Published in final edited form as:

Mol Ecol Resour. 2012 November ; 12(6): 1068–1078. doi:10.1111/1755-0998.12004.

Assessing group differences in biodiversity by simultaneously testing a user-defined selection of diversity indices

Philip Pallmann¹, Frank Schaarschmidt¹, Ludwig A. Hothorn¹, Christiane Fischer², Heiko Nacke², Kai U. Priesnitz³, and Nicholas J. Schork⁴

¹Institute of Biostatistics, Leibniz Universität Hannover, Herrenhäuser Strasse 2, 30419 Hannover, Germany

²Department of Genomic and Applied Microbiology, Institute of Microbiology and Genetics, Georg-August Universität Göttingen, Grisebachstrasse 8, 37077 Göttingen, Germany

³Bavarian State Research Center for Agriculture, Institute for Plant Protection, Lange Point 10, 85354 Freising, Germany

⁴Department of Molecular and Experimental Medicine, The Scripps Research Institute, and The Scripps Translational Science Institute, 3344 North Torrey Pines Court, La Jolla, CA 92037

Abstract

Comparing diversities between groups is a task biologists are frequently faced with, for example in ecological field trials or when dealing with metagenomics data. However, researchers often waver about which measure of diversity to choose since there is a multitude of approaches available. As Jost (2008) has pointed out, widely used measures such as the Shannon or Simpson index have undesirable properties which make them hard to compare and interpret. Many of the problems associated with the use of these “raw” indices can be corrected by transforming them into “true” diversity measures. We introduce a technique that allows the comparison of two or more groups of observations and simultaneously tests a user-defined selection of a number of “true” diversity measures. This procedure yields multiplicity-adjusted p -values according to the method of Westfall & Young (1993), which ensures that the rate of false-positives (type I error) does not rise when the number of groups and/or diversity indices is extended. Software is available in the R package “simboot”.

Keywords

metagenomics; Simpson index; Shannon entropy; bootstrap; multiple contrasts; Westfall-Young

Corresponding author: Philip Pallmann, Institute of Biostatistics, Leibniz Universität Hannover, Herrenhäuser Strasse 2, 30419 Hannover, Germany, Phone: +49 511 762 5821, Fax: +49 511 762 4966, pallmann@biostat.uni-hannover.de.

NJS and LAH developed the basic ideas. FS and PP elaborated the methodology. CF, HN and KUP provided data sets. PP wrote the paper.

Data accessibility

The R function “mcpHill” is included in the R package “simboot” and can be downloaded from CRAN (<http://CRAN.R-project.org/package=simboot>). The soil bacteria data and the predatory insects data are also in “simboot” as objects “Bacteria” and “predatGM”, respectively. The marine invertebrates data is available as object “saunders” in the R package “untb” (<http://CRAN.R-project.org/package=untb>).

Introduction

Many research projects in ecology, metagenomics, and genetics yield data sets with abundances of species (e.g., microbial species, RNA species, etc.). The goal of such studies is often to compare diversities between different groups, habitats or ecological niches. These studies commonly summarize observations in a diversity index. However, there exist many such indices, which inevitably creates a challenge of picking an appropriate or optimal index. Three of the most popular indices are the species richness index H_{SR} , the Shannon entropy index H_{Sh} and the Simpson concentration index H_{Si} (tab. 1) (Ricotta 1995; Magurran 2004). The Simpson index can be directly interpreted as probability that two randomly picked individuals belong to different species. Sometimes it is also expressed as Gini-Simpson index: $H_{GS}=1-H_{Si}$.

All these measures weight the two aspects of diversity – richness and evenness – differently. In other words, they all differ in their sensitivity to differences involving frequent and rare species. For example, H_{SR} strongly emphasizes rare species by weighting all species equivalently and irrespective of their frequency of occurrence, whereas H_{Si} emphasizes rather on common species (and is hence sometimes called dominance index). H_{Sh} weights each species according to its relative frequency in the population (Jost & Chao *in prep.*).

Beside the task of choosing among diversity measures, problems can arise from the use of “raw” Shannon or Simpson indices (Jost *et al.* 2010). The interpretation of both of these indices may lead to dangerously false conclusions since they often don’t behave in an intuitive way. For example, a very small decrease in H_{Si} might indicate a massive reduction of diversity when H_{Si} is close to unity (Jost 2006). In a fictitious population with 100 equi-frequent species, the disappearance of 50 of them would result in H_{Si} only dropping from 0.99 to 0.98. A more intuitive measure of diversity, however, should decrease by 50% if one half of the species is eliminated. Realistic data scenarios involve, of course, much more complicated combinations of richness and evenness, but the basic problem remains the same: in many cases it is wrong to associate the magnitude of the difference of H_{Si} values with the magnitude of the biological effect in terms of diversity.

In addition, “raw” indices are usually hard to interpret and compare. For instance, one could ask, is a population with $H_{Sh}=2.13$ more or less diverse than a population with $H_{Si}=0.83$? It is nearly impossible to answer this question at first glance, due to the fact that H_{Si} ranges between zero and unity, whereas H_{Sh} and especially H_{SR} can take on values larger than one. Although different indices are designed to emphasize different aspects of diversity, a certain degree of comparability would be advantageous.

