ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Royal Statistical Society
Statistics in Society
Series A
Journal of the Royal Statistical Society
Statistical Methodology
Series B
Journal of the Royal Statistical Society
Applied Statistics
Series C

Improved detection of changes in species richness in high-diversity microbial communities

Journal:	*Journal of the Royal Statistical Society*
Manuscript ID	JRSS-OA-SC-May-16-0104.R2
Manuscript Type:	Original Article - Series C
Date Submitted by the Author:	04-Nov-2016
Complete List of Authors:	Willis, Amy; Cornell University, Department of Statistical Science Bunge, John; Cornell University, Department of Statistical Science Whitman, Thea; University of Wisconsin Madison, Department of Soil Science
Keywords:	biodiversity, species richness, environmental gradient, microbial ecology, metaanalysis

Improved detection of changes in species richness in high-diversity microbial communities

Amy Willis
Cornell University, USA.
E-mail: adw96@cornell.edu
John Bunge
Cornell University, USA.
Thea Whitman
University of Wisconsin-Madison, USA.

Summary. Biodiversity is important for balance and function of a broad variety of ecosystems, and identifying factors that influence biodiversity can assist environmental management and maintenance. However, low-abundance taxa are often missing from ecosystem surveys. These rare taxa, which may be critical to the ecosystem function, are not accounted for in existing methods for detecting changes in species richness. We introduce a model for total (observed and unobserved) biodiversity that explicitly accounts for these rare taxa. Our method permits rigorous testing for both heterogeneity and biodiversity changes, and simultaneously improves Type I & II error rates compared to existing methods. To estimate model parameters we utilize the well-developed literature of meta-analysis. The problem of substantial low abundance taxa missing from samples is especially pronounced in microbiomes, which are the focus of our case studies.

Keywords: biodiversity, species richness, environmental gradient, microbial ecology, metaanalysis

1. Introduction

The species richness (number of distinct taxonomic groups) of a biological population is commonly used as a marker for ecosystem health (McDonald *et al.*, 2012; Li *et al.*, 2014; Karkman *et al.*, 2011; Lauber *et al.*, 2009; Gao *et al.*, 2013; Dethlefsen *et al.*, 2008). However, species richness is sensitive to changes in the ecosystem. Examples of affectors include temperature, time, biogeochemical conditions and anthropogenic factors (McDonald *et al.*, 2012; Li *et al.*, 2014; Karkman *et al.*, 2011; Lauber *et al.*, 2009; Gao *et al.*, 2013; Dethlefsen *et al.*, 2008). Because understanding the

mechanisms that may incite or accelerate changes in richness is crucial to sustaining ecosystem health, many micro- and macroecologists are interested in formally testing for richness changes in response to one or more covariates.

Ecosystem surveys rarely exhaustively sample the population under study. As a result, large components of biodiversity may be missing from samples. When comparing the same ecosystem longitudinally, or related ecosystems cross-sectionally, the observed species richness in each survey (number of different species observed in the survey) is usually positively correlated with number of samples in this survey. Hence unequal sample sizes amongst the surveys may lead to the false conclusion of higher richness in the larger samples. Thus, in order to compare across samples of different sizes while targeting the true total species richness in the ecosystem from which the sample was drawn, it is essential to account for the number of taxa missing from the samples, as well as the precision in predicting this number. The problem of substantial missing biodiversity is especially pronounced in many microbiomes, such as water, soil, gut and skin (Massana *et al.*, 2015; Fierer and Lennon, 2011; Dethlefsen *et al.*, 2008; Council *et al.*, 2016; Grice *et al.*, 2009).

Here we propose a method for modeling species richness that considers both the observed *and* unobserved members of the population. This allows us to draw conclusions about the population under study, rather than merely about the samples that were observed. A key advantage of the method is that it permits comparison across different sample sizes. Furthermore, rigourous inference regarding the effects of covariates on biodiversity is made possible, and adjustments for simultaneous inferences arise naturally. Finally, it provides the first inferential method for assessing homogeneity of samples with respect to total biodiversity.

Progressing point estimates and standard errors of species richness into a full inferential framework is the key development presented in this manuscript. Another important contribution is a simulation procedure for resampling from microbial ecology datasets, which we use to provide a realistic assessment of the advantages of the method. We demonstrate simultaneous improvements in Type I & II error rates compared to existing methods, and most importantly, re-analyze an influential gut microbiome study to rigorously test the claims of the authors (which were only based on an exploratory analysis). We conclude that a highly significant decrease in species richness of the human gut occurs in response to an antibiotic (concurring with the original paper), but observe a post-treatment recovery of the richness of the ecosystem (in opposition with the original paper). Furthermore, we use the method to analyze a soil field trial survey from a new angle, concluding homogeneity of biological replicates after accounting for missing taxa. Both ecosystems examined

here are characterized by medium-to-high latent diversity. Our simulation studies suggest that our method presents the greatest gains over existing methods in these cases.

We begin the paper by introducing our model for species richness and discussing computation and inference in Section 2, before estimating size and power improvements in Section 3. In Section 4 we apply the method to investigate homogeneity of biological replicates, and analyze a gut microbiome dataset in Section 5. We conclude with some model selection diagnostics, and discuss generalisations and software availability in Section 6.

2. Modelling total species richness

We propose a three-component hierarchical model for estimated species richness. The first component is an additive model, incorporating covariate information that is known or believed to influence true species richness. The second component captures the natural variability in species richness between different environments. The final component is an error term that accounts for the statistical error in estimating richness. The current standard approach to comparing species richness (a linear regression on the observed richness) critically fails to account for the latter component. We discuss our model, its estimation, and diagnostics below.

2.1. Model details

We wish to model the total species richness of m populations. Denote the total richness in the ith population, observed and unobserved, by $C_i, i = 1, \ldots, m$. Also associated with each population is a set of p covariates. We assume that species richness is a function of the covariates, but also a function of pure random variation, so that

$$C_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_p x_{i,p} + u_i,$$

where $x_{i,j}$ is the value of the jth covariate for the ith population, β_j is its coefficient ($j = 1, \ldots, p$), and u_i is a random variable representing the variation in richness not attributed to the covariates. We make the assumption that $u_1, \ldots, u_m$ are independent, identically distributed normal

random variables† with common variance σ_u^2. Additive nonlinear terms may be incorporated through the $x_{i,j}$'s as usual in a regression analysis.

