TITLE

specificity: an R package for generalized specificity analysis, with example analyses of Hawaiian foliar endophytic Fungi, human gut microbiome bacteria, and Antarctic glacier bacteria.

RUNNING TITLE

Generalized specificity analysis of microbiome data

AUTHORS

John L. Darcy[1], Anthony S. Amend[2], Sean O. I. Swift[2], Pacifica S. Sommers[3], Catherine A. Lozupone[1]

[1] Division of Biomedical Informatics and Personalized Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, USA
[2] School of Life Sciences, University of Hawai‘i at Mānoa, Honolulu, HI, USA
[3] Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, CO, USA

ABSTRACT

Understanding the factors that influence microbes' environmental distributions is important for determining drivers of microbial community composition. Species distributions are governed by parameters that influence their dispersal or survival. These include environmental variables like temperature and pH, and higher-dimensional variables like geographic distance and host species phylogenies. In microbial ecology, "specificity" is often described in the context of symbiotic or host parasitic interactions, but specificity can be more broadly used to describe the extent to which a species occupies a narrower range of an environmental variable than expected by chance. Using a standardization we describe here, Rao's (1982, 2010) Quadratic Entropy can be conveniently applied to calculate specificity of a focal species to many different environmental variables, including 1-dimensional variables (vectors), dissimilarity matrices, phylogenies, and ontologies. We present our R package "specificity" for performing the above analyses, and apply it to four real-life microbial datasets to demonstrate its application: fungi living within the leaves of native Hawaiian plants, bacteria from the human gut microbiome, bacteria living within Antarctic glacier ice, and the Earth Microbiome Project data set. Finally, we present an interactive visualization and data exploration tool for these analyses, available as a companion R-package called "specificity.shiny".

INTRODUCTION

The word "specificity" has uses across multiple disciplines. In ecology, and especially for microbes, "specificity" is often used in the context of symbiotic interactions; for example the specificity of a parasitic species may be the degree to which it associates with a narrow consortia of host species (Combes, 2001; Poulin *et al.*, 2011; Shefferson *et al.*, 2019). In pharmacology and biochemistry, specificity can describe the "narrowness of the range of substances with which an antibody or other agent acts" (Oxford English Dictionary, 2017). Synthesizing these definitions, we arrive at a general concept of specificity, where a feature (*e.g.* a species) is specific to some variable (*e.g.* elevation) when it occupies or is otherwise associated with a limited breadth of that variable.

This definition is consistent across multiple variable types. For example, a species that is found only across a narrow band of elevation, perhaps between 200 and 500 meters above sea level would have stronger specificity to elevation than a species that is found between sea level and 1000 meters. This is

similar to a parasite that is found only within a narrow clade of hosts; it has stronger host specificity than a parasite that is found across a much wider phylogenetic range (Poulin *et al.*, 2011). This concept can be expanded even farther to diversity of some co-occurring feature class; for example metabolomic diversity in the human gut microbiome. Under our definition, a microbe may have specificity to a narrow range of metabolomic compositions, as measured by the beta-diversity of metabolites found across samples.

Generalized specificity analysis has several benefits over modeling a microbe's relative abundance using a variable of interest, since specificity analysis has no underlying model. First, microbiome data are notoriously zero-inflated. Disregarding the difficulties in modeling such data, which certainly can be overcome (Zhang and Yi, 2020), zero-inflated data are perfect for the "specificity approach". This is because the alternative hypothesis of a specificity analysis includes cases where the focal species only occupies a limited range of the variable of interest, being absent (zero) everywhere else. A further consideration in modeling approaches is non-monotonic relationships between species and environmental variables. For example, a species may have specificity to intermediate elevations, so its density function of elevation would be non-monotonic, or even multimodal; and that's just one species. Within a microbiome dataset, species may be expected to run the gamut of distribution shapes and modalities. Variables of interest also present their own challenges to modeling, since variables may be vectors (e.g. elevation, pH), distance or dissimilarity matrices (e.g. geographic distance, beta-diversity), phylogenies, or even sample-type ontologies. The generalized specificity approach we present here does not produce a model, or answer the question "across what range of the variable does the species occur", instead we quantify the extent to which the species occupies a limited breadth of that variable *without* the need for such a model.

Meaningfully applying this general idea of specificity to multiple data types is challenging because of the different specificity metrics available to different kinds of data. With host phylogenetic data, specificity may be calculated as phylodiversity (Faith, 1992), or host phylogenetic entropy (Allen *et al.*, 2009), or host richness (Costello, 2016). However, with other data types these metrics are not useful – one cannot calculate phylogenetic entropy of elevation, for example. Per our definition above, specificity must be a measure of the breadth (i.e. heterogeneity, diversity) of an environmental variable occupied by the focal species. With a variable like elevation, a naive specificity metric may be as simple as the variance in elevation where the focal species is present, or weighted variance for a more intuitive approach. However, such a metric would not be applicable to phylogenetic data sets because it is limited to 1-dimensional data types (*i.e.* column vectors). Furthermore, we wanted our general idea of specificity to be useful for dissimilarity matrices. We found that Rao's Quadratic Entropy (Rao, 1982, 2010; Botta-Dukát, 2005) is a convenient diversity metric that can be applied to all abovementioned data, with a modicum of standardization (detailed in our Methods section).

Here, we present a software package written in R and C++ that implements a generalized specificity analysis. Our package, *specificity*, calculates specificity values for each species in a sample-by-species matrix, commonly referred to in microbiology as an "OTU table". We simulated species distributions with varying strengths of specificity, and used those simulated data to validate our implementation. Our simulations were also used to ensure that specificity is not sensitive to occupancy (i.e. in how many samples a species appears), which is a significant improvement compared to previous methods, especially those that used un-weighted (presence-absence) species data. Our simulations also confirmed that the specificity we calculate here is scale-invariant with regard to environmental / phylogenetic data, and also focal species abundance data. To illustrate how *specificity* can be used, we applied our software to four previously-published microbiological data sets, each from different environments:

fungi living within the leaves of native Hawaiian plants, human gut microbiome bacteria, bacteria living within Antarctic glaciers, and the global Earth Microbiome Project data set.

METHODS

RQE

We calculate specificity using Rao's (1982, 2010) metric. It is sometimes abbreviated FD_Q for quadratic functional diversity, but since we use the same mathematics in a non-functional context, here we simply refer to the metric as RQE (Rao's Quadratic Entropy), similar to the use of "QE" by its inventor. RQE is the sum of the lower-triangle of the elementwise product of two square matrices. In our use, the first matrix (*D*) is a dissimilarity matrix containing differences between samples (Figure 1). For example, in the context of phylogenetic specificity these differences are phylogenetic distances (*i.e.* cophenetic distances) between hosts. Samples from the same host species have 0 distance. The second matrix (*W*) contains all pairwise products of weights for the focal species. Given a column vector of species weights *p* from a site-by-species matrix ("OTU table"), W_{ij} contains the product of the abundances (weights) of the focal species at sample *i* and sample *j*: $p_i p_j$. Via $D°W$ (or $D_{ij}p_ip_j$ for a single pairwise product; Equation 1), we weight matrix *D* to up-weight distances between samples where the focal species occurred, and down-weight distances between samples where the focal species was absent in either. We use the term "weights" to describe *p* because the values within could be relative abundances or any other metric that describes the importance of a species within a sample. Conversely, we have chosen to focus this manuscript on "species", but note that *p* could be a vector of weights for any feature (a type of rock, a metabolite, *etc*).

$$RQE = \sum_{i=1}^{s-1} \sum_{j=i+1}^{s} D_{ij}\, p_i\, p_j \quad (1)$$

With RQE, a focal species with strong specificity has relatively high weights for low differences. This metric was originally developed for phylogenetic distances, but here we apply it to many different *D* matrices, including euclidean transformations of 1-dimensional data (e.g. pairwise elevational difference), or more complex 2-dimensional data like Bray-Curtis dissimilarity between host metabolomic profiles.

