MicroNiche: an R package for assessing microbial niche breadth and overlap from amplicon sequencing data

D R Finn, J Yu, Z E Ilhan, V M C Fernandes, C R Penton, R Krajmalnik-Brown, F Garcia-Pichel, T M Vogel

FEMS Microbiology Ecology, 96, 2020, fiaa131

doi: 10.1093/femsec/fiaa131
Advance Access Publication Date: 26 June 2020
Research Article

RESEARCH ARTICLE

MicroNiche: an R package for assessing microbial niche breadth and overlap from amplicon sequencing data

D. R. Finn[1,2,*], J. Yu[3,4], Z. E. Ilhan[4,5], V. M. C. Fernandes[3,4], C. R. Penton[3,6], R. Krajmalnik-Brown[4,5], F. Garcia-Pichel[4] and T. M. Vogel[2]

[1]School of Agriculture and Food Sciences, The University of Queensland, Brisbane 4072, Australia, [2]Environmental Microbial Genomics, Laboratoire Ampère, École Centrale de Lyon, Université de Lyon, Écully 69134, France, [3]School of Life Sciences, Arizona State University, Tempe, AZ 85287, USA, [4]Centre for Applied and Fundamental Microbiomics, Arizona State University, Tempe, AZ 85287, USA, [5]Swette Centre for Environmental Biotechnology; Biodesign Institute, Arizona State University, Tempe, AZ 85287, USA and [6]College of Integrative Sciences and Arts, Arizona State University, Mesa, AZ 85281, USA

*Corresponding author: School of Agriculture and Food Sciences, The University of Queensland, Brisbane 4072, Australia. E-mail: damien.finn@uqconnect.edu.au

One sentence summary: Mathematical indices to quantify niche properties were successfully adapted and applied to microbial taxa, and have been made publicly available as the R package 'MicroNiche'.

Editor: Rolf Kümmerli

ABSTRACT

Niche is a fundamental concept in ecology. It integrates the sum of biotic and abiotic environmental requirements that determines a taxon's distribution. Microbiologists currently lack quantitative approaches to address niche-related hypotheses. We tested four approaches for the quantification of niche breadth and overlap of taxa in amplicon sequencing datasets, with the goal of determining generalists, specialists and environmental-dependent distributions of community members. We applied these indices to *in silico* training datasets first, and then to real human gut and desert biological soil crust (biocrust) case studies, assessing the agreement of the indices with previous findings. Implementation of each approach successfully identified *a priori* conditions within *in silico* training data, and we found that by including a limit of quantification based on species rank, one could identify taxa falsely classified as specialists because of their low, sparse counts. Analysis of the human gut study offered quantitative support for *Bacilli*, *Gammaproteobacteria* and *Fusobacteria* specialists enriched after bariatric surgery. We could quantitatively characterise differential niche distributions of cyanobacterial taxa with respect to precipitation gradients in biocrusts. We conclude that these approaches, made publicly available as an R package (MicroNiche), represent useful tools to assess microbial environment-taxon and taxon-taxon relationships in a quantitative manner.

Keywords: niche theory; community ecology; proportional similarity index

INTRODUCTION

The niche is a fundamental concept with a long history of use in ecology (Leibold 1995). It was conceived as an integrating descriptor of an organism's place in a food web and the resources it depended upon (Elton 1927). It soon became associated with the principle of competitive exclusion and the notion that two

Received: 2 February 2020; **Accepted:** 24 June 2020

distinct species cannot coexist if they occupy the same niche, since one species would always be driven to extinction in that system (Volterra 1926; Gause 1932). In time, the definition of a niche was formalised to be an area (or volume) in multi-dimensional space formed by the relationship between a species and two (or more) environmental properties (Hutchinson 1957). By considering the niche as inherently dependent on the environment, the niche of two species that compete for the same resources cannot completely overlap in space (MacArthur 1958). In microbiology, the term niche is typically used to explain how environmental properties and species interactions determine the abundance and/or activity of a microorganism (Zhou and Ning 2017). Ultimately, the niche concept provides a framework for biologists to explain relationships between a species and its environment, species–species interactions, or both simultaneously. In search of a more quantitative approach, eventually ecologists devised means to measure a species' niche. The niche breadth (B_N) proposed by Levins (1968) is perhaps the first widely used example. It is a measurement of proportional similarity of resource use (p) of the ith resource given a variety of available resource states (R). If a species utilises all resources equally, B_N is 1 and this species is considered a generalist with a broad, non-discriminatory niche (MacArthur 1972). By contrast, if a species' B_N approaches 1/R, it is considered a specialist with a narrow, discriminatory niche:

$$\text{Levins' } B_N = \frac{1}{R}\sum_{i=1} p_i^2 \quad (1)$$

A limitation to Levins' B_N is the required assumption that all resources are equally abundant. This led to a refinement by Hurlbert (1978) to measure B_N whereby resource availability (r) of the ith resource could differ:

$$\text{Hurlbert's } B_N = \frac{1}{\sum_{i=1} \frac{p_i^2}{r_i}} \quad (2)$$

Feinsinger, Spears and Poole (1981) later simplified the concept by gauging a proportional similarity (PS) of resource use when resource availability differed as:

$$PS = 1 - 0.5\sum_{i=1} |p_i - r_i| \quad (3)$$

Unlike in Levins' B_N, that ranges from 1/R to 1, in the indices of Equations (2) and (3), a negative relationship between r and p approaches 0 and a positive relationship approaches 1. Levins' niche overlap (LO) is a pairwise comparison of the proportional use of resource r by species i and species j:

$$LO_{i,j} = \frac{\sum_{i,j=1}(p_{ir})(p_{jr})}{\sum_{i=1}(p_{ir}^2)} \quad (4)$$

An $LO_{i,j}$ of 1 indicates complete overlap between i and j, and as mentioned above, should never occur if i and j occupy the same niche.

With the routine adoption of Next-Generation Sequencing of amplicon marker genes and microbial taxon assignment pipelines, a challenge for the contemporary microbiologist lies in the appropriate application of statistics to test hypotheses. Many, if not all, of these amplicon-comparison studies are predicated on the notion that the environment dictates what microbial taxa can exist under a set of conditions, and that the presence and activity of specific taxa has consequences for the system as a whole. Examples include microbial surveys relevant for environmental geochemistry and biogeography (Martiny *et al.* 2006, Fierer *et al.* 2013), global climate change (Garcia-Pichel *et al.* 2013), human, animal or plant health (Penton *et al.* 2014, Ilhan *et al.* 2017), biological restoration (Velasco Ayuso *et al.* 2016) or industrial applications (Jung and Regan 2007). While microbiologists are familiar with the concept of niche, particularly the importance of the environment in shaping microbial communities (de Wit and Bouvier 2006), high-throughput metrics to quantify and compare niches of microbial taxa are currently lacking. The capacity to measure the relationship between two taxa and an environmental property, for example pH, and to definitively compare and state whether pH exerts a greater control over the distribution of one of the two taxa would be a useful metric. Equations (1)–(4) are proportional similarity indices which belong to the same family of indices as the Simpson index used to measure alpha diversity in microbial communities (Feinsinger, Spears and Poole 1981). We proposed to test the suitability of the above indices to measure the niche of microbial taxa in amplicon-sequencing datasets, whereby p_i and p_j were considered as the proportional abundance of taxa i and j, and r_i considered as the proportional value of any given environmental variable. R was considered as the sum of environments across which the niche was being measured. To do so, we first devised an *in silico* training dataset that included known generalists and specialists to test if our approach could correctly identify their niche on the basis of their B_N and PS values. Secondly, we tested if we could correctly apply these indices to real-life case studies of known outcome to confirm known relationships between the environment and a given taxon or between taxa pairs.

