randomization procedure used to generate β_{dev} (equation 1) should not be necessary. Second, because the pairwise Sørensen distance does not incorporate a null model, it cannot test when a region's β differs from that expected by chance. However, methods have been developed to apply the null-model approach to pairwise multivariate distances (Myers *et al.*, 2013). We test both null-model and standard pairwise distance approaches, but note that mean pairwise distance in a region is not indicative of non-random processes in and of itself. The third, related, difference is that the use of Sørensen distance provides an average measure of β between two plots (equation S1 in Appendix S2), and therefore does not estimate classical β that would be attained from sampling an entire region.

Although pairwise dissimilarity measures should not be sensitive to changes in sample coverage or estimated γ, they have a distinct limitation: they do not account for species co-occurrence among more than two sites (Diserud & Ødegaard, 2007). Multiple-site dissimilarity measures have been designed to overcome this limitation (Baselga *et al.*, 2007; Diserud & Ødegaard, 2007; Baselga, 2010). In this paper we focus on the multiple-site Sørensen index of Baselga (2010), which can be additively partitioned into both turnover and nestedness components; however, we examine an additional method (Diserud & Ødegaard, 2007) in Appendix S2. Although the multiple-site Sørensen index of Baselga (2010) and its components cannot be simplified with relation to β, as we have done with the other metrics, the concordance of these methods with classical β (Baselga, 2010, 2013) suggests they may be sensitive to γ. However, to our knowledge, the sensitivity of these measures to sampling effort and γ has not been tested, and the advantages of these measures suggest that they may be particularly useful for comparing β among regions.

Tests of β metrics

To test the ability of classical, null-model and multivariate metrics to compare β among regions, we tested for systematic dependences of these measures on sampling effort and the resulting estimate of γ. Any dependence of β metrics on sampling effort and γ would imply that tests of differences among regions will be either biased or scale dependent, as the magnitude of the differences will depend on γ and the sampling effort used in a given study. We used three datasets from plot-based plant surveys, and one dataset from surveys of microalgae in lakes, where the sample unit is an individual lake, to test performance of the metrics. The plant datasets included surveys of 1-m^2 plots in the Garry oak ecosystem meadows in British Columbia, Canada and Washington State, USA (Bennett *et al.*, 2012; Appendix S2), 1-m^2 plots in early successional fields in the Koffler scientific reserve in Ontario, Canada (B. Gilbert, unpublished data), and 50-m^2 forest plots near Montreal, Canada (Gilbert & Lechowicz, 2004). The diatom dataset used surveys from 492 lakes in eastern North America (see Ginn *et al.*, 2007; Bennett *et al.*, 2010 for details). For the sake of brevity, we present a subset of results from the meadow ecosystem and early successional datasets in the main text, but present full details on the datasets and results for the additional datasets in Appendix S2.

To test the sensitivity of each metric to sampling effort and γ, we calculated each metric for the full set of n sample units, and then repeated our analysis with incrementally smaller numbers down to five sampling units, using 100 random subsamples at each increment of sampling effort (see Appendix S2 for details). This allowed us to test if each metric showed any systematic trend with sampling effort and γ, and whether this in turn biased comparison among ecosystems. For each subset, we calculated classical (uncorrected) β and β_{dev} using the methods of Kraft *et al.* (2011). For multivariate pairwise distances, we used two common approaches: mean pairwise Sørensen and Jaccard distances, as well as mean and median distance to group centroids (Anderson *et al.*, 2006), using the correction of Stier *et al.* (2013). We also calculated the multiple-site Sørensen index, as per Baselga (2010), using the R package 'betapart' (Baselga & Orme, 2012), and the index proposed by Diserud and Ødegaard (Diserud & Ødegaard, 2007). In addition, we calculated an analogue of β_{dev} using the pairwise Sørensen distance, as per Myers *et al.* (2013). In all cases, mean β_{null} (equation 1) was calculated using 1000 randomizations of species distributions among plots (following Kraft *et al.*, 2011). Finally, we note that most β measures assume that species are measured without error. Although this is generally correct for complete censuses within plots used in plant studies, it is often incorrect when incomplete sampling within plots is used (Colwell & Coddington, 1994; Chao *et al.*, 2005). For these latter studies, appropriate estimators for α and γ should be used for classical β, and multivariate distances should be corrected using the Chao distance (Chao *et al.*, 2005). Because these methods have only been developed for pairwise distances, we restrict our exploration of the effect of these estimators on the Chao distance.