As such, a species with "perfect" specificity will always have RQE = 0. For example, consider a focal species S that can be found in habitats A, B, C, or D, with multiple samples collected for each category (Figure 1).

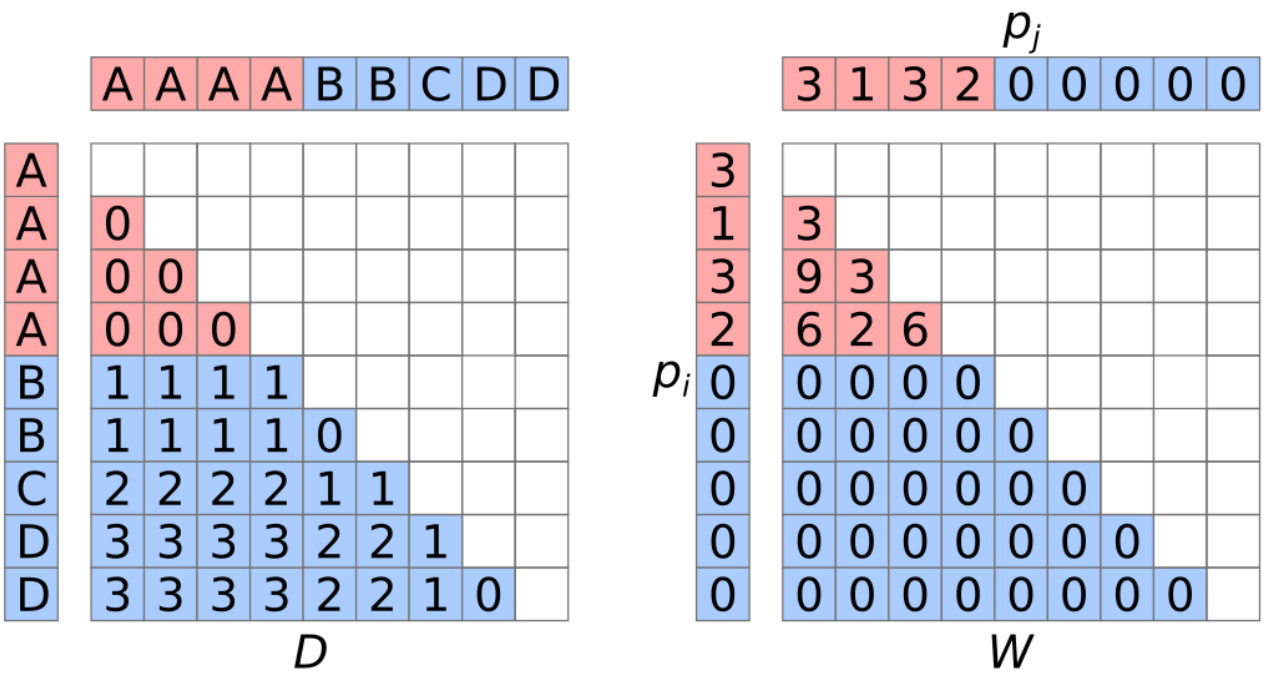

Figure 1: RQE as it applies to specificity. In this example, two matrices are shown, *D* and *W*. *D* is an environmental dissimilarity matrix, describing how different are several environment types, A through D, with multiple samples represented for each environment type. Matrix *W* is the pairwise product of species weights *p* (Equation 1). In this example, the focal species is perfectly specific to habitat A, which can be seen in *p*. Data corresponding to species detections are colored in red, and species absences in blue. The product $D°W$ ($=D_{ij}p_ip_j$) will be all zeroes for this example, because this example shows perfect specificity. Thus, the sum of that product, RQE, will be zero. If *p* had relatively small values instead of its zeroes, for example 0.25, those small values would still down-weight their corresponding larger differences in *D* and produce a signal of specificity, compared to random permutations of *p* which produce much higher RQE values.

If S is only found in samples from habitat A, matrices *D* and *W* will contain zeroes in opposite positions, resulting in RQE=0. Note that weights near zero can also act similarly to zero since this is a weighted metric. In this way, a focal species can occupy every single sample (all values of *p* are positive) and still have RQE near zero. This is important so that spurious species detections do not significantly contribute to specificity. For example, in a DNA sequencing experiment, small amounts of contamination may occur during DNA extraction or library preparation. The magnitude of that contamination is expected to be small compared to the signal in an actual sample, but may result in spurious species detections. However, because these contaminants would be expected to be rare in the sample, their weights would be low in samples where they are noise, and high in samples where they are signal.

Standardization

While empirical RQE is calculated as described above, it must be standardized in order to compare effect size between different variables and different species, because the metric's scale is dependent on the scales of *D* and *p*. Phylogenetic specificity approaches have previously used a standardized effect size (SES) approach (Poulin *et al.*, 2011), but we found that SES has unfortunate properties when used with our generalized approach to specificity. Critically, SES was highly sensitive to occupancy, which is the number of samples a species occupies. One would expect the strength of specificity to be lower when a species occupies more samples, because this means the species must occur in a broader range of habitat. However, SES counter-productively yields stronger specificity for more occupant species, and

also for species with more even distributions. This is because SES is standardized using a distribution of values generated with permuted species weights. If all weights are similar (high evenness), the standard deviation of that distribution will be small, leading to a strong SES. SES is also undesirable because for a given species it is tightly associated with that species' *P*-value (probability that empirical specificity is lower than null), enough so that a suggested remedy for yet other problems with SES is to use a probit transformed *P*-value in its place (Botta-Dukát, 2018).

We standardize RQE by leveraging the fact that for perfect specificity, empirical RQE (RQE_{emp}, Figure 1) equals zero. Our statistic, which we simply call *Spec*, ranges from -1 to 1, with 0 as the null hypothesis that species weights are randomly ordered with regard to sample identity. Similar to SES, RQE_{sim} is a vector of RQE values calculated using random permutations of *p*. The distribution of RQE_{sim} is used for calculating *P*-values, and its central tendency is defined as *Spec* = 0. This central tendency can be calculated several ways using our software, but the default is to use the mean of RQE_{sim} since in our testing, distributions of RQE_{sim} were unimodal and exhibited low skewness (not shown). The equation for *Spec* (Equation 2) is a piecewise function, with the two parts corresponding to specificity and generality, respectively. In the case of specificity, *Spec* simply scales RQE_{emp} relative to the center of RQE_{sim} such that perfect specificity returns -1, and the null hypothesis returns 0. In this case, "null hypothesis" refers to RQE_{emp} being the expected value of RQE_{sim}.

$$Spec = \begin{cases} RQE_{\text{emp}} \leq \overline{RQE_{\text{sim}}}, & \dfrac{RQE_{\text{emp}} - \overline{RQE_{\text{sim}}}}{\overline{RQE_{\text{sim}}}} \\ RQE_{\text{emp}} > \overline{RQE_{\text{sim}}}, & \dfrac{RQE_{\text{emp}} - \overline{RQE_{\text{sim}}}}{RQE_{\text{max}} - \overline{RQE_{\text{sim}}}} \end{cases} \quad (2)$$

