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

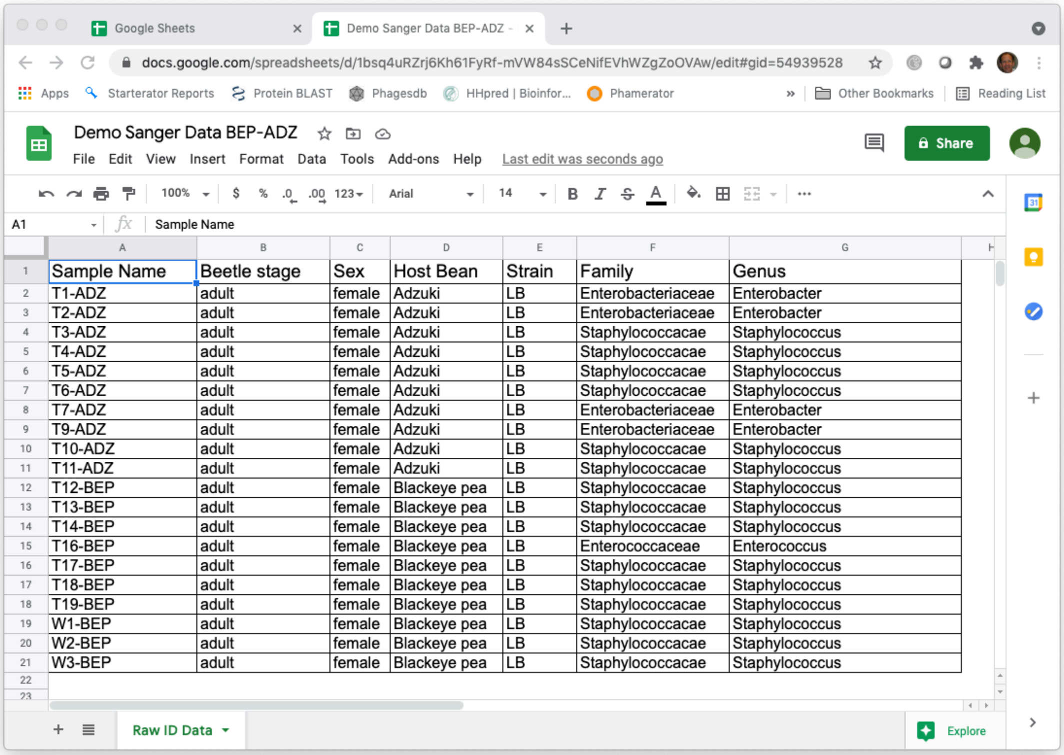
Introduction

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.

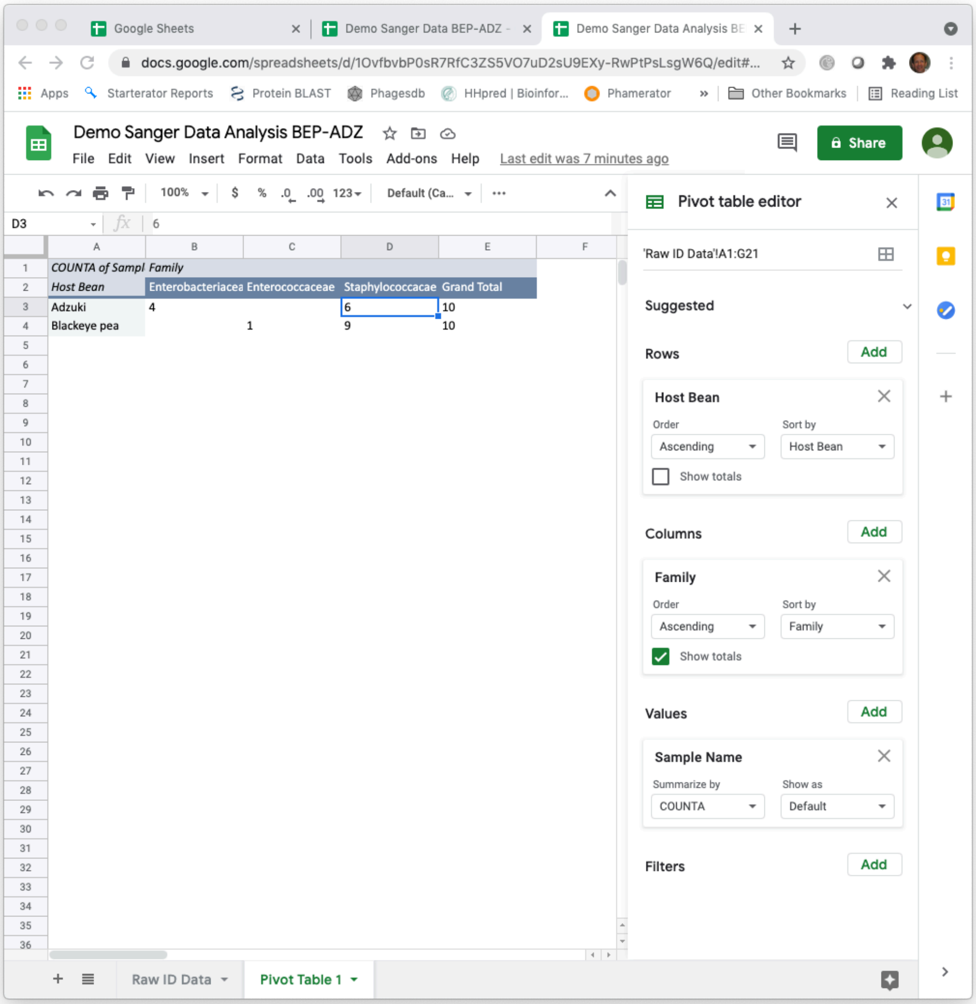
2. If your instructor shares a Google Sheets document with the class, be sure to **save your own copy of the document before beginning the data manipulation**.