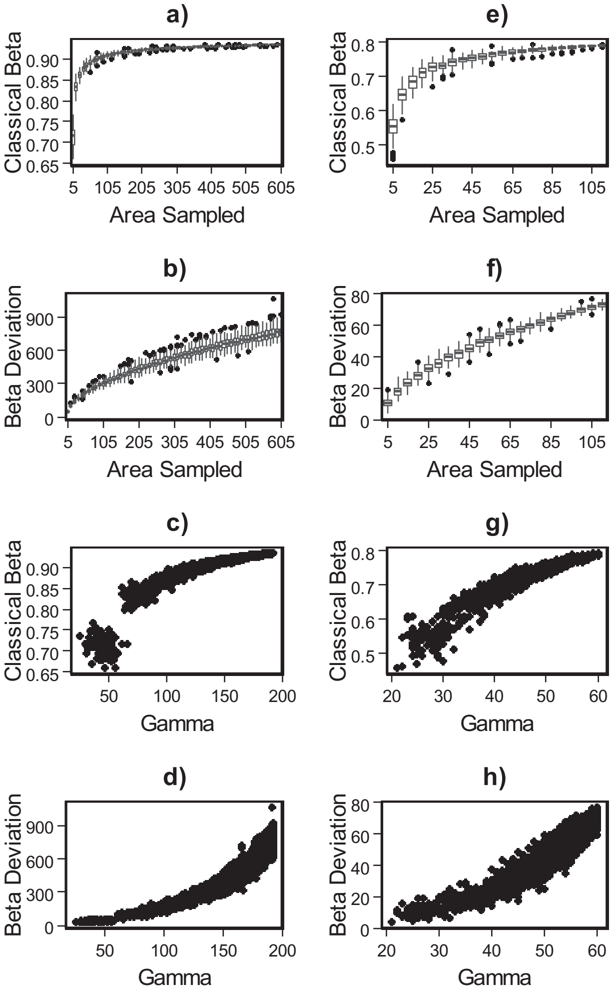

Figure 1 Observed classical β versus area sampled (1 m^2 plots) (a, e), β_{dev} versus area sampled (b, f), classical β versus γ among plots (c, g), and β_{dev} versus γ (d, h). Plots (a)–(d) are for a meadow ecosystem and (e)–(h) for an early successional ecosystem. The gap in classical β in (c) is due to a large increase in γ between 5 and 15 sub-sampled plots (mean γ difference 5–15 plots = 24.2 species, $T = 35.3$, $P < 0.0001$; mean α difference 5–15 plots = 0.32 species, T = 1.16, $P = 0.25$).

RESULTS

Metrics for comparing β among regions should ideally show no systematic change with sampling effort or γ, because if such a change occurs at different rates among regions, differences will be inconsistent. Standard measures of classical β unsurprisingly fail this criterion, as they increase with area sampled and therefore γ (Figs 1 & S1–S3 in Appendix S2). For example, mean classical β for the early successional and meadow ecosystems were 0.69 and 0.84, respectively, if only 15 1-m^2 plots were sampled in each study system, but 0.78 and 0.90 respectively if 75 plots were sampled. In other words, the difference in classical β between these ecosystems decreased by approximately 20% as a result of these increases in sampling effort and the resulting measured γ (Fig. 1).

The statistical correction for classical β, β_{dev}, was also sensitive to area sampled and the resulting γ, such that β_{dev} could not be reliably contrasted between ecosystems (Figs 1 & S1 in Appendix S2). For example, with a sampled area of 15 1-m^2 plots, β_{dev} was 25 and 99 for the early successional and meadow ecosystems, respectively, but this changed to 60 and 251 when 75 plots were sampled, corresponding to an increase in the difference in β_{dev} between sites of over 250%. This difference can also be seen by contrasting the effect of a change in γ in each region; increases in γ increased β_{dev} in both regions, but the increase was greatest in the ecosystem that contained more species (Fig. 1).

The increase in β_{dev} with sampling effort can be broken down into its component parts. The difference between observed classical β and mean β_{null} exhibited varying trends with sampling effort (Figs S4 & S5 in Appendix S2). However, the standard deviation (SD) of mean β_{null} always decreased at a rapid rate with sampling effort (Figs S4 & S5 in Appendix S2), causing β_{dev} to increase. Since γ was equal between observed and randomized data in each subsample, the decrease in the SD of mean β_{null} with sampling effort was itself driven by a decrease in the SD of the mean randomized α (the denominator of equation 2) with sampling effort (Figs 2 & S6 in Appendix S2).

To further explore why the β_{dev} method was sensitive to sampling effort, we generated simulated communities using lognormal abundance distributions and a range of specified regional species pools (γ_{true}), as per Kraft *et al.* (2011) to test the β_{dev} method (see Appendix S2 for details). These randomly generated data do not allow a direct test of the β_{dev} method because the numerator in equations 1 & 2 is meaningless, as it compares a randomly generated species richness with that obtained by randomizing the data (i.e. the expectation is zero; Fig. S7 in Appendix S2). However, these simulated data illustrate how the SD of mean randomized α changes as a function of sample size and γ. Consistent with our field data, the simulations showed that the SD of mean randomized α decreases with sampling effort and that the steepest slope was found in regions with higher γ_{true} (Fig. 2). These results were also found for other species abundance distributions, indicating that the patterns we see in our four tested datasets are likely to be representative of a broad range of communities. Because the slope of this relationship depends on the total number of species in a region (Fig. 2), and does not appear to change in a predictable way with sampled γ (unlike classical β), there does not appear to be a simple mathematical correction for β_{dev}.