The case of generality is slightly more complicated, since there is no intuitive maximum theoretical RQE value. However, for each *p* there exists an optimized permutation that yields the highest possible RQE value, RQE_{max}. We use a genetic algorithm (GA) with Population Based Incremental Learning (Baluja and Caruana, 1995) to search permutations of *p* that create RQE_{max}. Our GA begins with a population of surrogate vectors initialized via random permutations of *p* (default 150), and random swaps of *p* (a swap being the pairwise substitution of two values within the vector; default 150), and also *p* itself. Each generation, the GA calculates RQE for each vector in the population, then keeps some of the vectors with the highest RQE value (default 5). The next generation is composed of those kept vectors, and random swaps thereof until the total original population size is met (default 301). Our swapping algorithm can also use a stochastic number of swaps per vector per generation (including initialization), drawn at random from a user-defined set (default {1,1,2,3}). In addition to swapping, mutation can be performed by crossover via the PMX algorithm (Goldberg and Lingle, 1985), which is used because it incorporates both order and position of both parents, which is required for this problem. However, in our testing we found that crossover did not improve GA efficiency, so the default operation is not to perform PMX. The GA runs for a fixed number of generations (default 400), or until a number of generations have passed with no improvement (default 10). These parameters were chosen because they performed well on the data sets we analyze here, meaning that species reached the early termination condition.

Our GA is relatively computationally intensive, consuming the majority of computational time for a given specificity analysis even though it is only used for a minority of species. This is unlikely to be a concern on smaller data sets (i.e. a few hundred samples), but since many users may not be interested in "general" species, another option is to scale *Spec* for all species using the top half of Equation 2 instead. This is considerably faster, and the user can either discard "general" species as uninteresting, or

choose to interpret *Spec>0* within a ordinal framework (a brief analysis showed the results of this approach and those of the GA are strongly correlated; not shown).

Hypothesis testing

For the *Spec* calculation above, a *P*-value may be calculated as the proportion of RQE_{sim} values that are lower than RQE_{emp}. The default operation of our software is to adjust P-values calculated for different species from the same variable for multiple hypothesis testing by applying the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995).

Features of Spec

Spec captures signal of specificity in simulated 1D (vector), 2D (matrix), and phylogenetic data (Supplementary Figure S1, S2). It is insensitive to species occupancy (Supplementary Figures S3, S4) and is insensitive to the number of samples within a data set (Supplementary Figure S5). *Spec* is also scale-invariant, independently in regard to *p* and *D* inputs (Supplementary Figure S6). It is sensitive to multimodality, and multimodal species distributions are still detected as exhibiting significant specificity by *Spec* (Supplementary Figure S7).

Software implementation and availability

Our software, *specificity*, is available as an R package and can be downloaded from GitHub: https://github.com/darcyj/specificity. It is written in R and C++. *specificity* is free and open source software, licensed under the Gnu Public License (GPL). Installation is easily done using the "install_github" function of the R package *remotes* (Hester *et al.*, 2020). In addition to software, our GitHub repository contains thorough documentation, including guidance on installation, and a full tutorial vignette for using *specificity* with examples from an included data set.

Validation analyses

Species were simulated with varying levels of specificity by drawing from a normal distribution centered on an artificial "optimum" environmental location (e.g. elevation of 300 meters). Varying specificities were achieved by widening the standard deviation of that distribution, or by mixing the normal distribution with varying proportions of a uniform distribution. Multimodal specificity was simulated similarly by combining multiple distributions. Specificity of simulated species was analyzed using our software. Occupancy of simulated species was increased or decreased by randomly substituting simulated weights with zero, and specificity was analyzed across an occupancy gradient using that approach. Real data (see Endophyte analysis, below) were randomly downsampled to test the sensitivity of *Spec* to sample size. Real data were also re-ordered to create simulated high-specificity species that use empirical distributions of weights, and then those simulated species were subjected to a swapping algorithm that gradually introduced entropy into the species. The swapping algorithm takes a species weights vector, and swaps values for two positions in *p* (Equation 1). This was done recursively for 1000 generations (2 swaps per generation), saving the species vector each time. Our software was then run on all species vectors simulated this way.

Analysis of endophyte data

Data from Hawaiian foliar endophytic fungi (Darcy *et al.*, 2020) were downloaded from FigShare. These are illumina MiSeq data of the Internal Transcribed Spacer (ITS) region of fungal ribosomal RNA, from 760 samples collected from the leaves of native Hawaiian plants across five islands in the Hawaiian archipelago. This data set is also included in our R package. The features under investigation in this analysis were fungal OTUs. Data were transformed ("closed") using total sum scaling, and fungal OTUs present in fewer than 10 samples were excluded from specificity analysis. The remainder (416 OTUs) were run through our software using default settings except run in parallel using 20 CPU

cores. Variables used in this analysis were NDVI (an index of vegetation density), elevation, evapotranspiration, rainfall, host plant phylogeny, and geographic distance between sample sites. Results were visualized using our built-in visualization tool, also using default settings.

Analysis of Antarctic bacteria data

Data from Antarcic cryoconite hole bacteria (Sommers *et al.*, 2018) were obtained from the authors. Cryoconite holes are isolated melt pools on the surface of glaciers, caused by debris from nearby slopes falling onto the glacier, and then melting into its surface. These holes form discrete microbial communities that have been described as "natural microcosms" (Sommers, Porazinska, *et al.*, 2019). This data set comprises 90 samples across three adjacent glaciers, and features are bacterial (16S) and eukaryal (18S) ASVs. Eukaryote data were subset for algal ASVs, and algal beta-diversity was calculated as Bray-Curtis dissimilarity. Analyses on these data were run and visualized as above, with variables N (Nitrogen), P (Phosphorus), pH, geographic distance, fungal Bray-Curtis beta-diversity (calculated from 18s rDNA data), and algal beta-diversity. Bacterial associations with individual glaciers (*e.g.* "OTU4 is found predominantly on Canada Glacier") were computed using Dunn's test (Dunn, 1964), which is a nonparametric post-hoc test of difference in means.

Analysis of Human microbiome data

Data from Franzosa et al. (Franzosa *et al.*, 2019) were downloaded as supplemental data from the online version of the article. These data contain both gut microbiome data as well as corresponding metabolomic data, collected from 220 adults with Chrohn's disease, ulcerative coliutis, or healthy controls. Data were downloaded in a processed state, after the following procedures had been completed: microbiome data from this study were derived from metagenomic data, which were assigned taxonomy and grouped into OTUs using MetaPhlAn2 (Truong *et al.*, 2015), and excluded samples that did not meet a 0.1% relative abundance threshold in at least 5 samples. Metabolite data were measured using positive and negative ion mode LC/MS, and were reported as parts per million. Metabolite identities were assigned programatically, and were clustered into broad classes per the Human Metabolome Database (Wishart *et al.*, 2018). We subset the matrix of metabolome data by those classes, and used Euclidean distance to calculate the extent to which any two given samples differed in metabolomic composition within a given class (e.g. "how different is the composition of bile acids between sample A and sample B?"). Metabolite classes were excluded if they were totally absent in any sample, or if they contained fewer than 10 metabolites, which left 83 classes. *specificity* was used to calculate Spec for microbial OTUs to each metabolite class distance matrix.

Analysis of Earth Microbiome Project data

Data from the Earth Microbiome Project (EMP)(Thompson *et al.*, 2017) were compiled and downloaded from Qiita (Gonzalez *et al.*, 2018). These data comprise a global sampling of 16S ASVs produced by multiple studies. All of the studies followed a uniform protocol for collection, processing, and analysis of microbial data. A major component of the EMP is a rigid sample type ontology. The EMP Ontology (EMPO) was designed to categorically represent two main drivers of bacterial community composition: host association and salinity, for each sample that was collected. At the broadest level (EMPO1), samples were categorized as being host-associated or free living. At the intermediary level (EMPO2), samples were further divided into saline, non-saline, animal, plant, and fungus. The finest level (EMPO3) separated samples into 22 discrete substrate types (e.g. water (saline), plant corpus, animal distal gut).