Sample Name	Beetle stage	Sex	Host Bean	Strain	Family	Genus
T1-ADZ	adult	female	Adzuki	LB	Enterobacteriaceae	Enterobacter
T2-ADZ	adult	female	Adzuki	LB	Enterobacteriaceae	Enterobacter
T3-ADZ	adult	female	Adzuki	LB	Staphylococcacae	Staphylococcus
T4-ADZ	adult	female	Adzuki	LB	Staphylococcacae	Staphylococcus
T5-ADZ	adult	female	Adzuki	LB	Staphylococcacae	Staphylococcus
T6-ADZ	adult	female	Adzuki	LB	Staphylococcacae	Staphylococcus
T7-ADZ	adult	female	Adzuki	LB	Enterobacteriaceae	Enterobacter
T9-ADZ	adult	female	Adzuki	LB	Enterobacteriaceae	Enterobacter
T10-ADZ	adult	female	Adzuki	LB	Staphylococcacae	Staphylococcus
T11-ADZ	adult	female	Adzuki	LB	Staphylococcacae	Staphylococcus
T12-BEP	adult	female	Blackeye pea	LB	Staphylococcacae	Staphylococcus
T13-BEP	adult	female	Blackeye pea	LB	Staphylococcacae	Staphylococcus
T14-BEP	adult	female	Blackeye pea	LB	Staphylococcacae	Staphylococcus
T16-BEP	adult	female	Blackeye pea	LB	Enterococcaceae	Enterococcus
T17-BEP	adult	female	Blackeye pea	LB	Staphylococcacae	Staphylococcus
T18-BEP	adult	female	Blackeye pea	LB	Staphylococcacae	Staphylococcus
T19-BEP	adult	female	Blackeye pea	LB	Staphylococcacae	Staphylococcus
W1-BEP	adult	female	Blackeye pea	LB	Staphylococcacae	Staphylococcus
W2-BEP	adult	female	Blackeye pea	LB	Staphylococcacae	Staphylococcus
W3-BEP	adult	female	Blackeye pea	LB	Staphylococcacae	Staphylococcus

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.


Demo Sanger Data Analysis BEP-ADZ
File Edit View Insert Format Data Tools Add-ons Help
Last edit was 7 minutes ago
Share
Pivot table editor
Suggested
Raw ID Data
Pivot Table 1

COUNTA of Sampl	Family			
Host Bean	Enterobacteriacea	Enterococcaceae	Staphylococcaceae	Grand Total
Adzuki	4		6	10
Blackeye pea		1	9	10

