MicroNiche

An R package for measuring niche breadth and overlap indices of microbial taxa from amplicon sequencing data

Table of Contents

Introduction ***1***

Niche breadth ***2***

Levins' B_N index: "generalists *versus* specialists" 2

Relationships with environment: Hurlbert's B_N and Feinsinger's Proportional Similarity (PS) indices 2

Niche overlap ***3***

Levins' overlap: comparing similarity of distributions between two taxa 3

Proportional overlap: taxa distributions in relation to environment 4

Limit of Quantification ***4***

Null model testing ***7***

Running MicroNiche examples ***8***

Installing MicroNiche 8

The *in silico* training dataset 9

Calculating niche breadths 11

Calculating niche overlaps 14

References ***18***

Introduction

Niche theory is a fundamental concept in biology. It is the notion that the fitness of a population is intrinsically linked to its environment [1]. A quantitative approach to niche measurements allows for hypothesis testing of evolutionary processes, interspecies interactions (*i.e.* positive mutualism or antagonistic competition) and pressure that environmental boundaries exert on physiology. With the advent of Next-Generation sequencing technologies and bioinformatic pipelines for taxon assignment, a challenge for the modern microbial ecologist often lies in applied statistics to appropriately test hypotheses. This vignette describes a step-by-step guide for the MicroNiche R package, which is a collection of R functions to apply models of niche breadth and overlap to microbial amplicon sequencing data. Niche breadth (B_N) can be considered as a measurement of the distribution of a taxon either across environments or in relation to an environmental parameter, such as pH, salinity or precipitation [2]. Niche overlap can be considered as the (dis)similarity of two taxon's distributions across environments or in relation to an environmental parameter [3].

Niche breadth

Levins' B_N index: "generalists *versus* specialists"

Levins' B_N measures a taxon's distribution across environments, normalized by the number of environments being compared. Where a taxon is equally abundant across all environments, B_N is equal to 1. These taxa can be thought of as "generalists" that are not restricted to specific environments or conditions [4]. The closer to 0 a taxon's B_N becomes, the more unique this taxon is to a specific environment. These taxa can be thought of as "specialists", which are present only under specific conditions [4]. (Although note that Levins' B_N cannot be lower than 1/*R*, where *R* is the sum of different environments being compared). Levins' B_N is most useful when the specific environmental and/or ecological factors that affect a taxon's abundance are unknown, or to identify generalists. Levins' B_N is calculated as:

$$\text{Levins' } B_N = \frac{1}{R}\sum_{i=1} {p_i}^2 \qquad \text{Equation 1.}$$

whereby *p* is the proportional abundance of a taxon within the i^{th} environment, and as mentioned above, *R* is the sum of different environments being compared.

Relationships with environment: Hurlbert's B_N and Feinsinger's Proportional Similarity (PS) indices

The key difference between Levins' B_N and the following two indices is that they incorporate an environmental parameter. An example of an environmental parameter would be pH measured along a forest soil gradient, from acidic ($pH < 4$) to near-neutral ($pH \sim 7$) with the assumption that certain taxa will be acidophilic and others more neutrophilic. These models were initially designed by macro-ecologists to compare the distribution of a taxon in relation to the distribution of a resource it may depend on, *e.g.* a finch species in relation to seeds. However, as the 'resource' is transformed to a unit-less proportional abundance, mathematically it can be substituted with any environmental parameter. Hurlbert's B_N is calculated as:

$$\text{Hurlbert's } B_N = \frac{1}{\sum_{i=1} \frac{{p_i}^2}{r_i}} \qquad \text{Equation 2.}$$

whereby p is the proportional abundance of a taxon within the i^{th} environment, and r is the proportional abundance of an environmental parameter within the i^{th} environment. Feinsinger's PS is calculated as:

$$\text{PS} = 1 - 0.5\sum_{i=1} |p_i - r_i| \qquad \text{Equation 3.}$$

whereby the model parameters are the same as Hurlbert's B_N described above. Both of these models yield similar results, although arguably Feinsinger's PS is slightly more stringent. Hurlbert's B_N and PS range between 0 and 1, with 0 indicating an inverse relationship between the taxon and the environmental parameter (*i.e.* in the above example of forest soil pH gradient, an inverse relationship would be higher abundance of an acidophilic taxon as pH decreases relative to its abundance at neutral pH), 0.5 indicating no relationship, and 1 indicating a positive correlation between taxon abundance and the environmental parameter (*i.e.* a neutrophile's abundance being higher as pH increases relative to its abundance in more acidic soils). Comparisons of all three models are provided in Feinsinger *et al*., 1981.

Niche overlap

Levins' overlap: comparing similarity of distributions between two taxa

As the name implies, niche overlap indices measure the pairwise overlap of two taxa. Similar to Levins' B_N, Levins' Overlap (LO) considers the distribution of taxa across environments. An important thing to note with LO is that, for taxon *i* and *j*, their respective LO indices may differ. That is, $LO_{i,j}$ may not be equal to $LO_{j,i}$. For example, where the distribution of *i* completely overlaps with that of *j*, the $LO_{i,j}$ will be 1. In other words, where *i* is present, *j* is present. This will occur if *i* is a generalist that is distributed fairly evenly across environments. Where *j* overlaps poorly with *i*, particularly if *j* is a specialist present in only specific environments, the $LO_{j,i}$ will be close to 0. In other words, the absence of *j* does not indicate an absence of *i*. This may seem like a tricky concept (!) and Figure Nine, in the examples section of this document, endeavours to provide a visual example of this. The LO is calculated as:

$$\mathrm{LO}_{i,j} = \frac{\sum_{i,j=1}(p_{ir})(p_{jr})}{\sum_{i=1}(p_{ir}^{2})} \qquad \text{Equation 4.}$$

whereby p_i is the proportional abundance of taxon *i* in the r^{th} environment, and p_j is the proportional abundance of taxon *j* in the r^{th} environment. Detailed examples of LO and other niche overlap indices are presented in Ludwig and Reynolds, 1988.

Proportional overlap: taxa distributions in relation to environment

The proportional overlap (PO) index was developed here in order to compare the (dis)similarity of two taxa in relation to an environmental parameter. It simply calculates the Jaccard similarity coefficient of a pair of PS values as:

$$\mathrm{PO}_{i,j} = 1 - \left(\frac{X \cap Y}{X \cup Y}\right) \qquad \text{Equation 5.}$$

whereby *X* is the PS of taxon *i* and *Y* is the PS of taxon *j*. $\mathrm{PO}_{i,j}$ will approach 0 for taxa pairs that are inversely associated with each other. Again using the forest soil pH gradient example above, if *i* is a specialist most abundant under acidic conditions and if *j* is a specialist most abundant under near-neutral conditions, $\mathrm{PO}_{i,j}$ will approach 0. If *i* and *j* are both most abundant under acidic conditions, $\mathrm{PO}_{i,j}$ will approach 1. Please note that with this overlap index, $\mathrm{PO}_{i,j}$ and $\mathrm{PO}_{j,i}$ are identical.