Therefore it is advisable to transform “raw” indices into “true” diversities (Jost 2008), which all belong to one and the same mathematical family; for example, by regarding different measures as special cases of Hill’s general definition of diversity measures (Hill 1973), which is related to Rényi’s generalized entropy (Rényi 1961). The Hill number qD of order q for a population with S species and $\pi_s, s=1, \dots, S$ being the relative frequency of the s^{th} species is defined as

$${}^qD = \left(\sum_{s=1}^S \pi_s^q \right)^{1/(1-q)}$$

When q is 0, the measure is the H_{SR} itself. When q approaches unity, the measure results in the exponential of the Shannon entropy $\exp(H_{Sh})$, and the second order Hill number yields the reciprocal of the Simpson index $(H_{Si})^{-1}$ (tab. 1). The choice of q is, however, not limited

to these three values and not even to integers. The emphasis on rare species continuously increases with decreasing q and vice versa. For $q \rightarrow -\infty$ the Hill number approaches the reciprocal of the frequency of the rarest species, whereas it tends to the reciprocal of the frequency of the most common species when $q \rightarrow \infty$ (Hill 1973).

Hill numbers of any order q deserve to be called “true” diversities since they harbor – unlike “raw” H_{Si} or H_{Sh} – a couple of desirable features. Most importantly, they possess the doubling property (Hill 1973), which makes them intuitively understandable: when merging two equally large and equally distributed communities which don’t share any species, the “true” diversity doubles (MacArthur 1965; Jost 2008). Such diversity values are easily comparable because all “true” diversities are rooted in same “effective species” units. A “true” Shannon or “true” Simpson of, e.g., 13, can be directly interpreted as follows: this population consists of 13 “effective species” and may therefore be considered just as diverse as a population with 13 equally frequent species (Jost & Chao *in prep.*). Nevertheless, the transformed diversities maintain their tendency to attach more importance to rare ($q < 1$) or to common species ($q > 1$). Our interest here is to investigate diversity on the basis of a collection of different indices in order to accommodate rare as well as frequent species in relevant group comparisons. The choice of diversity measures to be used depends, of course, on the respective application. In agricultural ecology, for example, species from habitats surrounding the habitat under investigation may “accidentally” show up in a few samples. Hence they appear in the count as very rare species although they are not of particular interest for the agricultural ecosystem. Thus a Hill number of order $q > 1$, which focuses on abundant species, could be regarded as a proper choice. By contrast, in conservation biology rare species might be of substantially higher interest than the few common ones, so setting $q = 0$ or even $q < 0$ could be appropriate. However, such distinctions may not always be clear or reflect a consensus in a scientific field. In rather exploratory fields like metagenomics interest could be in abundant as well as rare species. In this setting inferences involving a range of several values of q may be most meaningful.

Inspecting more than one index simultaneously, however, results in a multiple testing problem and therefore demands multiplicity adjustment to make sure the desired type I error α (usually 5%) is maintained. A simple way to adjust α would be a division of α by the number of indices used, which is the widespread Bonferroni correction (Nelson 1989) known to be rather conservative. However, as closely related indices show a tendency to be highly correlated, it is reasonable to incorporate these correlations into the calculations and thereby reduce the conservativeness of the testing procedure. Here we present a method to circumvent the hard a priori choice of which index to use and simultaneously accommodate a multiplicity correction by explicitly modeling correlations between different indices applied to the data according to the method of Westfall & Young (1993).

Such a procedure has other advantages: When more than two groups of observations have to be compared, a common approach is to perform an analysis of variances (ANOVA). Indeed, ANOVA requires normally distributed data, which might be a questionable assumption for “raw” as well as for transformed diversity indices. Furthermore, assuming any kind of distribution – such as multinomial (Rogers & Hsu 2001) – is problematic for “raw” and transformed indices, in part because ecological count data sets tend to be highly overdispersed. Thus, an approach that is not based on any distributional assumptions should be the method of choice.

As an alternative to ANOVA, multiple contrast tests based on the resampling technique by Westfall & Young (1993) can be used to compare more than two groups. In this case a contrast matrix defines the comparisons one wishes to carry out. These can involve simple and well-known procedures analogous to Tukey’s all-pair (Tukey 1953) or Dunnett’s many-

to-one (Dunnnett 1955) procedure or the “GrandMean” comparison of each group to the mean of all groups (Dilba & Hothorn 2009) but could also involve user-defined contrasts. Our method gives multiplicity-adjusted p -values for each of the chosen Hill numbers and contrasts.

Methods

Data structure

The proposed procedure is suitable for data sets consisting of abundance data (species counts or relative abundances) represented in an $N \times S$ data matrix. The matrix consists of N rows listing the observation units $j=1, \dots, J_i$ within the different groups $i=1, \dots, I$ and S columns for the species. One additional column contains a factor variable assigning the N rows to two or more groups i . It is important to have multiple observations for each group in order to estimate variances for hypothesis testing.

Hill numbers and comparisons of interest

A set of K Hill numbers, qD_k , with index $k=1, \dots, K$ is chosen to construct diversity indices to be compared among groups. The requested comparisons are defined in an $M \times I$ contrast matrix C with $m=1, \dots, M$ comparisons of the I groups. The type of contrast is specified via a

priori coefficients c_{mi} . Note that $\sum_{i=1}^I c_{mi}=0$ for all $m=1, \dots, M$. The contrast matrix for a Tukey-like contrast (all possible pairwise comparisons) with $I=3$ groups is

$$C_{\text{Tukey}} = \begin{pmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & -1 & 1 \end{pmatrix}, \text{ for a Dunnnett-style comparison of two treatments to a control (many-}$$

to-one with $I=3$) it is $C_{\text{Dunnnett}} = \begin{pmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}$, for a comparison of each group versus the mean

of all $I=3$ groups it is $C_{\text{GrandMean}} = \begin{pmatrix} \frac{2}{3} & -\frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} \\ -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3} \end{pmatrix}$ and for a two-sample test it is simply $C_{\text{two}} = (-1 \ 1)$.

