**Colony-based Sanger Sequencing Analysis**

**Student Handout**

**Objectives**

* + Manipulate student collected datasets to conduct community-level ecological analyses
  + Use community-level data to address questions about insect microbiomes
  + Use Google Sheets to calculate community ecology variables
  + Compare microbial community using community ecology variables

**Introduction**

Your group or class has collected data on the microbial community of bean beetles based on 16S rRNA sequencing of individual bacterial colonies cultured from bean beetle homogenates plated on different media. Since only a small number of colonies were sequenced from each plate, the data do not represent the entire microbial community for a particular sample. However, comparisons may be made based on host bean species, sex, and other sample or experimental variables with the assumption that the collection of taxa from a given treatment are representative of the bacterial microbiome communities of bean beetles in that treatment. There are no independent samples within treatment groups in this kind of analysis, however, we do have an identification for each picked colony in the dataset to the level of genus.

*Dataset Creation*

1. You and your fellow students will have conducted BLASTn searches on the individual 16S rRNA gene sequences and entered the genus and family of each sample in a spreadsheet. This spreadsheet should be created in or uploaded to Google Sheets to begin this analysis.
2. If your instructor shares a Google Sheets document with the class, be sure to **save your own copy of the document before beginning the data manipulation**.

Graphical user interface, application, table, Excel

Description automatically generated

*Data manipulation*

1. We need to consolidate the data by the variable of interest, for example, for each host type, each sex, or two manipulation treatments by the bacterial taxa. The easiest way to do this is with the Pivot Table function in Google Sheets.
2. When clicked on a cell within the data, create a Pivot Table (Insert -> Pivot Table) in a new sheet. Make sure that the data source includes the top row, which has the column headings. Set the treatment(s) that you are interested in as the rows and the taxonomic level you are interested in as the columns. The Values should be a COUNTA of the Sample Name, as each row in the dataset represents a single sample.
3. Uncheck the box Show totals for Rows. Then copy the entire table and paste Special -> Values only in a new sheet.
4. Label this new sheet Community Data. You can add zeros to all of the empty cells in the Pivot Table but it is not necessary (You want to keep the Grand totals column and rename it Abundance).
5. Delete any extra rows at the top of the spreadsheet so cell A1 is Host Bean. The top row should be the names of the taxa.

Graphical user interface, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Calculating diversity indices

1. Species richness – the number of unique species in a sample
   1. Although you could manually count the number of cells with values greater than zero for each treatment, using the COUNTIF formula in Google Sheets is easier (e.g., =COUNTIF(range,”>0”)).

Graphical user interface, application

Description automatically generated

1. Simpson Index – the Simpson Index incorporates both species richness and species evenness.
   1. D = Ʃ(n/N)^2, where n=number of individuals of a particular species and N=total number of individuals in a sample. D increases as diversity decreases, which is counterintuitive.
   2. Reciprocal Simpson – 1/D
   3. Inverse Simpson – 1-D
   4. Using the total abundance for each treatment, calculate the proportions squared. Using the Google Sheets trick that $ before a column or row prevents the program from iterating when copying a formula makes this easy.

A screenshot of a computer

Description automatically generated with medium confidence

* 1. Calculate the sum of the proportions squared (=SUM(B6:D6) for the first row) to calculate the Simpson Index.

Graphical user interface

Description automatically generated

* 1. Calculate the reciprocal and inverse Simpson using formulas in Google Sheets.

Graphical user interface, application

Description automatically generated

1. Shannon-Weaver (Shannon-Weiner) Index – also incorporates species richness and species evenness
   1. H=-Ʃ*p*ln*p*, where p is the proportion of individuals of each species in a community (i.e., n/N).
   2. Using the grand totals for each treatment, calculate the proportions. Using the Excel trick that $ before a column or row prevents Excel from iterating when copying a formula makes this easy.
   3. Note that ln*p* is undefined if *p*=0, so you can use an “IF” statement in Excel to prevent the calculation of undefined values. For example, =IF(B2>0,(B2/$E2)\*LN((B2/$E2)),””)

A screenshot of a computer

Description automatically generated with medium confidence

* 1. Complete the calculation of the Shannon-Weaver Index by calculating the negative sum of each row (=-SUM(B10:D10) for the first row).

Graphical user interface

Description automatically generated

Calculating community similarity (distance)

Sometimes we are interested in how similar (or different) two communities are based on what species are present and the relative abundance of those species in the two communities. One of the most common measures of distance is the Bray Curtis Dissimilarity. Similarity can be measured as 1-BC.

[Bray Curtis Dissimilarity](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2018/03/bray.jpg)  
Where:

* i & j are the two communities or treatments
* Si is the total number of specimens counted on community i,
* Sj is the total number of specimens counted on community j,
* Cij is the sum of only the lesser counts for each taxon found in both communities

In Google Sheets, Si and Sj are just the total abundance for each community (in this case, each treatment). To calculate Cij, we need to find the taxa that are present in both samples and then find the minimum. We can use the following formula for the first taxon in our demo dataset:

=IF(AND(B2>0,B3>0),MIN(B2:B3),0)

Where B2 is the cell with the number of individuals of the taxon for one sample and B3 is the cell with the number of individuals of the same taxon for the other sample. The formula first checks that the number of individuals is greater than zero for both samples. If this is true, it finds the minimum. If not, it returns a value of 0. The formula can be copied for all of the taxa and then SUM can be used to add up the values to calculate Cij.

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Questions

1. Based on the diversity indices that you calculated, which treatment had the highest (lowest) diversity?
2. Does the answer depend on the measure of diversity that you use?
3. Do your answers to the questions above depend on the taxonomic level of analysis?
4. Do your conclusions based on the analysis of picked colony Sanger sequences agree with your conclusions based on the colony phenotype analysis? If they do not agree, what could cause the difference between these analyses?

This study is based on Blumer LS, Beck CW 2020. **Introducing community ecology and data skills with the bean beetle microbiome project.** Article 24 In: McMahon, K editor. Advances in Biology Laboratory Education Volume 4**1** . Publication of the 41st Conference of the Association for Biology Laboratory Education (ABLE) https://doi.org/10.37590/able.v41.art24

Last edited by LSB 25 May 2022.