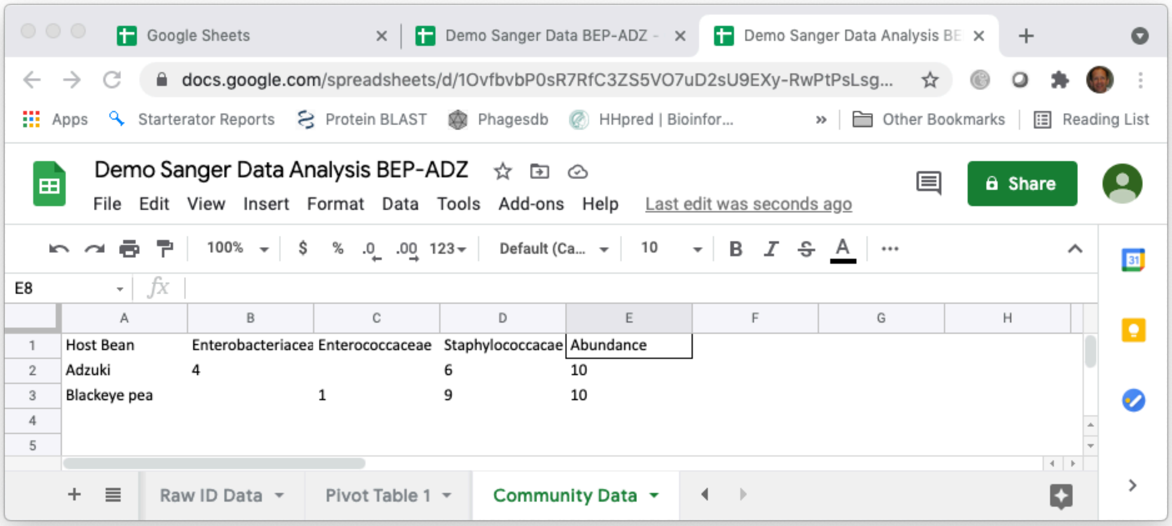
'Raw ID Data'!A1:G21

Rows: Host Bean — Order: Ascending; Sort by: Host Bean; ☐ Show totals

Columns: Family — Order: Ascending; Sort by: Family; ☑ Show totals

Values: Sample Name — Summarize by: COUNTA; Show as: Default

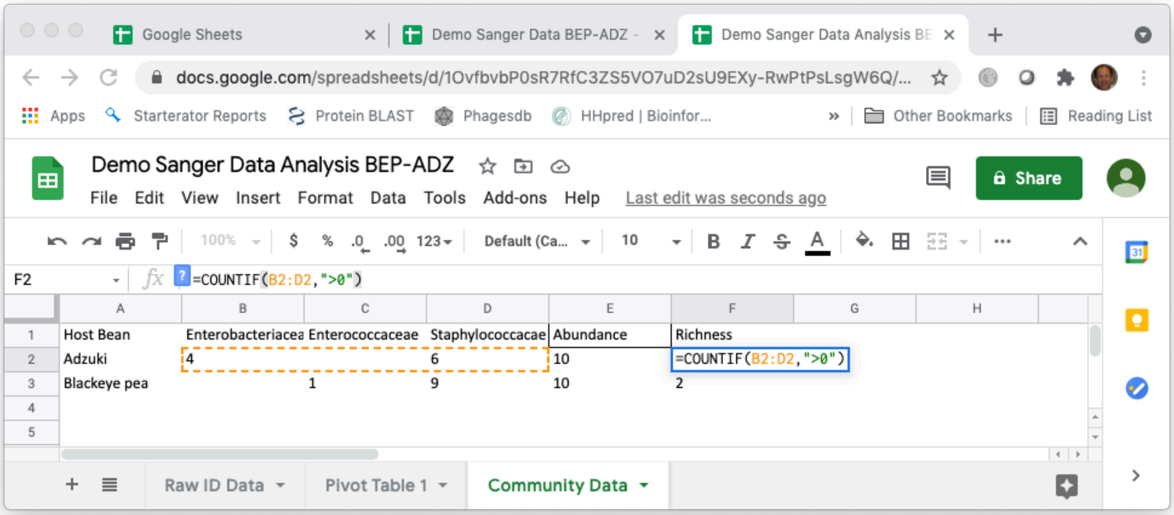
Filters


Demo Sanger Data Analysis BEP-ADZ
File Edit View Insert Format Data Tools Add-ons Help
Last edit was seconds ago
Share
Raw ID Data
Pivot Table 1
Community Data

Host Bean	Enterobacteriacea	Enterococcaceae	Staphylococcaceae	Abundance
Adzuki	4		6	10
Blackeye pea		1	9	10