Limit of Quantification

In analytical chemistry, the limit of quantification (LOQ) is considered as the lowest point by which an analyte can be quantified with confidence [5, 6]. Where analyses of microbial taxa from amplicon sequencing datasets are concerned, workflows typically do not consider an LOQ. Some workflows do call for the removal of ‘singleton’ taxa present in particularly low abundance [7] however this is arguably a rather arbitrary cut-off in that it does not consider statistical parameters of the dataset being analysed. The benefit of defining a LOQ is that it depends on the variance of measurements within your data. For example, if one were to measure pH along their forest soil gradient and found the variance to be particularly high (say, as a result of a faulty pH meter) an LOQ calculated for these measurements would result in a higher cut-off than an LOQ calculated based on measurements with low variance, good reproducibility and higher confidence (taken with a working pH meter). Figure One is a conceptual image of such an example reproduced from Eurachem, an international body concerned with defining best practices in analytical chemistry [5, 6]. Suppose that we suspect an inherent amount of variability in measurements of a sample where the measurand (*i.e.* variable) is absent, which is the null hypothesis (Figure One, blue curve). Similarly, an inherent amount of variability will be present in measurements of a sample where the measurand is present, which is the alternative hypothesis (Figure One, red curve). Where the two curves overlap is where a measurement is significantly greater than the mean of the null hypothesis (α) yet significantly lower than the mean of the alternative hypothesis (β). This is an area where it is uncertain if the measurement is true. The mid-point of this overlap is considered to be the detection decision boundary. In Figure One, the limit of detection (LOD) is calculated as being the mean of a normal distribution that forms a decision boundary at 1.65 *x* standard deviation from the mean of the null hypothesis.

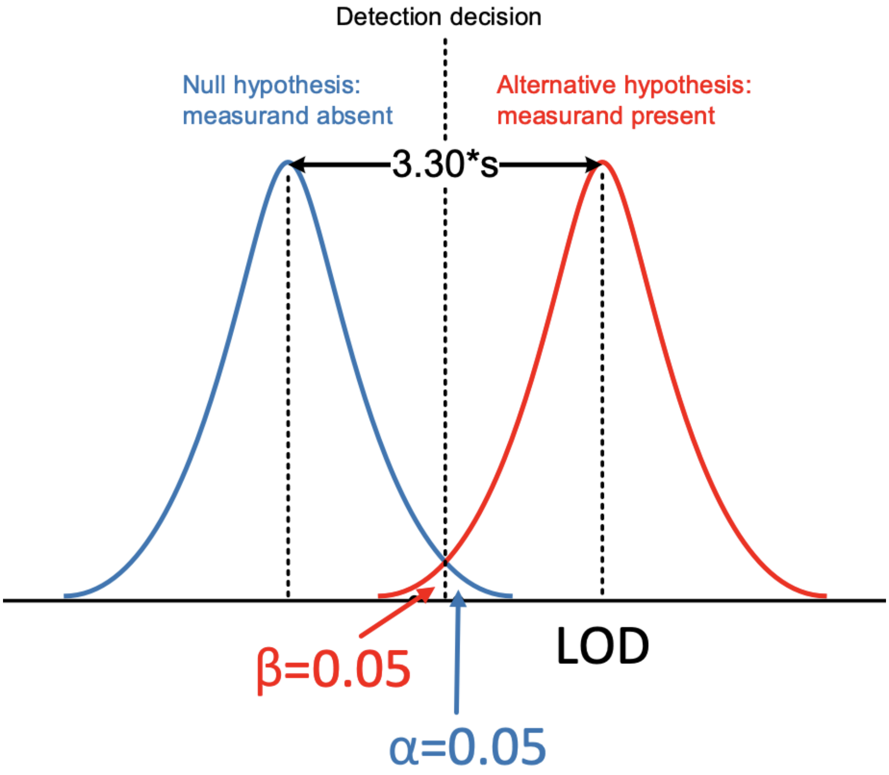

Detection decision
Null hypothesis: measurand absent
Alternative hypothesis: measurand present
3.30*s
LOD
β=0.05
α=0.05

Figure One: Conceptual diagram of defining a detection limit based on measurement variance.

In ecological analyses of microbial taxa from amplicon sequencing data, the distribution of discrete counts of taxa follow that of a negative binomial distribution as opposed to a normal distribution. Such a distribution is characterised by few highly abundant taxa and a long tail of taxa with decreasing abundance. When considered as log abundance *x* taxon rank, the distribution of taxa forms a lognormal distribution. This can be expressed as:

$$S(R) = S_0 e^{-a^2 R^2} \qquad \text{Equation 6.}$$

where the abundance of taxon S at rank R, S(R), decreases from the modal S, S_0, at an exponential rate dependent on *a* and R. The coefficient *a* is calculated as:

$$a = \sqrt{\frac{\ln(S_0)}{S_m} / R^2} \qquad \text{Equation 7.}$$

where S_0 and R is as above, and S_m is the lowest S. Further details regarding this lognormal model and other taxon rank models can be found in Ludwig and Reynolds, 1988. To calculate the LOQ of microbial taxa within a dataset, MicroNiche fits the above lognormal model to the user's data. The standard deviation of the model is calculated. Finally, using the conceptual approach demonstrated in Figure One, the LOQ is determined as the overlap between the null hypothesis (*i.e.* a taxon's mean abundance is 0) and any taxa within 1.65 *x* standard deviation of Equation 6. Figure Two is an example of applying the LOQ to an *in silico* training set.