Multivariate distances showed distinctly different patterns from β_{dev}. Mean pairwise Sørensen distance showed no trend with area sampled or γ (Figs 3 & S8 in Appendix S2; $R^2 < 0.001$; all $P > 0.13$, with over 1000 d.f. per test). Jaccard and Chao distances showed similar results (Figs S9–S12 in Appendix S2; all $P > 0.28$), as did centroid-based measures of mean Sørensen distance (Fig. S13 in Appendix S2). In contrast, the multiple-site Sørensen index of Baselga (2010) was sensitive to sampling effort and γ, showing a trend that was similar to classical β (Fig. 3). For example, the percentage difference between mean multiple-site Sørensen index in the successional and meadow ecosystems decreased from 25% for 5 samples to < 2% for 105 samples. Further analysis showed a strong correlation between classical β and multiple-site Sørensen index (Fig. S14 in Appendix S2). The multiple-site index of Diserud and Ødegaard (2007) was also sensitive to sampling effort and therefore γ (see Appendix S2 for details).

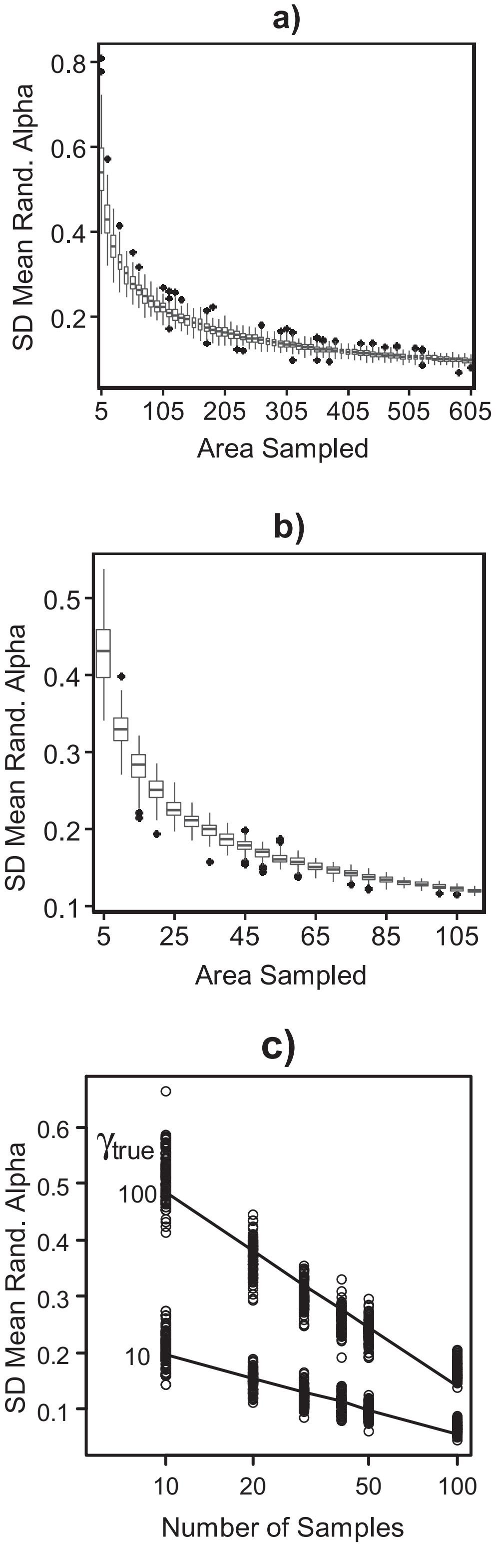

Figure 2 Standard deviation (SD) of mean randomized α versus sampling effort, for a meadow ecosystem (a), an early successional ecosystem (b), and simulated data as per Kraft *et al.* (2011) with specified species pools (γ_{true}) = 10 and 100 (c). Lines in (c) show a fitted linear relationships between the SD of mean randomized α and sampling effort for each specified species pool.