Calculating diversity indices

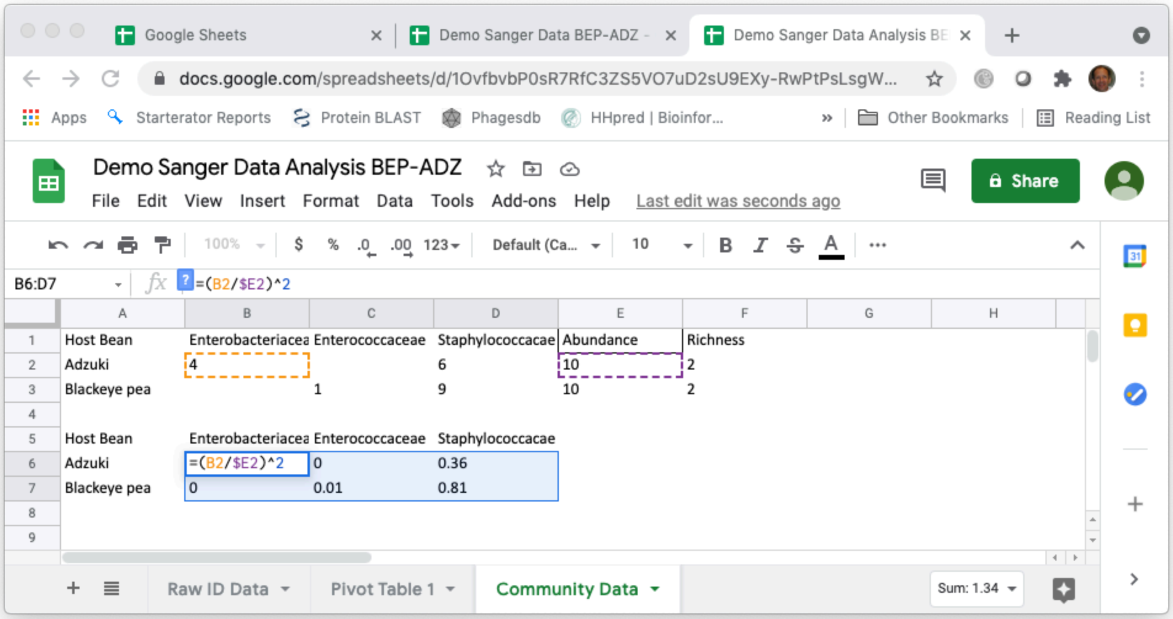
1. Species richness – the number of unique species in a sample
    a. Although you could manually count the number of cells with values greater than zero for each treatment, using the COUNTIF formula in Google Sheets is easier (e.g., =COUNTIF(range,”>0”)).



F2 =COUNTIF(B2:D2,">0")

	A	B	C	D	E	F
1	Host Bean	Enterobacteriacea	Enterococcaceae	Staphylococcacae	Abundance	Richness
2	Adzuki	4		6	10	=COUNTIF(B2:D2,">0")
3	Blackeye pea		1	9	10	2
4						
5						

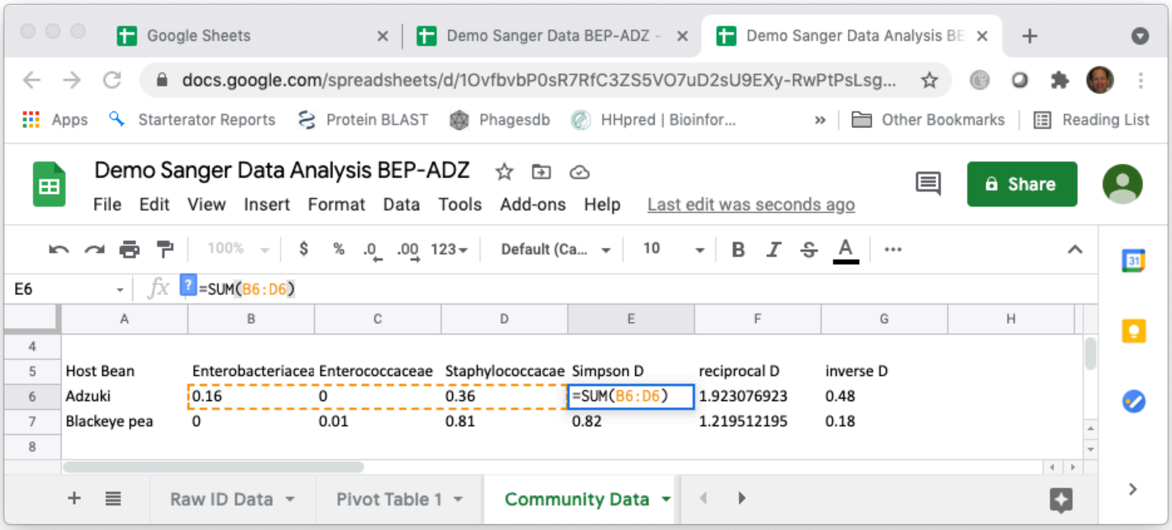
2. Simpson Index – the Simpson Index incorporates both species richness and species evenness.
    a. D = Σ(n/N)^2, where n=number of individuals of a particular species and N=total number of individuals in a sample. D increases as diversity decreases, which is counterintuitive.
    b. Reciprocal Simpson – 1/D
    c. Inverse Simpson – 1-D
    d. 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.