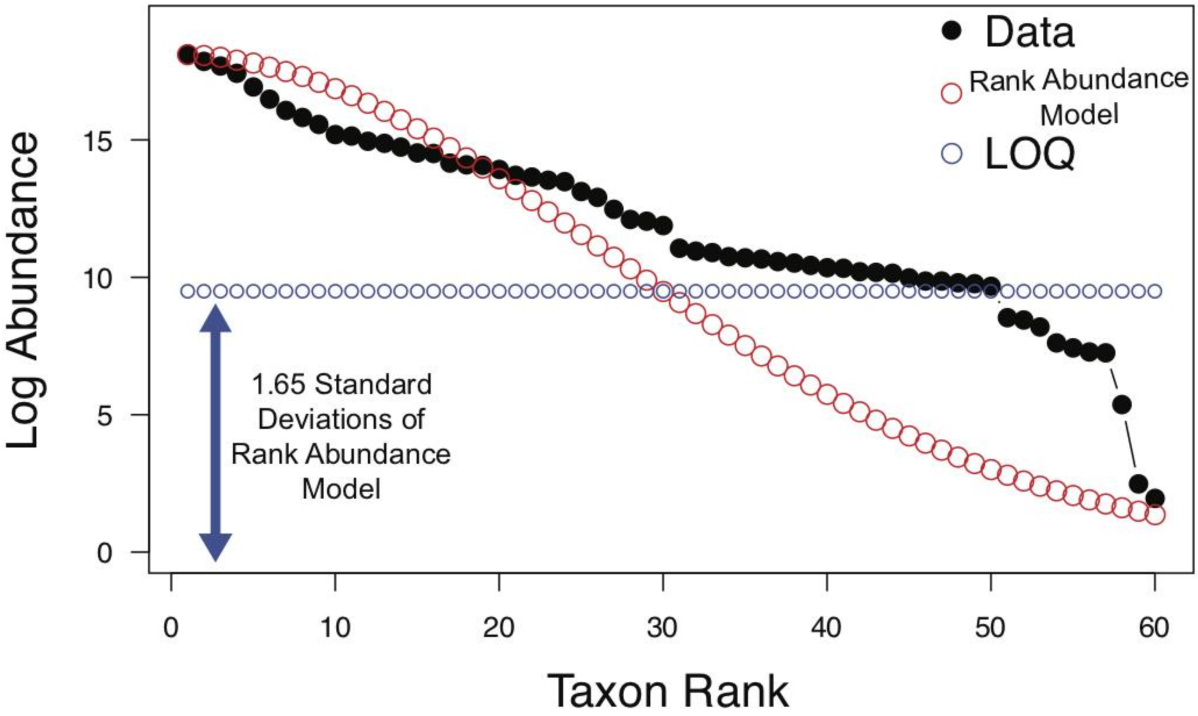


Figure Two: Calculating the LOQ based on a lognormal model typical of microbial taxa.

Here, the raw data is plotted in black, the model is plotted as red circles and the LOQ is plotted as blue circles. The point of overlap between the raw data and the LOQ is considered as the decision boundary and corresponds to the boundary given in Figure One. In MicroNiche, this boundary is calculated from user data. If a taxon falls below this decision boundary, it is considered below the LOQ and is flagged as such within niche breadth results produced by MicroNiche. The implementation of an LOQ was necessary to discern taxa being incorrectly identified as specialists (*i.e.* false positives). Therefore, any taxon flagged as being below the LOQ should be considered with caution. The default coefficient of 1.65 *x* standard deviations can be modified at the user's discretion, as described below in the 'Running MicroNiche Examples section.

Null model testing

Null models are a powerful tool to answer questions in ecology [8]. Essentially, a null model represents the null hypothesis and acts as a reference distribution of potential values for comparing a value of interest. Significance testing (that is, to derive a *p* value) can then be performed to test whether the occurrence of a value of interest is unlikely in relation to the null model. For niche breadth analyses, MicroNiche implements null model testing by generating a random normal distribution of 999 possible niche breadths under the supplied input parameters. This allows for a *p* value to be assigned to each taxon's niche breadth to determine whether it is significantly greater or lower than the mean of the null model. In addition, MicroNiche also provides Benjamin-Hochberg adjusted *p* values to account for false discovery rate inherent in large datasets, such as microbial amplicon sequencing. Figure Three is an example of a null model distribution provided by MicroNiche.

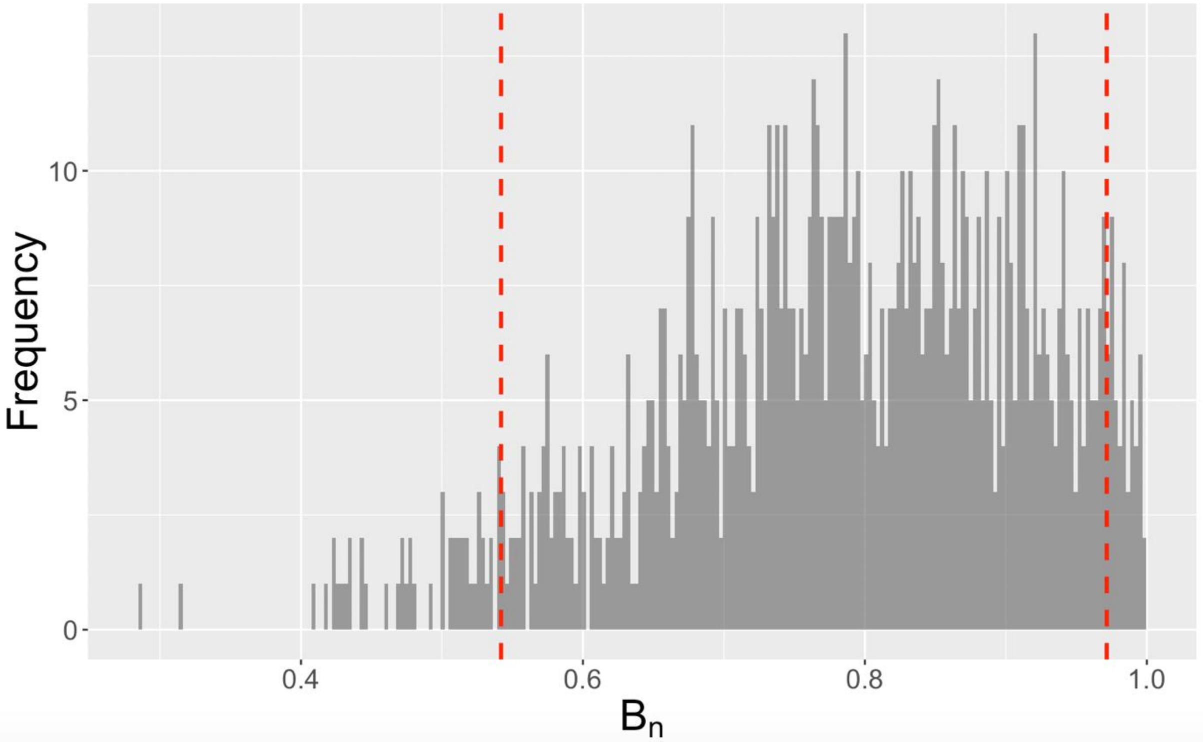


Figure Three: Example null model distribution comparing Levins' B_N across four environments on the *x* axis, and frequency of randomly generated B_N on the *y* axis. The red dotted lines indicate the 0.05 and 0.95 quantiles. In this example, a B_N of 0.42 would indicate a specialist, while a B_N of 0.98 would indicate a generalist.