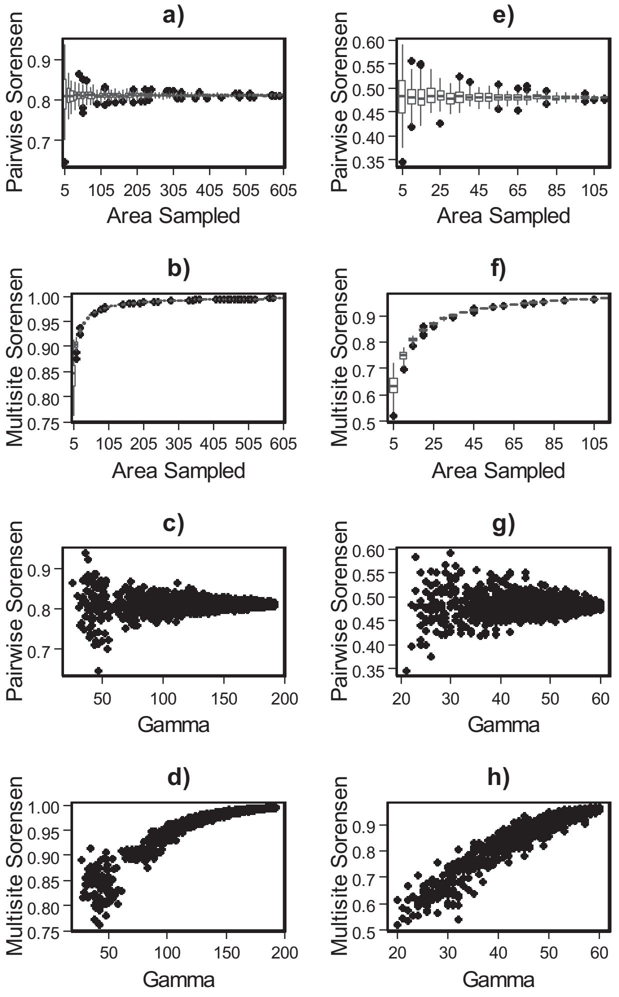

Figure 3 Mean pairwise Sørensen distance versus area sampled (m^2) (a, e), multiple-site Sørensen index versus area sampled (b, f), mean pairwise Sørensen distance versus γ (c, g) and multiple-site Sørensen index versus γ (d, h). Plots (a)–(d) are for a meadow ecosystem and (e)–(h) for an early successional ecosystem.

Our final analysis tested a randomization procedure similar to that used for β_{dev} to detect departures from randomness in the pairwise Sørensen distance. Interestingly, this analysis also created a significant trend with sampling effort and γ in three of four datasets (Figs S16 & S17 in Appendix S2). Moreover, the null pairwise Sørensen distance did not vary predictably with differences in γ. Mean pairwise Sørensen distances in the more diverse meadow ecosystem were higher than those of the early successional ecosystem, but this pattern was reversed in the mean null distances (compare Fig. 3 with Fig. S16 in Appendix S2). Together, these results suggest that the mean pairwise Sørensen distance should be preferred over a null-model Sørensen value when comparing β among regions. However, the relatively large spread of mean Sørensen values at small sample sizes (and low γ) suggests that the sample sizes required to find statistically significant differences among regions are likely much larger than those used in many studies (i.e. more than 25 plots per region).

DISCUSSION

Our mathematical consideration of β measures suggested that pairwise multivariate distances would provide scale-independent estimates of β but that statistical methods were necessary to correct for the dependence of classical β on γ. Our tests confirmed our prediction for the pairwise Sørensen distance and showed that classical β, even using a recently developed correction for variation in γ (β_{dev}; Kraft *et al.*, 2011), failed to correct for scale dependence and instead exhibited systematic bias with sampling effort and γ. Our tests also showed that multiple-site Sørensen indices (Diserud & Ødegaard, 2007; Baselga, 2010) were sensitive to sampling effort and γ. Depending on sampling effort, different predictions could be generated about how β differs among regions using either classical β, β_{dev} or the multiple-site Sørensen indices. An important question, then, is to determine when these metrics that rely on estimated γ are appropriate.

Previous research has challenged the β_{dev} method in particular, suggesting that this measure preserves drivers of β in the null model by preserving species abundance distributions (Qian *et al.*, 2013; Xu *et al.*, 2015), which may themselves result from various community assembly mechanisms (McGill *et al.*, 2007). More fundamentally, our results show that the null model does not remove the dependence of β on γ, as it was designed to do. The result is a metric that changes with sampling effort within a given region and has a different rate of change in different regions. Regions with more species appear to have higher rates of change. This inconsistency is likely to bias comparisons among regions with different numbers of species. The sensitivity of the β_{dev} method to sampling effort and γ suggests that results of previous studies using this method should be viewed with caution.