Suppose the goal of the experiment is to investigate which covariates do and do not alter total species richness, or equivalently, which elements of $\beta = (\beta_1, \ldots, \beta_p)^T$ are equal to zero. In order to answer this question, we take a sample of individuals from each of the m populations under study. We do not assume equal sample sizes or sampling depth, or that every taxon in each population was observed. Because we do not assume that every taxon in each population was observed, we do not know C_i exactly for any i: the total species richness is unknown for each of the populations under study. Consequently our inference about the β_j's requires accounting for error in estimation of the C_i's.

Based on each of our samples, we estimate C_i by $\hat{C}_i$ with standard error $\hat{\sigma}_i$. A large number of estimators for species richness have been developed, including under homogeneity models (the ubiquitous Chao1 estimate: Chao (1984)), mixed Poisson models (e.g. Bunge *et al.* (2012)), and non-mixed Poisson (Willis and Bunge, 2015). For estimators based on maximum likelihood or nonlinear regression, central limit theory ensures the asymptotic normality of estimates under the assumption of correct model specification‡ (Willis and Bunge, 2015). We therefore make the broadly reasonable assumption that, conditional on the value of C_i, the estimate $\hat{C}_i$ is normally distributed around C_i with standard deviation σ_i, that is,

$$\hat{C}_i | C_i = C_i + \epsilon_i,$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$, $i = 1, \ldots, m$. Unconditionally we then have the final model

$$\hat{C}_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_p x_{i,p} + u_i + \epsilon_i = \beta_0 + x_i^T \beta + u_i + \epsilon_i, \qquad (1)$$

†Independence is a very weak assumption in the microbial high-throughput sequencing case because most samples will be sequenced distinctly. However, since a primary objective will be to test if $\sigma_u^2 = 0$, the assumption of normality should be verified, and we provide a diagnostic procedure in Section 6.1. Identicality cannot be verified without replicates of every observed covariate combination, however if the other assumptions hold then the only consequence of non-identicality will be an inflated estimate of σ_u^2 and thus the tendency to conclude heterogeneity. In this way, the heterogeneity test can be considered to reflect both heterogeneity of richness *and* its variance: an interesting and unexpected consequence of this formulation.

‡If the species richness estimate under consideration has an asymptotic transformed normal distribution, this transformation should be applied and species richness modelled on the transformed scale. However, we are not aware of any rigourous (i.e. non-heuristic) results justifying this approach for currently available richness estimators.

where $x_i = (x_{i,1}, \ldots, x_{i,p})^T$. Since the only available information about σ_i is the standard error $\hat{\sigma}_i$ we substitute the latter for the former and henceforth refer only to σ_i. The results of our simulation studies (Sections 3.1, 3.2 and 3.3) suggest that this substitution is reasonable; however in Section 3.3 we propose an alternative approach that may improve robustness to this assumption. We refer the reader to Section 6.1 for a brief discussion on species richness estimation and cross-model comparison.

It is important to note that the stochastic nature of the estimated total diversity arises both from $u = (u_1, \ldots, u_m)^T$and $\epsilon = (\epsilon_1, \ldots, \epsilon_m)^T$, that is, through the inherent random variation of the C_i around $\beta_0 + x_i^T\beta$ and through the random variation of $\hat{C}_i$ around C_i. Procedures that model the observed diversity c_i as a linear function of the covariates effectively set $\hat{C}_i = c_i$ but treat $\sigma_i^2 = 0$, thus treating the sample as the population and unobserved diversity as null. As discussed previously, this can cause significant problems when sample sizes differ, because observed diversity correlating with sample size confounds the source of the elevated richness. Furthermore, modeling approaches based on relative frequencies cause problems when a small selection of the community greatly expands: lower relative frequencies of the rare taxa give the impression of lost diversity. However, they may not have been lost to the ecosystem (and the richness may be unchanged) but merely appear less frequently in samples due to the greater abundance of other taxa.

Given model (1) there are two main hypotheses of interest. The first is $H_0 : \sigma_u^2 = 0$; that is, the variation in the true species richnesses across the m populations is wholly attributable to the covariates $x_1, \ldots, x_p$ with no unexplained random variation. This hypothesis is often referred to as that of *homogeneity*. The alternative hypothesis of *heterogeneity*, $H_A : \sigma_u^2 > 0$, supposes that there is more variability in the diversity estimates than can be explained by sampling-based variation in the estimates alone, and that some other mechanism (which we ascribe to the random variables $u_1, \ldots, u_m$) contributes to the observed discordance of the estimated species richnesses. Possible interpretations of heterogeneity include model specification, missing predictors, or true biological heterogeneity between the ecosystems under study. While model misspecification may be diagnosed (see Section 6.1), the distinction between the other options may only be informed by the scientific literature concerning the ecosystem under study. Note that in the absence of covariates, the model simplifies to a weighted linear regression with the estimates weighted by the inverse of their variance estimates.

The second main hypothesis of interest is $H_0 : \beta_1 = \ldots = \beta_p = 0$, or alternatively, that none of covariates explains the variation in richness across populations. The alternative hypothesis is then that at least one of the covariates affects richness. If H_0 is rejected then interest focuses on

the covariates that do influence richness: which β_j are nonzero and what are their magnitudes? The relevant null hypothesis for the case of one variable is then $H_0 : \beta_j = 0$. Note that the usual regression interpretation of the coefficients applies and that β_j is the expected increase in the true diversity of any of the i populations for a one unit increase in $x_{i,j}$.

2.2. Computation and optimization

We now discuss estimation of the model parameters β and σ_u^2, and implementation of the stated hypothesis tests. The log-likelihood of our model is

$$l(\beta_0, \beta, \sigma_u^2 | x_1, \ldots, x_m, \hat{C}_1, \ldots, \hat{C}_m, \sigma_1^2, \ldots, \sigma_m^2) = -\frac{1}{2} \sum_{i=1}^{m} \left[\ln\left(\sigma_u^2 + \sigma_i^2\right) + \frac{(\hat{C}_i - \beta_0 - x_i^T \beta)^2}{\sigma_u^2 + \sigma_i^2} \right].$$

Maximum likelihood (ML) is a natural choice of parameter estimation technique due to its many asymptotic and finite sample optimality properties in standard settings (Casella and Berger, 2002; Godambe, 1960). However, in this application the choice to use ML is non-trivial because of the *boundary problem*: $\sigma_u^2 \geq 0$. This problem was studied by Crainiceanu and Ruppert (2004), who demonstrate the failure of the usual likelihood ratio test asymptotics when testing $\sigma_u^2 = 0$ against $\sigma_u^2 > 0$.