Running MicroNiche examples

Installing MicroNiche

As part of CRAN, MicroNiche can be installed within an R session simply by:

```
>install.packages('MicroNiche')
```

This package was originally built in R version 3.5.2 (2018-12-20) and is supported by CRAN under R version 3.6.2 (2019-12-12), and therefore should be viable under several versions of R from 3.5 and above. Please note that MicroNiche is dependent on the 'ggplots' and 'reshape2' R packages. It is also suggested to install the 'gplots' R package to visualise niche overlap results.

The *in silico* training dataset

Figure Four a) is a visual representation of the six taxon distributions across four environments tested by the *in silico* training dataset. The six distributions consist of 10 individual taxa per distribution. There are 10 independent samples per environment. The first distribution represent true generalists that have roughly equal counts generated from random normal distributions across the four environments. The second distribution represent specialists that linearly decrease from Environment One to Four. The third distribution are specialists that represent an exponentially decreasing abundance, where they are high in Environment One, approximately half as abundant in Environment Two, and mostly absent from the third and fourth environments. These abundances also decrease along an environmental gradient that is low in Environment One and increases across environments, represented at the bottom of Figure Four a). The fourth distribution are true specialists that are highly abundant only in Environment One. The fifth distribution are specialists that are equally abundant in two of the four environments yet have no relationship with the environmental gradient. Finally, the sixth distribution 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 are the long tail of sparse taxa frequently observed in microbial amplicon sequencing data and act as a control in MicroNiche to test the occurrence of Type I false positive errors. Figure Four b) is an example of the generalist and linearly decreasing distributions in Environment Four, with Taxon Abundance as the *y* axis and Sample Frequency as the *x* axis.

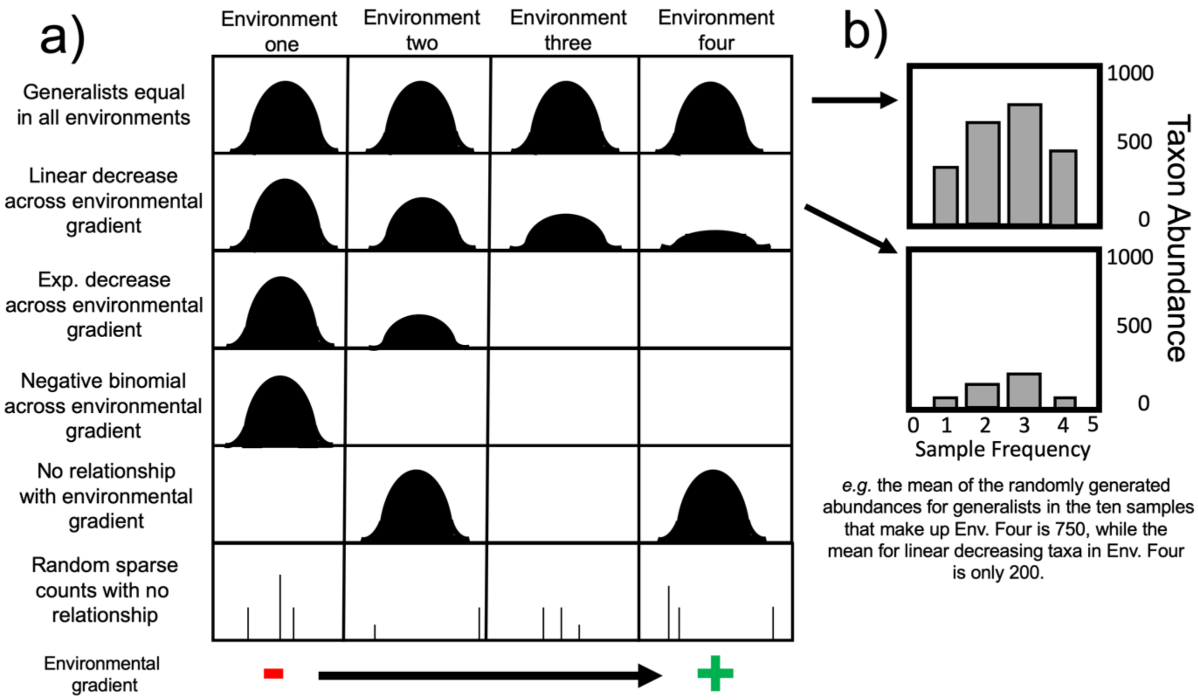
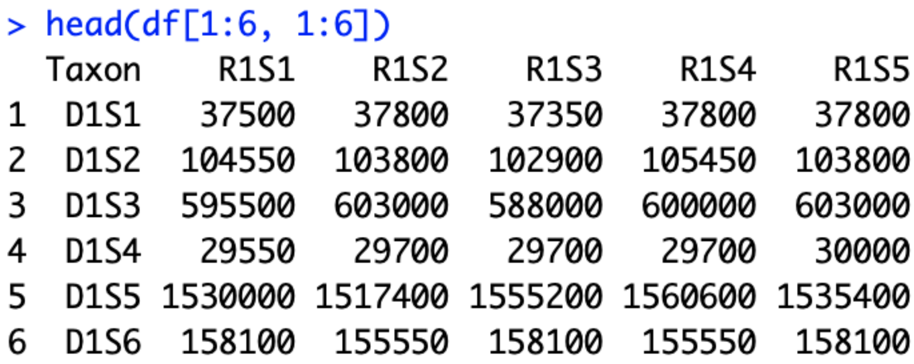

a)
Environment one
Environment two
Environment three
Environment four
Generalists equal in all environments
Linear decrease across environmental gradient
Exp. decrease across environmental gradient
Negative binomial across environmental gradient
No relationship with environmental gradient
Random sparse counts with no relationship
Environmental gradient
-
+
b)
1000
500
0
1000
500
0
Taxon Abundance
0 1 2 3 4 5
Sample Frequency
e.g. the mean of the randomly generated abundances for generalists in the ten samples that make up Env. Four is 750, while the mean for linear decreasing taxa in Env. Four is only 200.

Figure Four: The *in silico* training set devised for testing niche breadth and overlap models. a) the six distributions generated across four environments; and b) a representation of a taxon within Environment Four of the generalist and linearly decreasing distributions.