Because this data set is so large (28,842 samples and 309,469 ASVs), ASVs were excluded that were not present within at least 30 samples, samples were discarded if they had fewer than 5,000 reads, and samples without ontological data were discarded (leaving 25,188 samples and 7,014 ASVs). The EMP

ontology was transformed into a phylogeny using *specificity*'s "onto2nwk" function, which makes a cladogram within which all branch-lengths were set to 1. Specificity analysis was run using the ASV table and the ontological data. Taxonomy for individual species of interest was obtained using nucleotide BLAST (Zhang *et al.*, 2000) via the NCBI web portal, using the 16S ribosomal RNA sequence database as reference.

RESULTS AND DISCUSSION

Hawaiian endophyte specificity analysis

We found that foliar endophytic fungi (FEF) from within the leaves of native Hawaiian plants exhibited strong and statistically significant specificity to several environmental variables (Figure 2), including variables that were only weakly associated with FEF community composition (Darcy *et al.*, 2020). For example, in the original paper, rainfall and elevation were relatively weak predictors of FEF phylogenetic beta-diversity, but many FEF species show strong specificity to those variables in the analysis presented here. This reflects a fundamental difference between community-centric approaches (e.g. FEF community composition) vs. species-centric approaches like (e.g. specificity analysis). The signal of individual species is lost when a community is aggregated into a beta-diversity matrix or similar, and consequentially individual species within the community may even respond to environmental variables orthogonally to the community as a whole. Species that were strongly specific to rainfall or elevation are examples of this.

However, the aggregate strength of specificity as visualized in Figure 2 is still a community-level overview, albeit unweighted by species abundance. Here we see that many species have strong and statistically significant specificity to geographic location, which makes sense given the discrete spatial structure of the Hawaiian islands (Tipton *et al.*, 2021), and that these FEF communities only are spatially structured up to distances of 36 kilometers (Darcy *et al.*, 2020). But geographic specificity may be an artifact of specificity to other variables with strong geographic autocorrelation. For example, in Hawaii (and elsewhere), rainfall is a spatially structured phenomenon (Giambelluca *et al.*, 2013), with nearby areas experiencing dramatically different rainfall averages as a consequence of aspect and elevation. Thus, species that have strong specificity to rainfall likely also have strong specificity to geography, which is true in our analysis. Indeed, the same is true for elevation, albeit to a lesser extent.

One of the fungi with strongest specificity in our analysis was of the genus *Harknesia*, with closest BLAST match in the NCBI nucleotide database to *H. platyphyllae*, a eucalyptus pathogen. In our data set, this fungus was found on multiple hosts, including *Metrosideros polymorpha*, which is in the same family as eucalyptus (Myrtaceae). This *Harknesia* was found exclusively within the interior of Hawaii island, at elevations between 1700 and 2000 meters above sea level. Likely because of its strong geographic and elevational specificity, it also exhibited strong specificity to evapotranspiration, only being found in areas with 40 to 50 mm per year. Other fungi, such as an ASV most closely related to *Phaesosphaeria papayae*, exhibited strong specificity to elevation, evapotranspiration, and NDVI, but were found on multiple islands across the Hawaiian archipelago. Thus, while geographic specificity can appear as specificity to geographically autocorrelated variables, this is not always the case.

Notable generalists in the endophyte dataset include the genus Colletotrichum, a globally distributed genus of plant pathogenic and endophytic fungi. Almost all agricultural crops are impacted by members of Colletotrichum and it is considered a 'top ten' fungal pathogen for molecular plant pathological research (Dean *et al.*, 2012). Of the 9 ASVs identified as Colletotrichum, none showed specialization to plant host or geography. Recently, this genus has been subjected to close genomic scrutiny to provide insight into the genetic mechanisms behind host generalism and the activation of latent pathogenicity

(Gan *et al.*, 2013, 2016). Low specificity to geography and host within this dataset indicates that asymptomatic Colletotrichum species are widespread within the native Hawaiian flora.

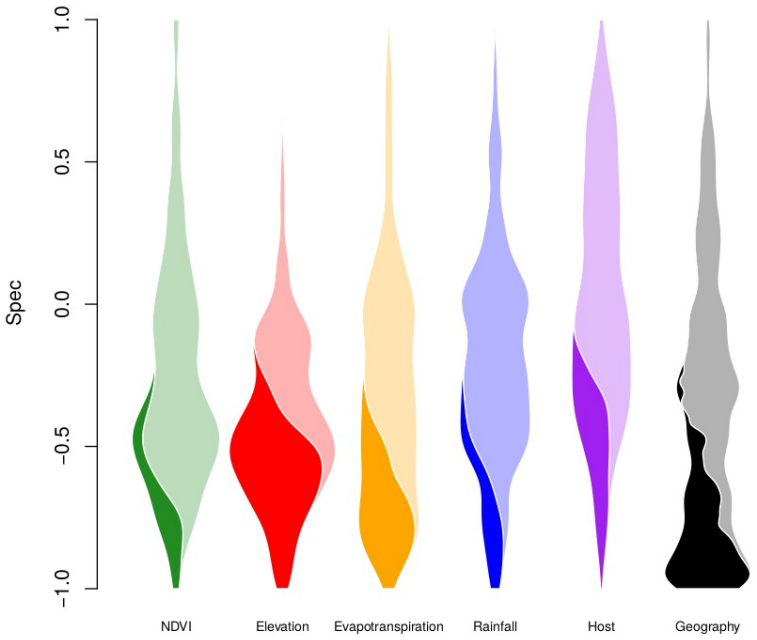

Figure 2: Specificity of Hawaiian Foliar Endophytic Fungi. In this plot, the *Spec* values for 416 fungal species are plotted as violins for different variables. Since the number of species is the same for each variable, each violin has the same total area. Violin area is divided between species with statistically significant specificity (dark) vs. species without (light).

Antarctic glacier bacteria specificity analysis

Similar to the FEF analysis above, bacteria living in cryoconite holes on glaciers in Antarctica's Taylor Valley exhibited strongest specificity to geographic distance (Figure 3). This data set spanned three glaciers: Canada, Taylor, and Commonwealth, with equal sampling on each, but geographic distance accounts for distances within glaciers as well. The strong geographic specificity observed here reflects bacteria that are deferentially abundant among glaciers, for example occupying only one or two of the three. The three glaciers, each flowing into a separate lake basin, are spaced along the 40-km length of the Taylor Valley, which stretches from the polar ice sheet to the coast. The cryoconite holes from the most inland glacier, Taylor Glacier, are more oligotrophic than those nearer the coast, on Commonwealth Glacier (Porazinska *et al.*, 2004). The more inland cryoconite holes also have the lowest diversity of bacteria, while the holes nearest the coast support the most diverse bacterial communities (Sommers *et al.*, 2019). Many bacterial species may therefore be specific to Commonwealth Glacier because it supports more species within its cryoconite holes than the other two glaciers. Besides the differences in bacterial richness among the glaciers, the composition of the bacterial community turns over among glaciers, with different sequence variants dominating each glacier (Sommers *et al.*, 2018). Biogeochemical differences within cryoconite holes among glaciers furthermore correspond with biogeochemical gradients along the valley in the surrounding soils (Porazinska *et al.*, 2004). The difference in dominant bacterial taxa on each glacier may primarily reflect (1) differences in which bacteria dominate the soils surrounding each glacier and therefore disperse onto each glacier, (2) a response to biogeochemical conditions within cryoconite holes on the glacier, or (3) an interaction of the two. Experimental microcosms manipulating dispersal and nutrient availability could help to parse out dominant controls on geographic specificity of bacteria in the future.