Fortunately, we can exploit the well-developed literature on meta-analysis to resolve these difficulties. Meta-analyses arise in many social and health sciences where a researcher wishes to pool a number of different studies to determine the presence of an overall effect. Each richness estimate fulfills the role of a study's effect estimate, the standard error of the richness estimate fulfills the role of the standard error of the effect estimate, and the m samples reflect m different studies to be pooled. A comprehensive treatment of meta-analyses is given by Demidenko (2004), who discusses both restricted maximum likelihood algorithms and also the best choice of hypothesis test in this nonstandard boundary case. We note also that in species richness comparison, as with meta-analyses, we only know the standard error in the estimates $\hat{\sigma}_i$ and not the true standard deviations σ_i. For this reason we base our choice of asymptotics on the results of Demidenko (2004) rather than those of Crainiceanu and Ruppert (2004). Thus our restricted ML procedure maximizes

$$l_R(\beta_0, \beta, \sigma_u^2) = -\frac{1}{2} \left\{ \sum_{i=1}^{m} \left[\ln\left(\sigma_u^2 + \sigma_i^2\right) + \frac{(\hat{C}_i - \beta_0 - x_i^T \beta)^2}{\sigma_u^2 + \sigma_i^2} \right] + \ln \left| \sum_{i=1}^{m} \frac{1 + x_i^T x_i}{\sigma_u^2 + \sigma_j^2} \right| \right\},$$

and we denote the maximizing values by $\hat{\beta}_0$, $\hat{\beta}$ and $\hat{\sigma}_u^2$. Unfortunately there do not exist closed form expressions for the estimates but we find that the range-restricted variable metric algorithm of Byrd *et al.* (1995) is a fast and stable maximization algorithm for our restricted likelihood. Our investigations suggest that the least squares estimates of (β_0, β) obtained by regressing the covariates on the richness estimates $\hat{C}_i$ are reasonable starting values for $(\hat{\beta}_0, \hat{\beta})$, and the empirical variance in the estimates $\hat{C}_i$ is a reasonable starting value for $\hat{\sigma}^2$.

2.3. Inference

Because there are no boundary complications for β, its hypothesis testing falls in the standard Wald-type framework. Inverting second derivatives of the restricted log-likelihood gives the variance estimate

$$\hat{\text{Var}}(\hat{\beta}) = (\mathbf{X}^T \hat{\mathbf{W}}^{-1} \mathbf{X})^{-1},$$

where $\mathbf{X} = (x_1^T, \ldots, x_m^T)^T$ and $\hat{\mathbf{W}} = \text{diag}(\hat{\sigma}_1^2 + \hat{\sigma}_u^2, \ldots, \hat{\sigma}_m^2 + \hat{\sigma}_u^2)$, which we use to make marginal inference about the effect of each predictor on species richness via the test statistic $\frac{\hat{\beta}_i}{\sqrt{[\hat{\text{Var}}(\hat{\beta})]_{ii}}}$, which is distributed approximately $\mathcal{N}(0,1)$. The global test of $H_0 : \beta_1 = \ldots = \beta_p = 0$ has test statistic

$$\hat{\beta}^T \mathbf{X}^T \hat{\mathbf{W}}^{-1} \mathbf{X} \hat{\beta},$$

which is distributed asymptotically according to a χ_p^2 distribution. Finally, we define our Q-statistic as

$$Q = \sum_{i=1}^{m} \frac{(\hat{C}_i - \hat{\beta}_0 - x_i^T \hat{\beta})^2}{\hat{\sigma}_i^2}.$$

Under the null hypothesis of homogeneity, Q follows a χ^2 distribution with $m - p - 1$ degrees of freedom.

3. Estimates of improvements to size and power

The proposed procedure, which we name `betta`, presents advantages in both Type I and Type II error rates in comparison to regression procedures, which we observe under simulation. The simulation methods underpinning both the size and power estimates (which we also believe to be a novel contribution) were designed to reflect data structures observed in microbial settings, and thus are intended to be realistic estimates of the method's advantages. The method's performance under negative binomial simulation structures is not shown here due to the limited ability of

this distribution to reflect microbial data structures. Furthermore, we examine two important questions in microbial ecology using the procedure: heterogeneity of soil communities, and dynamics of microbial communities in the human gut in response to an antibiotic. In the first instance, our analysis provides new insights on the community in question, and in the second instance we apply our method to rigorously confirm the conclusions of the original study by Dethlefsen *et al.* (2008).

3.1. Size of covariate tests

In order to compare the Type I error rate (statistical *size*) for the covariate test under a realistic high-diversity data structure, we redraw samples according to the distribution of operational taxonomic units (OTUs: the generalization of the concept of "species" to microbial ecology) in actual microbial datasets. We set up the simulation in the following way: choose an observed OTU table, and for each OTU in that sample, ascribe it the cell probability according to its relative abundance in the sample. Then, to mimic the differing sample sizes (numbers of observed OTUs) that high-throughput sequencing generates, we randomly choose a sample size based on the distribution of sample sizes across all OTU tables in the study. We then draw this number of samples from a multinomial distribution with the cell probabilities described above, use these draws to construct a frequency count table, estimate the total richness and its standard deviation, and calculate the sample richness. We repeat this 20,000 times, each time choosing a different sample size, and drawing a new multinomial sample of this sample size. In this way, we mimic two key features of microbial datasets: the relative distribution of both rare and common OTUs, and the differing numbers of OTUs observed in successive samples. Note that the actual population richness is irrelevant; resampling from an unchanging distribution is sufficient to evaluate the size of the test.