Multiplicity-adjusted p-values

The assumed Hill numbers used to construct the test are computed separately for each row so that “true” diversities are obtained for every single observation. In addition, the procedure largely follows the approach described by Westfall & Young (1993). Essentially, an empirical distribution function is gained by drawing data from the original data to create simulated data sets from which parameters can be estimated. This empirical distribution function replaces the unknown theoretical distribution function; i.e., nonparametric bootstrap (Efron & Tibshirani 1993).

The multiplicity-adjusted p -values for an upper-tailed test are calculated as follows:

1. Compute the desired Hill numbers ${}^q\hat{D}_{ijk}$ for each of the N rows as well as the group means ${}^q\bar{D}_{ik}$ for each of the K indices chosen.
2. Compute the residuals $\epsilon_{ijk} = {}^q\hat{D}_{ijk} - {}^q\bar{D}_{ik}$ and the estimated residual variance $\hat{\sigma}_k^2 = (\sum_{i=1}^I \sum_{j=1}^{J_i} (\epsilon_{ijk} - \bar{\epsilon}_{ik})^2) / ((\sum_{i=1}^I J_i) - I)$. The ϵ_{ijk} are stored in a matrix with $N = I \sum_{i=1}^I J_i$ rows and K columns.

3. Draw a random bootstrap sample ε_{ijk}^* from the residuals ε_{ijk} in the $N \times K$ matrix. Resampling (with replacement) is pursued only within the I groups such that all K residuals from a given row ij together may or may not build a new row ij^* in the $N \times K$ matrix of ε_{ijk}^* ; i.e., a set of K residuals of one observation ij must not be broken up. The bootstrapped data set should be of equal dimensions and should have the same structure of per group sample sizes J_i as the original $N \times K$ matrix.
4. Calculate the sample means $\bar{\varepsilon}_{ik}^* = (\sum_{j=1}^{J_i} \varepsilon_{ijk}^*) / J_i$ and the common variance $\hat{\sigma}_k^{2*} = (\sum_{i=1}^I \sum_{j=1}^{J_i} (\varepsilon_{ijk}^* - \bar{\varepsilon}_{ik}^*)^2) / ((\sum_{i=1}^I J_i) - I)$ from the bootstrapped data. Note that this assumes homoscedasticity.
5.
$$t_{mk}^* = \frac{\sum_{i=1}^I (c_{mi} \bar{\varepsilon}_{ik}^*)}{\hat{\sigma}_k^* \sqrt{\sum_{i=1}^I \frac{c_{mi}^2}{J_i}}}$$
 Compute the test statistic t_{mk}^* for each $m=1, \dots, M$ and each $k=1, \dots, K$.
6. Repeat steps 3 to 6 B times and store $\max_{m,k}(t_{mkb}^*)$ for every bootstrap step $b=1, \dots, B$.
7. Let
$$R_{mkb} = \begin{cases} 0: \max_{m,k}(t_{mkb}^*) \leq t_{mk} \\ 1: \max_{m,k}(t_{mkb}^*) > t_{mk} \end{cases}$$
8. The adjusted p -value for the m^{th} contrast and the k^{th} diversity index is given by
$$\tilde{p}_{mk} = (\sum_{b=1}^B R_{mkb}) / B.$$

Such a resampling-based adjusted p -value for an upper-tailed test can be interpreted as the probability that the largest test statistic $\max_{m,k}(t_{mkb}^*)$ in the resampled data set is larger than the observed test statistic t_{mk} in the original dataset whilst the complete null hypothesis (“diversity is equal in all groups”) is considered to be true (Westfall & Young 1993).

A lower-tailed test can be constructed with
$$R_{mkb} = \begin{cases} 0: \min_{m,k}(t_{mkb}^*) \geq t_{mk} \\ 1: \min_{m,k}(t_{mkb}^*) < t_{mk} \end{cases}$$
 after having stored $\min_{m,k}(t_{mkb}^*)$ in step 6. For a two-tailed test $\max_{m,k}|t_{mkb}^*|$ must be stored in step 6, and step 7 has to be replaced by
$$R_{mkb} = \begin{cases} 0: \max_{m,k}|t_{mkb}^*| \leq |t_{mk}| \\ 1: \max_{m,k}|t_{mkb}^*| > |t_{mk}| \end{cases}$$

Unlike conventional multiplicity adjustments (such as Bonferroni) this approach accounts for correlations among variables and tests and for distributional characteristics of the data and therefore gains more powerful and robust multiple testing procedures. Since this is a resampling method only an infinite number of bootstrapped data sets would give the exact p -value, thus the number of bootstrap replications B should be chosen to be sufficiently large (Westfall & Young 1993).

All calculations were done in R version 2.12.1 (R Development Core Team 2010). The bootstrapping was performed using the R package “boot” (Canty & Ripley 2012; Davison & Hinkley 2008), contrast matrices for multiple comparisons were built with “multcomp” (Hothorn *et al.* 2008). Our method is implemented as function “mcpHill” in the R package “simboot” (Scherer 2012).