Strong bacterial specificity to co-occuring algal communities was expected, given the strong correlation between bacterial and eukaryal diversity previously observed in this supraglacial system (Sommers *et al.*, 2018) and elsewhere (Darcy *et al.*, 2017). In our analysis, we found that specificity to algae was strongly negatively correlated with specificity to phosphorus (r=-0.69); even though those two variables were not strongly correlated with each other (Mantel correlation=-0.05). In other words, bacteria that are specific to algal community composition are not specific to sediment phosphorus concentration, and *vice-versa*. Using post-hoc tests, we found that bacteria with strong specificity to phosphorus concentration were predominantly associated with Taylor Glacier (but not exclusively), and that bacteria with strong specificity to algal community composition were predominantly found on Commonwealth Glacier.

Similarly, the correlation in *Spec* between geographic distance and algae (Figure 3B) highlights a feature of specificity analysis using *Spec*: when comparing specificity of the same features to two different variables, *Spec* is likely to be strongly correlated when the variables have a linear relationship. That linear relationship can be seen in the Mantel correlation between pairwise geographic distance and algal beta-diversity (r_M=0.52; Figure 3B). But with variables that are weakly correlated, *Spec* may or may not be correlated between the variables. For example, difference in phosphorus concentration is not correlated with algal beta-diversity, but bacterial specificity to those variables is correlated (Figure 3B).

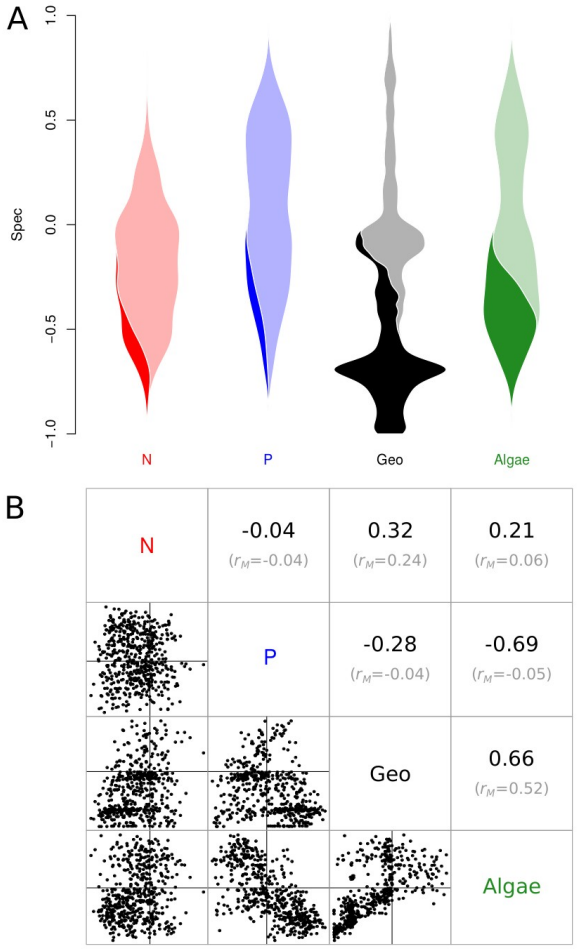

Figure 3: Specificity of Antarctic cryoconite hole bacteria. A: Violin plot of specificity to sediment nitrogen concentration (N, red), sediment phosphorus concentration (P, blue), geographic distance (Geo, black), and algal community composition (Algae, green). Violins are scaled and shaded identically to those in Figure 2. B: Pairwise *Spec* correlations. Correlation coefficients (r) for each pairwise comparison are shown in this subplot's upper triangle, with mantel correlation coefficients (r_M) shown below in gray. Mantel correlations

describe the relationship between the variables themselves, e.g. “is algal community dissimilarity correlated with geographic distance”. The *Spec* correlations shown above in black are the correlation values for the data plotted in the lower triangle of this subplot.

Human gut microbiome metabolomic composition specificity

In this analysis, we asked whether bacteria and archaea in the human gut microbiome had specificity to paired metabolomic data. We computed bacterial specificity to compositional dissimilarity of 83 different metabolomic classes, and of those 83 microbes showed statistically significant specificity to only 25 (Figure 4). The interpretation of this analysis is similar to that of the specificity to algal community composition above. Specificity to a certain composition of paired data is a more abstract concept than specificity to elevation or even to the EMP ontology, but this type of specificity makes intuitive ecological sense in the context of species-habitat association. Different microbes have different environmental needs, both within the human gut microbiome (Costello *et al.*, 2012) and elsewhere (Nemergut *et al.*, 2013). As such, those microbes can be expected to be found in environments that meet those needs. Similarly, microbes in the human gut influence their environment (Kelly *et al.*, 2015), and as such can be expected to be found in environments that are changed by their presence. Since different microbes interact with different sorts of metabolites, differential specificity to metabolite classes is an expected outcome.