To begin, load the package and attach the *in silico* training set data frame in R via:

```
>library(MicroNiche)
>data(df)
```

This example data is in the required format for MicroNiche to apply its functions. Specifically, the first column includes a named identifier for each taxon, and columns 2 – *n* are discrete counts of each taxon's abundance per sample. Here, the first taxon is identified as D_1S_1, and the individual samples are R_1S_1, R_1S_2, R_1S_3, etc. D identifies the taxon's distribution (as per Figure Four a) and R_1 through R_4 identifies Environment One through Four. Figure Five shows a subset of the training data.



```
> head(df[1:6, 1:6])
  Taxon    R1S1    R1S2    R1S3    R1S4    R1S5
1  D1S1   37500   37800   37350   37800   37800
2  D1S2  104550  103800  102900  105450  103800
3  D1S3  595500  603000  588000  600000  603000
4  D1S4   29550   29700   29700   29700   30000
5  D1S5 1530000 1517400 1555200 1560600 1535400
6  D1S6  158100  155550  158100  155550  158100
```

Figure Five: A subset of the training data with a taxon identifier in column 1 and discrete counts in columns 2 – *n*.

Calculating niche breadths

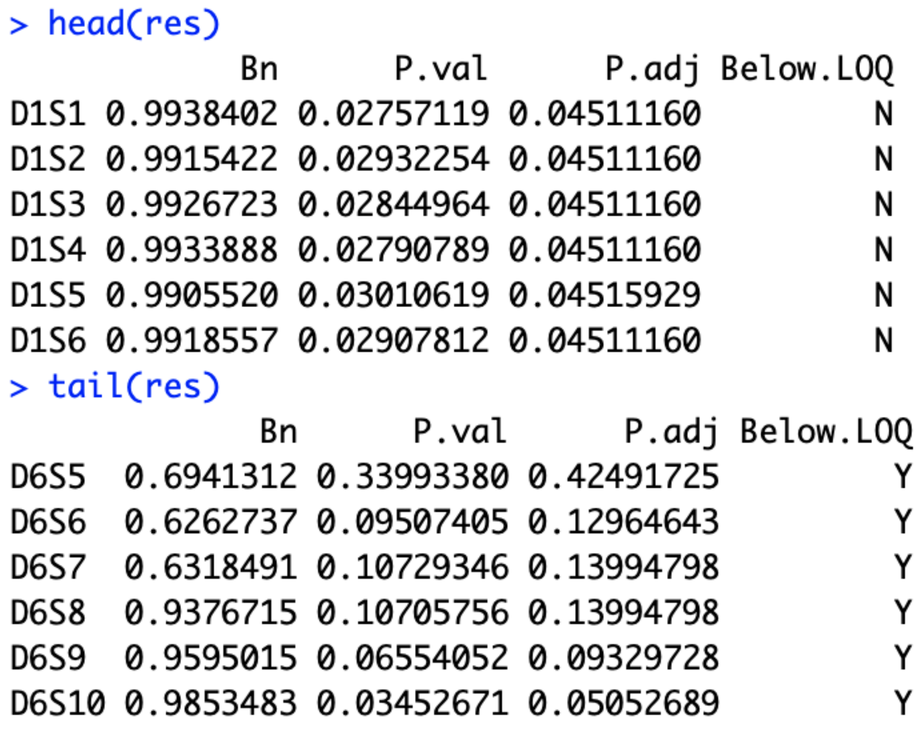
The first niche breadth index to test is Levins' B_N. As described above, this niche breadth does not incorporate environmental data. Instead, it only requires a matrix of taxa (rows) *x* samples (columns), the total number of environments to test, and a vector that informs the function which column matches with which environment. To calculate Levin's B_N on the *in silico* training data, perform the following in R:

```
>sampleInfo <- c(rep("R1",10), rep("R2",10), rep("R3",10), rep("R4",10))
>res <- levins.Bn(df, 4, sampleInfo)
```

Here, 'sampleInfo' is a vector of character strings of R_1 through to R_4, repeated 10 times each as there are 10 replicate samples per environment (R_1S_1 to R_1S_{10} and so forth for each R_N). We generate a data frame object 'res' that are the results of applying the function 'levins.Bn' to each taxon, a plot of the null distribution of 999 randomly generated B_N, and a plot of the taxon rank log abundances with the LOQ. View the results by:

```
>View(res)
```

This will show the individual results of applying Levin's B_N to the 60 taxa in the *in silico* training set. The first column is the B_N, the second column is a *p* value testing whether the taxon's B_N differs from the mean of the null distribution, the third column is a Benjamin-Hochberg adjusted *p* value and the fourth column is a warning to note which taxa fell below the LOQ and should be considered as false positives in regard to their B_N and *p* values. N indicates the taxon is not below the LOQ, whereas Y indicates the taxon is below the LOQ (Figure Two, blue dotted line). The generalist taxa (D_1S_1 to D_1S_{10}) are almost equally abundant in the four environments, and thus their B_N is almost 1 ($p < 0.05$). The D_2 group that decreases linearly across the four environments is neither proportionally equal (*i.e.* a generalist) nor aggregated strongly to specific environments (*i.e.* a specialist) and has a B_N of 0.83 toward the mean of the null distribution ($p = 0.6$). Due to the lack of environmental data, Levins' B_N considers the D_3 and D_5 groups to both be specialists with B_N of roughly 0.45 – 0.53 ($p < 0.05$). The D_4 group abundant in only one environment has the lowest B_N (0.25) and is very likely to be a specialist ($p < 0.001$). The final group, D_6, have a B_N that does not differ from the null in most cases ($p > 0.05$) except for D_6S_{10} ($p = 0.05$). Without the LOQ, we would expect this taxon to be a generalist with a B_N of 0.985. However, the LOQ informs us that due to the sparse and relatively low presence of this taxon in the dataset, it should be considered as a false positive. Figure Six is an example of the head and tail of 'res'.