Simulation study

A simulation study was set up to make sure that the desired α -level (type I error) is kept. For this purpose we generated data where all comparisons were under H_0 . Moreover, we created data settings with only a subset of all comparisons being under H_0 and the rest under the alternative H_A . That means that in these cases only some, but not all, comparisons were under H_0 , only for some, but again not all, of the indices that were tested.

Data sets were generated based on geometric series and Markov-chain Monte-Carlo sampling from Dirichlet-multinomial distributions, using the R package “MCMCpack” (Martin *et al.* 2011). The method was applied to data with several different simulated settings and parameter choices, and none of them indicated that the type I error exceeds its nominal level (data not shown). Therefore we conclude that the α -level is not inflated, so the adjusted p -values for single pairwise comparisons of groups for single indices can be considered valid and trustworthy.

Examples

Soil bacteria (two-sample test)—Two recent ecological metagenomics studies (Will *et al.* 2010; Nacke *et al.* 2011) yielded abundance data of soil bacteria from samples collected in German forest and grassland sites. Relative abundances of 18 bacterial phyla (including three candidate phyla) and five proteobacterial classes (alpha, beta, gamma, delta and epsilon) were in the data set, which is available from the R package “simboot” (Scherer 2010). There are 27 observations altogether, nine of which stem from forest and 18 from grassland plots. The bacteria’s relative abundances were determined by analyzing the V2–V3 region of the 16S rRNA gene via pyrosequencing-based DNA techniques (Nacke *et al.* 2011). One goal of these investigations was to unravel differences in bacterial diversity and community composition between the land use types forest and grassland.

The mosaic plot in fig. 1 itemizes the abundances of the single phyla and proteobacterial classes. It illustrates that around 80% of the entire bacterial population belong to the six most frequent phyla or classes, whereas more than half of the phyla appear very rarely and often only in few replicates. There is a quite consistent overall pattern: abundant phyla are frequent in nearly all plots, and infrequent phyla are similarly infrequent in all plots as well.

The rank/abundance plot (Whittaker 1965) depicts the patterns of relative abundances in forest and grassland plots, respectively (fig. 2). It shows that bacteria of 23 phyla and classes were found in grassland but only 18 in forest soils. There are a few phyla and classes predominating in the bacterial community of the forest soil, thus the forestal abundance curve is steeper than the one for grassland soils.

An initial way to study possible differences in diversity between the bacterial communities could involve a two-tailed test for integral Hill numbers of orders $-1 \leq q \leq 3$. This selection includes the transformed versions of the three familiar indices H_{SR} ($q=0$), H_{Sh} ($q \rightarrow 1$) and H_{Sj} ($q=2$) and in addition two indices that strongly emphasize rare ($q=-1$) and common ($q=3$) species, respectively. Multiplicity adjustment is required since there are five indices and hence five hypotheses to be tested.

Some of the indices are highly correlated, especially neighboring ones, but not all of them are correlated (fig. 3). This suggests that it makes sense to explore several indices simultaneously since they obviously describe different aspects of diversity. The procedure of Westfall & Young takes correlation structures into account, which reduces the impact of the multiplicity adjustment in comparison to Bonferroni’s method. The higher the correlations and less circular and the more ellipsoidal or even nearly linear the empirical distributions of bootstrapped test statistics (in this case for an upper-tailed test), the more points extend into

the dark gray area marking the loss of power in the method of Bonferroni compared to Westfall & Young (fig. 3).

The resulting adjusted p -values (tab. 2, second column) reveal significant differences in transformed H_{Sh} and H_{Si} but neither in H_{SR} nor any of the other two indices tested. We can conclude that soil bacterial communities of grassland plots are more diverse in terms of Shannon and Simpson index than those of forest soils.

A quite similar finding occurs when performing an upper-tailed test (tab. 2, fourth column), which tests the null hypothesis that bacterial diversity in grassland soils is not greater than in forest soils (and vice versa for the lower-tailed case): Testing one-sided hypotheses yields p -values that are large for $q=-1$ and 0 whereas they fall below the level of significance for $q=1, 2$ and 3. This indicates that the differences in bacterial diversity between forest and grassland sites are due to an effect in rather abundant phyla.

Note that it is possible to justify reducing the increment from $\Delta_q=1$ to $\Delta_q=0.5$. The range of q -values being covered remains the same ($-1 \leq q \leq 3$) but the selection tested rises from five (when $\Delta_q=1$) to nine hypotheses (when $\Delta_q=0.5$). This refinement provides additional information on four more values of q ($-0.5, 0.5, 1.5$ and 2.5) but in return results in larger p -values for corresponding values of q (tab. 2, third and fifth column). Thus, there is a significant difference in the Simpson index when performing a two-tailed test with $\Delta_q=1$ but the significance vanishes when Δ_q is reduced to 0.5. Corresponding p -values increase with growing number of indices tested. The price for extending the selection is, however, a small one compared to a Bonferroni-style multiplicity adjustment (fig. 4). When further refining the increment to $\Delta_q=0.1$ a perceivable, but not substantial, increase of the p -values can be observed (data not shown).

A closer look at the results of the two-tailed and especially the lower-tailed test (tab. 2, sixth and seventh column) provides a deeper insight into the data: The p -values are close to unity when q is positive but reach a minimum at $q=-0.5$. We have learned above that grassland plots are significantly more diverse than forest soils with respect to rather common phyla and classes. Now it is the other way around: forest plots appear to be more diverse than grasslands in terms of rare bacterial phyla – although this is not a statistically significant difference. But it shows that expanding the selection of indices as well as refining increments of q can be a useful tool to detect effects that would remain hidden if only “traditional” measures like H_{SR} , Shannon or Simpson index were used.