Although classical β and multiple-site Sørensen indices also showed strong trends with sampling effort, and thus estimated γ, these measures may be appropriate for developing scale-dependent measures of beta diversity among regions. For classical β, the differences among regions will depend on the scaling of SARs in each area – when these scalings can be estimated from the data, a scale-dependent measure of β is possible (see Appendix S2). Although we have not shown a similar mathematical relationship for the multiple-site Sørensen index of Baselga (2010), its strong correlation with classical β suggests that such a relationship may exist. In contrast, the change in β_{dev} with sampling effort follows a different pattern (Fig. 1), which probably reflects the effect of changing species abundance distributions as more sites are included in a region (Xu *et al.*, 2015) as well as the change in the denominator of the metric (Fig. 3). Because sampling-based changes in species abundance distributions differ among regions, a scale-dependent measure of β_{dev} may not be possible.

The pairwise multivariate metrics we tested, the Sørensen, Jaccard and Chao distances, provide measures of β that are invariant to sampling effort. This result was expected from our knowledge of how these metrics are calculated from sampling data – because they represent averages of all pairwise comparisons, they are not expected to change with sample size or γ. However, even though pairwise measures were invariant to sampling effort, the Sørensen-based null-model analogue of β_{dev} (cf. Myers *et al*., 2013) produced results that were sensitive to sampling effort. In other words, the mean pairwise Sørensen distance performs better than a null-model test that uses the Sørensen distance. In addition, unlike rarefaction techniques (e.g. Chao & Jost, 2012), multivariate metrics such as Sørensen distance do not depend on estimates of the number of individuals sampled or sample coverage, but can nonetheless be corrected if species sampling within sample units is incomplete (Chao *et al*., 2005).

While multivariate measures such as the Sørensen distance are not directly dependent on γ, they are, like classical β measures, sensitive to factors driving variation in α, such as spatial aggregation, species abundance distributions and densities of individuals per unit area (Chase & Knight, 2013). For example, a region with greater density of individuals per unit area will tend to have greater mean α than another, even if all species are shared between regions and relative abundances are the same (Gotelli & Colwell, 2001). Indeed, multiple factors including site productivity (and therefore density of individuals), the scale of environmental heterogeneity and species interactions, will influence α. However, these are often considered causal factors that structure both α and β, rather than confounding or nuisance factors (Cadotte & Fukami, 2005; Andrew *et al*., 2012; Germain *et al*., 2013).

A roadmap to comparing β among regions

There has been intense debate on how best to measure β (Tuomisto, 2010; Anderson *et al*., 2011). This debate has advanced theory and interpretations of ecological data, but always with Whittaker's original goal of providing a meaningful and comparable measure of diversity among sites or plots (Whittaker, 1960, 1972). Recent attempts to compare β among regions have introduced new complexities to the study of diversity, as regional differences in species pools and the accumulation of species with area make meaningful contrasts difficult.

Our study demonstrates that no single measure of β is perfect for a between-region comparison, but highlights the strengths and failures of different measures. The β_{dev} method fails due to an unforeseen dependence on γ that appears to bias differences among regions. Moreover, our inability to predict how β_{dev} changes with sampling effort makes it impossible to quantify or remove this bias. Unfortunately, this criticism applies to applications of this null-model approach to other measures of β, suggesting that the goal of comparing departures from randomness among regions cannot be achieved with current methods.

Of the other approaches tested, each is most appropriate for different questions, and we suggest that there are scenarios where they could be used in conjunction. The first approach that we recommend is the use of appropriate pairwise distance metrics. By removing dependence on γ, multivariate metrics such as the Sørensen distance eliminate a key source of uncertainty and can provide reliable measures of β. Pairwise multivariate approaches also offer additional advantages for measuring β that should be considered, such as allowing the contribution of individual sites or species to variation in community composition to be determined (Legendre & De Cáceres, 2013). Similarly, methods have been developed to relate multivariate β to environmental or spatial variables (e.g. Borcard *et al*., 2004; Ferrier *et al*., 2007), and to use multivariate β in conjunction with richness models to predict community composition (Mokany *et al*., 2011). These analyses can offer clues to community assembly mechanisms, although interpreting causality is not straightforward (Gilbert & Bennett, 2010; Anderson *et al*., 2011; Warton *et al*., 2012).

Despite these advantages of pairwise methods, they can diverge from multiple-site metrics and classical β because of their failure to consider co-occurrence in more than two samples at a time (Baselga, 2013). Classical β and multiple-site Sørensen indices can overcome this shortcoming of pairwise methods if their scale dependences can be accounted for. The known scale-dependence of classical β makes this measure the easiest to quantify across sampling efforts so long as SARs can be properly quantified (equations S8 & S9, Fig. S19 in Appendix S2). Although such a measure would be difficult to summarize, it benefits from incorporating estimates of γ beyond that found in two plots. Similarly, if the correct scaling can be modelled, the multiple-site Sørensen index may be used to generate scale-dependent measures of β, as this type of index offers important advantages over the standard pairwise index. In this case, determining how nestedness and turnover vary with scale would need to be explored further. Finally, because classical β and the multiple-site Sørensen index of Baselga (2010) appeared to asymptote at high levels of sample coverage (Fig. 3), areas with sufficient sampling to reach an asymptote may be directly compared with these scale-dependent measures.

