

Microbial Community Analysis Using Colony-based Sequencing Database

Student Handout

Objectives

- Manipulate large datasets to conduct community-level ecological analyses
- Use community-level data to address questions about insect microbiomes
- Use Excel or RStudio programs to calculate community ecology variables
- Compare microbial community using community ecology variables

Introduction

Microbiomes are the communities of microbes (bacteria, viruses, fungi and archaea) living symbiotically with all metazoans. In the past decade, both interest and research on microbiomes, including their implications for human health, have increased dramatically (Christian *et al.* 2015, Costello *et al.* 2012, McFall-Ngai *et al.* 2013, The Human Microbiome Consortium 2012, Young 2016). Insects have been used as model species to study the importance of microbiomes, because of their ease of use and the fact that microbial communities play diverse roles in insects (Engel and Moran 2013).

The data that are collected in any microbiome study consists of lists of the taxonomic units identified and their abundance. The same types of data are evaluated in an ecological community analysis, but now the communities are the collections of microbes that constitute different microbiomes. The community variables, “species” richness and relative abundance, are the same and the statistical methods used to compare communities, diversity and difference indices, also are the same. Perhaps the simplest measure of community structure used by ecologists is “species” or taxon richness, a count of the number of unique taxa in a sample. However, species richness does not consider the relative abundance of species in a community. Imagine two communities with five different species. In one community, all of the species have the same relative abundance. In the other community, one species dominates comprising 95% of individuals in the community. The other four species are very rare. Based on species richness as a measure of community structure, these two communities are the same, although they are clearly very different. As a result, ecologists use other species diversity indices that consider both the number of species and the relative abundance of species in a community. Two common indices are the Simpson Index and the Shannon-Weaver Index. Communities with greater numbers of species and higher evenness (i.e., similar relative abundance of species within a community) are considered more diverse. Finally, measures of species richness and species diversity do not consider the identity of species in a community. So, communities could have the same level of species diversity, but have completely different species. Measure of community similarity, such as the Bray-Curtis Index, compare the similarity (or dissimilarity) between two communities based on the identity of species in the communities, as well as their relative abundances. For more information on indices of species diversity and measures of community similarity, see Krebs (1999).

In this study, bean beetle gut microbiome data were collected by undergraduate students using the protocols developed by Cole *et al.* (2018). Three types of data were collected: colony phenotypes from cultured bacteria, 16s rRNA gene sequencing of specific bacterial colonies, and whole community 16s rRNA gene sequencing, but we will limit our analyses to the colony phenotype and colony-based 16s data.

Questions

Using data from the colony-based sequencing database and the analyses described below, answer the following questions.

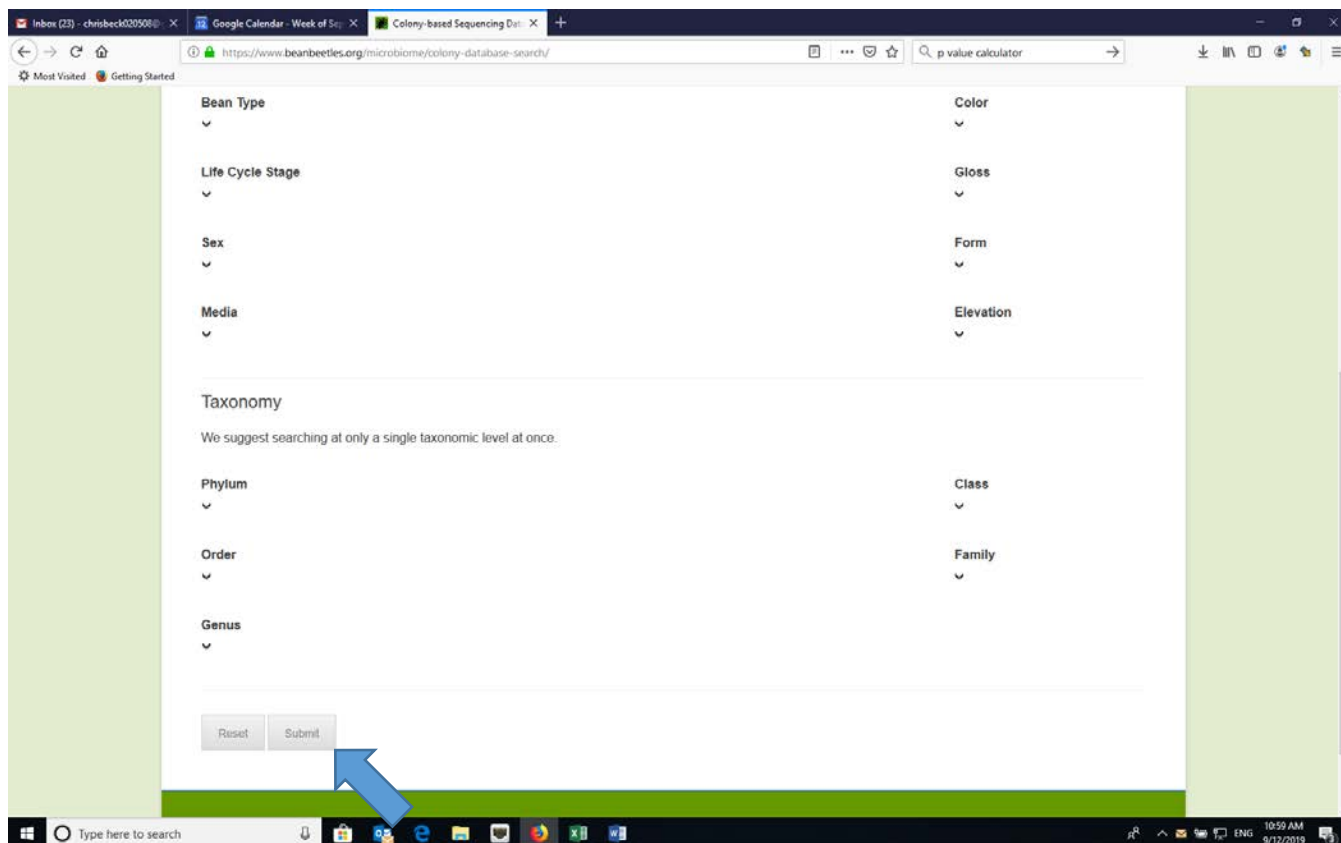
1. Which taxa are most prevalent in the bacterial communities in bean beetles?
2. Do the most prevalent taxa vary based on host bean type?
3. Based on the diversity indices that you calculated, which treatment had the highest (lowest) diversity?
4. Does the answer depend on the measure of species (taxon) diversity that you use?
5. Is there a relationship between number of samples and taxonomic diversity? If so, what might explain this?
6. Which communities are most similar (different)?
7. Do your answers to the questions above depend on the taxonomic level of analysis?

Database description

This database contains data for 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 are sequenced from each plate, the data do not represent the entire microbial community for a particular sample. However, qualitative comparisons can be made based on bean host species, sex of beetle, and other variables.

Access the database at <https://www.beanbeetles.org/microbiome/colony-database-search/>.

The database allows you to limit your search by bean host type, sex, life cycle stage, media on which bacteria were grown, colony phenotype, and bacterial taxonomy. Since we are interested in making comparisons between bacterial communities based on host species and sex, we want to download the entire database. Clicking “Submit” without limiting the search will allow you to view all of the data.