To illustrate this point, if only the “raw” Shannon entropy were calculated and inference were made with a standard t -test (ignoring that the normality assumption is doubtful) only for this single measure, the p -value for the null hypothesis of equal bacterial diversity in both forest and grassland soils would be 0.022. This is even a little larger than the multiplicity-adjusted p -value of 0.019 yielded by our method. And what is more, the significant difference in H_{Si} , which suggests that the soil types differ mainly due to an effect concerning frequent phyla, would remain concealed.

The situation would be even worse if one only picked the species richness as a measure of biodiversity: testing the difference in H_{SR} between grassland and forest soils with a Wilcoxon rank sum test gives a p -value of 0.610. This is very similar to the p -value of 0.627 provided by our approach; however, assessing only H_{SR} completely fails to find the clear and significant effect in frequent phyla.

Marine invertebrates (two-sample test)—The diversity of the invertebrate fauna inhabiting holdfasts of the kelp *Ecklonia radiata* is considered to be a meaningful indicator

of marine pollution and environmental change (Anderson *et al.* 2005a). The data set is part of the R package “unth” (Hankin 2007), rich in zeroes and contains abundances of 176 species colonizing 40 holdfasts altogether, 20 of which were exposed to waves and 20 were sheltered. They were collected from eight sites on New Zealand’s Northeastern coast, four of which were wave-exposed and four were sheltered (Anderson *et al.* 2005b). This clustering of 40 kelp holdfasts in eight habitats (and their spatial distribution) was ignored for the following data analysis since it was not documented in the available data set.

The mosaic plot (fig. 5) looks rather confusing, not least because of the multiplicity of rare and extremely rare species. There is no consistent pattern like in the bacterial data (fig. 1): for example, *Ventojassa*, which is the most abundant species in the whole data set, hardly appears in sheltered habitats at all. However, there are some replications of wave-exposed holdfasts in which *Ventojassa* is scarcely found.

The distribution of the relative frequencies in the rank/abundance plot (fig. 6) reveals that there are two particularly common species living on wave-exposed holdfasts whereas sheltered holdfasts don’t harbor any species being that predominant. Therefore the distribution curve of the fauna inhabiting unexposed holdfasts is less steep, which becomes manifest in significant p -values for small values of q . When testing the null hypothesis that biodiversity on wave-exposed holdfasts is not greater compared to sheltered ones, p -values are small for q from -1 to 0.5 and large especially for q from 2 to 3 (tab. 3), which suggests that the diversities of the invertebrate populations mainly differ in the rather infrequent species. Hence the essential differences are to be found in the long “tail” of rare species, and the significant p -values of the upper-tailed test indicate that the fauna on wave-exposed holdfasts is more diverse than on holdfasts in sheltered habitats.

The correlations between the test statistics for different values of q (fig. 7) are not quite as high as in the soil bacteria data set (fig. 3). There is no gain in biologically relevant information when further refining the increments Δ_q (data not shown).

Predatory insects (four-sample multiple contrast test)—The potential impact of growing genetically modified crops on ecologically important beneficial insects is a hotly debated topic in European agriculture. In 2005 a field trial was set up in Germany to investigate the abundances of predatory non-target insects under four different cultivating methods randomized over eight complete blocks: a genetically modified corn variety (“GM”), a near-isogenic line being almost identical to the GM variety except for the transformation (“S1”) and two other conventional varieties (“S2” and “S3”). The predators sampled in each of the 32 plots were counted and classified to the species level. Their abundances are listed in the blinded data set “predatGM” in the R package “simboot” (Scherer 2010).

According to the mosaic plot (fig. 8) effects of the blocks appear to be more striking than effects of the groups. “Sp12”, the most abundant species in the whole data set, is for example apparently most frequent in block 8, which is the block where most predators were caught. “Sp27” doesn’t occur at all in some of the observations, is fairly infrequent in a few ones but the most abundant species in block 6 for “S2”. Thus, single replicates can show quite different patterns compared to the rest. This can be explained by the fact that insects are, of course, not uniformly distributed over the cornfield. There might have been, for instance, a cluster of “Sp27” individuals anywhere close to the traps in block 6 for “S2”.

However, the rank/abundance curves show hardly any differences between the four groups (fig. 9) except for the total species richness which ranges from 18 (“GM”) to 25 (“S3”). The

only tendency that might be detected from the curves is that a little fewer abundant and more infrequent species can be found in group “S3”.

At first the effect of the block was eliminated by an alignment: After calculating the favored Hill numbers of orders $-1 \leq q \leq 3$ the block means, which corresponds to each respective Hill number, were subtracted. A Dunnett-like multiple testing procedure with the contrast matrix

$C_{Dunnett} = \begin{pmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}$ was carried out to compare “GM” versus each of the three conventional varieties. None of them yields any significant p -values (tab. 4), especially the comparison of “GM” and “S1”, which yields p -values mostly close to unity. This is not surprising since both originate from the same varieties. Also the shapes of their abundance curves are very similar – except for the longer “tail” of very rare species for “S1”.