Apart from the types of analyses that ecologists should use, our data give clear recommendations about sampling effort. In particular, all metrics considered are highly variable at small sample sizes (fewer than 25 plots in our datasets), and thus have little power if sampling effort is low. As a result, the power of regional comparisons should be considered carefully, as a null result may reflect sampling effort rather than biology. We note that these estimates are also sensitive to variation related to the spatial grain of sampling. Community assembly mechanisms may differ across spatial grains (Whittaker *et al*., 2001; Chase & Knight, 2013), and thus patterns in diversity and species distributions may change considerably as the spatial grain of samples changes (Tuomisto & Ruokolainen, 2012). The power of the analyses would also be likely to change with the spatial grain of samples.

CONCLUSION

The sensitivity of the null-model approach to variations in γ suggests that analogous problems may occur in other areas of community ecology. In particular, several statistical tests used in community ecology rely on randomizations of a species matrix to develop a statistical null model or standardized effect size. The randomization approach is especially important in community ecology because there is no relevant statistical null for species assemblages, and this null is therefore often constructed from randomized data. Tests using this approach range from community phylogenetic statistics (Webb, 2000; Kraft *et al.*, 2007) to null models for species co-occurrence and nestedness (Gotelli & McCabe, 2002; Ulrich & Gotelli, 2007).

The challenge arises when an effect size increases with sampling (numbers of plots or species), and the rate of change in effect size depends on the characteristics of the ecosystem, such as the regional species pool. In these cases, developing standard measures across ecosystems becomes difficult. Although previous research has attempted to statistically control for the importance of sampling effort on the measured effect size of individual tests (e.g. Gotelli & McCabe, 2002), the importance of sampling effort for the measured effect is often highly variable among ecosystems (e.g. Ulrich & Gotelli, 2007). The approaches used in these cases may be valid; however, the results from our exploration of β_{dev} suggest that further study should be conducted to ensure that this is indeed the case.

ACKNOWLEDGEMENTS

This work was supported by the Australian Research Council Centre of Excellence for Environmental Decisions, the Natural Sciences and Engineering Research Council of Canada and the University of Toronto Connaught Fund. We thank John Dwyer, Nathan Kraft and Felipe Micali Nuvoloni for valuable discussions.

REFERENCES

Anderson, M.J., Ellingsen, K.E. & McArdle, B.H. (2006) Multivariate dispersion as a measure of beta diversity. *Ecology Letters*, **9**, 683–693.

Anderson, M.J., Crist, T.O., Chase, J.M., Vellend, M., Inouye, B.D., Freestone, A.L., Sanders, N.J., Cornell, H.V., Comita, L.S., Davies, K.F., Harrison, S.P., Kraft, N.J.B., Stegen, J.C. & Swenson, N.G. (2011) Navigating the multiple meanings of β diversity: a roadmap for the practicing ecologist. *Ecology Letters*, **14**, 19–28.

Andrew, M.E., Wulder, M.A., Coops, N.C. & Baillargeon, G. (2012) Beta-diversity gradients of butterflies along productivity axes. *Global Ecology and Biogeography*, **21**, 352–364.

Baselga, A. (2010) Partitioning the turnover and nestedness components of beta diversity. *Global Ecology and Biogeography*, **19**, 134–143.

Baselga, A. (2013) Multiple site dissimilarity quantifies compositional heterogeneity among several sites, while average pairwise dissimilarity may be misleading. *Ecography*, **36**, 124–128.

Baselga, A. & Orme, C.D.L. (2012) betapart: an R package for the study of beta diversity. *Methods in Ecology and Evolution*, **3**, 808–812.

Baselga, A., Jiménez-Valverde, A. & Niccolini, G. (2007) A multiple-site similarity measure independent of richness. *Biology Letters*, **3**, 642–645.

Bennett, J.R., Cumming, B.F., Ginn, B.K. & Smol, J.P. (2010) Broad-scale environmental response and niche conservatism in lacustrine diatom communities. *Global Ecology and Biogeography*, **19**, 724–732.

Bennett, J.R., Dunwiddie, P.W., Giblin, D.E. & Arcese, P. (2012) Native versus exotic community patterns across three scales: roles of competition, environment and incomplete invasion. *Perspectives in Plant Ecology, Evolution and Systematics*, **14**, 381–392.

Borcard, D., Legendre, P., Avois-Jacquet, C. & Tuomisto, H. (2004) dissecting the spatial structure of ecological data at multiple scales. *Ecology*, **85**, 1826–1832.

Cadotte, M.W. & Fukami, T. (2005) Dispersal, spatial scale, and species diversity in a hierarchically structured experimental landscape. *Ecology Letters*, **8**, 548–557.