MATERIALS AND METHODS

The in silico training dataset

Initially, an *in silico* training dataset was designed to mimic typical datasets derived from microbial amplicon sequencing studies. Specifically, data were discrete counts of observed taxa within each sample. The training dataset was generated in, and all statistics mentioned throughout this manuscript were performed in, the R computing language, version 3.5.2 (R Core Team 2013). Figure 1(A) is a visual representation of the six taxon distributions and four environments considered in the *in silico* training dataset. The six distributions (D_1 to D_6) consist of 10 individual taxa (S_1 to S_{10}) per distribution. Each Environment consists of 10 independent samples. The first distribution (D_1) represents true generalists that have roughly equal counts generated from random normal distributions across the four environments. The second distribution (D_2) are taxa present in all environments yet decrease in a slow, linear fashion from Environments One to Four. This follows an inverse relationship with a mock environmental gradient that increases from Environments One to Four, shown below the six distributions in Fig. 1(A). The third distribution (D_3) represent specialists that decrease exponentially across the environments, that are high in Environment One, approximately half as abundant in Environment Two and mostly absent from Three and Four. These abundances also reflect an inverse relationship with the environmental gradient. The fourth distribution (D_4) are true specialists present only in

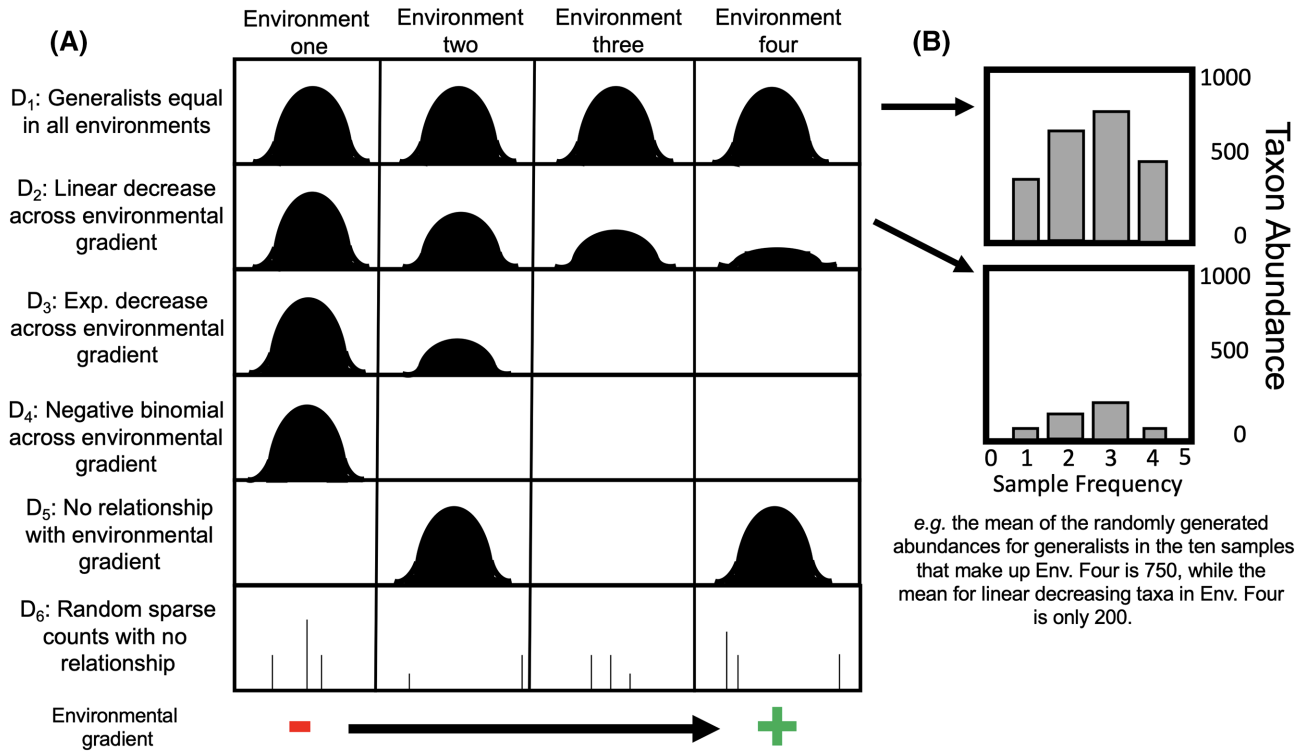

Figure 1. **(A)** Graphical representation of the *in silico* training dataset. Taxon distributions (D_1 through D_6) are rows and Environments (one through four) are columns. A total of ten independent taxa were generated per distribution, and ten independent samples were included per environment. The size of the normal distributions within each distribution/environment reflects the size of the random normally distributed counts generated for each distribution/environment. A mock environmental gradient was generated to increase from Environment One to Four. **(B)** A representation of taxon abundance of D_1 generalists and D_2 linear decreasing distributions in the ten samples that comprise Environment Four.

Environment One, which has the lowest values of the environmental gradient. The fifth distribution (D_5) represent specialists that have no relationship with the environmental gradient via being equally abundant in both Environments Two and Four. Furthermore, D_3 and D_5 represent taxa that only partially overlap in their distributions across the four environments, while D_4 and D_5 do not overlap. Finally, the sixth distribution (D_6) represent sparse counts of taxa that have no relationship with the environmental gradient or with any of the four environments. Taxa that fall into this category represent the long tail of sparse taxa frequently observed in microbial amplicon sequencing data. Figure 1(B) conceptualises a comparison of a taxon's abundance between the first and second distributions in Environment Four, where the mean abundance of the randomly generated taxon in D_1 is greater than D_2 in the ten samples that make up Environment Four. The *in silico* training dataset is available within the MicroNiche R package.

Index validation

Equations (1)–(3) were coded as functions within R and applied to the *in silico* training dataset as described (Levins 1968; Hurlbert 1978; Feinsinger, Spears and Poole 1981). To support hypothesis testing of taxon B_N and PS results (i.e. identify generalists and specialists) null model testing was incorporated into R functions for Levins' and Hurlbert's B_N, and PS. Briefly, a B_N and PS index was generated for 999 randomly generated taxon distributions dependent on the sum of environments in which generalists/specialists are being tested (i.e. the *R* parameter of Equation (1). A *P* value was derived based on the mean and 3 standard deviations of the null model as described (Benjamini and Hochberg 1995) to calculate the probability that each taxon's B_N/PS differed from the mean. Taxa *p* values were Benjamin–Hochberg adjusted to account for false discovery rate (Benjamini and Hochberg 1995). Equation (4) was also coded as an R function and applied as described (Ludwig and Reynolds 1988). Values for $LO_{i,j} \times LO_{j,i}$ were plotted as heatmaps with the 'gplots' package in R (Warnes *et al.* 2019).

Finally, a limit of quantification (LOQ) was employed to identify taxa yielding Type I Errors (i.e. falsely being identified as generalists/specialists) due to low counts and sparse, random distribution across samples. The LOQ was defined as taxa below a 'decision boundary' calculated from the distribution of taxa within the dataset and a 95% certainty that these taxa will fall within a null distribution where mean taxon abundance is 0 (Eurachem Working Group 2012). Simply put, it is not certain that taxa below the LOQ have an abundance greater than zero. The standard deviation of the null distribution was calculated from the lognormal rank distribution of taxa within the dataset (Ludwig and Reynolds 1988), specifically expressed as:

$$S(R) = S_0 e^{-a^2 R^2} \quad (5)$$

where the log abundance of taxon S at rank R is dependent on the modal S, S_0 and exponential decrease in abundance is dependent on *a* and rank R. The coefficient *a* is defined as:

$$a = \sqrt{\frac{\ln S_0}{S_m}} \Big/ R^2 \quad (6)$$

Where the parameters are as above, and S_m is the lowest taxon abundance of S. Determination of the LOQ was incorporated into each of the above functions. All functions are available within the MicroNiche R package. A step-by-step guide for installation, reproduction of these functions, null model testing and the calculation of the LOQ is available within the MicroNiche vignette (github.com/DamienFinn/MicroNiche_Vignette). The MicroNiche R package can be downloaded and installed via the Comprehensive R Archive Network (CRAN).