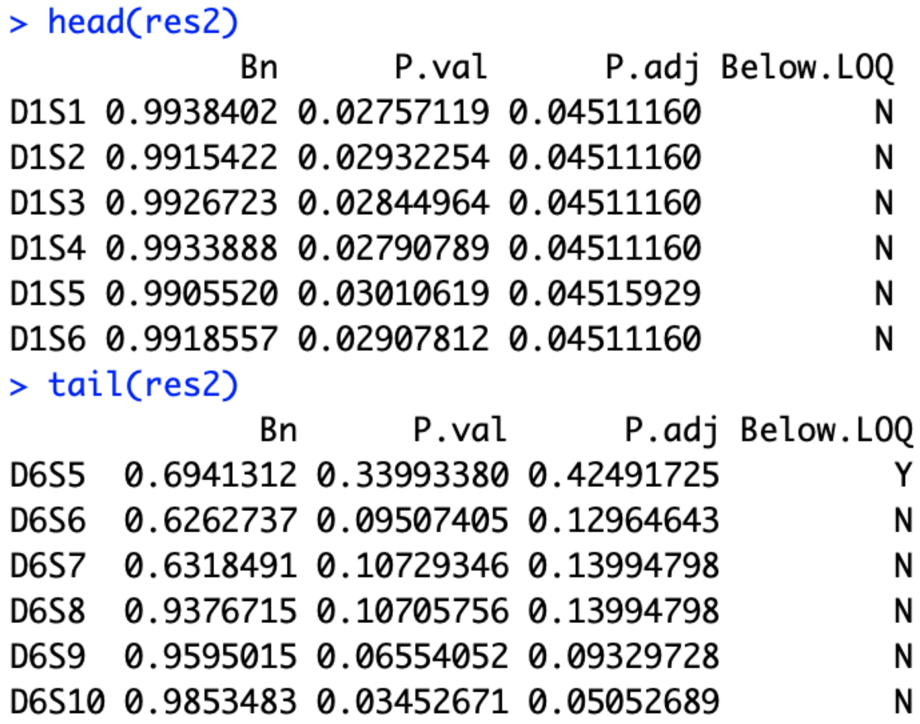
```
> head(res)
            Bn      P.val      P.adj Below.LOQ
D1S1 0.9938402 0.02757119 0.04511160         N
D1S2 0.9915422 0.02932254 0.04511160         N
D1S3 0.9926723 0.02844964 0.04511160         N
D1S4 0.9933888 0.02790789 0.04511160         N
D1S5 0.9905520 0.03010619 0.04515929         N
D1S6 0.9918557 0.02907812 0.04511160         N
> tail(res)
             Bn      P.val      P.adj Below.LOQ
D6S5  0.6941312 0.33993380 0.42491725         Y
D6S6  0.6262737 0.09507405 0.12964643         Y
D6S7  0.6318491 0.10729346 0.13994798         Y
D6S8  0.9376715 0.10705756 0.13994798         Y
D6S9  0.9595015 0.06554052 0.09329728         Y
D6S10 0.9853483 0.03452671 0.05052689         Y
```

Figure Six: Results of Levin's B_N performed on the *in silico* training set.

It is possible to modify the LOQ if the user justifiably considers it to be excluding sparse, low-count taxa of biological importance. To lower the LOQ, alter the q parameter as follows:

```
>res2 <- levins.Bn(df, 4, sampleInfo, q = 1)
```

Figure Seven shows the head and tail of the results ‘res2’ that have a lower stringency for flagging taxa as below the LOQ.



```
> head(res2)
            Bn      P.val      P.adj Below.LOQ
D1S1 0.9938402 0.02757119 0.04511160         N
D1S2 0.9915422 0.02932254 0.04511160         N
D1S3 0.9926723 0.02844964 0.04511160         N
D1S4 0.9933888 0.02790789 0.04511160         N
D1S5 0.9905520 0.03010619 0.04515929         N
D1S6 0.9918557 0.02907812 0.04511160         N
> tail(res2)
             Bn      P.val      P.adj Below.LOQ
D6S5  0.6941312 0.33993380 0.42491725         Y
D6S6  0.6262737 0.09507405 0.12964643         N
D6S7  0.6318491 0.10729346 0.13994798         N
D6S8  0.9376715 0.10705756 0.13994798         N
D6S9  0.9595015 0.06554052 0.09329728         N
D6S10 0.9853483 0.03452671 0.05052689         N
```

Figure Seven: Results of Levin’s B_N performed on the *in silico* training set with reduced stringency of the LOQ.

Under these parameters D_6S_{10} can be considered as a generalist. However, given that a generalist in group D_1, such as D_1S_1, is present as 1.5 million counts across the 40 samples, and a specialist D_5S_5 is present as 63.5 thousand counts across the 40 samples, the user must thoughtfully consider whether the relatively low count of D_6S_{10} at 5 thousand counts across 40 samples is both: a) a true reflection of its abundance *in situ* given potential primer amplification bias, sequencing error and any taxon annotation error during bioinformatics; and b) whether the potentially erroneous abundance of D_6S_{10} has biological relevance to the hypotheses proposed by the user’s study.

Hurlbert's B_N and Feinsinger's PS can be calculated if a vector of a numerical environmental measurement is supplied. Here, we consider the gradient as a pH gradient that increases from acidic in R_1 to basic in R_4. Note the order of the values must correspond to the order of samples given in 'sampleInfo'. Do so as:

```
>pH.grad <- c(2.1, 2.2, 2, 1.9, 2.1, 1.8, 1.9, 2, 2.1, 1.9, 3.5, 3.6, 3.5,
            3.4, 3.6, 3.5, 3.5, 3.4, 3.7, 3.4, 6.6, 6.5, 6.4, 6.8, 7, 6.6,
            6.8, 6.9, 7, 7.1, 8, 8.2, 7.9, 8.1, 7.8, 7.9, 8.3, 8.2, 8.1, 7.9)
```

And now to calculate Hurlbert's B_N of taxa in relation to the pH gradient:

```
>res <- hurlberts.Bn(df, 4, sampleInfo, pH.grad)
```

And Feinsinger's PS:

```
>res <- feinsingers.PS(df, 4, sampleInfo, pH.grad)
```

View the results as above. Note that Hurlbert's B_N gives quite different values to Levin's B_N – only taxa from D_3 and D_4 significantly differ from the null, with B_N approximately 0.1 – 0.17. By looking at Figure Four we see that D_3 and D_4 both increase in abundance as the pH decreases. Interestingly, the relationship between the linearly decreasing D_2 and pH increase does not elicit a strong *p* value (0.08) and as shown with D_3 and D_4, this model may be most effective when taxa distributions are non-linear. Feinsinger's PS is much the same. This emphasises the differences in the models (Levins' *versus* Hurlbert's / Feinsinger's), and that their application is dependent on the hypothesis being tested.