Armed with realistic redraws reflecting microbial datasets, we partition the 20,000 sets of richness estimates into 2,000 samples of 10 replicates, and create a covariate unrelated to richness for which to test for falsely significant relationships. That is, for each replicate we randomly ascribe a value across the grid $\{10, 20, \ldots, 100\}$, such that every covariate is ascribed exactly one frequency count table. We then compare `betta` against a simple linear regression of the observed richness (denoted c_i for sample i) on the covariate, which is the most common method in the literature for evaluating changes in richness (Li *et al.*, 2014; Sun *et al.*, 2011; Lauber *et al.*, 2009; Dethlefsen *et al.*, 2008; Bordes *et al.*, 2011; Gao

Table 1. Empirical Type I error rates for significance levels of $\alpha = (0.01, 0.05, 0.10)$ for the $\beta_1 = 0$ for 20,000 homogeneous redraws from the Whitman *et al.* (2016) dataset, and 8,000 from the Dethlefsen *et al.* (2008) dataset, each partitioned into samples of 10 and 20 replicates

Dataset	Model	$n = 10$	$n =$
Whitman *et al.* (2016)	`betta` (`breakaway`)	(0.016, 0.028, 0.039)	(0.006, 0.0
Whitman *et al.* (2016)	Regression on c	(0.006, 0.048, 0.093)	(0.013, 0.0
Dethlefsen *et al.* (2008)	`betta` (`CatchAll`)	(0.017, 0.063, 0.098)	(0.021, 0.0
Dethlefsen *et al.* (2008)	Regression on c	(0.016, 0.058, 0.100)	(0.011, 0.0

et al., 2013)§. To compare the effect of more replicates, we also consider the partition of 1,000 samples of 20 replicates modeled across the grid $\{5, 10, \ldots, 100\}$. In both cases, we compare the performance of `betta` in modeling a richness estimate suited to the resampled data structure: `breakaway` for high-diversity cases and `CatchAll` for medium-diversity cases; see Section 6.1.

The Type I error rates for resamples from a particular sample of the Whitman *et al.* (2016) dataset (see Supplementary Materials) are shown in Table 1. For each partition and richness measure, we show the empirical error rate (proportion of null hypotheses determined significant) for levels of significance of 1%, 5% and 10%, noting that we expect less than 1%, 5% and 10% Type 1 error rates. For 5 out of 6 combinations, `betta` has lower Type I error rates compared to the regression procedures, with the error rates consistently halved. The improvement is greatest for less stringent thresholds: for a 10% level of significance `betta` reduces the Type I error rate by a factor of 5.

The analogous table for the Dethlefsen *et al.* (2008) dataset is shown in Table 1, though in accordance with the experimental design of this dataset, the covariate under assessment is categorical rather than continuous (for each partition, half of the samples are assigned Category A, and the other half category B). In this case, we see that `betta` remains either accurate or conservative: it consistently maintains a Type I error rate equal to or less than claimed. The improvement is less pronounced than in the high-diversity case of Table 1, and ranges between a small loss in size and a small improvement in size when compared with the regression procedure. Repeating the simulation with the continuous and discrete covariates reversed for the two datasets suggests that the differences between the two datasets shown in Table 1 are due to the differing data structures (that is, high versus medium rare diversity observed in the samples) and not to testing a continuous versus categorical covariate (see Supplementary Appendix). We conclude overall improvements in the

§Another common approach is a regression on the Chao1 index (Newsham *et al.*, 2015; Timling *et al.*, 2014).

Type I error rate for our method, with the improvement most pronounced in the high-diversity case and more stringent levels of significance.

Before proceeding to compare Type II error rates, we emphasise that an appropriate richness estimation method is essential to the performance of `betta`. Choosing an overly restrictive estimator in a high-diversity case (e.g. `CatchAll` or Chao1) leads to artificially small standard errors due to model misspecification. In this case, `betta` has reduced ability to detect no relationship between a covariate and richness (H_0), because the variability in richness estimates is falsely deflated. In the same way, choosing a highly flexible model in a medium or low-diversity case leads to reduced ability to detect a true relationship between a covariate and richness (H_A), because an inflated measure of variability overwhelms the true richness differences. It is for this reason that the above analyses were only conducted using richness estimators appropriate to the data structure. The authors maintain transparency with respect to the poor performance of `betta` when modeling inappropriate estimates of richness and encourage practitioners to utilise the recommendations in the literature and in Section 6.1 with respect to the richness estimate appropriate to their data structure.

3.2. Power of covariate tests

In order to examine the ability of `betta` to detect true changes in richness (statistical *power*), we must introduce a richness gradient into the model. To maintain the realistic data structure of the size simulations, and to reflect that by its nature biodiversity loss almost always affects rare species (Chapin III *et al.*, 2000), we introduce this gradient along the rare species.

For the Whitman *et al.* (2016) dataset examined in the previous section, we create 1% more multinomial categories and assign them each the same relative weight as the OTUs observed as singletons in the original data set (effectively creating 1% more rare species). We ascribe this sample the covariate $x_{1,1} = 1$. We repeat this for 2% more multinomial categories for $x_{1,2} = 2$, and so forth up to 20% more multinomial categories with $x_{1,20} = 20$. We repeat this 1,000 times to have 1,000 datasets with richness gradients and thus 1,000 p-values for which to assess the power of the test that $\beta_1 = 0$, and repeat with halving the partition to have 2,000 datasets with $n = 10$ each. The results, shown in Table 2, show that `betta` is capable of enormous improvements in Type II error rate (the complement of power), the advantage being the most pronounced when the sample size is small and the desired Type I error rate is low. The improvement ranges between 1.8-fold to 18-fold improvement in power.

To evaluate the power in the medium-diversity case, and with a cate-

Table 2. Empirical power (rate of correct H_A detection) at significance levels of $\alpha = (0.01,$ null hypothesis of $\beta_1 = 0$ for inhomogeneous resamples mimicking the cell probabilities of the dataset, with a continuous richness gradient introduced. The same is shown for the Dethlefse with a 10% increase in number of species.

Dataset	Model	$n = 10$	$n =$
Whitman *et al.* (2016)	`betta` (`breakaway`)	(0.210, 0.295, 0.362)	(0.303, 0.5(
Whitman *et al.* (2016)	Regression on c	(0.012, 0.076, 0.133)	(0.092, 0.2
Dethlefsen *et al.* (2008)	`betta` (`CatchAll`)	(0.926, 0.950, 0.959)	(0.958, 0.9
Dethlefsen *et al.* (2008)	Regression on c	(0.044, 0.143, 0.230)	(0.094, 0.2(

gorical covariate, we resample 20,000 datasets with 10% more rare multinomial categories than the Dethlefsen *et al.* (2008) dataset resampled in Section 3.1, and model the richnesses of 10 of the original resamples and 10 of the higher richness resamples, investigating the significance of the difference according to `betta` and to a regression on the observed richness. We note that even for this small increase in richness `betta` is extremely powerful, with greater than 92% power to detect the change in richness. In comparison, regression methods never exceed 39% power under this sampling scenario.