Case studies

The first case study considered the effect of two different bariatric surgeries on the composition of the human gut microbiome (Ilhan *et al.* 2017). Sequence data is available through the National Centre for Biotechnology Information Short Read Archive (SRA) as Bioproject PRJNA321731. Briefly, this study involved 54 individuals in four categorical groups: (1) a control, non-obese group, (2) an obese control group before Roux-en-Y gastric bypass surgery (Pre-RYGB), (3) post Roux-en-Y gastric bypass surgery (RYGB) and (4) post laparoscopic adjustable gastric banding (LAGB). Data was obtained from amplicon sequencing of the V4-V6 region of the prokaryote universal 16S rRNA gene performed on the Illumina MiSeq platform with paired-end sequencing of 300 bp. The second case study considered the effect of precipitation on the composition of cyanobacteria in arid soil biocrusts (Fernandes *et al.* 2018). Sequence data is available through the SRA as Bioproject PRJNA394792. The subset of data from this study analysed here involved 40 samples from two separate grasslands in New Mexico, USA, termed "Black Grama" and "Blue Grama" sites. A total of two environmental conditions were compared at each location, a control with 100% precipitation and an artificial drought condition with 33% precipitation. Precipitation for each site was measured between the years 1999–2017 and 2002–2017 for Black and Blue Grama sites, respectively. Total precipitation across all years was used as the environmental property for niche breadth calculations to capture historical effects of taxon enrichment under varying precipitation. Amplicon sequencing of the V4 region of the 16S rRNA gene was performed on the Illumina MiSeq platform with paired-end sequencing of 150 bp. Both amplicon sequencing datasets were re-analysed from raw-data using Vsearch and Uchime in the Qiime2 pipeline (Bolyen *et al.* 2019) to denoise, *de novo* cluster at 99% identity and remove chimeras with base parameters. The Silva132 16S rRNA gene database was used to annotate operational taxonomic units (OTUs) at the 99% identity level (Quast *et al.* 2013). Prior to analysis, the bariatric surgery and biocrust datasets were rarefied to 11 000 and 36 000 sequences per sample, respectively, with the Vegan package in R (Oksanen *et al.* 2013). The use of rarefaction to normalise sequence depth prior to niche breadth analyses is emphatically recommended to improve the accuracy of taxa comparisons between samples and reduce the occurrence of Type I specialist taxa Errors. Niche breadth and overlap indices of OTUs were measured with the MicroNiche R package, with a particular focus on Classes previously identified as being enriched under RYGB treatment (Ilhan *et al.* 2017) or cyanobacterial taxa differentially affected under precipitation conditions (Fernandes *et al.* 2018). Here, all MicroNiche functions were applied to OTU data after taxon assignment as discrete counts from Qiime2 outputs. However, please note that these functions can be readily applied to individual OTUs that have not been grouped by taxon assignment. While the use of these functions on amplicon sequence variants (ASVs) as specific markers of individual taxa (Callahan, McMurdie and Holmes 2017) was not tested here, conceptually this would be entirely feasible, although the stringency of the LOQ may need to be manually optimised as increased variance in ASV abundance may push the LOQ high enough to misclassify certain taxa as Type I Errors. Optimisation of the LOQ in MicroNiche functions is discussed within the Vignette, and can be performed if users consider the LOQ to be unjustifiably stringent. For more detail regarding the format of data required for the MicroNiche functions, including access to the training dataset as an example of what exactly is required and step-by-step guidance on applying the various functions, please refer to the Vignette companion piece. For this study, Levins' B_N and LO were applied to the bariatric surgery dataset to test whether specific taxa were enriched post-bariatric surgery, and Hurlbert's B_N and PS applied to the biocrust dataset to test whether precipitation affected the distribution of specific cyanobacterial taxa. Finally, Hurlbert's B_N and PS results for unweighted *versus* weighted OTUs were compared. Unweighted cyanobacterial OTUs were considered as relative abundances (%). For the weighted data, cyanobacterial OTUs were weighted by total 16S rRNA gene copies per cm^{-2} of biocrust, as determined by quantitative polymerase chain reaction of the prokaryote community (Fernandes *et al.* 2018).

RESULTS

Index validation

Each of the three niche breadth indices were applied to the taxa within the *in silico* training dataset. These included D_1 generalists equally abundant across the four environments, D_2 that decreased linearly across the four environments, D_3 that decreased exponentially across the four environments, D_4 that were present in only one environment, D_5 that were equally present in Environments Two and Four and had no relationship with a mock environmental gradient, and finally D_6 that were sparse and had no relationship with either the four environments or the environmental gradient. Figure 1 (Supporting Information) shows the taxon rank distribution of the *in silico* training dataset. Figure 2 shows results of null model testing of Levins' B_N, Hurlbert's B_N and PS on the *in silico* training dataset. Table 1 summarises the results and Benjamin–Hochberg *P* values for the three indices applied to the six taxa distributions. Levins' B_N successfully identified D_1 as generalists ($B_N = 0.99$, $P = 0.045$). This group represents cosmopolitan microbial taxa that are evenly distributed across all environments. Levins' B_N identified D_3, D_4 and D_5 as specialists (B_N 0.25–0.52, $P = 0$–0.013). These groups represent microbial taxa that are not evenly distributed across many environments, and that are proportionally more abundant in specific environments. As can be visualised with the null models in Fig. 2, these values either fall below the fifth (specialists) or above the 95th (generalists) quantiles highlighted as red lines of the Levins' B_N null model in Fig. 2.

Hurlbert's B_N and PS only identified D_3 and D_4 as including taxa that had a relationship with the environmental gradient ($B_N = 0.1$–0.17 and PS = 0.1–0.29, $P < 0.001$). These groups not only represent microbial taxa that are most abundant in specific environments, but also are most abundant when the mock environmental gradient is relatively low. In both cases, the low B_N and PS values approaching 0 indicate a negative relationship between these groups and the environmental property (i.e. taxon abundance of D_3 increases as the property decreases). The linearly decreasing D_2 did not have a negative relationship with

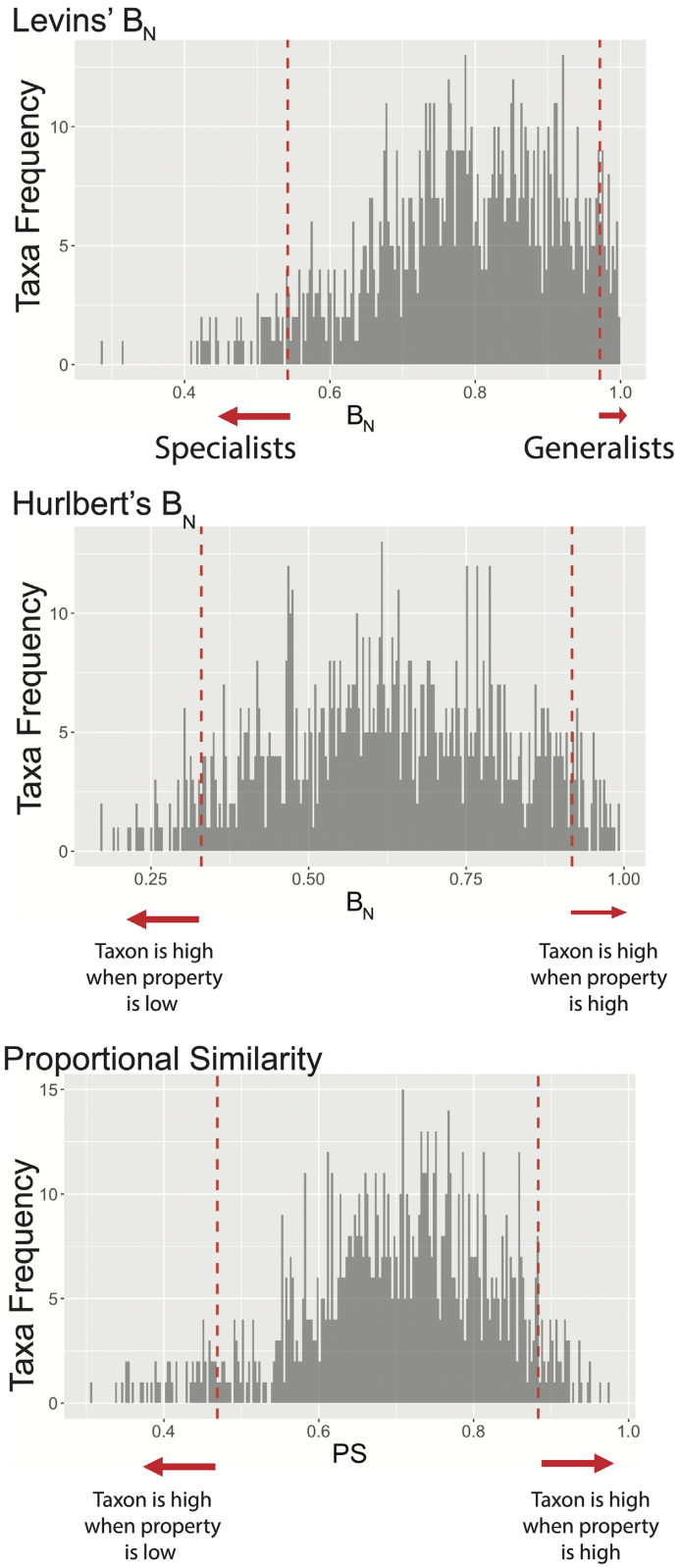

Figure 2. Null model distributions generated from applying Levins' B_N, Hurlbert's B_N and PS to the *in silico* training dataset. Null distributions were calculated from 999 randomly generated taxon distributions. Red dotted lines indicate the fifth and 95th quantiles. Specialists are taxa with Levins' B_N below the fifth quantile, while generalists are taxa with Levins' B_N above the 95th quantile (marked beneath the Levins B_N null model). Taxa that are high when an environmental property is low have a Hurlbert's B_N or Proportional Similarity value below the fifth quantile of those particular null models. Conversely, taxa that are high when an environmental property is high have a Hurlbert's B_N or Proportional Similarity value above the 95th quantile of those null models. Both of these cut-off points are also marked for Hurlbert's B_N and Proportional Similarity null models.

Table 1. Validating the three niche breadth indices against the *in silico* training dataset. Taxa below the LOQ are labelled as Y, while those not below are N. *n* per taxa = 10. Significant *P* values are highlighted as: $P < 0.05$ (*); $P = 0.001$ (**); $P < 0.001$(***).