The comparisons of “GM” versus “S2” and “S3”, respectively, don’t lead to significant p -values either. However, it appears that diversity concerning more abundant species ($q=2$, $q=3$) tends to be larger in “GM”. By contrast, the diversity measure of order $q=-1$, which mainly accounts for very rare species, tends to indicate higher diversity in “S2” and “S3”, respectively, compared to “GM”. The smallest p -value for a one-sided test is 0.546 for testing the hypothesis that the species richness ${}^0D=H_{SR}$ is larger in “S3” than in “GM”. All these effects are, however, far from being significant.

Note: With this statistical analysis we have not proven that GM corn is innocuous for predatory insects since we have tested for superiority and not for non-inferiority (“absence of evidence is not evidence of absence” (Altman & Bland 1995)).

Discussion

The great advantage of the proposed procedure is that a researcher doesn’t have to commit a priori to a particular diversity measure. Looking at several diversity indices simultaneously can provide insights that would remain concealed when only viewing one arbitrary measure of diversity, and there is not even a limitation to the frequently used H_{SR} , H_{Sh} and H_{Si} . This is appropriate because of the multiplicity adjustment, so one can’t be accused of “hunting for p -values”.

The conversion of “raw” indices into “true” diversities is a highly advisable step to ensure ease of interpretability, mainly by means of their doubling property and their use of comparable “effective species” units. It also facilitates the simultaneous investigation of several diversity measures by unifying them in one and the same mathematical family of indices. These Hill numbers represent a variety of useful diversity indices for all kinds of applications for which testing a wide range or strongly emphasizing rare ($q<0$) and abundant species ($q>2$) is appropriate.

Taking distributional characteristics of the data and correlation structures of the different Hill numbers into account diminishes the conservativeness of the testing procedure. Since the test statistics of neighboring Hill numbers tend to be highly correlated (fig. 3, fig. 7), the price (in terms of higher p -values) for multiple inference is rather small (fig. 4), not least because we often obtain little additional information when analyzing closely related indices. Furthermore, our method can be easily expanded to comparisons between more than two groups.

It is important to keep in mind that this is a resampling-based procedure and therefore the resulting p -values are subject to variation. They are influenced by the initial value as well as by the number of bootstrap replications, which is required to be large in order to yield

trustworthy p -values. $B > 5000$ can be recommended as a rather secure choice for stable p -values in most cases.

One of the next steps could be to modify the introduced method in a way that it also allows to adjust for covariates or secondary factors such as soil properties in field trials or age and sex in clinical metagenomics studies.

To summarize, our approach advances the investigation of biodiversity in two ways: it ameliorates practical interpretation by decomposing and simultaneously evaluating the effects of rare and common species, and it provides statistically valid and powerful inference for user-defined sets of diversity measures.

Acknowledgments

LAH is supported by the German Science Foundation (DFG) grant HO 1687/9-1.

CF and HN thank the managers of the three exploratories, Swen Renner, Sonja Gockel, Andreas Hemp, Martin Gorke and Simone Pfeiffer for their work in maintaining the plot and project infrastructure, and Markus Fischer, Elisabeth Kalko, Eduard Linsenmair, Dominik Hessenmöller, Jens Nieschulze, Daniel Prati, Ingo Schöning, François Buscot, Ernst-Detlef Schulze and Wolfgang W. Weisser for their role in setting up the Biodiversity Exploratories project. The work has been partly funded by the DFG Priority Program 1374 “Infrastructure-Biodiversity-Exploratories” (DA 374/6-1). Field work permits were issued by the responsible state environmental offices of Baden-Württemberg, Thüringen, and Brandenburg (according to § 72 BbgNatSchG).

NJS is supported in part by NIH grants: 5 UL1 RR025774, 5 U01 DA024417, 5 R01 HL089655, 5 R01 DA030976, 5 R01 AG035020, 1 R01 MH093500, 2 U19 AI063603, 2 U19 AG023122, 5 P01 AG027734 as well as the Stand Up To Cancer Foundation, the Price Foundation and Scripps Genomic Medicine.

The authors thank the Subject Editor and three anonymous referees for careful reading and helpful comments on the manuscript.

References

- Altman DG, Bland JM. Statistics notes: Absence of evidence is not evidence of absence. *British Medical Journal*. 1995; 311:485. [PubMed: 7647644]
- Anderson MJ, Diebel CE, Blom WM, Landers TJ. Consistency and variation in kelp holdfast assemblages: Spatial patterns of biodiversity for the major phyla at different taxonomic resolutions. *Journal of Experimental Marine Biology and Ecology*. 2005; 320:35–56.
- Anderson MJ, Connell SD, Gillanders BM, et al. Relationships between taxonomic resolution and spatial scales of multivariate variation. *Journal of Animal Ecology*. 2005; 74:636–646.
- Canty A, Ripley B. boot: Bootstrap R (S-Plus) functions. R package version. 2012; 1:3–4.
- Davison, AC.; Hinkley, DV. Bootstrap methods and their applications. Cambridge: Cambridge University Press; 1997.
- Djira GD, Hothorn LA. Detecting relative changes in multiple comparisons with an overall mean. *Journal of Quality Technology*. 2009; 41:60–65.
- Dunnnett CW. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*. 1955; 50:1096–1121.
- Efron, B.; Tibshirani, RJ. An introduction to the bootstrap. New York: Chapman & Hall; 1993.
- Hankin RKS. Introducing untb, an R package for simulating ecological drift under the unified neutral theory of biodiversity. *Journal of Statistical Software*. 2007; 22:1–15.
- Hill MO. Diversity and evenness: a unifying notation and its consequences. *Ecology*. 1973; 54:427–432.
- Hothorn T, Bretz F, Westfall P. Simultaneous inference in general parametric models. *Biometrical Journal*. 2008; 50:346–363. [PubMed: 18481363]
- Jost L. Entropy and diversity. *OIKOS*. 2006; 113:363–375.