3.3. Size and power of the homogeneity test

We now turn our attention to evaluating the homogeneity test with respect to size and power. Note that no method for homogeneity determination of true species richness exists in the literature, and thus comparisons similar to Sections 3.1 and 3.2 are not possible for this test. As a result, we focus only on the Type I and II error rates. For the same size and power resamples from Sections 3.1 and 3.2, we evaluate the Type I and Type II error rates for the null hypothesis that the samples are homogeneous with respect to richness in Table 3. We observe that for the high-diversity dataset of Whitman *et al.* (2016), the Type I error rate of the test is conservative, that is, for an α-level test we observe less than an α rate of error. Furthermore, for this data structure the power is very high: 71% for $\alpha = 0.01$ with 10 data points and 99% for 20 data points. For the medium-diversity datasets of Dethlefsen *et al.* (2008), the power is even higher: 77% for $\alpha = 0.01$ with 10 data points. However, this is at the expense of size, with Type I error rates up to double than controlled for. The explanation for this arises not from the richness comparison method, but from the richness estimation method. For the homogeneous replicates, the average of the standard errors should match the standard deviation of the estimates (by definition). For the Dethlefsen *et al.* (2008) dataset, the mean absolute deviation of the estimates was 64.34 while the median of the standard errors was 58.20. Thus in this case `CatchAll` understates

Table 3. Empirical size and power of the homogeneity test of `betta`. The size estimates were from the same sampling scheme as in Section 3.1, and the power estimates as in Section covariate information was modeled in this simulation. Each cell shows results for significance $\alpha = (0.01, 0.05, 0.10)$.

Dataset	$n = 10$	$n = 20$
Size: Whitman *et al.* (2016)	(0.005, 0.006, 0.007)	(0.003, 0.003, 0.003)
Size: Dethlefsen *et al.* (2008)	(0.074, 0.136, 0.184)	(0.126, 0.183, 0.244)
Power: Whitman *et al.* (2016)	(0.709, 0.735, 0.755)	(0.990, 0.994, 0.996)
Power: Dethlefsen *et al.* (2008)	(0.774, 0.895, 0.935)	(0.968, 0.993, 0.997)

the true variability of its richness estimates, leading to inflated confidence and thus inflated risk of false determination of non-homogeneity. By comparison, for the Whitman *et al.* (2016) dataset, `breakaway` overstates its standard errors, with a median absolute deviation of 126.24 but a median standard error determination of 189.10. This comparison provides a full explanation for the tendency of the method to favor H_0 (homogeneity) under `breakaway` and to favor H_A under `CatchAll`. We advise practitioners to consider the results of the simulations carefully in conjunction with their own analysis of homogeneity determinations in the medium-diversity case. In particular, in order to maintain a Type I error rate of 7.4%, we recommend rejecting at $\alpha = 0.01$ when approximately 10 data points were obtained. Alternatively, the practitioner could resample from one of their own data structures in the same way as was performed above to find the appropriate α level for their number of samples and desired Type I error rate. The `R` code used to generate Table 3 is available (see [companion data files]) to facilitate such investigations.

Biological replicates are ideal for estimating the standard deviations of richness estimates because they incorporate variability due to both environmental and sequencing sources (Willis, 2016). Technical (sequencing) replicates, while only dealing with the latter source of variability, can also provide some information on estimate variability. We encourage the use of true, biological replication to empirically confirm standard errors. However, this may not be practical nor even possible under some experimental designs. In these cases, if the practitioner is skeptical of the standard errors produced by the richness procedure, a parametric bootstrap approach can provide some information on its plausibility. Resampling from a multinomial distribution with cell probabilities equal to the sample's empirical taxa weights and passing the generated frequency count tables to the same estimator should give a collection of estimates whose standard deviation is close to the standard error of the original sample's richness estimate. Bootstrap theory gives us that this procedure should underestimate the standard error: a larger standard deviation of this collection compared to the standard error of the original sample suggests

that errors are underestimated, and that the hypothesis of heterogeneity may be favored. The standard deviation of the resamples may be substituted for the standard error to correct for this.

4. Application to homogeneity of soil communities

To illustrate our test for species richness homogeneity of a highly diverse microbial environment, we investigate true replicates from a soil field trial. Soil microbial communities are perhaps the most species-rich of all studied environments on Earth (Fierer and Lennon, 2011). Housing complex interfaces between the hydrosphere, atmosphere, lithosphere, and biosphere, soils exhibit extreme microscale heterogeneity in potential microbial habitats (Nunan *et al.*, 2002; Totsche *et al.*, 2010), which may support the persistence of microbial species diversity (Lozupone and Knight, 2007). The complexity of these communities poses considerable challenges for diversity analysis and thus provide an interesting test case for the homogeneity hypothesis.

Whitman *et al.* (2016) extracted, amplified, and sequenced (Illumina MiSeq) bacterial 16S DNA with soils from a field trial with no amendments, with pyrogenic organic matter additions, and with fresh biomass additions. We analyze field replicates within different plots of the same field (under an randomized complete block design) that were sampled within the hour and without amendments to the soil. For each Day 1 sample with no amendment, `breakaway` (Willis and Bunge, 2014) was used to estimate the total microbial OTU richness in the soil due to the high-diversity nature of the data (singleton dominance). Confidence intervals for the estimates may be seen in Figure 1. The `breakaway` algorithm failed to converge for one sample (Sample S026; see Supplementary Data), which was thus excluded from the analysis.

Because no (measured) covariates characterize differences between the samples, we fit an intercept-only model for estimates of species richnesses. Our method fails to reject the hypothesis that species richness is homogeneous between samples ($p = 0.169$). Note that if only observed richness is considered, the samples appear to have different richnesses (Figure 1), though no inferential method for determining this was previously available. Thus by accounting for the high variability across the samples we note that the samples are not distinct after the taxa that eluded detection are considered.