Dist.	Levins' B_N	BH *p* val.	Hurlbert's B_N	BH *p* val.	PS	BH *p* val.	Below LOQ
D_1	0.99 ± 0.001	0.045 ± 0*	0.73 ± 0.005	0.33 ± 0.01	0.79 ± 0.003	0.436 ± 0	N
D_2	0.83 ± 0	0.608 ± 0	0.44 ± 0	0.08 ± 0	0.57 ± 0	0.303 ± 0	N
D_3	0.42 ± 0.003	0.003 ± 0**	0.17 ± 0.001	0 ± 0***	0.29 ± 0.004	0 ± 0***	N
D_4	0.25 ± 0.002	0 ± 0***	0.1 ± 0	0 ± 0***	0.1 ± 0.004	0 ± 0***	N
D_5	0.52 ± 0.01	0.013 ± 0.003*	0.49 ± 0.01	0.23 ± 0.03	0.59 ± 0.01	0.359 ± 0.04	N
D_6	0.75 ± 0.149	0.228 ± 0.19	0.66 ± 0.135	0.53 ± 0.33	0.7 ± 0.089	0.7 ± 0.316	Y

the environmental gradient ($B_N = 0.44$ and $PS = 0.57$, $P > 0.05$). As with Levins' B_N, the null models and highlighted quantiles for Hurlbert's B_N and PS (Fig. 2) offer useful aids to visualise where these values fall in relation to the null distribution.

The LOQ was successful in identifying D_6 as being the sparse, relatively low count taxa. This group represents the long tail of sparse, low count taxa often seen in microbial amplicon sequencing datasets. By being below the LOQ, these taxa could be identified as Type I Errors in that, depending on the individual taxon, they inappropriately gave positive results for one or more of the three indices.

Figure 3 is a heatmap of niche overlap measured via LO, with $LO_{i,j}$ on the y axis and $LO_{j,i}$ on the x axis. The generalists present in all four environments, D_1, overlapped with all other groups and had LO values approaching 1. Again, this group represents cosmopolitan microbial taxa that are equally present across all environments, and therefore will overlap with all other taxa. The linearly decreasing D_2, which were also present in all environments, overlapped with all other groups but less so than D_1. D_3 and D_5, which represented largely non-overlapping specialists, had poor overlap values with each other approaching 0.3. D_4 and D_5 had overlap values of 0–0.03, as D_4 was present in only one environment, in which D_5 was absent. Overlap values for D_6 varied considerably from 0.1 to 0.57. In order to assist in identifying group D_6 as Type I Errors in LO analyses, MicroNiche applies an asterisk to the names of taxa below the LOQ when plotting user data (please refer to the MicroNiche Vignette). To aid in interpretation of how LO is determined and why $LO_{i,j}$ values are not necessarily equal to $LO_{j,i}$ values, a conceptual diagram that compares overlap between two generalists and a generalist and specialist is provided as Figure 2 (Supporting Information).

Bariatric surgery case study

The purpose of the bariatric surgery case study was to confirm that Levins' B_N and LO could identify *Bacilli*, *Gammaproteobacteria* and *Fusobacteria* as specialists enriched under RYGB treatment (Ilhan *et al.* 2017). Figure 4(A) is a stacked bar chart comparing distributions of prokaryote Classes across gut microbiomes of Control, Pre-RYGB, RYGB and LAGB individuals. Together, *Bacteroidia* and *Clostridia* dominated all gut microbiomes, regardless of treatment (71.5–97%). Enriched under the RYGB treatment were: Bacilli (0.6–8% in RYGB *versus* 0–1.4% in non-RYGB), Fusobacteria (0–11% in RYGB *versus* 0–0.3% in non-RYGB) and *Gammaproteobacteria* (0.5–13% in RYGB *versus* 0–0.5% in non-RYGB). Interestingly, while *Bacteroidia* and *Clostridia* appeared to be generalists with consistently high LO values across Classes (Fig. 4(B)), their abundances varied sufficiently to lower their Levins' B_N below the threshold to be classified as generalists ($P > 0.05$, Table 2). *Bacilli*, *Fusobacteria* and *Gammaproteobacteria* were all classified as strong specialists ($P < 0.001$, Table 2). Figure 4(B)

Table 2. Levins' B_N results for each Class in the bariatric surgery case study. Significant *P* values are highlighted as: $P < 0.05$ (*); $P = 0.001$ (**); $P < 0.001$(***).

Taxa	Levins' B_N	BH *p* val.	Below LOQ
Actinobacteria	0.629	0.240	Y
Coriobacteriia	0.762	0.810	Y
Bacteroidia	0.838	0.644	N
Flavobacteriia	0.424	0.001**	Y
Melainabacteria	0.498	0.007**	Y
Bacilli	0.334	0.000***	N
Clostridia	0.923	0.276	N
Erysipelotrichia	0.897	0.353	N
Negativicutes	0.864	0.509	N
Fusobacteriia	0.258	0.000***	N
Alphaproteobacteria	0.488	0.006**	Y
Betaproteobacteria	0.925	0.276	N
Deltaproteobacteria	0.858	0.519	N
Gammaproteobacteria	0.416	0.000***	N
Saccharibacteria	0.667	0.353	Y
Synergistia	0.254	0.000***	Y
Mollicutes	0.744	0.703	Y
Verrucomicrobiae	0.896	0.353	N
Unassigned	0.874	0.471	Y

also demonstrated strong overlap of *Bacilli* and *Gammaproteobacteria*, although *Fusobacteria*, highly enriched in only two individuals, did not particularly overlap with any other Classes. Several other Classes had B_N values low enough to be specialists (*Flavobacteria*, *Melainabacteria*, *Alphaproteobacteria* and *Synergistia*), however, all of these groups were below the LOQ and, thus, the calculated Levins' B_N should be interpreted with caution. Table 1 (Supporting Information) shows results of Levins' B_N applied to this dataset at the Genus level to identify potential specialists within *Bacilli* and *Gammaproteobacteria*. These genus-level specialists had Levins' $B_N < 0.56$ and BH $P < 0.05$. Specialist *Bacilli* included: Uncultured *Carnobacteriaceae*, *Enterococcus*, *Lactobacillus* and *Streptococcus*. Specialist *Gammaproteobacteria* included: *Enterobacter*, *Escherichia-Shigella*, *Klebsiella*, Uncultured *Enterobacteriaceae* and *Haemophilus*. Figure 3 (Supporting Information) is a collection of box plots showing that all of these genera are specifically enriched in the RYGB treatment.

Biocrust case study

The biocrust case study sought to validate the use of Hurlbert's B_N and PS in quantitatively demonstrating that precipitation exerts a greater effect on the abundance of *Microcoleus steenstrupii* than *M. vaginatus* (Fernandes *et al.* 2018). Figure 5(A) is a stacked bar chart of unweighted cyanobacterial OTUs as relative abundance (%) across the biocrust samples. Imposed drought