Downloading Data

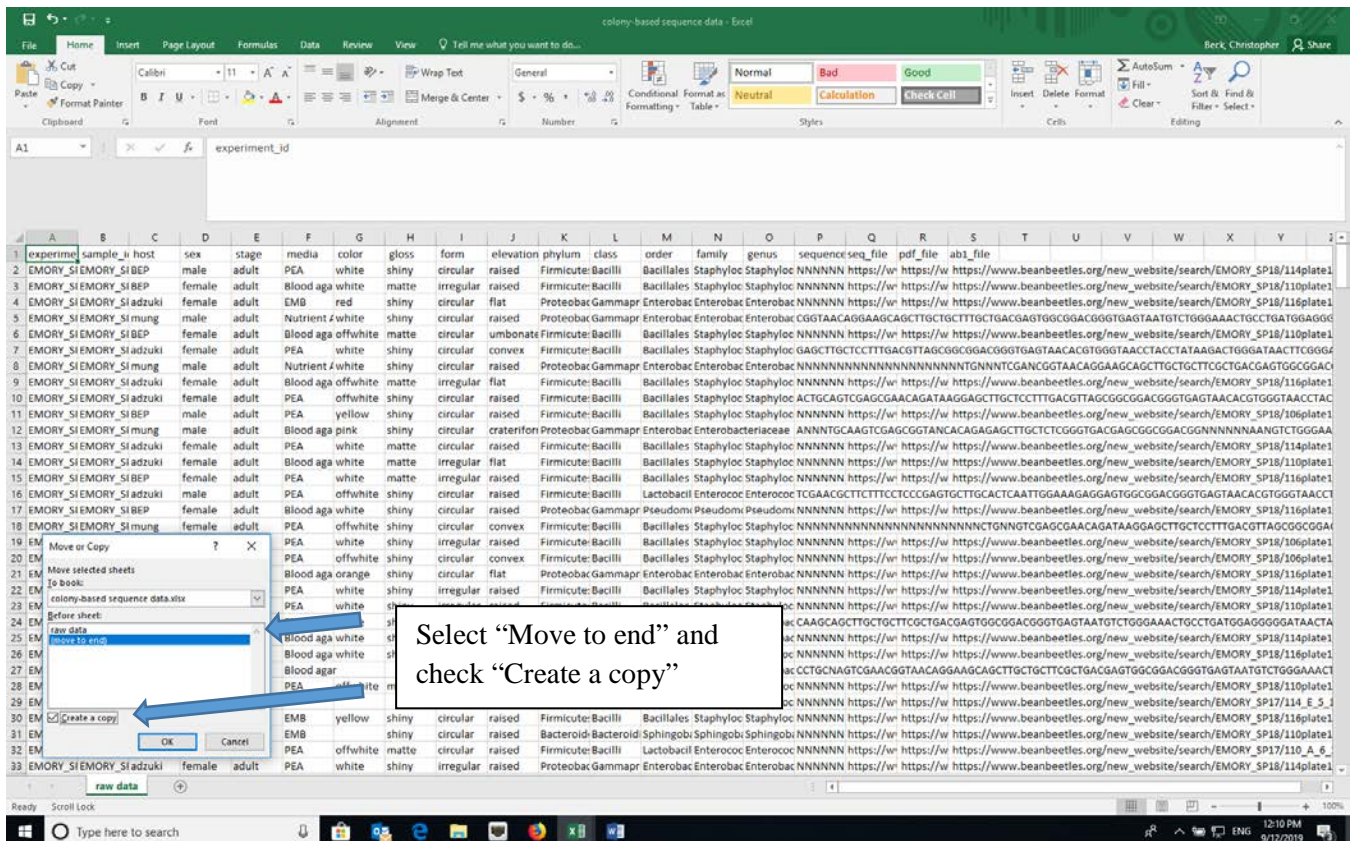
While we can view the data on the website, we want to download the data to manipulate. Click the download link to download a csv file with the data. Then, save the file as an Excel file (name the file “colony-based sequence data”) and rename the tab “raw data.”

Double click tab and rename as "raw data"

Data Reduction

1. Make a copy of the raw data in a new sheet using the sheet copy function in Excel (right click on the tab and select “Move or copy” and rename the tab (“reduced raw data”).

The screenshot displays an Excel spreadsheet titled "colony-based sequence data - Excel". The spreadsheet contains a table with 33 rows of data. The columns are labeled as follows: A (experim), B (sample_id), C (host), D (sex), E (stage), F (media), G (color), H (gloss), I (form), J (elevation), K (phylum), L (class), M (order), N (family), O (genus), P (sequence), Q (seq_file), R (pdf_file), and S (abi_file). The data rows start with "EMORY_SI EMORY_SI BEP" and continue with various sample IDs and host names. A right-click context menu is open over the "raw data" tab at the bottom, with the "Move or Copy..." option selected. A text box overlay with a blue arrow pointing to the "Move or Copy..." option reads: "Right click on tab and select “Move or Copy”".



2. In the “reduced raw data” sheet, delete any columns that we don’t need, such as the colony phenotype (color, gloss, form, elevation) and sequence data columns. The “reduced raw data” sheet is the data source if you choose to analyze these data in RStudio. Additional data manipulation and formatting (below) is required if you choose to analyze these data in Excel.

Data manipulation

1. We need to consolidate the data for each host species, each sex, or the combination of the two by the bacterial taxa. The easiest way to do this is with the Pivot Table function in Excel.
2. When clicked on a cell within the data, create a Pivot Table (Insert -> Pivot Table OR Data -> Summarize with Pivot Table). Make sure that the data source includes the top row (the cell range should include “\$A\$1”), which has the column headings. Click OK to create the pivot table in a new worksheet and label the tab “pivot table”.

colony-based sequence data - Excel

Beck, Christopher Share

The Excel ribbon is visible with the following tabs and icons:

- File**: Save, Open, Recent, Print, etc.
- Home**: Font, Paragraph, Styles, etc.
- Insert**: Tables, Charts, Links, etc.
- Page Layout**: Themes, Background Images, etc.
- Formulas**: Calculation Groups, etc.
- Data**: Data Tools, etc.
- Review**: Proofing, Changes, etc.
- View**: Views, etc.
- Help**: Tell me what you want to do...
- Store**: My Add-ins, etc.
- My Add-ins**: Bing Maps, People Graph, etc.
- Recommended Charts**: Charts, etc.
- PivotChart**: PivotChart, etc.
- 3D Map**: 3D Map, etc.
- Line**: Line, Column, Win/Loss, etc.
- Slicer**: Slicer, Timeline, etc.
- Hyperlink**: Hyperlink, etc.
- Text Box**: Text Box, etc.
- Header & Footer**: Header & Footer, etc.
- WordArt**: WordArt, etc.
- Signature**: Signature, etc.
- Object**: Object, etc.
- Equation**: Equation, etc.
- Symbol**: Symbol, etc.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	experime	sample_id	host	sex	stage	media	phylum	class	order	family	genus															
2	EMORY_SI	EMORY_SI	BEP	male	adult	PEA	Firmicute: Bacilli	Bacillales	Staphylococ	Staphylococcus																
3	EMORY_SI	EMORY_SI	BEP	female	adult	Blood aga	Firmicute: Bacilli	Bacillales	Staphylococ	Staphylococcus																
4	EMORY_SI	EMORY_SI	adzuki	female	adult	EMB	Proteobac	Gammapr	Enterobac	Enterobac	Enterobacter															
5	EMORY_SI	EMORY_SI	mung	male	adult	Nutrient /	Proteobac	Gammapr	Enterobac	Enterobac	Enterobacter															
6	EMORY_SI	EMORY_SI	BEP	female	adult	Blood aga	Firmicute: Bacilli	Bacillales	Staphylococ	Staphylococcus																
7	EMORY_SI	EMORY_SI	adzuki	female	adult	PEA	Firmicute: Bacilli	Bacillales	Staphylococ	Staphylococcus																
8	EMORY_SI	EMORY_SI	mung	male	adult	Nutrient /	Proteobac	Gammapr	Enterobac	Enterobac	Enterobacter															
9	EMORY_SI	EMORY_SI	adzuki	female	adult	Blood aga	Firmicute: Bacilli	Bacillales	Staphylococ	Staphylococcus																
10	EMORY_SI	EMORY_SI	adzuki	female	adult	PEA	Firmicute: Bacilli	Bacillales	Staphylococ	Staphylococcus																
11	EMORY_SI	EMORY_SI	BEP	male	adult	PEA	Firmicute: Bacilli	Bacillales	Staphylococ	Staphylococcus																
12	EMORY_SI	EMORY_SI	mung	male	adult	Blood aga	Proteobac	Gammapr	Enterobac	Enterobac	Enterobacteriaceae															
13	EMORY_SI	EMORY_SI	adzuki	female	adult	PEA	Firmicute: Bacilli	Bacillales	Staphylococ	Staphylococcus																
14	EMORY_SI	EMORY_SI	adzuki	female	adult	Blood aga	Firmicute: Bacilli	Bacillales	Staphylococ	Staphylococcus																
15	EMORY_SI	EMORY_SI	BEP	female	adult	PEA	Firmicute: Bacilli	Bacillales	Staphylococ	Staphylococcus																
16	EMORY_SI	EMORY_SI	adzuki	male	adult	PEA	Firmicute: Bacilli	Lactobacil	Enterococ	Enterococcus																
17	EMORY_SI	EMORY_SI	BEP	female	adult	Blood aga	Proteobac	Gammapr	Pseudomi	Pseudomi	Pseudomonas															
18	EMORY_SI	EMORY_SI	mung	female	adult	PEA	Firmicute: Bacilli	Bacillales	Staphylococ	Staphylococcus																
19	EMORY_SI	EMORY_SI	BEP	male	adult	PEA	Firmicute: Bacilli	Bacillales	Staphylococ	Staphylococcus																
20	EMORY_SI	EMORY_SI	BEP	male	adult	PEA	Firmicute: Bacilli	Bacillales	Staphylococ	Staphylococcus																
21	EMORY_SI	EMORY_SI	adzuki	female	adult	Blood aga	Proteobac	Gammapr	Enterobac	Enterobac	Enterobacter															
22	EMORY_SI	EMORY_SI	adzuki	male	adult	PEA	Firmicute: Bacilli	Bacillales	Staphylococ	Staphylococcus																
23	EMORY_SI	EMORY_SI	BEP	male	adult	PEA	Firmicute: Bacilli	Bacillales	Staphylococ	Staphylococcus																
24	EMORY_SI	EMORY_SI	mung	male	adult	EMB	Proteobac	Gammapr	Enterobac	Enterobac	Enterobacter															
25	EMORY_SI	EMORY_SI	adzuki	male	adult	Blood aga	Proteobac	Gammapr	Enterobac	Enterobac	Enterobacter															
26	EMORY_SI	EMORY_SI	adzuki	female	adult	Blood aga	Firmicute: Bacilli	Bacillales	Staphylococ	Staphylococcus																
27	EMORY_SI	EMORY_SI	adzuki	male	adult	Blood aga	Proteobac	Gammapr	Enterobac	Enterobac	Enterobacter															
28	EMORY_SI	EMORY_SI	adzuki	male	adult	PEA	Firmicute: Bacilli	Lactobacil	Enterococ	Enterococcus																
29	EMORY_SI	EMORY_SI	adzuki	female	adult	Firmicute: Bacilli	Bacillales	Staphylococ	Staphylococcus																	
30	EMORY_SI	EMORY_SI	BEP	female	adult	EMB	Firmicute: Bacilli	Bacillales	Staphylococ	Staphylococcus																
31	EMORY_SI	EMORY_SI	adzuki	female	adult	EMB	Bacteroid: Bacteroidi	Sphingobi: Sphingobi	Sphingobacterium																	
32	EMORY_SI	EMORY_SI	mung	male	adult	PEA	Firmicute: Bacilli	Lactobacil	Enterococ	Enterococcus																
33	EMORY_SI	EMORY_SI	adzuki	female	adult	PEA	Proteobac	Gammapr	Enterobac	Enterobac	Enterobacter															

Ready Scroll Lock

12:18 PM 9/12/2019

colony-based sequence data - Excel

File Home Insert Page Layout Formulas Data Review View Tell me what you want to do... Beck, Christopher Share

PivotTable Recommended Tables PivotTables Pictures Online Pictures Illustrations My Add-ins Add-ins Ring Maps People Graph Recommended Charts Charts PivotChart 3D Map Tours Sparklines Slicer Timeline Hyperlink Text Box Header & Footer WordArt Signature Line Object Equation Symbol

B1 sample_id

Create PivotTable

Choose the data that you want to analyze

☒ Select a table or range

Table/Range: reduced raw data: \$A\$1:\$X\$303

☐ Use an external data source

Connection name:

☐ Use this workbook's Data Model

Choose where you want the PivotTable report to be placed

☒ New Worksheet

☐ Existing Worksheet

Location:

Choose whether you want to analyze multiple tables

☐ Add this data to the Data Model

OK Cancel

Make sure that the cell range includes \$A\$1

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
2	EMORY_SI	EMORY_SI	BEP	male	adult	PEA	Firmicute: Bacilli	Bacillales	Staphylococ	Staphylococcus																
3	EMORY_SI	EMORY_SI	BEP	female	adult	Blood aga	Firmicute: Bacilli	Bacillales	Staphylococ	Staphylococcus																
4	EMORY_SI	EMORY_SI	adzuki	female	adult	EMB	Proteobac	Gammapr	Enterobac	Enterobac	Enterobacter															
5	EMORY_SI	EMORY_SI	mung	male	adult	Nutrient /	Proteobac	Gammapr	Enterobac	Enterobac	Enterobacter															
6	EMORY_SI	EMORY_SI	BEP	female	adult	Blood aga	Firmicute: Bacilli	Bacillales	Staphylococ	Staphylococcus																
7	EMORY_SI	EMORY_SI	adzuki	female	adult	PEA	Firmicute: Bacilli	Bacillales	Staphylococ	Staphylococcus																
8	EMORY_SI	EMORY_SI	mung	male	adult	Nutrient /	Proteobac	Gammapr	Enterobac	Enterobac	Enterobacter															
9	EMORY_SI	EMORY_SI	adzuki	female	adult	Blood aga	Firmicute: Bacilli	Bacillales	Staphylococ	Staphylococcus																
10	EMORY_SI	EMORY_SI	adzuki	female	adult	PEA	Firmicute: Bacilli	Bacillales	Staphylococ	Staphylococcus																
11	EMORY_SI	EMORY_SI	BEP	male	adult	PEA	Firmicute: Bacilli	Bacillales	Staphylococ	Staphylococcus																
12	EMORY_SI	EMORY_SI	mung	male	adult	Blood aga	Proteobac	Gammapr	Enterobac	Enterobac	Enterobacteriaceae															
13	EMORY_SI	EMORY_SI	adzuki	female	adult	PEA	Firmicute: Bacilli	Bacillales	Staphylococ	Staphylococcus																
14	EMORY_SI	EMORY_SI	adzuki	female	adult	Blood aga	Firmicute: Bacilli	Bacillales	Staphylococ	Staphylococcus																
15	EMORY_SI	EMORY_SI	BEP	female	adult	PEA	Firmicute: Bacilli	Bacillales	Staphylococ	Staphylococcus																
16	EMORY_SI	EMORY_SI	adzuki	male	adult	PEA	Firmicute: Bacilli	Lactobacil	Enterococ	Enterococcus																
17	EMORY_SI	EMORY_SI	BEP	female	adult	Blood aga	Proteobac	Gammapr	Pseudomi	Pseudomi	Pseudomonas															
18	EMORY_SI	EMORY_SI	mung	female	adult	PEA	Firmicute: Bacilli	Bacillales	Staphylococ	Staphylococcus																
19	EMORY_SI	EMORY_SI	BEP	male	adult	PEA	Firmicute: Bacilli	Bacillales	Staphylococ	Staphylococcus																
20	EMORY_SI	EMORY_SI	BEP	male	adult	PEA	Firmicute: Bacilli	Bacillales	Staphylococ	Staphylococcus																
21	EMORY_SI	EMORY_SI	adzuki	female	adult	Blood aga	Proteobac	Gammapr	Enterobac	Enterobac	Enterobacter															
22	EMORY_SI	EMORY_SI	adzuki	male	adult	PEA	Firmicute: Bacilli	Bacillales	Staphylococ	Staphylococcus																
23	EMORY_SI	EMORY_SI	BEP	male	adult	PEA	Firmicute: Bacilli	Bacillales	Staphylococ	Staphylococcus																
24	EMORY_SI	EMORY_SI	mung	male	adult	EMB	Proteobac	Gammapr	Enterobac	Enterobac	Enterobacter															
25	EMORY_SI	EMORY_SI	adzuki	male	adult	Blood aga	Proteobac	Gammapr	Enterobac	Enterobac	Enterobacter															
26	EMORY_SI	EMORY_SI	adzuki	female	adult	Blood aga	Firmicute: Bacilli	Bacillales	Staphylococ	Staphylococcus																
27	EMORY_SI	EMORY_SI	adzuki	male	adult	Blood aga	Proteobac	Gammapr	Enterobac	Enterobac	Enterobacter															
28	EMORY_SI	EMORY_SI	adzuki	male	adult	PEA	Firmicute: Bacilli	Lactobacil	Enterococ	Enterococcus																
29	EMORY_SI	EMORY_SI	adzuki	female	adult	Firmicute: Bacilli	Bacillales	Staphylococ	Staphylococcus																	
30	EMORY_SI	EMORY_SI	BEP	female	adult	EMB	Firmicute: Bacilli	Bacillales	Staphylococ	Staphylococcus																
31	EMORY_SI	EMORY_SI	adzuki	female	adult	EMB	Bacteroid: Bacteroidi	Sphingobi: Sphingobi	Sphingobacterium																	
32	EMORY_SI	EMORY_SI	mung	male	adult	PEA	Firmicute: Bacilli	Lactobacil	Enterococ	Enterococcus																
33	EMORY_SI	EMORY_SI	adzuki	female	adult	PEA	Proteobac	Gammapr	Enterobac	Enterobac	Enterobacter															
34	EMORY_SI	EMORY_SI	adzuki	female	adult	PEA	Proteobac	Gammapr	Enterobac	Enterobac	Enterobacteriaceae															