5. Application to richness changes in the human gut

We now demonstrate the method's applicability in determining the effect of antibiotics on gut microbiome richness. Dethlefsen *et al.* (2008) em-

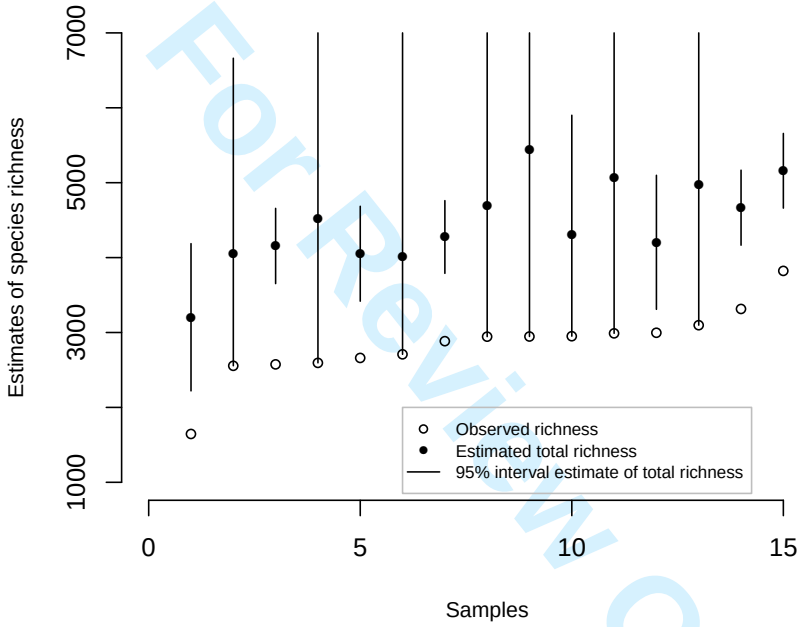

Fig. 1. Here we observe estimates of species richness for 15 biological replicates (no covariate effects) from the Whitman *et al.* dataset. Open symbols indicate observed richness and closed symbols indicate estimated total richness with 95% confidence intervals. This ordering of the replicates (increasing with the value of observed richness) suggests inconsistent levels of biodiversity across samples when only observed richness is considered. However, when total richness is considered along with its standard errors, there is insufficient evidence reject the null hypothesis of richness homogeneity ($p = 0.169$).

ployed pyrosequencing to obtain rRNA sequences from the guts of three human subjects before, during, and after a course of ciprofloxacin. They observed that the treatment led to an overall decrease in the observed richness of the microbiota communities but were unable to test this formally. `CatchAll` (Bunge *et al.*, 2012) was used to estimate the total microbial OTU richness due to the doubleton and tripleton dominance. No outliers were excluded. We fit our model for richness estimates with fixed treatment (pre-treatment, during treatment and post-treatment) and random patient effects (Figure 2) using the implementation `betta_random`. We conclude that treatment is highly significant in decreasing richness ($p = 0.027$), reducing richness by 494 species on average. However, we find that there is no significant post-treatment effect ($p = 0.955$) and that richness recovers to pre-treatment levels after 4 weeks. This concurs with the visual conclusions of Dethlefsen *et al.* (2008), but we emphasize that this methodology provides a formal approach to testing their hypotheses. In contrast, a mixed effects linear regression on the observed richness suggests that only 351 species are lost due to the antibiotic, and that a difference can be concluded with far more confidence ($p = 0.0004$). This highlights that failing to account for uncertainty in species richness estimation leads to overstated confidence in tests for covariate influences. Note that within-patient heterogeneity of the gut microbiome (Davenport *et al.*, 2014; Lu *et al.*, 2014; Wu *et al.*, 2011) may be observed in Figure 2.

6. Discussion

We conclude that for high-diversity datasets, the greatest gain of the proposed methodology is with respect to its ability to correctly detect no change in richness, where the Type I error rate of the method is up to 5 times lower than the only currently available method. This is simultaneously achieved with power improvements, usually by a factor of around 3. Conversely, for medium-diversity datasets, the greatest gain is with respect to detection of true changes in richness, where for even small changes in richness the power can exceed 92%. Similarly, this is achieved with little to no loss in size. Both of these results are highly intuitive: large numbers of latent species destabilize richness estimates, thus increasing their variability and standard errors. When this variability is accounted for, richness differences that are small compared to the precision in estimation are correctly detected as attributable to sampling variability. When variability in estimation is small, it is easier to detect changes, and the improvement in power is attributable to corrections for differing sample sizes across the covariate gradient. Overall it is clear that depending on the data structure either size or power can be greatly im-

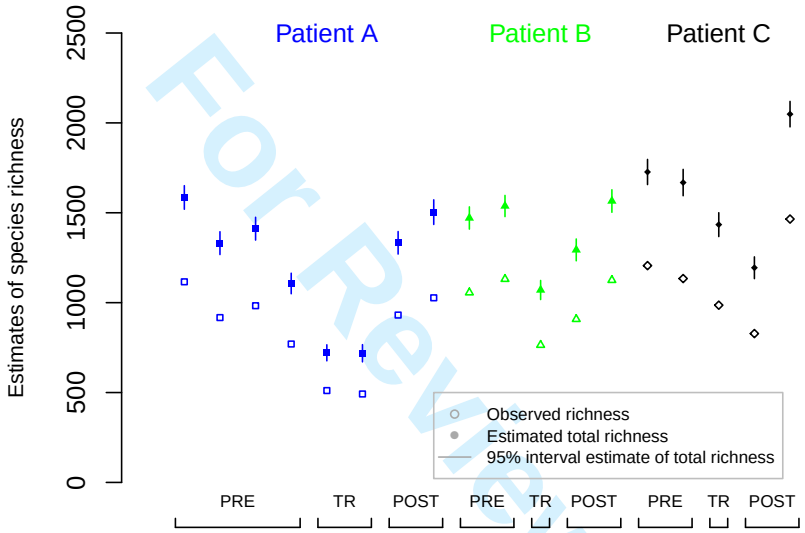

Fig. 2. We observe estimates of species richness of the gut microbiomes of three different human subjects before (PRE), during (TR) and after (POST) a course of ciprofloxacin (Dethlefsen *et al.*, 2008). Each shape indicates a different patient, with open symbols indicating observed richness and closed symbols indicating estimated total richness. We conclude that the drug reduces gut richness by 494 taxa on average ($p = 0.027$), however there are no longer significant differences in pre- and post- treatment levels ($p = 0.955$) after 4 weeks. By comparison, a mixed effects regression on observed species richness implies an average loss of only 351 taxa, but much greater confidence in the difference ($p = 0.0004$) due to the omission of uncertainty in estimation.

proved with almost no loss to the other. Furthermore, we have proposed a first formal test for homogeneity. It is particularly useful for comparing biological or technical replicates, and can be used to assess whether the experimental and computational procedures from sampling to final data output give consistent results.

Nevertheless, due to the hierarchical nature of the model, a number of assumptions must be verified in order to produce valid results. The choice of species richness estimator, and perhaps more importantly the estimate of its standard deviation, is the most serious modeling choice in this method. Too much confidence in the precision of the richness estimate can induce false heterogeneity. Furthermore, over fitting, under fitting, or an inappropriate covariance structure across the samples can also compromise the validity of conclusions. Some practical guidelines are available below. We conclude with some comments of the potential applicability of the method to modeling other α-diversity or community composition metrics, and statistical developments necessary for this generalization.