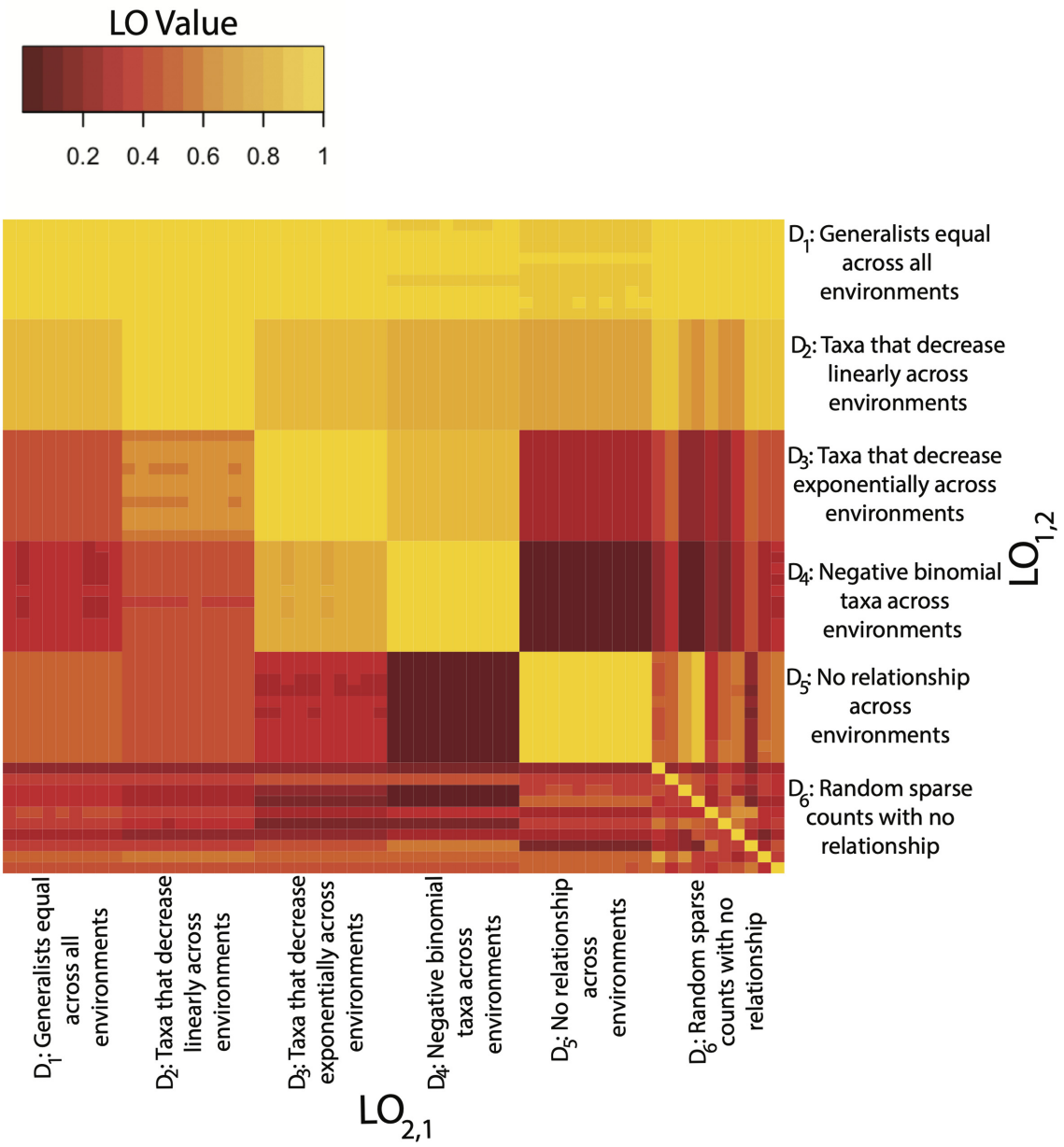

Figure 3. Heatmap of Levins' Overlap performed on the *in silico* training dataset. The scale bar shows LO values between 0 and 1. D_1 generalists have relatively high $LO_{1,2}$ values than other non-uniform distributions. D_1 generalists have varying $LO_{2,1}$ values dependent on how abundant taxa within distributions D_2 through D_6 are in the four environments. Lowest values are between D_4 and D_5 that do not overlap in any of the four environments.

conditions had a greater effect on the Black Grama biocrusts, with the relative abundance of cyanobacteria dropping from 41.8–71.3% in controls to 1.4–66.8% in imposed drought conditions. Biocrusts in the Blue Grama site showed little variation, from 10.6–36.3% in controls to 15–41.9% in imposed drought conditions. As relative proportions of the cyanobacterial population, *M. vaginatus* dominated all biocrusts and differed little between Black Grama control and drought (24–95% and 14–91%, respectively) conditions. *M. vaginatus* also dominated the Blue Grama biocrusts under drought conditions (53–98% and 85–95% for control and drought, respectively). *M. steenstrupii* decreased in all biocrusts under drought conditions, from 0.4–7% to 0–4.5% in Black Grama control and drought, respectively, and from 0.3–11% to 0.1–2.5% in Blue Grama control and drought, respectively. Figure 5(B) is a stacked bar chart of weighted cyanobacterial OTUs as 16S rRNA gene copies cm^{-2} of biocrust. The negative effect of Black Grama drought on all cyanobacteria is more evident here. *M. vaginatus* populations differed markedly from being stable in the control at 1.9×10^4–4.5×10^5, to unstable under drought at 100–3.9×10^5 16S rRNA gene copies cm^{-2}. The *M. vaginatus*-dominated Blue Grama biocrusts were more resilient to imposed drought conditions. Here, *M. vaginatus* populations were 8.9×10^4–1×10^6 and 1.1×10^5–7.2×10^5 16S rRNA gene copies cm^{-2} under control and drought, respectively. As above, *M. steenstrupii* was more affected by drought. Populations under control conditions were 1.1×10^3–3.1×10^4 and 3.1×10^3–1.8×10^4 16S rRNA gene copies cm^{-2} in Black and Blue Grama, respectively. Populations under drought conditions were 0–2.3×10^3 and 582–5.3×10^3 16S rRNA gene copies cm^{-2}, respectively. These results are essentially identical to

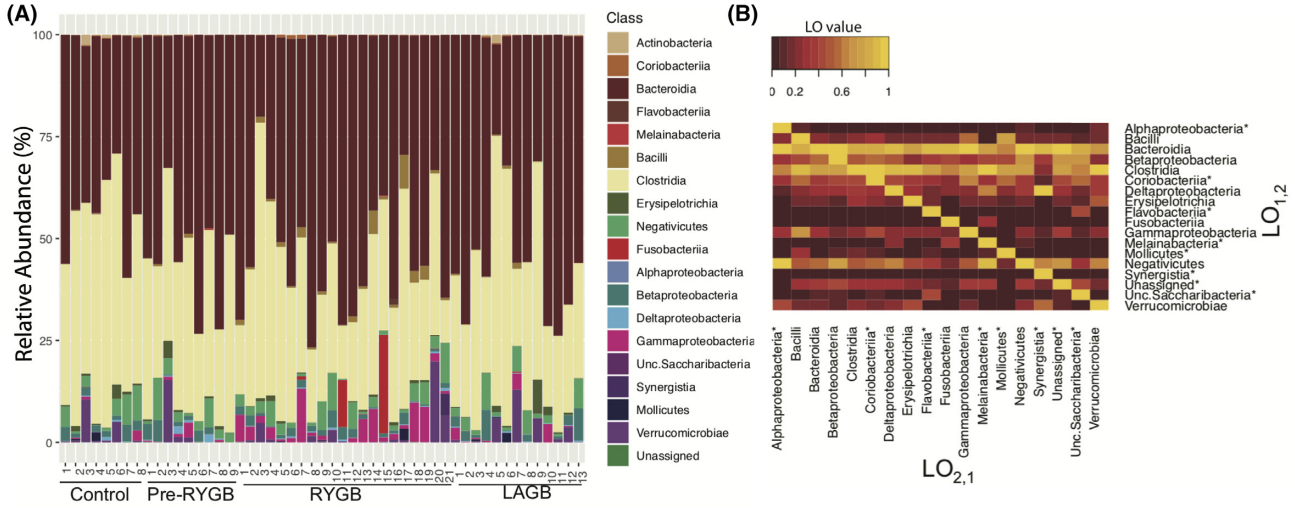

Figure 4. (A) Stacked bar chart of Class relative abundances across individuals in the bariatric surgery case study. Minor Classes that make up less than 5% of the population are grouped as Other. *Bacteroidia* and *Clostridia* are fairly dominant, cosmopolitan members across four treatments whereas *Bacilli* and *Gammaproteobacteria* are enriched under the RYGB treatment. **(B)** Heatmap of Levins' Overlap performed on the bariatric surgery case study. The scale bar shows LO values between 0 and 1. Taxa below the LOQ are noted with an asterisk. Bacteroidia and Clostridia have high $LO_{1,2}$ values reflective of their cosmopolitan nature across treatments. *Bacilli* and *Gammaproteobacteria* have relatively high $LO_{1,2}$ and $LO_{2,1}$ values as they are both enriched under RYGB treatment.

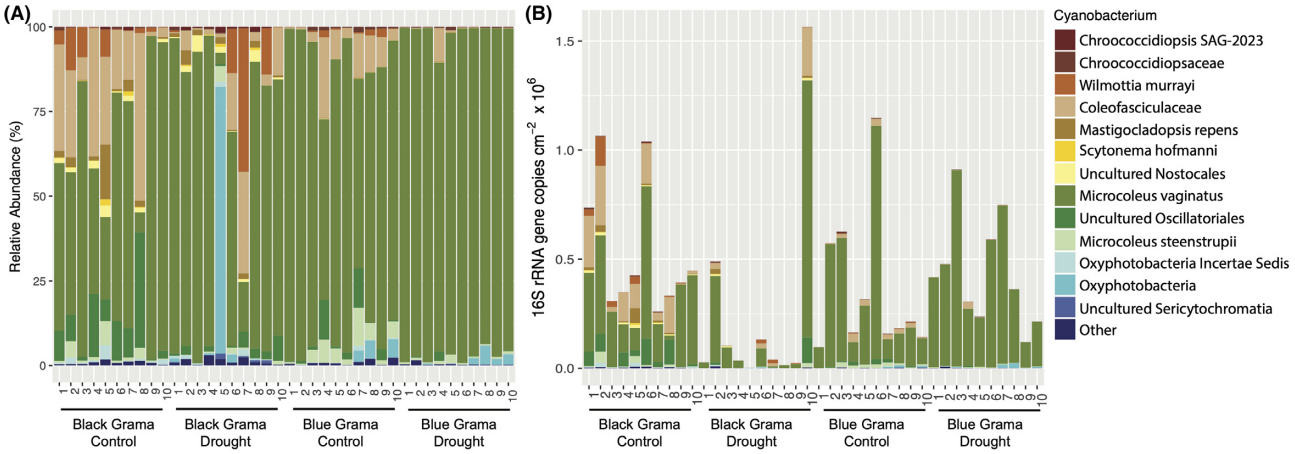

Figure 5. (A) Stacked bar chart of unweighted *Cyanobacteria* relative abundances across samples in the desert biocrust case study. Both Black and Blue Grama sites, under Control and Drought conditions, are dominated by *Microcoleus vaginatus*, while *M. steenstrupii* is enriched in Control relative to Drought. **(B)** Stacked bar chart of *Cyanobacteria* weighted by total 16S rRNA gene copies cm^{-2} x 10^6 across samples in the biocrust case study. Weighting the relative abundances emphasises the detrimental effect of Drought on non-*M. vaginatus Cyanobacteria* at the Black Grama site. Minor OTUs that make up less than 5% of the total Cyanobacterial population are grouped as Other.