Google Sheets
Demo Sanger Data BEP-ADZ -
Demo Sanger Data Analysis BE
docs.google.com/spreadsheets/d/1OvfbvbP0sR7RfC3ZS5VO7uD2sU9EXy-RwPtPsLsgW...
Apps
Starterator Reports
Protein BLAST
Phagesdb
HHpred | Bioinfor...
Other Bookmarks
Reading List
Demo Sanger Data Analysis BEP-ADZ
File Edit View Insert Format Data Tools Add-ons Help
Last edit was seconds ago
Share
100%
Default (Ca...
10
Raw ID Data
Pivot Table 1
Community Data
Sum: 1.34

B6:D7 =(B2/$E2)^2

	A	B	C	D	E	F
1	Host Bean	Enterobacteriacea	Enterococcaceae	Staphylococcacae	Abundance	Richness
2	Adzuki	4		6	10	2
3	Blackeye pea		1	9	10	2
4						
5	Host Bean	Enterobacteriacea	Enterococcaceae	Staphylococcacae		
6	Adzuki	=(B2/$E2)^2	0	0.36		
7	Blackeye pea	0	0.01	0.81		

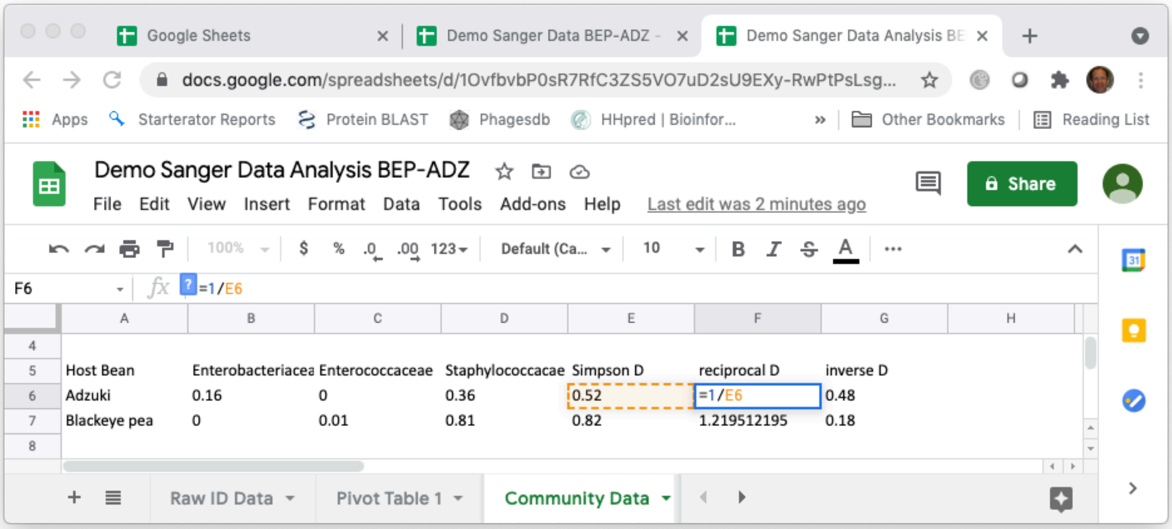
e. Calculate the sum of the proportions squared (=SUM(B6:D6) for the first row) to calculate the Simpson Index.


Google Sheets
Demo Sanger Data BEP-ADZ -
Demo Sanger Data Analysis BE
docs.google.com/spreadsheets/d/1OvfbvbP0sR7RfC3ZS5VO7uD2sU9EXy-RwPtPsLsg...
Apps
Starterator Reports
Protein BLAST
Phagesdb
HHpred | Bioinfor...
Other Bookmarks
Reading List
Demo Sanger Data Analysis BEP-ADZ
File Edit View Insert Format Data Tools Add-ons Help
Last edit was seconds ago
Share
100%
Default (Ca...
10
Raw ID Data
Pivot Table 1
Community Data

E6 =SUM(B6:D6)

	A	B	C	D	E	F	G
4							
5	Host Bean	Enterobacteriacea	Enterococcaceae	Staphylococcacae	Simpson D	reciprocal D	inverse D
6	Adzuki	0.16	0	0.36	=SUM(B6:D6)	1.923076923	0.48
7	Blackeye pea	0	0.01	0.81	0.82	1.219512195	0.18
8							

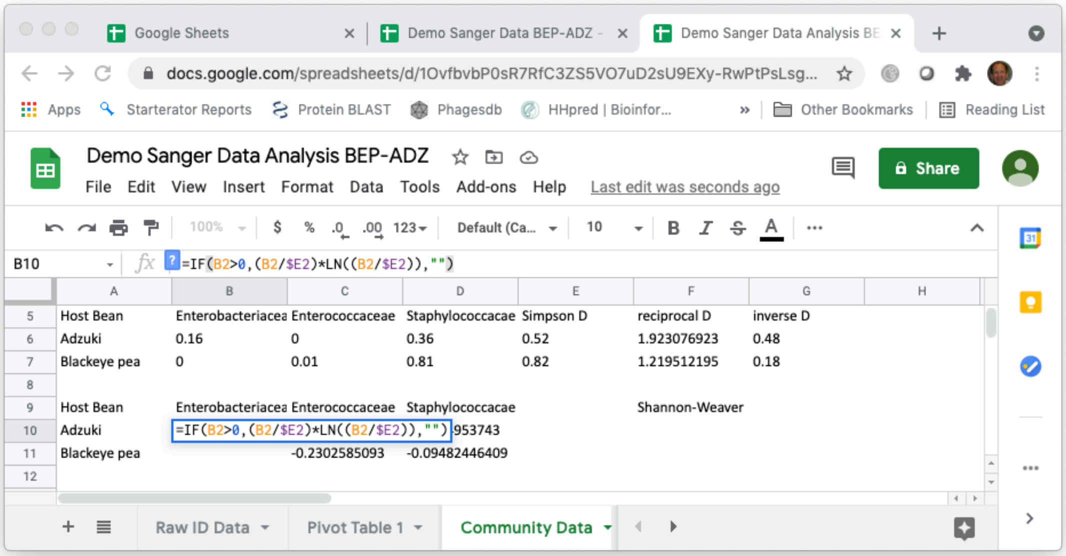
f. Calculate the reciprocal and inverse Simpson using formulas in Google Sheets.