raw data reduced raw data

Enter Scroll Lock

Type here to search

12:20 PM 9/12/2019

- Set the treatment(s) that you are interested (for example, host species) in as the rows and the bacterial taxonomic level you are interested in as the columns. The Values should be a COUNT of the sample_id, as each row in the dataset represents a single sample.

Count of sample_id

Count of sample_id	Column Labels													
Row Labels	Acinetobacter	Bacillus	Burkholderia	Caballeronia	Paraburkholderia	Corynebacterium	Enterobacter	Enterococcus	Escherichia-Shigella	Klebsiella	Paenibacillus	Pseudomonas	Ralstonia	Sphingomonas
adzuki	1	1												
BEP						1	1	27	1	1			4	1
mung								56	10				4	
pigeon						1		9			1	1	5	
Grand Total	1	1			1	2	147	31	1	2	1	17	1	

Drag treatment of interest (e.g., host) to ROWS, taxonomic level (e.g., genus) to COLUMNS, and sample_id to VALUES

- You can add zeros to all of the empty cells in the Pivot Table using the Options menu and remove the Grand totals for Columns using the Options menu or the Design tab depending of your version of Excel. (You want to keep the Grand totals for rows to calculate diversity indices.)

colony-based sequence data - Excel

PivotTable Tools: Analyze, Design

PivotTable Name: PivotTable1

Active Field: Count of sample_id

Field Settings: Field Settings

PivotTable Options

PivotTable Name: PivotTable1

Layout & Format: Layout, Totals & Filters, Display, Printing, Data, Alt Text

Layout

☐ Merge and center cells with labels

When in compact form indent row labels: 1 (character(s))

Display fields in report filter area: Down, Then Over

Report filter fields per column: 0

Format

☐ For error values show:

☒ For empty cells show: 0

☒ Autofill column widths on update

☒ Preserve cell formatting on update

OK Cancel

PivotTable Fields

Choose fields to add to report:

experiment_id
sample_id
host
sex
stage
media
phylum
class
order
family
genus

MORE TABLES...

Drag fields between areas below:

FILTERS: genus

COLUMNS: host

ROWS: Count of sam...

VALUES: Count of sam...

Defer Layout Update UPDATE

Count of sample_id	Column Labels	Bacillus	Burkholderia-Caballeronia-Paraburkholderia	Corynebacterium 1	Enter
adzuki	Acinetobacter	1	1	0	5
BEP		0	0	1	1
mung		0	0	0	0
pigeon		0	0	0	1
Grand Total		1	1	1	7

colony-based sequence data - Excel

PivotTable Tools: Analyze, Design

PivotTable Name: PivotTable1

Active Field: Count of sample_id

Field Settings: Field Settings

PivotTable Options

PivotTable Name: PivotTable1

Layout & Format: Layout, Totals & Filters, Display, Printing, Data, Alt Text

Grand Totals

☒ Show grand totals for rows

☐ Show grand totals for columns

Filters

☐ Subtotal filtered page

☐ Allow multiple filters per field

Sorting

☒ Use Custom Lists when sorting

OK Cancel

Unselect checkbox for "Show grand totals for columns"

PivotTable Fields

Choose fields to add to report:

experiment_id
sample_id
host
sex
stage
media
phylum
class
order
family
genus

MORE TABLES...

Drag fields between areas below:

FILTERS: genus

COLUMNS: host

ROWS: Count of sam...

VALUES: Count of sam...

Defer Layout Update UPDATE

Count of sample_id	Column Labels	Bacillus	Burkholderia-Caballeronia-Paraburkholderia	Corynebacterium 1	Enter
adzuki	Acinetobacter	1	1	0	5
BEP		0	0	1	1
mung		0	0	0	0
pigeon		0	0	0	1

- You can remove the “blanks” column using the Column labels dropdown (located at upper left of the sheet) and unselecting “blank”.

The screenshot shows an Excel spreadsheet with a PivotTable. The PivotTable is named 'Count of sample_id' and is located in the range A3:L33. The PivotTable Fields task pane is open on the right, showing the 'Columns' section with 'genus' selected. The 'Rows' section has 'host' selected. The 'Values' section has 'Count of sam...' selected. The 'Filters' section is empty. The PivotTable data is as follows:

Count of sample_id	Burkholderia	Caballeronia	Paraburkholderia	Corynebacterium	1	Enterobacter	Enterococcus	Escherichia-Shigella	Klebsiella	Paenibacillus	Pseudomonas	Ralstonia	Sph...
0	3	55	20	1	0	1	0	0	4	1			
1	1	27	1	1	0	1	0	0	4	0			
0	0	56	10	0	1	1	1	4	0				
0	1	9	0	0	0	0	0	5	0				

The PivotTable Fields task pane shows the following fields:

- Columns: genus
- Rows: host
- Values: Count of sam...
- Filters: (empty)

The 'Column Labels' dropdown in the PivotTable is set to 'genus'. The '(blank)' checkbox in the task pane is unchecked. A text box with the text 'Unselect checkbox for "blanks"' is overlaid on the task pane.

- If you selected more than one treatment for the rows, you can get the treatment data to repeat for each sample. In the Design tab, select “Report Layout” and choose “Show in Tabular form” and “Repeat All Item Labels”.

- Copy and paste (as values) the pivot table to a new worksheet and remove any extra rows at the top. The top row should now have the taxa names. Name this tab “analysis”. Conduct the community ecology analyses that follow in Excel on the “analysis” sheet.