those reported by Fernandes *et al.* (2018) other than for the actual names of the taxonomic assignments. Fernandes used a specialised taxonomic database to improve the assignment of cyanobacterial taxa over that provided by the Silva database. For our work, the actual nomenclature of taxa is irrelevant, and automated Silva-based assignments were considered sufficient.

For the unweighted data, numerous Cyanobacterial taxa were sufficiently abundant in the total prokaryote community to pass the LOQ including: *M. vaginatus*, *M. steenstrupii*, *Wilmottia murrayi*, *Mastigocladopsis repens*, unknown *Coleofasciculaceae*, uncultured *Oscillatoriales* and several uncultured '*Oxyphotobacteria*' (which is equivalent to Cyanobacteria (Garcia-Pichel *et al.* 2019)). All Cyanobacteria were above the LOQ when weighted data was considered. The distribution of *M. vaginatus* was found to be independent of precipitation for both Hurlbert's B_N and Feinsinger's PS when unweighted and weighted data were considered. Conversely, *M. steenstrupii* was shown to be dependent on precipitation. This relationship was stronger when weighted data was considered (Table 3, $B_N = 0.976$, $p = 0.002$ *versus* $B_N = 0.925$, $p = 0.016$ for weighted and unweighted data, respectively). These conclusions fully match the conclusions of the original study. The weighted analysis also found the Silva-unassignable cyanobacterium (likely a *Leptolyngbya* sp.) to be dependent on precipitation, while the unweighted analysis did not. Interestingly, an unknown *Chroococcidiopsaceae* that was below the LOQ and found to be dependent on precipitation in the unweighted analysis was not found to be dependent on precipitation in the weighted analysis. These refinements were not addressed in the original analysis. Feinsinger's PS was generally more stringent than Hurlbert's B_N, and the PS index did not find that *M. steenstrupii* was associated with increasing precipitation for either weighted or unweighted data.

Table 3. Hurlbert's B_N and Feinsinger's PS results for major Cyanobacterial taxa in the BSC case study. Each index was measured with unweighted and weighted OTU counts, and compared. Significant Benjamin–Hochberg adjusted p values are highlighted as: $p < 0.05$ (*); $p = 0.001$ (**); $p < 0.001$(***).

	Unweighted			Weighted		
	B_N	BH p val.	Below LOQ	B_N	BH p val.	Below LOQ
Chroococcidiopsis SAG-2023	0.358	0.039*	Y	0.282	0.004**	N
Chroococcidiopsaceae	0.885	0.040*	Y	0.826	0.090	N
Wilmottia murrayi	0.585	0.868	N	0.540	0.541	N
Coleofasciculaceae	0.612	0.967	N	0.654	0.829	N
Mastigocladopsis repens	0.483	0.334	N	0.520	0.449	N
Scytonema hofmanni	0.474	0.300	Y	0.536	0.525	N
Uncultured Nostocales	0.510	0.480	N	0.521	0.449	N
Microcoleus vaginatus	0.787	0.245	N	0.664	0.776	N
uncultured Oscillatoriales	0.557	0.742	N	0.651	0.843	N
Microcoleus steenstrupii	0.925	0.016*	N	0.976	0.002**	N
Oxyphotobacteria.Incertae.Sedis	0.535	0.627	N	0.566	0.680	N
Oxyphotobacteria	0.493	0.388	N	0.285	0.004**	N
uncultured *Sericytochromatia*	0.522	0.548	Y	0.320	0.010*	N
	PS	BH p val	Below LOQ	PS	BH p val	Below LOQ
Chroococcidiopsis.SAG.2023	0.560	0.503	Y	0.474	0.147	N
Chroococcidiopsaceae	0.889	0.204	Y	0.814	0.492	N
Wilmottia murrayi	0.580	0.621	N	0.570	0.485	N
Coleofasciculaceae	0.622	0.864	N	0.644	0.853	N
Mastigocladopsis repens	0.516	0.273	N	0.527	0.284	N
Scytonema hofmanni	0.491	0.185	Y	0.562	0.440	N
Uncultured Nostocales	0.519	0.281	N	0.525	0.276	N
Microcoleus vaginatus	0.834	0.468	N	0.751	0.820	N
uncultured Oscillatoriales	0.571	0.562	N	0.639	0.823	N
Microcoleus steenstrupii	0.888	0.204	N	0.934	0.094	N
Oxyphotobacteria.Incertae.Sedis	0.554	0.480	N	0.586	0.573	N
Oxyphotobacteria	0.622	0.864	N	0.495	0.169	N
uncultured *Sericytochromatia*	0.606	0.785	Y	0.512	0.227	N

DISCUSSION

Levins' B_N for measuring niche breadth

The capacity to measure niche indices of microbial taxa is of value to test hypotheses regarding environment-taxon and taxon-taxon ecological relationships. While amplicon sequencing has its limitations in identifying a causative, physiological mechanism of how the environment determines the abundance of a taxon, it can demonstrate relationships between a taxon and its environment. This can be used to infer the underlying physiological mechanism defining a taxon's niche and could provide support for more detailed analyses in the future on a taxon of interest to demonstrate such a physiological mechanism. Here, we report the application of four different niche indices (Equations (1)–(4)) that address differing niche-related hypotheses.

Levins' B_N successfully identified generalist and specialist taxa from the *in silico* training dataset (Table 1) and specialists in the bariatric surgery case study (Table 2). The identification of taxa as either generalists or specialists can be used to infer their respective physiology (MacArthur 1972). Generalists, with their non-discriminatory niches, are equally proportional across vastly different environments (Levins' B_N approaching 1) and must employ one or more functional traits that enables their colonisation and persistence within those environments. Conversely, specialists do discriminate between environments (Levins' B_N approaching 1/R) and, therefore, must either be dispersal limited (Stegen *et al.* 2013) or restricted by environmental properties or competition between taxa (MacArthur 1958). The strong specialist Levins' B_N results for *Bacilli*, *Gammaproteobacteria* and *Fusobacteria* was in agreement with previous work (Ilhan *et al.* 2017). Ilhan *et al.* (2017) identified these groups as being uniquely enriched after RYGB surgery. A more detailed look into specialists within these Classes enriched in the RYGB treatment, not performed in the original study, identified opportunistic *Lactobacillus*, *Streptococcus*, *Enterobacter*, *Escherichia* and *Klebsiella* as the specialist genera responsible for results observed at the level of Class. This surgery was found to be most effective at reducing body mass index ($kg\,m^{-2}$) post-bariatric surgery, and the enrichment of *Bacilli*, *Gammaproteobacteria* and *Fusobacteria* were linked to changes in short chain fatty acids, amino acids and sugars in faecal profiles of RYGB patients. The classification of these Classes (and Genera) as specialists provides strong quantitative support that physiological changes post-RYGB surgery drastically alter the gut environment in a manner not seen in controls, pre-RYGB or post LAGB surgery. These changes are extensive and include increased intestinal transit, altered bile acid metabolism, increased localised pH and bypassing of the stomach that can act as a filter for removing ingested microorganisms (le Roux *et al.* 2006). These changes are evidently necessary to support the activity and persistence of these specialists.