Calculating niche overlaps

In the following examples, it is suggested to install the package 'gplots' to utilise the 'heatmap.2' function. This is a simple way to visually compare the results of the niche overlap models. Begin by performing LO with the 'levins.overlap' function:

```
>res <- levins.overlap(df)
```

Note that to calculate LO, no sample or environmental information is necessary. This is because the LO is purely a comparison of the distribution of taxa *i* and *j* across all samples. To visualise the results, perform the following:

```
>my_palette <- colorRampPalette(c(“darkred”, "firebrick1", "orange", "gold"))
>rownames(res) <- res[,1]
>res <- res[,-1]
>heatmap.2(as.matrix(res), density.info = c("none"), cexCol = 1, cexRow = 1, dendrogram = "none", trace = c("none"), col = my_palette, Colv = F, Rowv = F, xlab = expression(LO[21]),ylab = expression(LO[12]))
```

Figure Eight is a plot of LO applied to the *in silico* training set. $LO_{1,2}$ is the *y* axis, and $LO_{2,1}$ is the *x* axis. The first thing to note is that the D_1 group on $LO_{1,2}$ overlap with all other taxa. This is because, as the generalists, the D_1 group are present in all samples while the other taxa groups are less consistent. This is apparent by considering the D_1 group on the $LO_{2,1}$ axis – despite its linear decrease, D_2 is also present in all samples as D_1 (Figure Four), and thus the second highest $LO_{2,1}$ values for D_1 are with this group. D_5 are present in roughly half the samples as D_1 and have the next strongest $LO_{2,1}$ values of roughly 0.5. The poorest overlaps are with the highly specialised group D_4 and D_5 that have almost no overlap in their distributions. Because the poor evenly distributed D_4 and D_3 share a high abundance in environment R_1, D_4 has quite high overlap scores (0.83) with this specific group. The taxa flagged as being below the LOQ (*i.e.* D_6) tend toward low, inconsistent overlap with the other groups. MicroNiche attaches an asterisk (*) to the names of taxa below the LOQ.

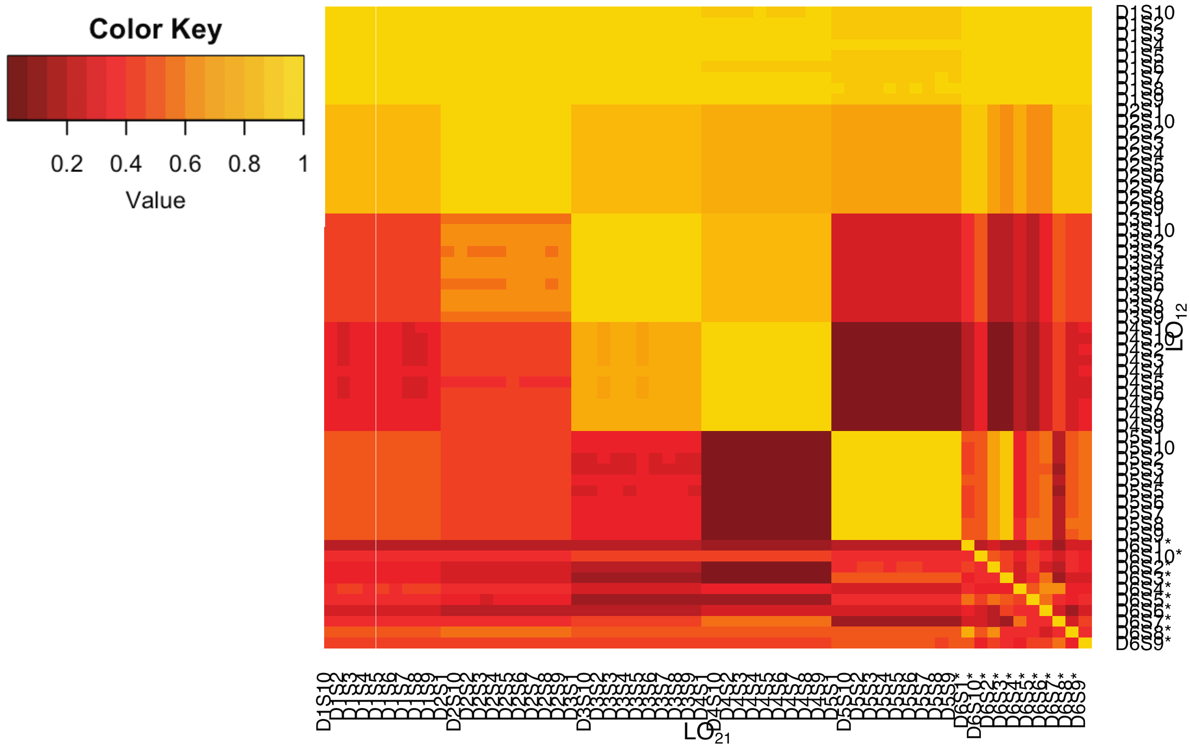
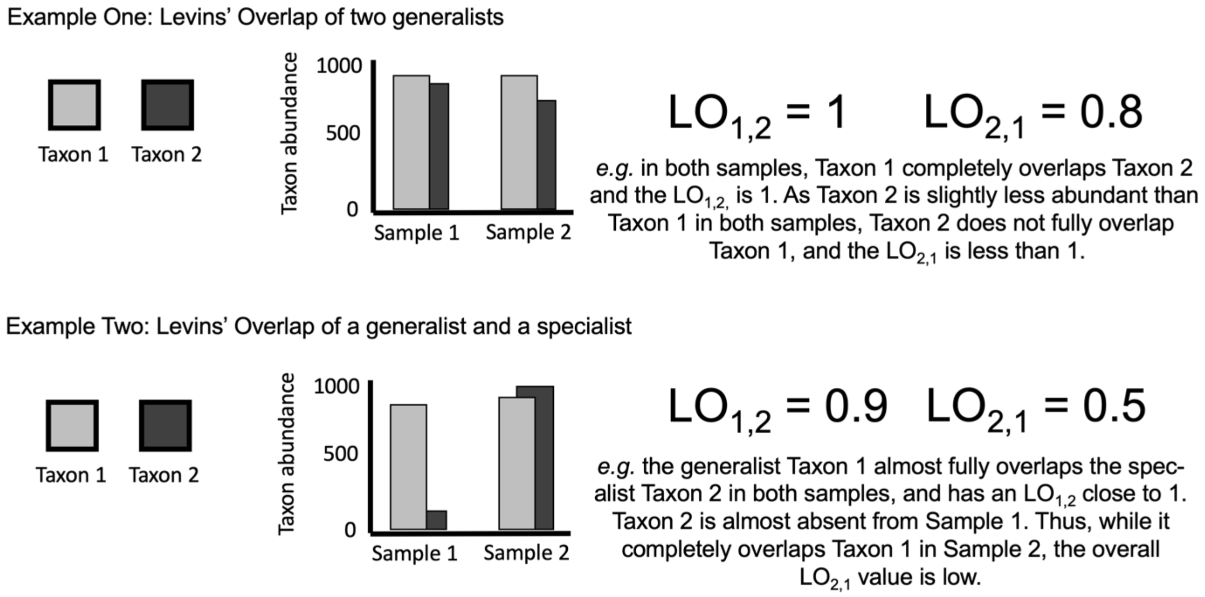


Figure Eight: Heatmap of LO performed on the *in silico* training set.