- Jost L. G_{ST} and its relatives do not measure differentiation. *Molecular Ecology*. 2008; 17:4015–4026. [PubMed: 19238703]
- Jost L, DeVries P, Walla T, et al. Partitioning diversity for conservation analyses. *Diversity and Distributions*. 2010; 16:65–76.
- Jost L, Chao A. Diversity analysis: a fresh approach. (*in prep.*) From URL: [http://www.loujost.com/Statistics and Physics/Diversity and Similarity/SampleChapter.pdf](http://www.loujost.com/Statistics%20and%20Physics/Diversity%20and%20Similarity/SampleChapter.pdf).
- MacArthur RH. Patterns of species diversity. *Biological Reviews of the Cambridge Philosophical Society*. 1965; 40:510–533.
- Magurran, AE. *Measuring biological diversity*. Malden, MA: Blackwell Publishing; 2004.
- Martin AD, Quinn KM, Park JH. MCMCpack: Markov chain Monte Carlo in R. *Journal of Statistical Software*. 2011; 42:1–21.
- Nacke H, Thürmer A, Wollherr A, et al. Pyrosequencing-based assessment of bacterial community structure along different management types in German forest and grassland soils. *PLoS One*. 2011; 6:e17000. [PubMed: 21359220]
- Nelson PR. Multiple comparisons of means using simultaneous confidence-intervals. *Journal of Quality Technology*. 1989; 21:232–241.
- R Development Core Team. Vienna, Austria: R Foundation for Statistical Computing; 2010. R: A language and environment for statistical computing. URL: <http://www.R-project.org/>.
- Rényi, A. On measures of entropy and information. In: Neyman, J., editor. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. Vol. 1. Berkeley: University of California Press; 1961. p. 547–561.
- Ricotta C. Through the jungle of biological diversity. *Acta Biotheoretica*. 1995; 53:29–38. [PubMed: 15906141]
- Rogers JA, Hsu JC. Multiple comparisons of biodiversity. *Biometrical Journal*. 2001; 43:617–625.
- Scherer R. simboot: Simultaneous inference for diversity indices. R package version 0.1–6. 2012
- Shannon CE. A mathematical theory of communication. *Bell System Technical Journal*. 1948; 27:379–423.
- Simpson EH. Measurement of diversity. *Nature*. 1949; 163:688.
- Tukey, JW. The problem of multiple comparisons. In: Braun, HI., editor. (1994) *The collected works of John W. Tukey*. VIII. Multiple comparisons. New York, NY: Chapman and Hall; 1953. Unpublished manuscript, reprinted
- Westfall PH, Young SS. On adjusting p -values for multiplicity. *Biometrics*. 1993; 49:941–944.
- Westfall, PH.; Young, SS. New York: John Wiley & Sons, Inc.; 1993. *Resampling-based multiple testing: Examples and methods for p -value adjustment*.
- Whitaker RH. Dominance and diversity in land plant communities. *Science*. 1965; 147:250–260. [PubMed: 17788203]
- Will C, Thürmer A, Wollherr A, et al. Horizon-specific bacterial community composition of German grassland soils, as revealed by pyrosequencing-based analysis of 16S rRNA genes. *Applied and Environmental Microbiology*. 2010; 76:6751–6759. [PubMed: 20729324]

Figure 1.

Mosaic plot of the soil bacteria data set. The columns are ordered by group (forest and grassland) and represent the single replications. The heights of the boxes within each column are proportional to the relative frequencies of the phyla or proteobacterial classes in each replication. The phyla and classes are ordered by decreasing overall abundance. The names of phyla or classes whose total frequency of occurrence falls below 0.02 are replaced by dots. Dashed lines indicate phyla or classes to be missing in this replication.

Figure 2. Rank/abundance plot of the soil bacteria data set. The relative abundances of the phyla and proteobacterial classes are on a logarithmic scale. All replications of the groups (forest and grassland) were summed up.

Figure 3.

Correlation structures of the soil bacteria data set. Empirical distributions of 1000 bootstrap test statistics for integral Hill numbers of orders q from -1 to 3 are plotted in pairs. The respective marginal upper 5% quantiles (corresponding to an upper-tailed test) are shaded in light gray, the overlapping area in dark gray. r^2 is the Pearson correlation coefficient.

Figure 4.

Increase of the two-sided 95%-quantile of the soil bacteria data set with multiplicity correction according to Bonferroni and Westfall & Young, respectively, when raising the number of hypotheses tested. The Bonferroni quantiles were estimated from the empirical bootstrap distribution. For $-1 \leq q \leq 3$ the numbers of 2, 3, 5, 9, 41 and 81 indices correspond to $\Delta_q = 4, 2, 1, 0.5, 0.1$ and 0.05 ($B=99,999$).

Figure 5.

Mosaic plot of the marine invertebrates data set. The columns are ordered by group (wave-exposed and sheltered) and represent the single replications. The widths of the columns are proportional to the number of individuals in each replication. The heights of the boxes within each column are proportional to the relative frequencies of the species in each replication. The species are ordered by decreasing overall abundance. Only the names of the ten most abundant species are listed, the others are replaced by dots. Dashed lines indicate species to be missing in this replication.