Calculating diversity indices

1. Species (taxon) richness – the number of unique species (taxa) 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 Excel is easier (e.g., =COUNTIF(range,">0")). Where "range" is the cell range in the datasheet, for example "C2:M2", a single row or treatment.

host	Acinetob	Bacillus	Burkholder	Coryneb	Enterobac	Enterococ	Escherich	Klebsiella	Paenibac	Pseudom	Raistonia	Sphingob	Staphyloc	Stenotroph	Grand Tot.	Richness
adzuki	1	1	0	5	55	20	0	1	0	4	1	1	64	1	154	15
BEP	0	0	1	1	27	1	1	0	0	4	0	0	50	2	87	10
mung	0	0	0	0	56	10	0	1	1	4	0	0	22	0	94	12
pigeon	0	0	0	1	9	0	0	0	0	5	0	0	7	0	22	7

2. Simpson Index – the Simpson Index incorporates both species (taxon) richness and species (taxon) evenness.

- $D = \sum (n/N)^2$, where n =number of individuals of a particular species (taxon) and N =total number of individuals in a sample. D increases as diversity decreases, which is counterintuitive. A reciprocal or inverse index would be more intuitive and are easily calculated.
- Reciprocal Simpson = $1/D$ and scales so the maximum value is the species richness of a community.
- Inverse Simpson = $1-D$ and scales to a maximum value of 1.0.
- Create a new data array below the original using the same row labels (treatment variables) and the same column labels (bacterial taxa).

colony-based sequence data - Excel

Beck, Christopher Share

File Home Insert Page Layout Formulas Data Review View Tell me what you want to do...

Clipboard Font Alignment Number Styles Editing

Calibri 11 A A

General Conditional Formatting Table

Normal Bad Good Neutral Calculation Check Cell

Insert Delete Format

AutoSum Fill Clear Sort & Find & Filter Select

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	host	Acinetob	Bacillus	Burkholder	Coryneb	Enterobac	Enterococ	Escherich	Klebsiella	Paenibaci	Pseudom	Ralstonia	Sphingobi	Staphyloc	Stenotroph	Grand Tot	Richness									
2	adzuki	1	1	0	5	55	20	0	1	0	4	1	1	64	1	154	11									
3	BEP	0	0	1	1	27	1	1	0	0	4	0	0	50	2	87										
4	mung	0	0	0	0	56	10	0	1	1	4	0	0	22	0	94										
5	pigeon	0	0	0	1	9	0	0	0	0	5	0	0	7	0	22										
6																										
7																										
8	host	Acinetob	Bacillus	Burkholder	Coryneb	Enterobac	Enterococ	Escherich	Klebsiella	Paenibaci	Pseudom	Ralstonia	Sphingobi	Staphyloc	Stenotroph	Grand Tot	Richness									
9	adzuki																									
10	BEP																									
11	mung																									
12	pigeon																									
13																										
14																										
15																										
16																										
17																										
18																										
19																										
20																										
21																										
22																										
23																										
24																										
25																										
26																										
27																										
28																										
29																										
30																										
31																										
32																										
33																										

raw data Sheet2 analysis reduced raw data

Ready Scroll Lock

2:04 PM 9/12/2019

- e. To calculate the proportion squared for each taxa, use the grand totals for each treatment. Using the Excel trick that \$ before a column or row prevents Excel from iterating when copying a formula makes this easy. For example, $= (C2/\$P2)^2$. Copy the formula across the row and then down.

colony-based sequence data - Excel

Home Insert Page Layout Formulas Data Review View Tell me what you want to do...

Clipboard Font Alignment Number Styles Editing

COUNTIF $= (B2/SP2)^2$

	host	Acinetobacte	Bacillus	Burkholder	Corynebaci	Enterobac	Enterococ	Escherichi	Klebsiella	Paenibaci	Pseudom	Ralstonia	Sphingobi	Staphyloc	Stenotroph	Grand Tot	Richness
1	adzuki	1	1	0	5	55	20	0	1	0	4	1	1	64	1	154	11
3	BEP	0	0	1	1	27	1	1	0	0	4	0	0	30	2	87	8
4	mung	0	0	0	0	56	10	0	1	1	4	0	0	22	0	94	6
5	pigeon	0	0	0	1	9	0	0	0	0	5	0	0	7	0	22	4

8 host Acinetobacte Bacillus Burkholder Corynebaci Enterobac Enterococ Escherichi Klebsiella Paenibaci Pseudom Ralstonia Sphingobi Staphyloc Stenotrophomonas

9 adzuki $= (B2/SP2)^2$ 4.22E-05 0.001054 0.127551 0.016866 0 4.22E-05 0 0.000675 4.22E-05 4.22E-05 0.17271 4.22E-05

10 BEP

11 mung

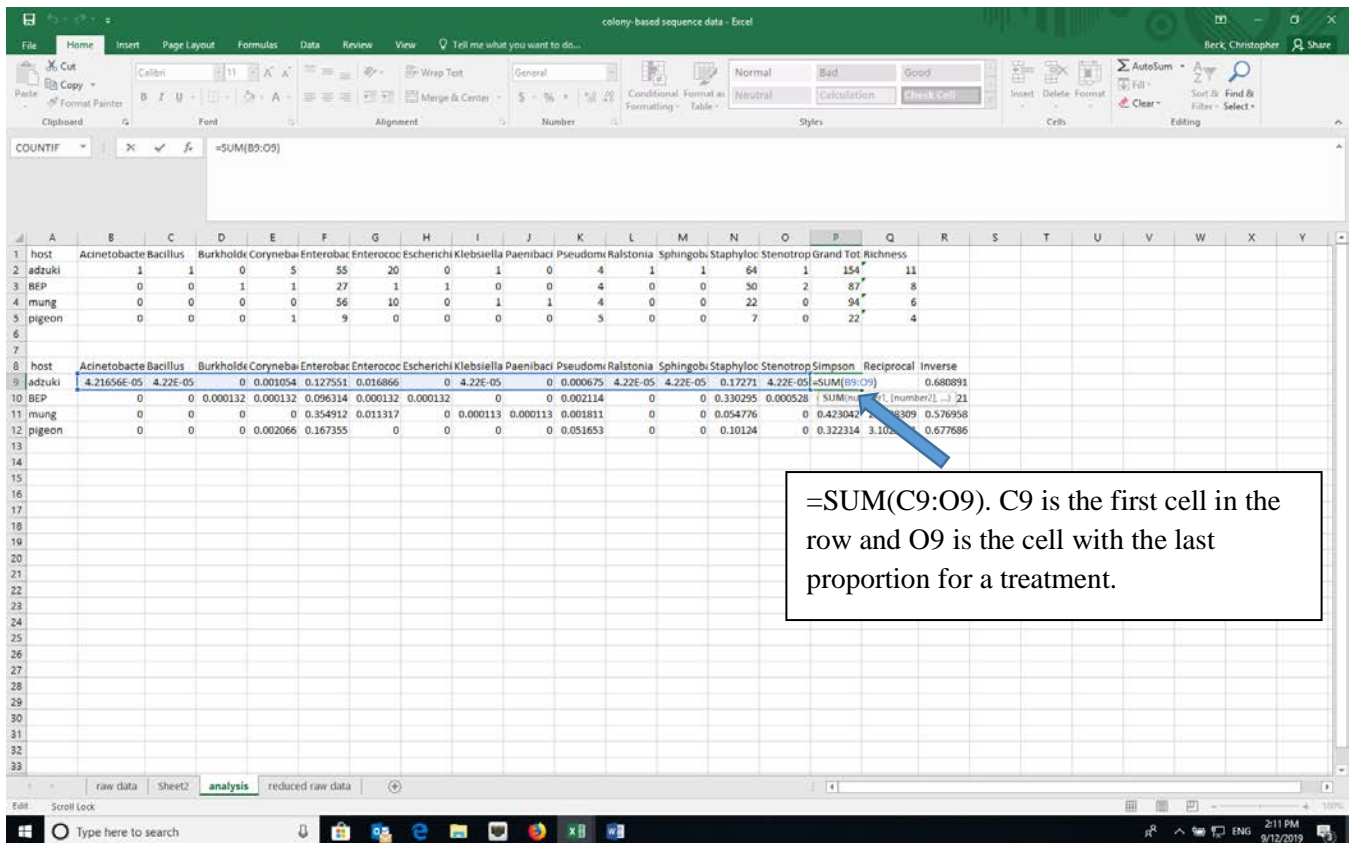
12 pigeon

$= (B2/SP2)^2$. B2 is the cell with the abundance of the first taxon and P2 is the cell with the grand total for a treatment. The \$ prevents the column identifier from changing. Copy the formula across the row and then down.

raw data Sheet2 analysis reduced raw data

Type here to search 2:07 PM 9/12/2019

- f. Calculate the sum of the proportions squared ($=SUM$ in Excel) to calculate the Simpson Index.



g. Calculate the reciprocal (e.g., $=1/P9$) and inverse Simpson (e.g., $=1-P9$) using formulas in Excel.