6.1. Model selection and diagnostics

The methodology is sensitive to the design matrix $\mathbf{X}$, and method of estimating C. Perusal of the richness estimates, and more importantly, their standard errors, is essential to ensure that the model is not overfit (with respect to predictors) and heterogeneity is not falsely concluded. One exploratory approach to diagnosing possible outliers and points of influence is to plot the estimated richness with error bars at ± 2 standard errors. This technique derives from, and is limited by, the assumption that estimated richnesses are normally distributed around the true richness with standard deviation equal to the estimated standard error. The visual diagnostic for a point of influence is a tight interval (small estimated error), especially one centered far from the overall mean. While it is tempting to notice the large intervals in this type of plot, in fact these types of points do not exert a large influence on the model because their variability is captured in the large local error σ_i^2 rather than affecting the estimate of the global error σ_u^2.

While visual diagnostics of this nature can assist with model selection and formulating appropriate hypotheses, the graphical procedure described above suffers ofrom the problem of simultaneous inference (when performing multiple α-level tests, the probability of making at least one Type 1 error is greater than α). For this reason, "testing" multiple 95% confidence intervals for overlap has an exaggerated probability of Type 1 errors. For this reason we advocate the mixed model procedure described above for inference, which does not require multiple testing corrections,

though the exploratory confidence intervals may assist with diagnosing model misspecification in either the richness estimates or the predictors.

Some guidelines for appropriate richness estimators are available Bunge *et al.* (2014), though the authors consistently find that for high-diversity settings (the authors propose singleton-to-doubleton ratios of 1.5 or higher), the `breakaway` estimator (Willis and Bunge, 2015) functions best with respect to plausibility of both estimates and standard errors. In medium- and low-diversity settings (singleton-to-doubleton ratios of 1.5 or lower), `CatchAll` (Bunge *et al.*, 2012) functions well due to its stability, though as discussed below, its standard errors appear conservative in some medium-diversity settings. Note that in microbiome studies, the extent to which rare reads are discarded heavily affects richness estimates, and robustness of results to the quality control parameters should be thoroughly investigated. Furthermore, since richness estimates are often highly sensitive to the chosen model, we encourage `betta`'s users to only compare estimates obtained from the same estimation procedure.

Finally, one assumption of our model is the normality of the u_i's. Under this assumption, the distribution of $\left\{\frac{\hat{C}_i - \hat{\beta}_0 - x_i^T\hat{\beta}}{\hat{\sigma}_i}\right\}_{i=1}^{n}$ should be approximately normal. Thus perusal of histograms and qqplots of these values should display approximate symmetry and no large outliers. Note that inferential tests for normality generally have poor power, and thus visual diagnostics should also be utilized.

6.2. Generalization to other diversity indices

While outside of the scope of this paper, we wish to briefly discuss the generalization of this method to other diversity indices. Any one-dimensional summary statistic which estimates a population parameter and provides an estimate of its standard deviation can be modeled using the methodology described above. The asymptotics of the tests remain valid provided the distribution of the estimate is approximately normal around the true parameter with standard deviation close to its standard error. The main difficulty with extending this type of analysis to evenness indices is that standard errors in the estimates are rarely available, and bootstrap errors understate true sampling variability (Kulesa *et al.*, 2015). There is emerging statistical research on estimators and error estimates for evenness indices (Zhang and Zhou, 2010; Zhang, 2012), and the authors hope that as the literature develops further the same analysis will be possible for a broad variety of diversity measures of interest.

6.3. Integrating richness estimation and comparison

The approach of this paper was to estimate species richness for each sample, and then compare the estimates and uncertainties across covariate gradients. A potential avenue for future research would be to integrate the covariates into species richness estimation, and then assess differences in richness by assessing differences in the estimated model parameters. Developing this approach and comparing it to that presented here is an interesting question but beyond the scope of this paper. We hypothesize that the increased stability would likely come at the expense of computational efficiency, because the dimension of the necessary parameter space is substantially larger. A Bayesian approach to investigating covariate effects on diversity indices was recently investigated by Arbel *et al.* (2016), though the extension to incorporate unobserved species is nontrivial and we leave it to future research.

6.4. Closing remarks

This method is the first inferential procedure for investigating homogeneity and response to covariates of species richness that focuses on the target of interest (population richness) rather than sample richness. This eliminates issues arising from different sample sizes or sampling depths, because precision is already reflected in the standard errors of the richness estimates. Allowing broad comparisons across microbial communities sampled to different depths, this procedure is capable of demonstrating factors affecting biodiversity and illuminating the presence or absence of heterogeneity across different ecosystems and processing pipelines. Parallels with meta-analysis were used to inform the parameter estimation procedure, but the key innovation of the paper was accounting for species richness estimate randomness rather than modifications to the meta-analysis framework.

The methodology, called `betta`, is available from CRAN via the `R` package `breakaway`. The random effects implementation is called `betta_random`. Sample workflows are available as Supplementary Material. Inquiries and extension requests are welcomed and should be directed to the corresponding author.

Acknowledgements

The authors are very grateful to two anonymous referees and the associate editor for many helpful suggestions that improved both the structure and exposition of the manuscript.

References

Arbel, J., Mengersen, K., and Rousseau, J. (2016) Bayesian nonparametric dependent model for partially replicated data: the influence of fuel spills on species diversity. *Annals of Applied Statistics*, In Press.

Bordes, F., Guégan, J. F., and Morand, S. (2011). Microparasite species richness in rodents is higher at lower latitudes and is associated with reduced litter size. *Oikos*, **120**(12), 1889–1896.

Bunge, J., Woodard, L., Böhning, D., Foster, J. A., Connolly, S., and Allen, H. K. (2012). Estimating population diversity with CatchAll. *Bioinformatics*, **28**(7), 1045–1047.

Bunge, J., Willis, A., and Walsh, F. (2014). Estimating the number of species in microbial diversity studies. *Annual Review of Statistics and Its Application*, **1**.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, **16**(5), 1190–1208.

Casella, G. and Berger, R. L. (2002). *Statistical inference*, volume 2. Duxbury Pacific Grove, CA.

Chao, A. (1984). Nonparametric Estimation of the Number of Classes in a Population. Scandinavian Journal of Statistics, **11**(4), 265–270.