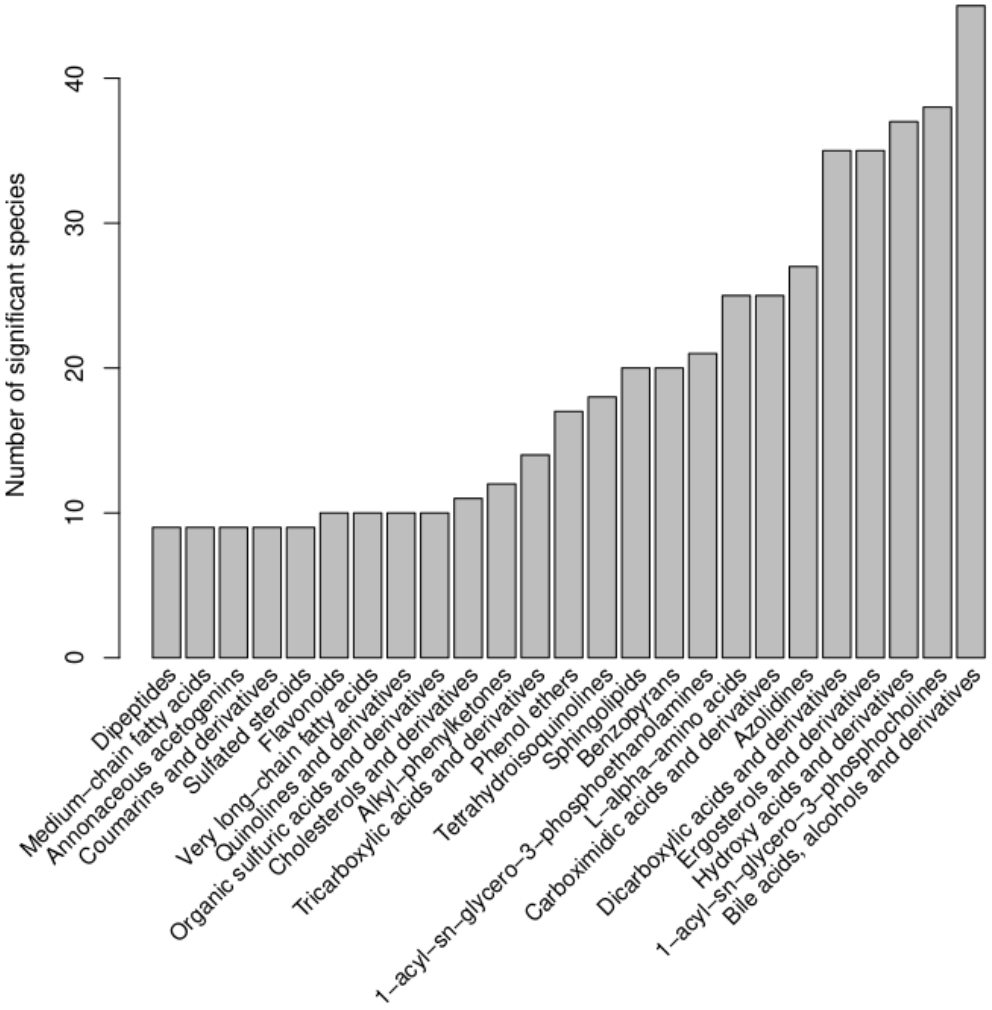

Figure 4: Count of bacterial (and archaeal) species that exhibited significant specificity to metabolomic composition in human gut microbiome data. Each bar represents the number of species (out of 201 total) that showed significant specificity to the composition of a given metabolomic class. Ordering of metabolomic classes on the horizontal axis is arbitrary.

We found that more microbial species had significant specificity to the composition of bile acids, alcohols, and derivatives, compared to other metabolomic classes (Figure 4). This result is not surprising, since bile acids interact with the gut microbiome, and are also manipulated by it (Ridlon *et al.*, 2014). Furthermore, the experimental design for these data contained subjects with Crohn's disease, with ulcerative coliutis, and healthy controls; and bile acids play a significant role in both Crohn's disease and ulcerative colitis (Tiratterra *et al.*, 2018). Microbes in this analysis could be specific to either of those two conditions and their plausibly co-occurring bile acids, alcohols, and derivatives, or to subclasses thereof. Composition of this metabolomic class did not strongly correlate with composition of other classes; its highest Mantel correlation was with Benzopyrans (r_M=0.46).

Species with the strongest specificity to bile acids *etc.* were *Bacteroides plebeius,* an unclassified *Methanobrevibacter* species, *Odoribacter laneus, Methanosphaera stadtmanae*, and *Ruminococcus callidus* (all *Spec* < -0.60), although many other species showed significant specificity to this metabolomic class as well. *B. plebeius* was initially isolated on bile media (Kitahara *et al.*), and was found to be associated with primary bile acids (as opposed to secondary bile acids) in patients with pediatric Crohn's disease (Connors *et al.*, 2020). *Methanobrevibacter* sp (likely *M. smithii*) and *Methanosphaera stadtmanae* (Archaea) are the predominant methanogens found in the human gut (Brown *et al.*, 2018). *M. smithii* is known to grow in the presence of bile salts (Miller *et al.*, 1982), and may be a biomarker against IBD or for IBD remission (Ghavami *et al.*, 2018). The metabolomic class with the second most specific microbes was 1-acyl-sn-glycero-3-phosphocholines (also called 2-lysolecithins). These compounds are derivatives of phosphatidylcholine, which is used as a treatment for ulcerative colitis, but is also found naturally in some foods (Kokkinidis *et al.*, 2017). This finding is different than showing some bacterial species' relative abundances correspond to the amount of phosphatidylcholine derivatives; instead this analysis focuses on the composition of phosphatidylcholine derivatives; albeit with the amount of those compounds as a component since Euclidean distance was used.

In our analysis, we asked which metabolomic classes had the most microbial species specific to them (Figure 4). The results of this type of analysis are intended to mirror common beta-diversity analyses used in microbial ecology, which ask to what extent variables explain differences in microbiome community structure (Sommers *et al.*, 2018; Darcy *et al.*, 2020; Shaw *et al.*) However, more complex questions can be asked of these data, using the results of *specificity* as a starting point for feature set reduction or variable selection. For example, given an individual bacterial species of interest, the variables to which it is specific may be used in a random forests model to predict its presence. For the purpose of variable selection, *specificity* is very lightweight when used with only the top half of Equation 2 (option: `denom_type=sim_center`). This mode produces the same P-values as the more comprehensive mode, and produces the same *Spec* values for any species with *Spec*<0. In addition to variable selection, *specificity* has application in detecting species that may be common lab contaminants, as shown in our EMP analysis below.

Earth Microbiome Project (EMP) ontological specificity

As expected, the vast majority (6,909 / 7,014) of bacterial ASVs we analyzed within the EMP data set exhibited significant and strong specificity to the EMP ontology (Figure 5). Given the distinct community-level differences in microbiomes across this ontology (Thompson *et al.*, 2017), it is not a surprise that most microbial species exhibit the same pattern. Instead, the species that buck this trend are of interest as potential cosmopolitan taxa, or possible bioinformatic failures. The sequence data obtained from the EMP data set are only 91 base-pairs in length, and even though they were clustered

as exact sequence variants (Callahan *et al.*, 2016, 2), it is possible that environmentally divergent ecotypes (Cohan, 2002) were clustered together into the same ASV in this way.

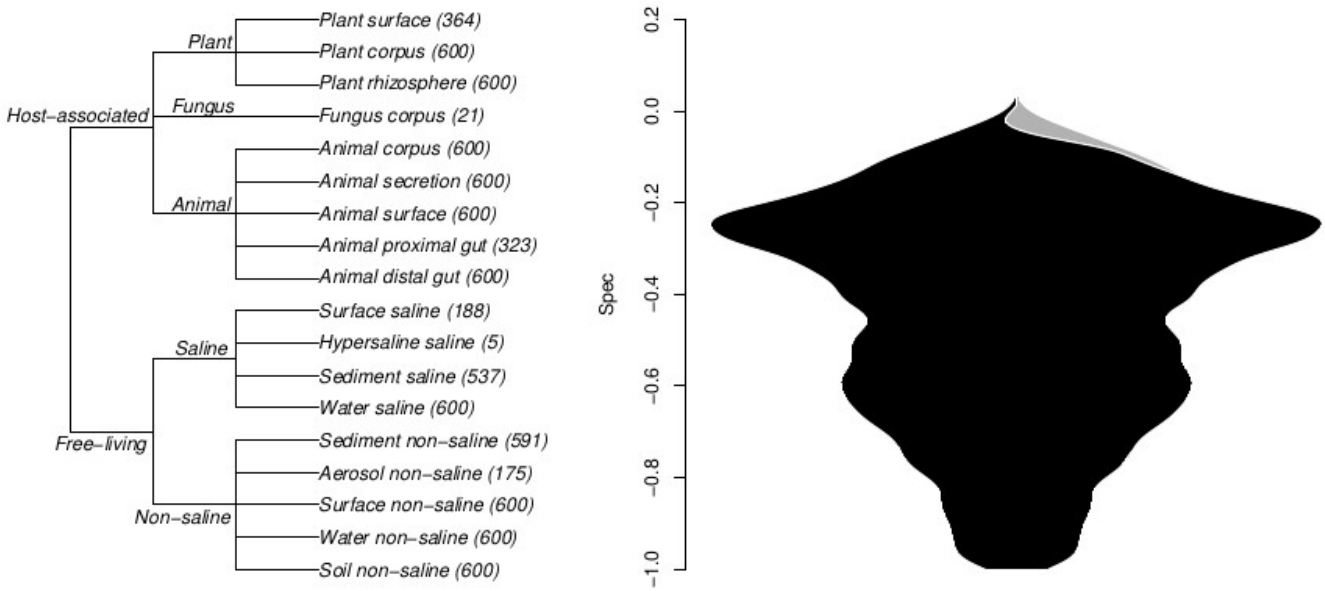

Figure 5: EMP sample type ontology visualized as a cladogram (left), and distribution of *Spec* values for micobial specificity thereto (right). Within the cladogram, all brach-lengths are set to 1 for calculation of a cophenetic distance matrix, which is used as *D* when calculating *Spec* (Equations 1 and 2). Numbers in parentheses are sample counts. An overwhelming majority of microbes within the EMP data set showed statistically significant specificity to sample type ontology (black, right), which was expected given the previously reported strong signal in clustering of community composition by sample type (Thompson *et al.*, 2017). Very few microbes did not show significant specificity (gray).