3. Shannon-Weaver (Shannon-Weiner) Index – also incorporates species (taxon) richness and species (taxon) evenness

- $H = -\sum p \ln p$, where p is the proportion of individuals of each bacterial taxon in a community (i.e., n/N).
- Create a new data array below the original using the same row labels (treatment variables) and the same column labels (species).
- Using the grand totals for each treatment, calculate the proportions ($p \ln p$). Using the Excel trick that \$ before a column or row prevents Excel from iterating when copying a formula makes this easy.
- Note that $\ln p$ is undefined if $p=0$, so you can use an "IF" statement in Excel. For example, $=IF(B2>0,(B2/$P2)*LN((B2/$P2)),"")$

colony-based sequence data v2 - Excel

File Home Insert Page Layout Formulas Data Review View Tell me what you want to do...

Clipboard Font Alignment Number Styles Editing

COUNTIF =IF(B2>0,(B2/\$P2)*LN((B2/\$P2)), "")

host	Acinetobacte	Bacillus	Burkholder	Coryneb	Enterobac	Enterococ	Escherich	Klebsiella	Paenibaci	Pseudom	Ralstonia	Sphingobi	Staphyloc	Stenotro	Simpson	Reciprocal	Inverse
adzuki	1	1	0	5	55	20	0	1	0	4	1	1	64	1	154	11	
BEP	0	0	1	1	27	1	1	0	0	4	0	0	30	2	87	8	
mung	0	0	0	0	56	10	0	1	1	4	0	0	22	0	94	6	
pigeon	0	0	0	1	9	0	0	0	0	5	0	0	7	0	22	4	
host	Acinetobacte	Bacillus	Burkholder	Coryneb	Enterobac	Enterococ	Escherich	Klebsiella	Paenibaci	Pseudom	Ralstonia	Sphingobi	Staphyloc	Stenotro	Simpson	Reciprocal	Inverse
adzuki	4.21656E-05	4.22E-05	0	0.001054	0.127551	0.016866	0	4.22E-05	0	0.000675	4.22E-05	0.17271	4.22E-05	0.319109	3.1337209	0.680891	
BEP	0	0	0.000132	0.000132	0.096314	0.000132	0.000132	0	0	0.002114	0	0	0.330295	0.000528	0.429779	2.3267753	0.570221
mung	0	0	0	0	0.354912	0.011317	0	0.000113	0.000113	0.001811	0	0	0.054776	0	0.423042	2.3638309	0.576958
pigeon	0	0	0	0.002066	0.167355	0	0	0	0	0.051653	0	0	0.10124	0	0.322314	3.1025641	0.677686
host	Acinetobacte	Bacillus	Burkholder	Coryneb	Enterobac	Enterococ	Escherich	Klebsiella	Paenibaci	Pseudom	Ralstonia	Sphingobi	Staphyloc	Stenotro	Shannon		
adzuki	=IF(B2>0,(B2/\$P2)*LN((B2/\$P2)), "")				-0.36772	-0.26509		-0.03271		-0.09482	-0.03271	-0.03271	-0.36491	-0.03271	1.400077		
BEP	=IF(logical_test, [value_if_true], [value_if_false])				-0.36313	-0.05133	-0.05133			-0.14159			-0.31832	-0.08673	1.115101		
mung					-0.30856	-0.23837		-0.04833	-0.04833	-0.13434			-0.33989		1.11783		
pigeon					-0.1405	-0.36565				-0.33673			-0.36436		1.207243		