Chapin III, F. S., Zavaleta, E. S., Eviner, V. T., Naylor, R. L., Vitousek, P. M., Reynolds, H. L., Hooper, D. U., Lavorel, S., Sala, O. E., Hobbie, S. E., *et al.* (2000). Consequences of changing biodiversity. *Nature*, **405**(6783), 234–242.

Council, S. E., Savage, A. M., Urban, J. M., Ehlers, M. E., Skene, J. P., Platt, M. L., Dunn, R. R., and Horvath, J. E. (2016). Diversity and evolution of the primate skin microbiome. In *Proceedings of the Royal Society B*, volume 283. The Royal Society.

Crainiceanu, C. M. and Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **66**(1), 165–185.

Davenport, E. R., Mizrahi-Man, O., Michelini, K., Barreiro, L. B., Ober, C., and Gilad, Y. (2014). Seasonal variation in human gut microbiome composition. *PloS One*, **9**(3), e90731.

Demidenko, E. (2004). *Mixed Models Theory and Applications*. Wiley-Interscience.

Dethlefsen, L., Huse, S., Sogin, M. L., and Relman, D. A. (2008). The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16s rRNA sequencing. *PLoS Biology*, **6**(11).

Fierer, N. and Lennon, J. T. (2011). The generation and maintenance of diversity in microbial communities. *American Journal of Botany*, **98**(3), 439–448.

Fisher, R. A., Corbet, S., and Williams, C. B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology*, **12**, 42–58.

Gao, W., Weng, J., Gao, Y., and Chen, X. (2013). Comparison of the vaginal microbiota diversity of women with and without human papillomavirus infection: a cross-sectional study. *BMC Infectious Diseases*, **13**(1), 271.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics*, pages 1208–1211.

Grice, E. A., Kong, H. H., Conlan, S., Deming, C. B., Davis, J., Young, A. C., Bouffard, G. G., Blakesley, R. W., Murray, P. R., Green, E. D., *et al.* (2009). Topographical and temporal diversity of the human skin microbiome. *Science*, **324**(5931), 1190–1192.

Karkman, A., Mattila, K., Tamminen, M., and Virta, M. (2011). Cold temperature decreases bacterial species richness in nitrogen-removing bioreactors treating inorganic mine waters. *Biotechnology and Bioengineering*, **108**(12), 2876–2883.

Kulesa, A., Krzywinski, M., Blainey, P., and Altman, N. (2015). Sampling distributions and the bootstrap: The bootstrap can be used to assess uncertainty of sample estimates. *Nature methods*, **12**(6), 477–478.

Lauber, C. L., Hamady, M., Knight, R., and Fierer, N. (2009). Soil pH as a predictor of soil bacterial community structure at the continental scale: a pyrosequencing-based assessment. *Applied and Environmental Microbiology*.

Li, F., Kwon, Y.-S., Bae, M.-J., Chung, N., Kwon, T.-S., and Park, Y.-S. (2014). Potential impacts of global warming on the diversity and distribution of stream insects in South Korea. *Conservation Biology*, **28**(2), 498–508.

Lozupone, C. A. and Knight, R. (2007). Global patterns in bacterial diversity. *Proceedings of the National Academy of Sciences*, **104**(27), 11436–11440.

Lu, H.-P., Lai, Y.-C., Huang, S.-W., Chen, H.-C., Hsieh, C.-h., and Yu, H.-T. (2014). Spatial heterogeneity of gut microbiota reveals multiple bacterial communities with distinct characteristics. *Scientific Reports, Nature Publishing Group*, **4**.

Massana, R., Gobet, A., Audic, S., Bass, D., Bittner, L., Boutte, C., Chambouvet, A., Christen, R., Claverie, J.-M., Decelle, J., *et al.* (2015). Marine protist diversity in European coastal waters and sediments as revealed by High Throughput Sequencing. *Environmental Microbiology*, **17**(10), 4035–4049.

McDonald, K. W., McClure, C. J., Rolek, B. W., and Hill, G. E. (2012). Diversity of birds in eastern North America shifts north with global warming. *Ecology and Evolution*, **2**(12), 3052–3060.

Nunan, N., Wu, K., Young, I., Crawford, J., and Ritz, K. (2002). In situ spatial patterns of soil bacterial populations, mapped at multiple scales, in an arable soil. *Microbial Ecology*, **44**(4), 296–305.

Newsham, K.K., Hopkins, D.W., Carvalhais, L.C., Fretwell, P.T., Rushton, S.P., O'Donnell, A.G. and Dennis, P.G. (2015). Relationship between soil fungal diversity and temperature in the maritime Antarctic. *Nature Climate Change*, **6**, 182–186.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sun, J., Li, X., Wang, X., Lv, J., Li, Z., and Hu, Y. (2011). Latitudinal pattern in species diversity and its response to global warming in permafrost wetlands in the Great Hing'an Mountains, China. *Russian Journal of Ecology*, **42**(2), 123–132.

Timling, I., Walker, D.A., Nusbaum, C., Lennon, N.J., and Taylor, D.L. (2014). Rich and cold: Diversity, distribution and drivers of fungal communities in patterned-ground ecosystems of the North American Arctic. *Molecular Ecology*, **23**(13), 3258–3272.

Totsche, K. U., Rennert, T., Gerzabek, M. H., Kögel-Knabner, I., Smalla, K., Spiteller, M., and Vogel, H.-J. (2010). Biogeochemical interfaces in soil: the interdisciplinary challenge for soil science. *Journal of Plant Nutrition and Soil Science*, **173**(1), 88–99.

Whitman, T., Pepe-Ranney, C., Enders, A., Koechli, C., Campbell, A., Buckley, D. H., and Lehmann, J. (2016). Dynamics of microbial community composition and soil organic carbon mineralization in soil following addition of pyrogenic and fresh organic matter. *The ISME Journal*.

Willis, A. and Bunge, J. (2014). Package breakaway. *R package*.

Willis, A. and Bunge, J. (2015). Estimating diversity via frequency ratios. *Biometrics*, **71**, 1042–1049.

Willis, A. (2016). Species richness estimation with high diversity but spurious singletons. *arXiv:1604.02598*.

Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.-Y., Keilbaugh, S. A., Bewtra, M., Knights, D., Walters, W. A., Knight, R., *et al.* (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science*, **334**(6052), 105–108.

Zhang, Z. (2012). Confidence intervals for Simpson's diversity index. Unpublished manuscript.

Zhang, Z. and Zhou, J. (2010). Re-parameterization of multinomial distributions and diversity indices. *Journal of Statistical Planning and Inference*, **140**(7), 1731–1738.