One such highly distributed ASV was found in almost all EMP ontology categories except for saline, hypersaline, and non-saline water (*Spec*=-0.13, *P*=0.35). It was 100% identical to multiple *Actinomadura* species, including *A. algeriensis* from Saharan soil (Lahoum *et al.*, 2016), the human pathogen *A. madurae (Izri et al.*, 2020), the root endophyte *A. syzygii* (Rachniyom *et al.*, 2015), *A. maheshkhaliensis* from mangrove rhizosphere soil (Ara *et al.*, 2008), and *A. apis* from honeybees (Promnuan *et al.*, 2011), and others. These environment types span the EMP ontology, suggesting that the highly distributed ASV in question is a spurious combination of multiple *Actinomadura* species, each of which would likely have a strong specificity signal if analyzed independently. This example illustrates that like with every analysis, the results can only be as good as the input data. Users with very short-read marker gene data are likely already aware of this limitation, so we will not belabor the point.

Interactive data visualization

In addition to using our *specificity* R package to calculate *Spec* and produce the figures shown above, we also used its companion package, *specificity.shiny*, to explore data and identify interesting features (Figure 6). With this tool, users can easily create interactive visualizations from *specificity* analyses, and share them over the internet. *specificity.shiny* was used in the preparation of this manuscript, to share results between authors.

CONCLUSION

Our R package, *specificity*, enables specificity analysis of microbiome data in the context of multiple variable types. Here, we've shown examples of specificity to geographic variables like elevation and rainfall (Figure 2), host phylogenetic specificity (Figure 2), specificity to co-occurring microbial community structure (Figure 3) and metabolomic structure (Figure 4), and specificity to sample type ontology (Figure 5). Our validation analyses show that our statistic, *Spec*, performs intuitively and is sensitive to specificity in both empirical and simulated data (Figures S1-S7). Our companion package, *specificity.shiny*, can be used to explore results and collaborate on specificity analyses, which was done by the authors on the four example analyses we presented here. Both *specificity* and *specificity.shiny* are available from the authors' GitHub repository, along with installation instructions and a tutorial vignette.

SOFTWARE AND DATA AVAILABILITY

R packages *specificity* and *specificity.shiny* can be downloaded from GitHub: https://github.com/darcyj/specificity, https://github.com/darcyj/specificity.shiny. Code and data to replicate the analyses shown here can be found on GitHub as well: https://github.com/darcyj/specificity_analyses.

REFERENCES

Allen, B., Kon, M., and Bar-Yam, Y. (2009) A new phylogenetic diversity measure generalizing the shannon index and its application to phyllostomid bats. *American Naturalist*.

Ara, I., Matsumoto, A., Abdul Bakir, M., Kudo, T., Omura, S., and Takahashi, Y. (2008) Actinomadura maheshkhaliensis sp. nov., a novel actinomycete isolated from mangrove rhizosphere soil of Maheshkhali, Bangladesh. *J Gen Appl Microbiol* **54**: 335–342.

Baluja, S. and Caruana, R. (1995) Removing the genetics from the standard genetic algorithm. In *Machine Learning Proceedings 1995*. Elsevier, pp. 38–46.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* **57**: 289–300.

Botta-Dukát, Z. (2018) Cautionary note on calculating standardized effect size (SES) in randomization test. *Community Ecology*.

Botta-Dukát, Z. (2005) Rao's quadratic entropy as a measure of functional diversity based on multiple traits. *Journal of Vegetation Science*.

Brown, J.R.-M., Flemer, B., Joyce, S.A., Zulquernain, A., Sheehan, D., Shanahan, F., and O'Toole, P.W. (2018) Changes in microbiota composition, bile and fatty acid metabolism, in successful faecal microbiota transplantation for Clostridioides difficile infection. *BMC Gastroenterology* **18**: 131.

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., and Holmes, S.P. (2016) DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods* **13**: 581–583.

Cohan, F.M. (2002) What are bacterial species? *Annual Reviews in Microbiology* **56**: 457–487.

Combes, C. (2001) Parasitism : the ecology and evolution of intimate interactions ; translated by Isaure de Buron and Vincent A. Connors ; with a new foreword by Daniel Simberloff.

Connors, J., Dunn, K.A., Allott, J., Bandsma, R., Rashid, M., Otley, A.R., et al. (2020) The relationship between fecal bile acids and microbiome community structure in pediatric Crohn's disease. *ISME J* **14**: 702–713.

Costello, E.K., Stagaman, K., Dethlefsen, L., Bohannan, B.J., and Relman, D.A. (2012) The application of ecological theory toward an understanding of the human microbiome. *Science* **336**: 1255–1262.

Costello, M.J. (2016) Parasite Rates of Discovery, Global Species Richness and Host Specificity. In *Integrative and Comparative Biology*.

Darcy, J.L., King, A.J., Gendron, E.M.S., and Schmidt, S.K. (2017) Spatial autocorrelation of microbial communities atop a debris-covered glacier is evidence of a supraglacial chronosequence. *FEMS microbiology ecology*.

Darcy, J.L., Swift, S.O.I., Cobian, G.M., Zahn, G.L., Perry, B.A., and Amend, A.S. (2020) Fungal communities living within leaves of native Hawaiian dicots are structured by landscape-scale variables as well as by host plants. *Molecular Ecology*.

Dean, R., Van Kan, J.A., Pretorius, Z.A., Hammond-Kosack, K.E., Di Pietro, A., Spanu, P.D., et al. (2012) The Top 10 fungal pathogens in molecular plant pathology. *Molecular plant pathology* **13**: 414–430.

Dunn, O.J. (1964) Multiple comparisons using rank sums. *Technometrics* **6**: 241–252.

Faith, D.P. (1992) Conservation evaluation and phylogenetic diversity. *Biological Conservation* **61**: 1–10.

Franzosa, E.A., Sirota-Madi, A., Avila-Pacheco, J., Fornelos, N., Haiser, H.J., Reinker, S., et al. (2019) Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nature microbiology* **4**: 293–305.

Gan, P., Ikeda, K., Irieda, H., Narusaka, M., O'Connell, R.J., Narusaka, Y., et al. (2013) Comparative genomic and transcriptomic analyses reveal the hemibiotrophic stage shift of Colletotrichum fungi. *New Phytologist* **197**: 1236–1249.

Gan, P., Narusaka, M., Kumakura, N., Tsushima, A., Takano, Y., Narusaka, Y., and Shirasu, K. (2016) Genus-Wide Comparative Genome Analyses of Colletotrichum Species Reveal Specific Gene Family Losses and Gains during Adaptation to Specific Infection Lifestyles. *Genome Biology and Evolution* **8**: 1467–1481.

Ghavami, S.B., Rostami, E., Sephay, A.A., Shahrokh, S., Balaii, H., Aghdaei, H.A., and Zali, M.R. (2018) Alterations of the human gut Methanobrevibacter smithii as a biomarker for inflammatory bowel diseases. *Microbial Pathogenesis* **117**: 285–289.