Chao, A. & Jost, L. (2012) Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. *Ecology*, **93**, 2533–2547.

Chao, A., Chazdon, R.L., Colwell, R.K. & Shen, T.-J. (2005) A new statistical approach for assessing similarity of species composition with incidence and abundance data. *Ecology Letters*, **8**, 148–159.

Chase, J.M. & Knight, T.M. (2013) Scale-dependent effect sizes of ecological drivers on biodiversity: why standardised sampling is not enough. *Ecology Letters*, **16**, 17–26.

Colwell, R.K. & Coddington, J.A. (1994) Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **345**, 101–118.

Colwell, R.K., Chao, A., Gotelli, N.J., Lin, S.-Y., Mao, C.X., Chazdon, R.L. & Longino, J.T. (2012) Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *Journal of Plant Ecology*, **5**, 3–21.

Crist, T.O., Veech, J.A., Gering, J.C. & Summerville, K.S. (2003) Partitioning species diversity across landscapes and regions: a hierarchical analysis of alpha, beta, and gamma diversity. *The American Naturalist*, **162**, 734–743.

Diserud, O.H. & Ødegaard, F. (2007) A multiple-site similarity measure. *Biology Letters*, **3**, 20–22.

Ferrier, S., Manion, G., Elith, J. & Richardson, K. (2007) Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. *Diversity and Distributions*, **13**, 252–264.

Germain, R.M., Johnson, L., Schneider, S., Cottenie, K., Gillis, E.A. & MacDougall, A.S. (2013) Spatial variability in plant predation determines the strength of stochastic community assembly. *The American Naturalist*, **182**, 169–179.

Gilbert, B. & Bennett, J.R. (2010) Partitioning variation in ecological communities: do the numbers add up? *Journal of Applied Ecology*, **47**, 1071–1082.

Gilbert, B. & Lechowicz, M.J. (2004) Neutrality, niches, and dispersal in a temperate forest understory. *Proceedings of the National Academy of Sciences USA*, **101**, 7651–7656.

Ginn, B.K., Cumming, B.F. & Smol, J.P. (2007) Diatom-based environmental inferences and model comparisons from 494 northeastern North American lakes. *Journal of Phycology*, **43**, 647–661.

Gotelli, N.J. & Colwell, R.K. (2001) Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters*, **4**, 379–391.

Gotelli, N.J. & McCabe, D.J. (2002) Species co-occurrence: a meta-analysis of J. M. Diamond's assembly rules model. *Ecology*, **83**, 2091.

Harms, K.E., Wright, S.J., Calderón, O., Hernández, A. & Herre, E.A. (2000) Pervasive density-dependent recruitment enhances seedling diversity in a tropical forest. *Nature*, **404**, 493–495.

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

Kraft, N.J.B., Comita, L.S., Chase, J.M., Sanders, N.J., Swenson, N.G., Crist, T.O., Stegen, J.C., Vellend, M., Boyle, B., Anderson, M.J., Cornell, H.V., Davies, K.F., Freestone, A.L., Inouye, B.D., Harrison, S.P. & Myers, J.A. (2011) Disentangling the drivers of β diversity along latitudinal and elevational gradients. *Science*, **333**, 1755–1758.

Lande, R. (1996) Statistics and partitioning of species diversity, and similarity among multiple communities. *Oikos*, **76**, 5–13.

Legendre, P. & De Cáceres, M. (2013) Beta diversity as the variance of community data: dissimilarity coefficients and partitioning. *Ecology Letters*, **16**, 951–963.

Legendre, P. & Legendre, L. (2012) *Numerical ecology*, 3rd edn. Elsevier, Amsterdam.

Legendre, P., Borcard, D. & Peres-Neto, P.R. (2005) Analyzing beta diversity: partitioning the spatial variation of community composition data. *Ecological Monographs*, **75**, 435–450.

Lessard, J.-P., Belmaker, J., Myers, J.A., Chase, J.M. & Rahbek, C. (2012) Inferring local ecological processes amid species pool influences. *Trends in Ecology and Evolution*, **27**, 600–607.

McGill, B.J., Etienne, R.S., Gray, J.S., Alonso, D., Anderson, M.J., Benecha, H.K., Dornelas, M., Enquist, B.J., Green, J.L., He, F., Hurlbert, A.H., Magurran, A.E., Marquet, P.A., Maurer, B.A., Ostling, A., Soykan, C.U., Ugland, K.I. & White, E.P. (2007) Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecology Letters*, **10**, 995–1015.

Mokany, K., Harwood, T.D., Overton, J.M., Barker, G.M. & Ferrier, S. (2011) Combining α- and β-diversity models to fill gaps in our knowledge of biodiversity. *Ecology Letters*, **14**, 1043–1051.