F6 =1/E6

	A	B	C	D	E	F	G	H
4								
5	Host Bean	Enterobacteriacea	Enterococcaceae	Staphylococcacae	Simpson D	reciprocal D	inverse D	
6	Adzuki	0.16	0	0.36	0.52	=1/E6	0.48	
7	Blackeye pea	0	0.01	0.81	0.82	1.219512195	0.18	
8								

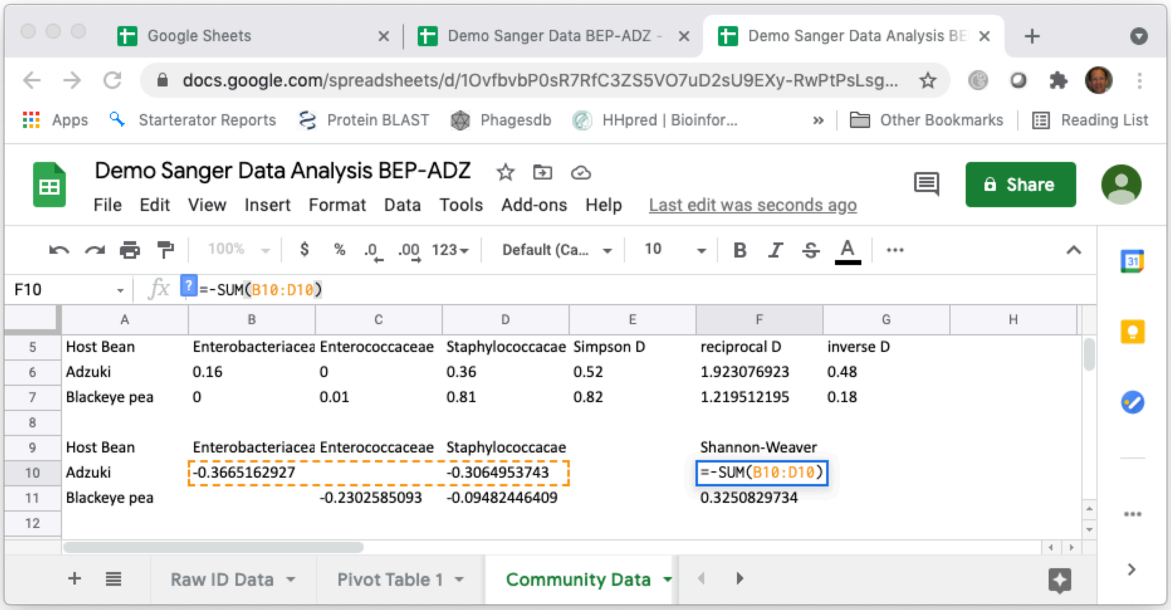
3. Shannon-Weaver (Shannon-Weiner) Index – also incorporates species richness and species evenness
    a. H=-Σ*p*ln*p*, where p is the proportion of individuals of each species in a community (i.e., n/N).
    b. Using the grand totals for each treatment, calculate the proportions. Using the Excel trick that $ before a column or row prevents Excel from iterating when copying a formula makes this easy.
    c. 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)),"")



B10 =IF(B2>0,(B2/$E2)*LN((B2/$E2)),"")

	A	B	C	D	E	F	G	H
5	Host Bean	Enterobacteriacea	Enterococcaceae	Staphylococcacae	Simpson D	reciprocal D	inverse D	
6	Adzuki	0.16	0	0.36	0.52	1.923076923	0.48	
7	Blackeye pea	0	0.01	0.81	0.82	1.219512195	0.18	
8								
9	Host Bean	Enterobacteriacea	Enterococcaceae	Staphylococcacae		Shannon-Weaver		
10	Adzuki	=IF(B2>0,(B2/$E2)*LN((B2/$E2)),"")		953743				
11	Blackeye pea		-0.2302585093	-0.09482446409				
12								