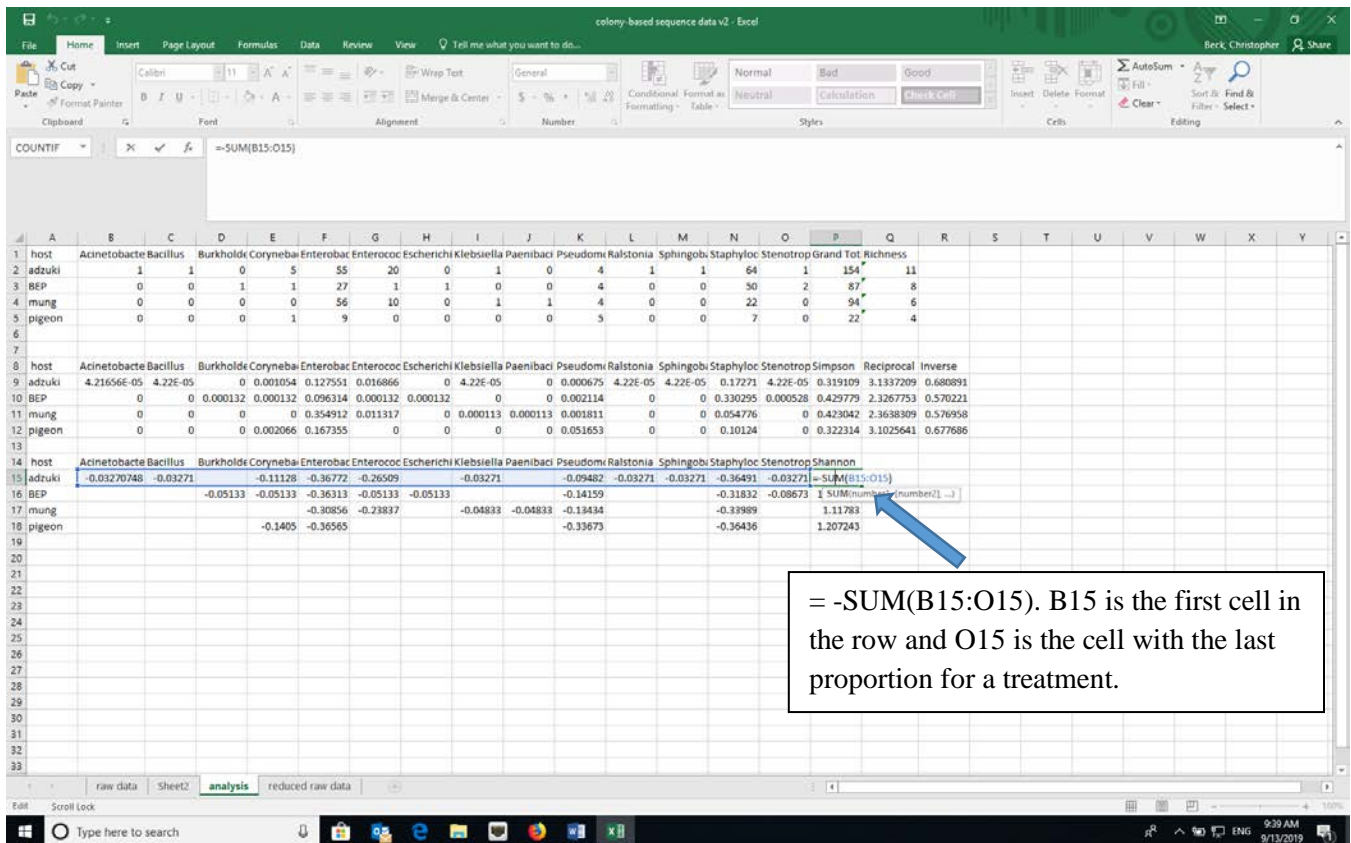
raw data Sheet2 analysis

Type here to search

9:32 AM 9/13/2019

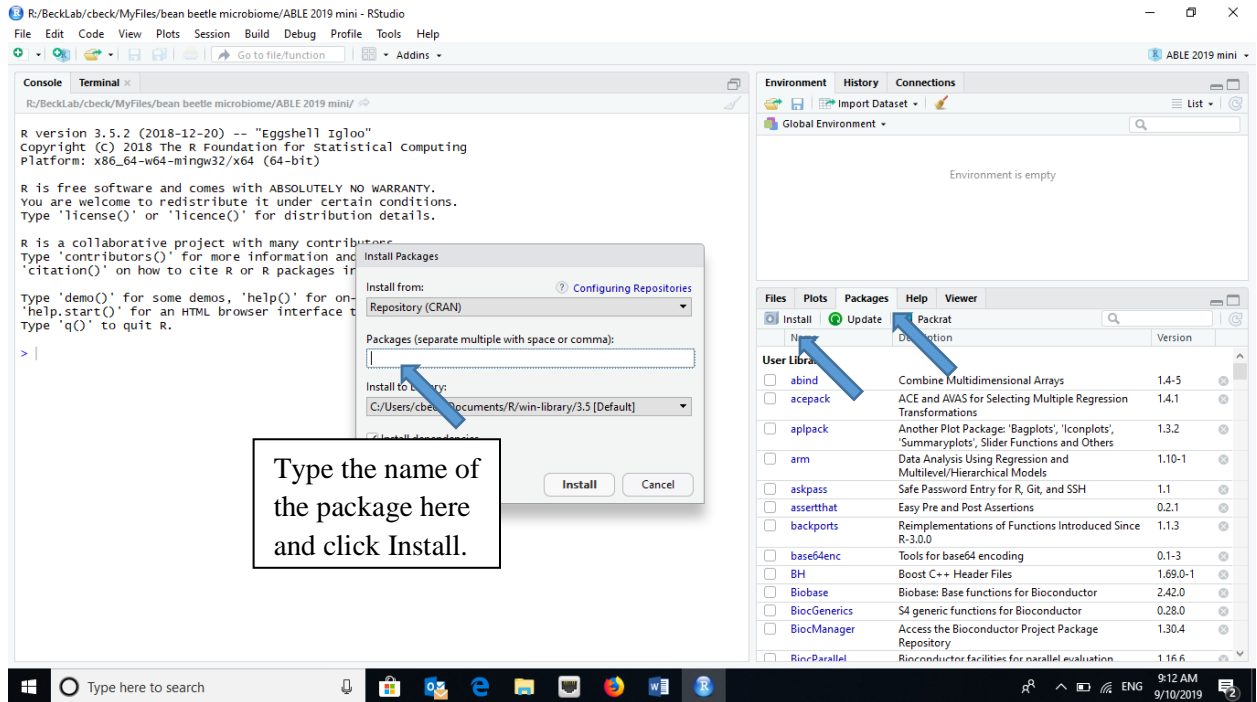
=IF(B2>0,(B2/\$P2)*LN((B2/\$P2)), ""). B2 is the cell with the abundance of the first taxon and P2 is the cell with the grand total for a treatment. The \$ prevents the column identifier from changing. Copy the formula across the row and then down.

- e. Calculate the negative sum of the proportions ($p \ln p$) (=SUM in Excel for each row, a different microbial community) to calculate the Shannon-Weaver Index.

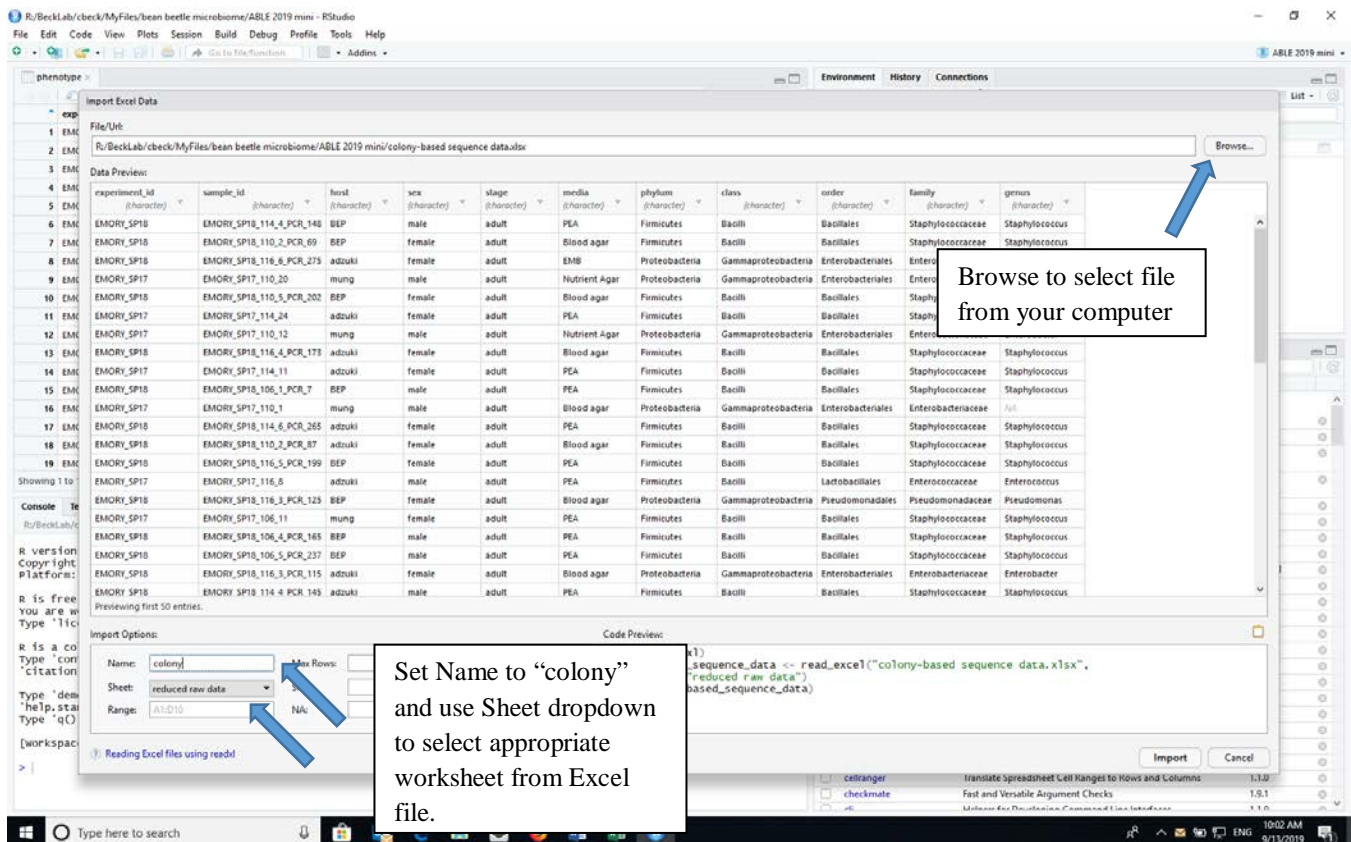
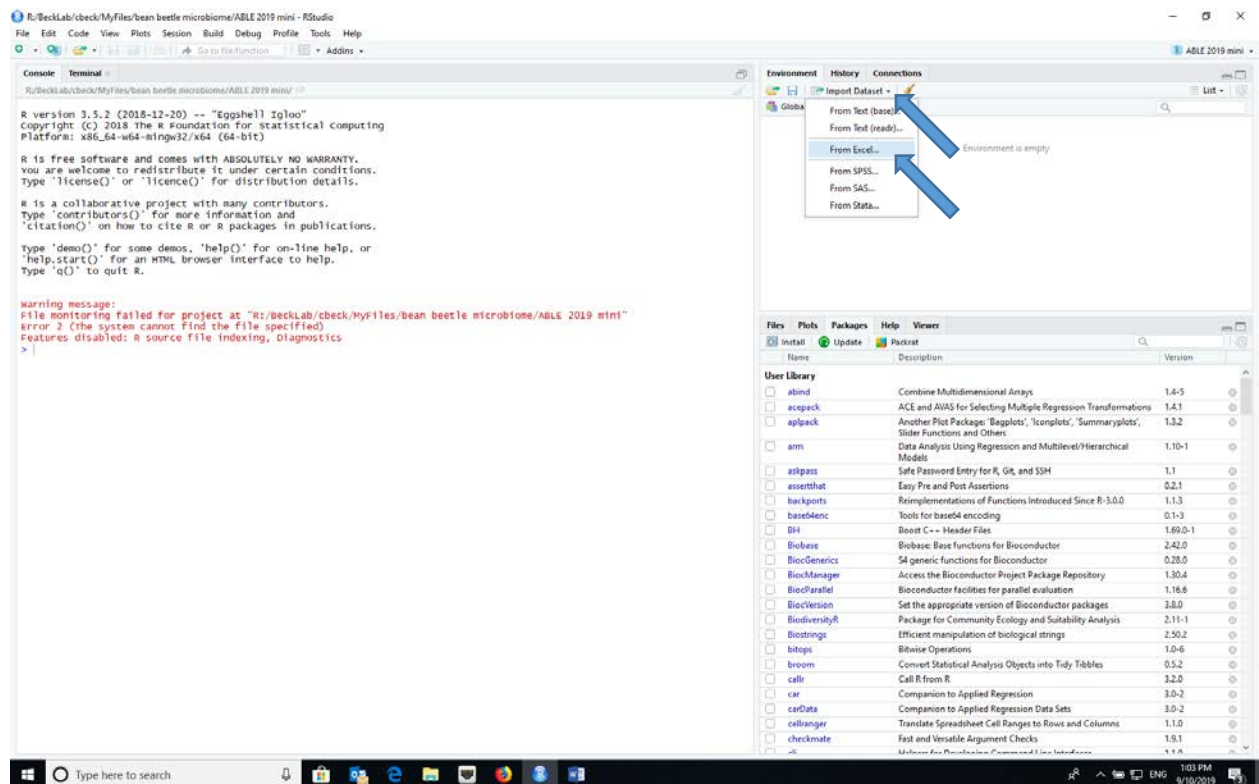