Levins' Overlap for comparing taxa co-occurrence

The LO is useful for identifying relationships between taxa. The strength of this approach does not lie in classifying a taxon as either a generalist or specialist, but rather in comparing taxa

pairs that compete for a shared resource more (or less) effectively within certain environments (MacArthur 1958) or a taxon that lacks a functional trait necessary to colonise and persist in certain environments (Hutchinson 1957). In environmental microbial ecology, an example would be methane oxidising bacteria of the genera *Methylocaldum* and *Methylococcus* that preferentially compete for methane in disturbed soil environments, whereas *Methylobacter* and *Methylomonas* preferentially compete for methane in undisturbed soil environments (Ho *et al.* 2013). The low LO values between $D_3 \times D_5$ and $D_4 \times D_5$ (Fig. 3) in conjunction with their significant, low Levins' B_N (Table 1) can be used to infer that these are specialists that may compete for a shared resource, but if so, this competition separates their populations in space. Alternatively, these groups may not have the physiological capacity to colonise and persist in the same environments. The high $LO_{1,2}$ values of D_1 support their nature as generalists that must occupy different niches than the specialist groups. The non-symmetrical $LO_{2,1}$ values of D_1 generalists is not necessarily intuitive, and Figure 2 (Supporting Information) aims to clarify this. As can be seen in Fig. 3, the D_2 linear decreasing group has the second highest $LO_{2,1}$ values with D_1. All other distributions, many of which are missing taxa in certain environments or are random, sparse counts, have relatively low $LO_{2,1}$ values with D_1. Thus, while the D_1 generalists overlap all other distributions because of their consistently high, cosmopolitan presence (yielding high $LO_{1,2}$ values) the other distributions vary in how well they overlap with D_1, following a gradient of $D_2 > D_5 > D_3 > D_4 > D_6$ (yielding variable $LO_{2,1}$ values dependent on the taxon being compared to D_1).

In the bariatric surgery case study, *Bacteroidia*, *Clostridia*, *Negativicutes* and *Betaproteobacteria* all had high $LO_{1,2}$ scores across other Classes reflective of their cosmopolitan nature across treatments (Fig. 4(A)). In the examples in Figure 2 (Supporting Information), these Classes are reflective of the generalist 'Taxon 1' that has consistently high $LO_{1,2}$ values regardless of being compared to another generalist or a specialist. None of these were classified as 'true' generalists, with Levins' B_N values of 0.838–0.925 barely falling short of the significance threshold imposed by the null model distribution. As relative abundances/proportions of populations are non-Euclidean, there is the potential for certain technical 'quirks' of the mathematics—substantial increases of specialists in RYGB treatments results in unequal proportions of *Bacteroidia* across the four environments. This may be why *Bacteroidia* are not sufficiently equal across the four environments to meet the significance threshold. This did not pose a problem with the *in silico* training dataset (Table 1), however, where the generalist D_1 populations were all substantially larger than others. Thus, as with all significance testing and p value assignment, care should be exercised in interpretation of the relevance of a p value. By pairing the Levins' B_N values with the visual LO heatmaps, it is possible to demonstrate that the niche of *Bacteroidia* clearly overlaps well with the majority of other Classes. In terms of specialists, *Fusobacteria* did not particularly overlap with any other Class. This is reflective of its highly scattered distribution of enrichment in two RYGB individuals. Interestingly, *Bacilli* and *Gammaproteobacteria* demonstrated relatively high overlap with each other. Consequently, we can speculate that these two Classes must have (different?) trait(s) that explain their persistence in individuals post RYGB surgery as direct competitors cannot overlap strongly.

Measuring how the environment determines niche with Hurlbert's B_N and PS

The two indices (Equations (2) and (3)) are of value to associate the importance of an environmental property in dictating a taxon's niche. Hurlbert's B_N and PS were both successful in identifying a relationship between non-linear specialist D_3 and D_4 groups with the gradient ($p < 0.001$) (Table 1). The values of Hurlbert's B_N and PS approaching 0 indicated that the abundance of D_3 and D_4 was greatest when the environmental property was relatively low. Conversely, values approaching 1 would have indicated that the abundance of D_3 and D_4 were greatest when the environmental property was relatively high. That these indices could not identify the linearly decreasing D_2 as having a relationship with the environmental gradient was surprising. Indeed, the abundance of D_2 was almost perfectly linearly correlated with the mock environmental gradient ($R^2 = 0.99$, $p < 0.001$ as determined by linear regression). An interesting finding of this study was therefore that Hurlbert's B_N and PS are perhaps better suited for testing a relationship between an environmental property and taxa that have a non-linear relationship, and that linear methods (*e.g.* regression or Pearson coefficients) could be used in conjunction with Hurlbert's B_N and PS.

In the biocrust case study, community composition differed between Black and Blue Grama sites (Fig. 5(A) and (B) as unweighted and weighted stacked bar charts, respectively). The larger, more diverse population of Cyanobacteria in the Black Grama control biocrust has been described as a result of being a mature, late-succession stage biocrust relative to the Blue Grama control (Couradeau *et al.* 2016). Fernandes *et al.* described highly contrasting responses of these two biocrusts to drought, with *M. vaginatus* dominated Blue Grama biocrust minorly affected by drought relative to Black Grama biocrust where a greater diversity of cyanobacterial taxa, such as *Scytonema* spp. and those in the *M. steenstrupii* complex, were established. Hurlbert's B_N supported this conclusion strongly—*M. vaginatus* was not affected by precipitation while *M. steenstrupii*'s distribution was strongly dependent on increasing precipitation (Table 3). Furthermore, by measuring and comparing Hurlbert's B_N between many taxa, we see that the distribution of four taxa (*M. repens*, Uncultured Nostocales, *W. murrayi* and *S. hofmanni*) are almost completely independent of precipitation (Weighted Hurlbert's B_N of 0.52–0.54), even moreso than *M. vaginatus* (Weighted Hurlbert's B_N of 0.664). Thus, we can hypothesise that *M. vaginatus* and several other cyanobacterial taxa, and not *M. steenstrupii*, are equipped with one or more functional traits that enable resistance to adverse effects of drought, although further experimental work in culture would be required to demonstrate this.

Fernandes *et al.* also found *Scytonema* to be affected by precipitation while we did not. The results here may differ as Fernandes *et al.* assigned cyanobacterial taxa from a curated Cyanobacteria database whereas the Silva132 database may have been unsuccessful in identifying specific *Scytonema* OTUs. Of further note was a discrepancy between Hurlbert's B_N and PS outcomes for *M. steenstrupii*—while B_N found a positive relationship with precipitation, PS did not. The PS was, in general, more stringent than B_N. Additionally, while the unadjusted p values for unweighted and weighted *M. steenstrupii* were below 0.05 (data not shown) the Benjamin–Hochberg adjusted P values were greater than 0.05. Thus, by increasing the higher stringency of PS results further by adjusting for false discovery rate we would

draw the conclusion that *M. steenstrupii* is unaffected by precipitation. As mentioned above, this emphasises the need for caution when interpreting *p* values (Colquhoun 2014).

Finally, the relationship between *M. steenstrupii* and precipitation was stronger when considering weighted abundance as 16S rRNA gene copies cm^{-2} of biocrust as opposed to relative abundance (%). The weighted data also identified novel cyanobacteria as being dependent on precipitation, whilst simultaneously not predicting a relationship between an uncultured *Chroococcidiopsaceae* (below the LOQ) and precipitation. This is because the weighted abundances more accurately reflect the detrimental effect of drought across the community and on specific taxa, such as *M. steenstrupii*. When only considering relative abundance (%), an increase in the abundance of one taxon may be due to the absence of another taxon rather than a true increase in OTU abundance. This point is made poignantly by the *Oxyphotobacteria* in Black Grama Drought sample 5–the absence of many taxa give the impression that *Oxyphotobacteria* abundance is 'high' at 80% in Fig. 5(A), whereas in reality the *Cyanobacterial* population is almost non-existent in this sample (Fig. 5(B)). In macroecology, the benefits of weighting the relative abundance of a taxon within a sample by the sum of individuals in that sample is well described (Griffin-Nolan *et al.* 2018). The results here certainly support the application of weighted OTU analyses in microbial ecology, particularly when investigating how the environment determines the abundance of taxa.

Some considerations for niche-based studies

The case studies to validate the four niche metrics were chosen as they had qualitatively demonstrated niche-differentiation at the taxonomic level of Class (bariatric surgery case study) and level of Species (biocrust case study). These can be considered as high and low taxonomic rank. An important consideration for niche-based studies is the potential underlying physiological traits that dictate a taxon's niche, and at what level of taxonomic rank these traits are likely evident. For example, Cyanobacteria derive the majority of their carbon from photosynthesis (Garcia-Pichel *et al.* 2019) while methanogenic Archaea are obligate anaerobes that perform the final step of fermentation, the reduction of carbon dioxide to methane (Garcia, Patel and Ollivier 2000). If one were to sample microbial communities in a water column from the surface to the anaerobic sediment, one could reasonably expect to find more Cyanobacteria at the surface and more methanogens in the anaerobic sediment, which could be considered their respective niches. Therefore, at this scale, niche differentiation can be observed at the high taxonomic rank of phylum. However, other, more specific physiological traits may only be evident at the species or OTU level. For example, the biocrust study analysed here demonstrated drought resistance in *M. vaginatus* and not *M. steenstrupii*, at low taxonomic rank. Similarly, certain (micro)niches may only become evident depending on the sampling scale, for example relationships between specific taxa and gradients of oxygen or water potential in soil aggregates. Such a relationship will not be seen when applying the niche indices to sequencing data derived from DNA extracted from bulk soil. Ultimately the usefulness of applying the four metrics to demonstrate niche-differentiation will depend on the user's hypothesis and experimental design.