Mori, A.S., Shiono, T., Koide, D., Kitagawa, R., Ota, A.T. & Mizumachi, E. (2013) Community assembly processes shape an altitudinal gradient of forest biodiversity. *Global Ecology and Biogeography*, **22**, 878–888.

Myers, J.A., Chase, J.M., Jiménez, I., Jørgensen, P.M., Araujo-Murakami, A., Paniagua-Zambrana, N. & Seidel, R. (2013) Beta-diversity in temperate and tropical forests reflects dissimilar mechanisms of community assembly. *Ecology Letters*, **16**, 151–157.

Novotny, V., Miller, S.E., Hulcr, J., Drew, R.A.I., Basset, Y., Janda, M., Setliff, G.P., Darrow, K., Stewart, A.J.A., Auga, J., Isua, B., Molem, K., Manumbor, M., Tamtiai, E., Mogia, M. & Weiblen, G.D. (2007) Low beta diversity of herbivorous insects in tropical forests. *Nature*, **448**, 692–695.

Olden, J.D. & Rooney, T.P. (2006) on defining and quantifying biotic homogenization. *Global Ecology and Biogeography*, **15**, 113–120.

Qian, H., Wang, X. & Zhang, Y. (2012) Comment on 'Disentangling the drivers of β diversity along latitudinal and elevational gradients'. *Science*, **335**, 1573.

Qian, H., Chen, S., Mao, L. & Ouyang, Z. (2013) Drivers of β-diversity along latitudinal gradients revisited. *Global Ecology and Biogeography*, **22**, 659–670.

Stegen, J.C., Lin, X., Konopka, A.E. & Fredrickson, J.K. (2012) Stochastic and deterministic assembly processes in subsurface microbial communities. *ISME Journal*, **6**, 1653–1664.

Stegen, J.C., Freestone, A.L., Crist, T.O., Anderson, M.J., Chase, J.M., Comita, L.S., Cornell, H.V., Davies, K.F., Harrison, S.P., Hurlbert, A.H., Inouye, B.D., Kraft, N.J.B., Myers, J.A., Sanders, N.J., Swenson, N.G. & Vellend, M. (2013) Stochastic and deterministic drivers of spatial and temporal turnover in breeding bird communities. *Global Ecology and Biogeography*, **22**, 202–212.

Stier, A.C., Geange, S.W., Hanson, K.M. & Bolker, B.M. (2013) Predator density and timing of arrival affect reef fish community assembly. *Ecology*, **94**, 1057–1068.

Tscharntke, T., Steffan-Dewenter, I., Kruess, A. & Thies, C. (2002) Contribution of small habitat fragments to conservation of insect communities of grassland–cropland landscapes. *Ecological Applications*, **12**, 354–363.

Tuomisto, H. (2010) A diversity of beta diversities: straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity. *Ecography*, **33**, 2–22.

Tuomisto, H. & Ruokolainen, K. (2012) Comment on 'Disentangling the drivers of β diversity along latitudinal and elevational gradients'. *Science*, **335**, 1573.

Ulrich, W. & Gotelli, N.J. (2007) Null model analysis of species nestedness patterns. *Ecology*, **88**, 1824–1831.

Warton, D.I., Wright, S.T. & Wang, Y. (2012) Distance-based multivariate analyses confound location and dispersion effects. *Methods in Ecology and Evolution*, **3**, 89–101.

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

Whittaker, R. (1960) Vegetation of the Siskiyou mountains, Oregon and California. *Ecological Monographs*, **30**, 279–338.

Whittaker, R. (1972) Evolution and measurement of species diversity. *Taxon*, **21**, 213–251.

Whittaker, R.J., Willis, K.J. & Field, R. (2001) Scale and species richness: towards a general, hierarchical theory of species diversity. *Journal of Biogeography*, **28**, 453–470.

Xu, W., Chen, G., Liu, C. & Ma, K. (2015) Latitudinal differences in species abundance distributions, rather than spatial aggregation, explain beta-diversity along latitudinal gradients. *Global Ecology and Biogeography*, **24**, 1170–1180.

Zhang, H., Gilbert, B., Zhang, X. & Zhou, S. (2013) Community assembly along a successional gradient in sub-alpine meadows of the Qinghai-Tibetan Plateau, China. *Oikos*, **122**, 952–960.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's web-site.

Appendix S1 References for data and supporting information.
Appendix S2 Additional description of methods and supporting figures.

BIOSKETCHES

Joseph Bennett is an assistant professor in the Institute of Environmental Science and Department of Biology at Carleton University, Ottawa. Joseph's research specializes in conservation biology and spatial ecology.

Benjamin Gilbert is an assistant professor in the Department of Ecology and Evolutionary Biology at the University of Toronto. Benjamin's research specializes in processes that structure species coexistence and diversity, and resulting patterns of diversity.

Editor: Andres Baselga