Giambelluca, T.W., Chen, Q., Frazier, A.G., Price, J.P., Chen, Y.L., Chu, P.S., et al. (2013) Online rainfall atlas of Hawai'i. *Bulletin of the American Meteorological Society* **94**: 313–316.

Goldberg, D.E. and Lingle, R. (1985) Alleles, loci, and the traveling salesman problem. In *Proceedings of an international conference on genetic algorithms and their applications*. Carnegie-Mellon University Pittsburgh, PA, pp. 154–159.

Gonzalez, A., Navas-Molina, J.A., Kosciolek, T., McDonald, D., Vázquez-Baeza, Y., Ackermann, G., et al. (2018) Qiita: rapid, web-enabled microbiome meta-analysis. *Nature methods* **15**: 796–798.

Hester, J., Csardi, G., Wickham, H., Chang, W., Morgan, M., and Tenenbaum, D. (2020) remotes: R Package Installation from Remote Repositories, Including "GitHub."

Izri, A., Aljundi, M., Billard-Pomares, T., Fofana, Y., Marteau, A., Ferreira, T.G., et al. (2020) Molecular identification of Actinomadura madurae isolated from a patient originally from Algeria; observations from a case report. *BMC Infectious Diseases* **20**: 829.

Kelly, C.J., Zheng, L., Campbell, E.L., Saeedi, B., Scholz, C.C., Bayless, A.J., et al. (2015) Crosstalk between Microbiota-Derived Short-Chain Fatty Acids and Intestinal Epithelial HIF Augments Tissue Barrier Function. *Cell Host Microbe* **17**: 662–671.

Kitahara, M., Sakamoto, M., Ike, M., Sakata, S., and Benno, Y. 2005 Bacteroides plebeius sp. nov. and Bacteroides coprocola sp. nov., isolated from human faeces. *International Journal of Systematic and Evolutionary Microbiology* **55**: 2143–2147.

Kokkinidis, D.G., Bosdelekidou, E.E., Iliopoulou, S.M., Tassos, A.G., Texakalidis, P.T., Economopoulos, K.P., and Kousoulis, A.A. (2017) Emerging treatments for ulcerative colitis: a systematic review. *Scandinavian journal of gastroenterology* **52**: 923–931.

Lahoum, A., Bouras, N., Mathieu, F., Schumann, P., Spröer, C., Klenk, H.-P., and Sabaou, N. (2016) Actinomadura algeriensis sp. nov., an actinobacterium isolated from Saharan soil. *Antonie Van Leeuwenhoek* **109**: 159–165.

Miller, T.L., Wolin, M.J., de Macario, E.C., and Macario, A.J. (1982) Isolation of Methanobrevibacter smithii from human feces. *Applied and Environmental Microbiology* **43**: 227–232.

Nemergut, D.R., Schmidt, S.K., Fukami, T., O'Neill, S.P., Bilinski, T.M., Stanish, L.F., et al. (2013) Patterns and processes of microbial community assembly. *Microbiology and Molecular Biology Reviews* **77**: 342–356.

Oxford English Dictionary (2017) Oxford English Dictionary Online. *Oxford English Dictionary*.

Porazinska, D.L., Fountain, A.G., Nylen, T.H., Tranter, M., Virginia, R.A., and Wall, D.H. (2004) The Biodiversity and Biogeochemistry of Cryoconite Holes from McMurdo Dry Valley Glaciers, Antarctica. *Arctic, Antarctic, and Alpine Research* **36**: 84–91.

Poulin, R., Krasnov, B.R., and Mouillot, D. (2011) Host specificity in phylogenetic and geographic space. *Trends in Parasitology*.

Promnuan, Y., Kudo, T., Ohkuma, M., and Chantawannakul, P. (2011) Actinomadura apis sp. nov., isolated from a honey bee (Apis mellifera) hive, and the reclassification of Actinomadura cremea subsp. rifamycini Gauze et al. 1987 as Actinomadura rifamycini (Gauze et al. 1987) sp. nov., comb. nov. *Int J Syst Evol Microbiol* **61**: 2271–2277.

Rachniyom, H., Matsumoto, A., Indananda, C., Duangmal, K., Takahashi, Y., and Thamchaipenet, A. (2015) Actinomadura syzygii sp. nov., an endophytic actinomycete isolated from the roots of a jambolan plum tree (Syzygium cumini L. Skeels). *Int J Syst Evol Microbiol* **65**: 1946–1949.

Rao, C.R. (1982) Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology*.

Rao, C.R. (2010) Quadratic entropy and analysis of diversity. *Sankhya A*.

Ridlon, J.M., Kang, D.J., Hylemon, P.B., and Bajaj, J.S. (2014) Bile Acids and the Gut Microbiome. *Curr Opin Gastroenterol* **30**: 332–338.

Shaw, L., Ribeiro, A.L.R., Levine, A.P., Pontikos, N., Balloux, F., Segal, A.W., et al. The Human Salivary Microbiome Is Shaped by Shared Environment Rather than Genetics: Evidence from a Large Family of Closely Related Individuals. *mBio* **8**: e01237-17.

Shefferson, R.P., Bunch, W., Cowden, C.C., Lee, Y.I., Kartzinel, T.R., Yukawa, T., et al. (2019) Does evolutionary history determine specificity in broad ecological interactions? *Journal of Ecology*.

Sommers, P., Darcy, J.L., Gendron, E.M.S., Stanish, L.F., Bagshaw, E.A., Porazinska, D.L., and Schmidt, S.K. (2018) Diversity patterns of microbial eukaryotes mirror those of bacteria in Antarctic cryoconite holes. *FEMS Microbiology Ecology* **94**:.

Sommers, P., Darcy, J.L., Porazinska, D.L., Gendron, E.M.S., Fountain, A.G., Zamora, F., et al. (2019) Comparison of Microbial Communities in the Sediments and Water Columns of Frozen Cryoconite Holes in the McMurdo Dry Valleys, Antarctica. *Frontiers in Microbiology* **10**: 65.

Sommers, P., Porazinska, D.L., Darcy, J.L., Zamora, F., Fountain, A.G., and Schmidt, S.K. (2019) Experimental cryoconite holes as mesocosms for studying community ecology. *Polar Biology* **42**: 1973–1984.

Thompson, L.R., Sanders, J.G., McDonald, D., Amir, A., Ladau, J., Locey, K.J., et al. (2017) A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**: 457–463.

Tipton, L., Zahn, G.L., Darcy, J.L., Amend, A.S., and Hynson, N.A. (2021) Hawaiian Fungal Amplicon Sequence Variants Reveal Otherwise Hidden Biogeography. *Microbial Ecology* 1–10.

Tiratterra, E., Franco, P., Porru, E., Katsanos, K.H., Christodoulou, D.K., and Roda, G. (2018) Role of bile acids in inflammatory bowel disease. *Ann Gastroenterol* **31**: 266–272.

Truong, D.T., Franzosa, E.A., Tickle, T.L., Scholz, M., Weingart, G., Pasolli, E., et al. (2015) MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature methods* **12**: 902–903.

Wishart, D.S., Feunang, Y.D., Marcu, A., Guo, A.C., Liang, K., Vázquez-Fresno, R., et al. (2018) HMDB 4.0: the human metabolome database for 2018. *Nucleic acids research* **46**: D608–D617.

Zhang, X. and Yi, N. (2020) NBZIMM: negative binomial and zero-inflated mixed models, with application to microbiome/metagenomics data analysis. *BMC Bioinformatics*.

Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. (2000) A greedy algorithm for aligning DNA sequences. *J Comput Biol* **7**: 203–214.