A further consideration to note is that calculating niche indices based on amplicon sequencing data is dependent on certain traits being linked to an OTU. This suggests that testing niche-related hypotheses with OTU data is likely to be more successful when physiological traits necessary for that niche are encoded in the organism's chromosome. Mobile genetic elements that convey unique niches, for example *Escherichia coli* strain-specific antibiotic resistance plasmids (de Been *et al.* 2014), that are inconsistently linked to the gene being surveyed (*e.g.* 16S rRNA gene), will not yield appropriate niche metrics.

CONCLUSION

The concept of a taxon's niche is a useful mechanism to describe the relationship between a taxon and its environment and between a taxon and other taxa within a community. Presented here is a thorough investigation of several niche-related indices in measuring various metrics of microbial taxa from amplicon sequencing data. These indices are applicable to testing distinct niche-related hypotheses, including: (a) whether a taxon can be considered a generalist or specialist; (b) the extent by which the niches of two taxa overlap and (c) the role of an environmental property in determining the spatial distribution of a taxon. The case studies analysed here were chosen as previous work had demonstrated niche-differentiation between taxa. The application of the four indices to these case studies confirmed their utility in quantitatively comparing niche between microbial taxa and drawing biologically-relevant conclusions. The incorporation of an LOQ to identify taxa being incorrectly identified as specialists (based on *a priori* assumptions) was entirely novel to this work, not utilised in macroecological studies, and was a consequence of the nature of microbial amplicon sequencing data. Null models were also incorporated to support hypothesis testing (i.e. deriving a P value). We have made these indices available as the R package MicroNiche. It was concluded that these niche metrics hold value for investigating the ecology of microbial taxa, however, the interpretation of *p* values derived from these indices should always be considered with caution and in conjunction with other methodologies testing microbial physiology (i.e. in pure cultures or laboratory enrichments) to support biologically-relevant conclusions.

SUPPLEMENTARY DATA

Supplementary data are available at *FEMSEC* online.

ACKNOWLEDGMENTS

The authors wish to thank the Federation of European Microbiological Societies (FEMS) for funding D.R. Finn through the FEMS Research and Training Grant (FEMS-GO-2019–503).

Conflicts of interests. None declared.

REFERENCES

Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Statist Soc Ser A* 1995;**57**:289–300.

Bolyen E, Rideout JR, Dillon MR *et al.* Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol* 2019;**37**:852–7.

Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic unitts in marker-gene data analysis. *ISME J* 2017;**11**:2639–43.

Colquhoun D. An investigation of the false discovery rate and the misinterpretation of p-values. *R Soc Open Sci* 2014;**1**:140216. DOI: 10.1098/rsos.140216.

Couradeau E, Karaoz U, Lim HC *et al*. Bacteria increase arid-land soil surface temperature through the production of sunscreens. *Nat Commun* 2016;**7**:10373.

de Been M, Lanza VF, de Toro M *et al*. Dissemination of cephalosporin resistance genes between *Escherichia coli* strains from farm animals and humans by specific plasmid lineages. *PLos Genet* 2014;**10**:e1004776.

de Wit R, Bouvier T. 'Everything is everywhere, but, the environment selects'; what did Baas Becking and Beijerinck really say? *Environ Microbiol* 2006;**8**:755–8.

Elton C. Animal ecology, *Sidgwick and Jackson*. London, England, 1927.

Eurachem Working Group. Quantifying uncertainty in analytical measurement, 6/29/2020. In: Ellison SLR, Williams A (eds) *Eurachem / CITAC Guide CG 4*. Eurachem, Torino, Italy, 2012.

Feinsinger P, Spears EE, Poole RW. A simple measure of niche breadth. *Ecology* 1981;**62**:27–32.

Fernandes VMC, Lima NMM, Roush D *et al*. Exposure to predicted precipitation patterns decreases population size and alters community structure of cyanobacteria in biological soil crusts from the Chihuahuan Desert. *Environ Microbiol* 2018;**20**:259–69.

Fierer N, Ladau J, Clemente JC *et al*. Reconstructing the microbial diversity and function of pre-agricultural tallgrass Prairie soils in the United States. *Science* 2013;**342**:621–4.

Garcia-Pichel F, Loza V, Marusenko Y *et al*. Temperature drives the continental-scale distribution of key microbes in topsoil communities. *Science* 2013;**340**:1574–7.

Garcia-Pichel F, Zehr JP, Bhattacharya D *et al*. What's in a name? The case of Cyanobacteria. *J Phycol* 2019. DOI:10.1111/jpy.12934.

Garcia JL, Patel BKC, Ollivier B. Taxonomic, phylogenetic and ecological diversity of methanogenic Archaea. *Anaerobe* 2000;**6**:205–26.

Gause GF. Experimental studies on the struggle for existence I. Mixed population of two species of yeast. *J Exp Biol* 1932;**9**:389–402.

Griffin-Nolan RJ, Bushey JA, Carroll CJW *et al*. Trait selection and community weighting are key to understanding ecosystem responses to changing precipitation regimes. *Funct Ecol* 2018;**32**:1746–56.

Ho A, Kerckhof FM, Luke C *et al*. Conceptualizing functional traits and ecological characteristics of methane-oxidizing bacteria as life strategies. *Environ Microbiol Rep* 2013;**5**:335–45.

Hurlbert SH. The measurement of niche overlap and some relatives. *Ecology* 1978;**59**:67–77.

Hutchinson GL, Concluding remarks, *Cold Spring Harb Sym*. 1957;**22**:415–27.

Ilhan ZE, DiBaise JK, Isern NG *et al*. Distinctive microbiomes and metabolites linked with weight loss after gastric bypass, but not gastric banding. *ISME J* 2017;**11**:2047–58.

Jung S, Regan JM. Comparison of anode bacterial communities and performance in microbial fuel cells with different electron donors. *Appl Microbiol Biotechnol* 2007;**77**:393–402.

Leibold MA. The niche concept revisited: mechanistic models and community context. *Ecology* 1995;**76**:1371–82.

le Roux CW, Aylwin SJB, Batterham RL *et al*. Gut hormone profiles following bariatric surgery favor an anorectic state, facilitate weight loss, and improve metabolic parameters. *Ann Surg* 2006;**243**:108–14.

Levins R. *Evolution in Changing Environments*. Princeton University Press, Princeton, NJ, USA, 1968.

Ludwig J, Reynolds J. *Statistical Ecology*. Wiley Intersciences, Hoboken, NJ, USA, 1988.

MacArthur RH. *Geographical Ecology: Patterns in the Distribution of Species*. Harper and Row, New York, NY, USA, 1972.

MacArthur RH. Population ecology of some warblers of northeastern coniferous forests. *Ecology* 1958;**39**:599–619.

Martiny JBH, Bohannan BJM, Brown JH *et al*. Microbial biogeography: putting microorganisms on the map. *Nat Rev Microbiol* 2006;**4**:102–12.

Oksanen J, Guillaume Blanchet F, Kindt R *et al*. Vegan: Community Ecology Package. R package version 2.0-10. http://CRAN.R-project.org/package=vegan. 2013.

Penton CR, Gupta VVSR, Tiedje JM *et al*. Fungal community structure in disease suppressive soils assessed by 28S LSU gene sequencing. *PLoS One* 2014;**9**: DOI:10.1371/journal.pone.0093893.

Quast C, Pruesse E, Yilmaz P *et al*. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* 2013;**41**:D590–6.

R Core Team. R: A language and environment for statistical computing. *R Foundation for statistical computing*, Vienna, Austria, 2013.

Stegen JC, Lin X, Fredrickson JK *et al*. Quantifying community assembly processes and identifying features that impose them. *ISME J* 2013;**7**:2069–79.

Velasco Ayuso S, Giraldo Silva A, Nelson C *et al*. Microbial nursery production of high-quality biological soil crust biomass for restoration of degraded dryland soils. *Appl Environ Microbiol* 2016;**83**:e02179–02116.

Volterra V. Vartazioni e fluttuaziom del numero d'individui in specie animali conviventi. *Memoirs d'Accademia de Lincei* 1926;**6**:1–36.

Warnes GR, Bolker B, Bonebakker L *et al*. gplots: various R programming tools for plotting data. https://cran.r-project.org/web/packages/gplots/index.html. 2019.

Zhou J, Ning D. Stochastic community assembly: does it matter in microbial ecology? *Microbiol Mol Biol Rev* 2017;**81**:1–32.