Data Manipulation in R

1. Open RStudio and create a new project using the New Project option under File and select for the new project to be in an existing folder where your data are.
2. Install the following packages either using the Packages tab in RStudio or the command `install.packages("name_of_package")` in the console. Note that BiodiversityR requires QuartzX on a Mac. If you are using a MacOS and don't have QuartzX, install it first and restart your computer before install these packages.
 - a. dplyr
 - b. reshape2
 - c. vegan
 - d. BiodiversityR
 - e. ggplot2



3. Load the packages listed above by clicking the checkboxes for the appropriate packages in the Packages tab or the command `library("name_of_package")` in the console.
4. Import the dataset "reduced raw data" (dataset without the extra metadata that you created in the Excel section) into RStudio.



5. Attach the imported dataset ("colony") to the dataframe using the attach command in the console (`attach(colony)`)
6. Create a community matrix (named "community" in this example) for a particular treatment. This example assumes that you are doing the analysis at the genus level. This can be changed to other taxonomic levels using the appropriate variable name

```
>community<-table(host,genus)
```

7. If you want to look at two factors at the same time, creating the community matrix is a little more complicated. The first command calculates the count of each genus by each sex and host combination and drops any missing values. The second command creates a community matrix.

```
> community_2 <- colony %>% count(sex,host,genus) %>% drop_na()
```

```
> comm2<-dcast(community_2, sex+host~genus, value.var = "n", fun.aggregate = sum)
```

"genus" in both command lines may be whatever taxon level you wish to evaluate in the dataset. For example, it could be changed to "family" or "order".

Calculating diversity indices

Note: "community" is the name of the community matrix

1. Species Richness

```
> diversityresult(community,index="richness",method="each site")
```

2. Simpson

```
> diversityresult(community,index="Simpson",method="each site")
```

This calculates the inverse Simpson described above

```
> diversityresult(community,index="inverseSimpson",method="each site")
```

This calculates the reciprocal Simpson described above. (confusing that it is called in the inverseSimpson)

3. Shannon

```
> diversityresult(community,index="Shannon",method="each site")
```

Calculating community similarity (distance)

Sometimes we are interested in how similar (or different) two communities are based on what species (taxa) are present and the relative abundance of those species (taxa) 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 samples,
- S_i is the total number of specimens counted in sample i ,
- S_j is the total number of specimens counted in sample j ,
- C_{ij} is the sum of only the lesser counts for each taxa found in both sites.

Although Bray-Curtis Dissimilarity is often used in community ecology, it is not robust to incomplete sampling of the community (all taxa are not sampled) or unbalanced sampling (all treatments are not equally sampled). An alternative is the Morista-Horn Index of Dissimilarity (1- C_H). Morista-Horn Index of Similarity is

$$C_H = \frac{2 \sum_{i=1}^{D_{12}} \frac{X_i}{n} \frac{Y_i}{m}}{\sum_{i=1}^{D_1} \left(\frac{X_i}{n}\right)^2 + \sum_{i=1}^{D_2} \left(\frac{Y_i}{m}\right)^2}$$

Where

- D_1 =number of taxa in sample 1
- D_2 =number of taxa in sample 2
- D_{12} =number of taxa in shared in both communities
- X_i =number of individuals of taxon i in sample 1
- Y_i =number of individuals of taxon i in sample 2
- n =total number of individuals in sample 1
- m =total number of individuals in sample 2

So that X_i/n and Y_i/m are proportion of individuals of taxon i in each of the samples (communities).

To produce a matrix of all of the pair-wise distances between samples using the Bray Curtis index of distance, use the following command.

```
> vegdist(community, method="bray", binary=FALSE, diag=FALSE, upper=FALSE)
```

To produce a matrix of all of the pair-wise distances between samples using the Morista-Horn index of distance.

```
> vegdist(community, method="horn", binary=FALSE, diag=FALSE, upper=FALSE)
```

Cited References

- Christian N, Whitaker BK, Clay K. 2015. Microbiomes: unifying animal and plant systems through the lens of community ecology theory. *Front. Microbiol.* 6:1–15.
- Cole MF, Acevedo-Gonzalez T, Gerardo NM, Harris EV, Beck CW. 2018. Effect of diet on bean beetle microbial communities. Article 3 In: McMahon K, editor. *Tested studies for laboratory teaching*. Volume 39. Proceedings of the 39th Conference of the Association for Biology Laboratory Education (ABLE).
- Costello EK, Stagaman K, Dethlefsen L, Bohannan BJM, Relman DA. 2012. The application of ecological theory toward an understanding of the human microbiome. *Science*. 336:1255–1262
- Engel P, Moran NA. 2013. The gut microbiota of insects – diversity in structure and function. *FEMS Microbiol. Rev.* 37:699-735.
- Krebs CJ. 1999. *Ecological Methodology*, 2nd edition. New York: Benjamin Cummings.
- McFall-Ngai M, Hadfield MG, Bosch TCG, Carey HV, Domazet-Lošo T, Douglas AE, Dubilier N, Eberl G, Fukami T, Gilbert SF, Hentschel U, King N, Kjelleberg S, Knoll AH, Kremer N, Mazmanian SK, Metcalf JL, Nealson K, Pierce NE, Rawls JF, Reid A, Ruby EG, Rumpho M, Sanders JG, Tautz D, Wernegreen JJ. 2013. Animals in a bacterial world, a new imperative for the life sciences. *Proc. Natl. Acad. Sci. U.S.A.* 110:3229–3236.
- The Human Microbiome Consortium. 2012. Structure, function and diversity of the healthy human microbiome. *Nature* 486: 207-214.
- Young E. 2016. *I Contain Multitudes: The Microbes Within Us and a Grand View of Life*. New York: HarperCollins Publishers.

This study is based on Blumer LS, Beck CW 2020. **Introducing community ecology and data skills with the bean beetle microbiome project**. *Advances in Biology Laboratory Education* 41.