As described on Page 3, $LO_{1,2}$ and $LO_{2,1}$ are not necessarily equal! Figure Nine is a conceptual diagram that (hopefully) provides a simplified explanation as to why this is.


Example One: Levins' Overlap of two generalists

$LO_{1,2} = 1$ $LO_{2,1} = 0.8$

e.g. in both samples, Taxon 1 completely overlaps Taxon 2 and the $LO_{1,2,}$ is 1. As Taxon 2 is slightly less abundant than Taxon 1 in both samples, Taxon 2 does not fully overlap Taxon 1, and the $LO_{2,1}$ is less than 1.

Example Two: Levins' Overlap of a generalist and a specialist

$LO_{1,2} = 0.9$ $LO_{2,1} = 0.5$

e.g. the generalist Taxon 1 almost fully overlaps the specialist Taxon 2 in both samples, and has an $LO_{1,2}$ close to 1. Taxon 2 is almost absent from Sample 1. Thus, while it completely overlaps Taxon 1 in Sample 2, the overall $LO_{2,1}$ value is low.

Figure Nine: A conceptual diagram comparing two examples of taxa with differing LO values.

Finally, the PO is simply a (dis)similarity measure of a pair of taxon's PS values, ranging from 0 (completely different) to 1 (identical). PO is derived from the 'proportional.overlap' function:

```
> res <- proportional.overlap(df, sampleInfo, pH.grad)
```

As Feinsinger's PS is applied to each taxon, sample and environmental information are necessary. Figure Ten is a plot of PO applied to the *in silico* training set, visualised with the 'heatmap.2' function. Here, $PO_{1,2}$ and $PO_{2,1}$ are identical. The generalist D_1 group are most different to the environmental specialists D_3 and D_4. As neither the D_1, D_2 nor D_5 groups had any relationship with the environmental gradient, their respective PO values remain relatively high at 0.84 – 0.86. This should be kept in mind when considering the PO results, as both PS and PO values should be considered in tandem to identify taxa/environment and taxa/taxa relationships of potential interest.

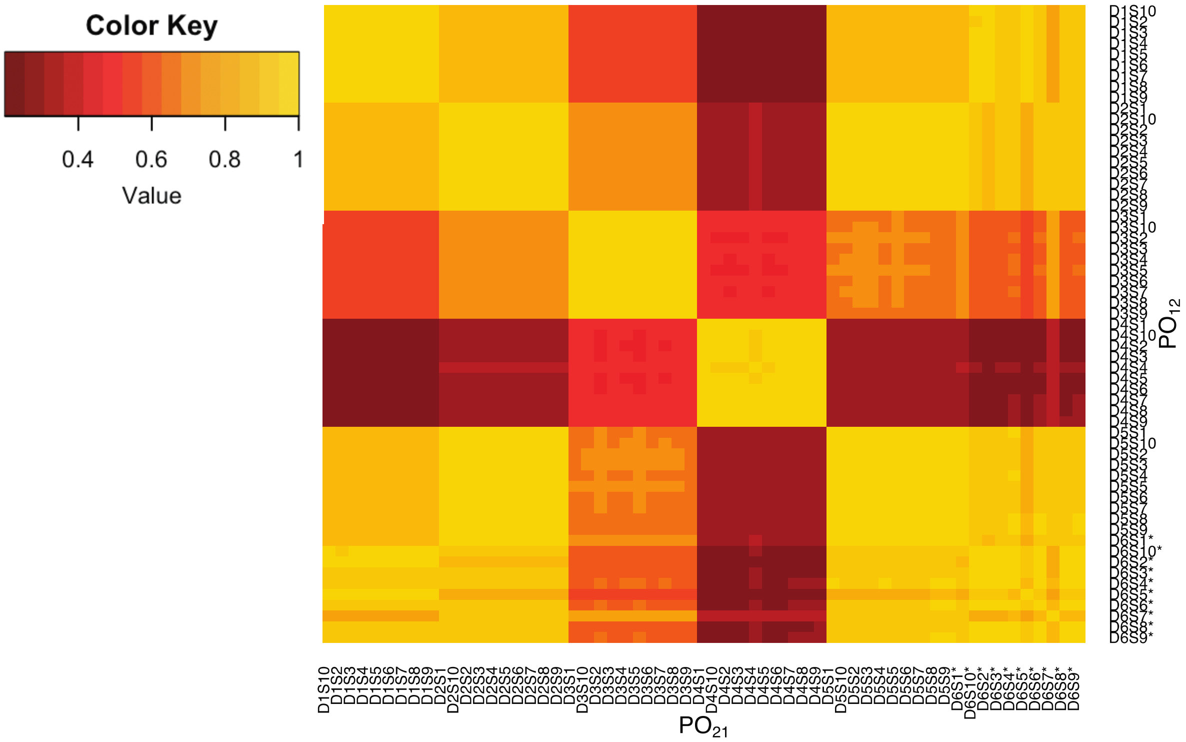


Figure Ten: Heatmap of PO performed on the *in silico* training set.

References

1. Liebold, M.A., *The niche concept revisited: mechanistic models and community context.* Ecology, 1995. **76**: p. 1371-1382.
2. Feinsinger, P., E.E. Spears, and R.W. Poole, *A simple measure of niche breadth.* Ecology, 1981. **62**(1): p. 27-32.
3. Ludwig, J. and J. Reynolds, *Statistical Ecology. Wiley Intersciences, Hoboken, New Jersey, USA.* 1988.
4. MacArthur, R.H., *Geographical ecology: patterns in the distribution of species. Harper and Row, New York, New York, USA.* 1972.
5. Theodorsson, E., *Limit of detection, limit of quantification and limit of blank.* Eurachem Professional Workshop, October 2015, 2015.
6. Group, E.C.W., *Quantifying uncertainty in analytical measurement, Third Ed. (Eds. SLR Ellison, A Williams). Eurachem / CITAC Guide CG 4. Eurachem, Torino, Italy.* 2012.
7. Bolyen, E., et al., *QIIME 2: Reproducible, interactive, scalable and extensible microbiome data science.* PeerJ Preprints, 2018. **6**: p. e27295v2.
8. Weiher, E. and P.A. Keddy, *Ecological assembly rules: perspectives, advances, retreats (eds. E Weiher, P Keddy). Cambridge University Press, Cambridge, UK.* 2000.