Figure 6.

Rank/abundance plot of the marine invertebrates data set. The relative abundances of the species are on a logarithmic scale. All replications of the groups (wave-exposed and sheltered) were summed up.

Figure 7.

Correlation structures of the marine invertebrates data set. Empirical distributions of 1000 bootstrap test statistics for integral Hill numbers of orders q from -1 to 3 are plotted in pairs. The respective upper 5% quantiles are shaded in light gray, the overlapping area in dark gray. r^2 is the Pearson correlation coefficient.

Figure 8.

Mosaic plot of the predatory insects data set. The columns are ordered by group (“GM”, “S1”, “S2” and “S3”) and represent the single replications. The numbers above the columns denote the blocks. The widths of the columns are proportional to the number of individuals in each replication. The heights of the boxes within each column are proportional to the relative frequencies of the species in each replication. The species are ordered by decreasing overall abundance. The (abbreviated) names of species whose total frequency of occurrence falls below 0.02 are replaced by dots. Dashed lines indicate species to be missing in this replication.

Figure 9.

Rank/abundance plot of the predatory insects data set. The relative abundances of the species are on a logarithmic scale. All replications of the groups (“GM”, “S1”, “S2” and “S3”) were summed up.

Tab. 1

Three of the most commonly used diversity indices, their appropriate transformation according to Jost (2008) and their interpretation. S denotes the total number of species in a community, π_s is the frequency of the s^{th} species.

Diversity index	Hill no.	Transformation (Jost)	Interpretation
Species richness $H_{SR} = \sum_{s=1}^S \pi_s^0$	$q=0$	${}^0D = \sum_{s=1}^S \pi_s^0 = H_{SR}$	Ignores proportional abundances; emphasis on rare species
Shannon entropy $H_{Sh} = - \sum_{s=1}^S \pi_s \log \pi_s$	$q \rightarrow 1$	${}^1D = \exp(- \sum_{s=1}^S \pi_s \log \pi_s) = \exp(H_{Sh})$	Weights species according to their proportional abundances
Simpson index $H_{Si} = \sum_{s=1}^S \pi_s^2$	$q=2$	${}^2D = (\sum_{s=1}^S \pi_s^2)^{-1} = (H_{Si})^{-1}$	Dominance index; emphasis on abundant species

Tab. 2

Multiplicity-adjusted p -values for indices from $q=-1$ to $q=3$, performed as a two-tailed, an upper-tailed and a lower-tailed test with increments $\Delta_{q=1}$ and $\Delta_{q=0.5}$ for the soil bacteria data set ($B=9999$).

q	two-tailed $\Delta_q = 1$	two-tailed $\Delta_q = 0.5$	upper-tailed $\Delta_q = 1$	upper-tailed $\Delta_q = 0.5$	lower-tailed $\Delta_q = 1$	lower-tailed $\Delta_q = 0.5$
-1	0.588	0.627	0.976	0.986	0.304	0.335
-0.5		0.276		0.998		0.142
0	0.627	0.663	0.356	0.367	0.981	0.985
0.5		0.007		0.004		1
1	0.019	0.024	0.012	0.010	1	1
1.5		0.042		0.020		1
2	0.047	0.056	0.025	0.025	1	1
2.5		0.063		0.028		1
3	0.060	0.068	0.029	0.031	1	1

Multiplicity-adjusted p -values for indices from $q=-1$ to $q=3$, performed as a two-tailed, an upper-tailed and a lower-tailed test with increments $\Delta_{q=1}$ and $\Delta_{q=0.5}$ for the marine invertebrates data set ($B=9999$).

Tab. 3

q	two-tailed $\Delta_{q=1}$	two-tailed $\Delta_{q=0.5}$	upper-tailed $\Delta_{q=1}$	upper-tailed $\Delta_{q=0.5}$	lower-tailed $\Delta_{q=1}$	lower-tailed $\Delta_{q=0.5}$
-1	0.007	0.008	0.004	0.005	1	1
-0.5		0.009		0.005		1
0	0.011	0.014	0.007	0.007	1	1
0.5		0.035		0.019		1
1	0.157	0.161	0.076	0.079	0.998	0.998
1.5		0.409		0.207		0.986
2	0.569	0.585	0.297	0.307	0.969	0.967
2.5		0.689		0.372		0.952
3	0.735	0.752	0.404	0.416	0.935	0.940

Tab. 4

Multiplicity-adjusted p -values for indices from $q=-1$ to $q=3$, performed for a Dunnett-type many-to-one contrast with “GM” as control and as a two-tailed, an upper-tailed and a lower-tailed test with increment $\Delta_q=1$ for the predatory insects data set ($B=9999$).

	q	two	upper	lower		q	two	upper	lower		q	two	upper	lower
S1	-1	1	0.946	0.995	S2	-1	0.911	0.661	1	S3	-1	0.974	0.770	1
vs.	0	1	0.924	0.997	vs.	0	1	0.924	0.997	vs.	0	0.811	0.546	1
GM	1	1	0.993	0.957	GM	1	0.970	1	0.729	GM	1	1	0.996	0.937
	2	1	0.995	0.940		2	0.947	1	0.678		2	0.979	1	0.760
	3	0.972	0.994	0.950		3	0.973	1	0.740		3	0.972	1	0.736