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


Google Sheets
Demo Sanger Data BEP-ADZ
Demo Sanger Data Analysis BE
docs.google.com/spreadsheets/d/1OvfbvbP0sR7RfC3ZS5VO7uD2sU9EXy-RwPtPsLsg...
Apps
Starterator Reports
Protein BLAST
Phagesdb
HHpred | Bioinfor...
Other Bookmarks
Reading List
Demo Sanger Data Analysis BEP-ADZ
File Edit View Insert Format Data Tools Add-ons Help
Last edit was seconds ago
Share
100%
$ % .0 .00 123
Default (Ca...
10
B I S A
Raw ID Data
Pivot Table 1
Community Data

F10 =-SUM(B10:D10)

	A	B	C	D	E	F	G	H
5	Host Bean	Enterobacteriacea	Enterococcaceae	Staphylococcacae	Simpson D	reciprocal D	inverse D	
6	Adzuki	0.16	0	0.36	0.52	1.923076923	0.48	
7	Blackeye pea	0	0.01	0.81	0.82	1.219512195	0.18	
8								
9	Host Bean	Enterobacteriacea	Enterococcaceae	Staphylococcacae		Shannon-Weaver		
10	Adzuki	-0.3665162927		-0.3064953743		=-SUM(B10:D10)		
11	Blackeye pea		-0.2302585093	-0.09482446409		0.3250829734		
12								

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.

$$BC_{ij} = 1 - \frac{2C_{ij}}{S_i + S_j}$$

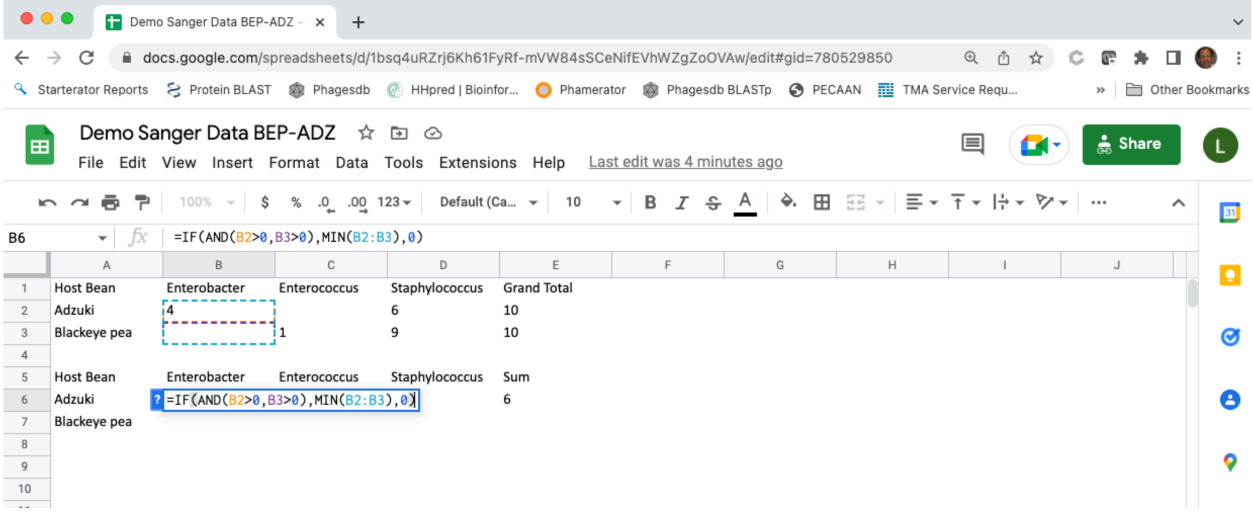
Where:

- i & j are the two communities or treatments
- S_i is the total number of specimens counted on community i,
- S_j is the total number of specimens counted on community j,
- C_{ij} is the sum of only the lesser counts for each taxon found in both communities

In Google Sheets, S_i and S_j are just the total abundance for each community (in this case, each treatment). To calculate C_{ij}, 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:

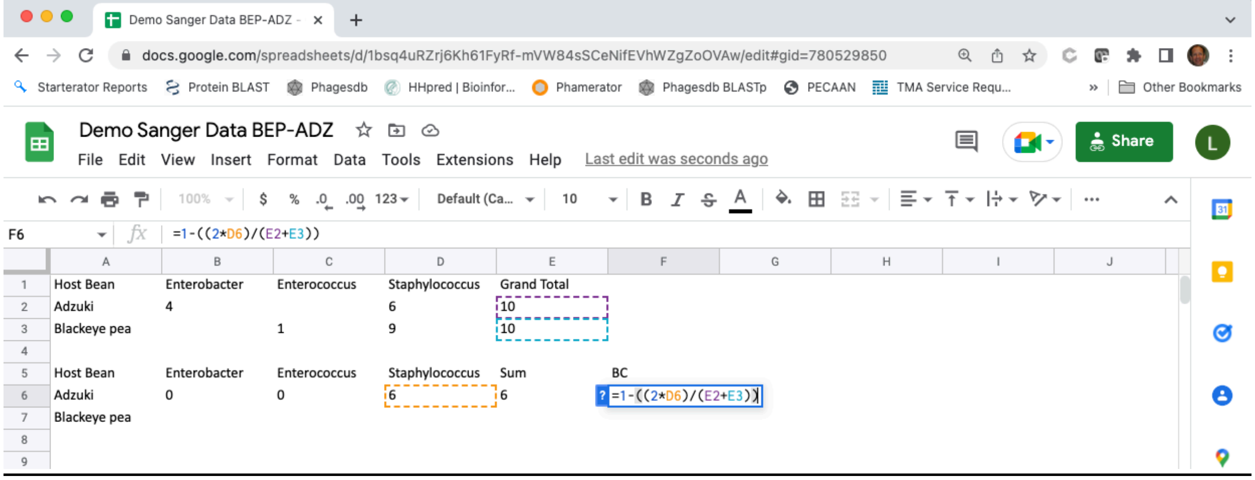
=IF(AND(B2>0,B3>0),MIN(B2:B3),0)

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 C_{ij}.



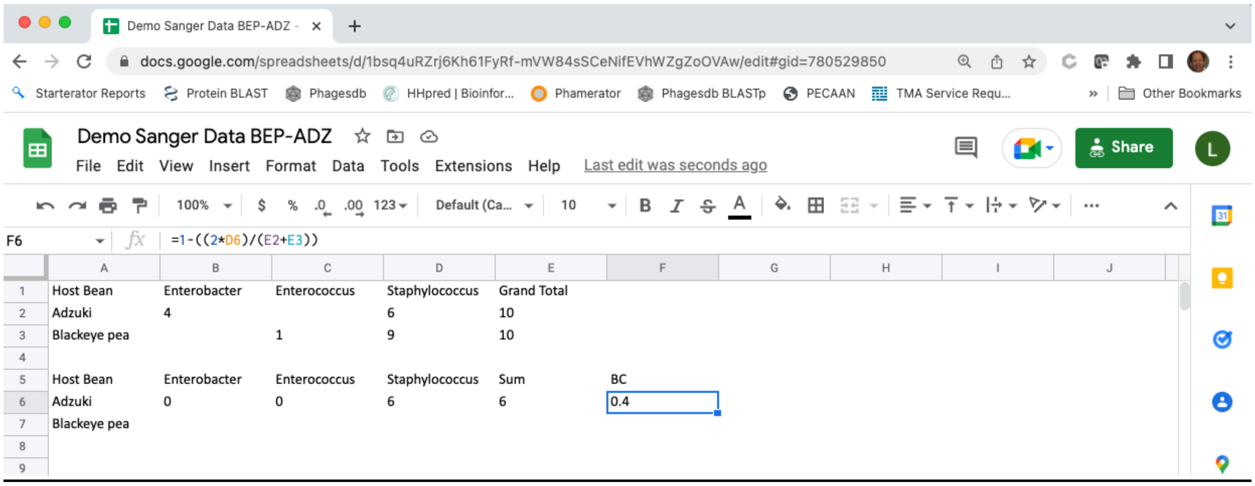
B6 =IF(AND(B2>0,B3>0),MIN(B2:B3),0)

	A	B	C	D	E
1	Host Bean	Enterobacter	Enterococcus	Staphylococcus	Grand Total
2	Adzuki	4		6	10
3	Blackeye pea		1	9	10
4					
5	Host Bean	Enterobacter	Enterococcus	Staphylococcus	Sum
6	Adzuki	=IF(AND(B2>0,B3>0),MIN(B2:B3),0)			6
7	Blackeye pea				
8					
9					
10					



F6 =1-((2*D6)/(E2+E3))

	A	B	C	D	E	F
1	Host Bean	Enterobacter	Enterococcus	Staphylococcus	Grand Total	
2	Adzuki	4		6	10	
3	Blackeye pea		1	9	10	
4						
5	Host Bean	Enterobacter	Enterococcus	Staphylococcus	Sum	BC
6	Adzuki	0	0	6	6	=1-((2*D6)/(E2+E3))
7	Blackeye pea					
8						
9						



F6 =1-((2*D6)/(E2+E3))

	A	B	C	D	E	F
1	Host Bean	Enterobacter	Enterococcus	Staphylococcus	Grand Total	
2	Adzuki	4		6	10	
3	Blackeye pea		1	9	10	
4						
5	Host Bean	Enterobacter	Enterococcus	Staphylococcus	Sum	BC
6	Adzuki	0	0	6	6	0.4
7	Blackeye pea					
8						
9						

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.