Advances in Biology Laboratory Education
Publication of the Association for Biology Laboratory Education
Volume 41, Article 24, 2020

Introducing Community Ecology and Data Skills with the Bean Beetle Microbiome Project

Lawrence S. Blumer[1] and Christopher W. Beck[2]

[1]Morehouse College, Department of Biology, 830 Westview Drive SW, Atlanta GA 30314, USA
[2]Emory University, Department of Biology, 1510 Clifton Road, Atlanta GA 30033, USA
(**lawrence.blumer@morehouse.edu; christopher.beck@emory.edu**)

In this study, gut microbiome data from bean beetles, *Callosobruchus maculatus*, are used to perform community ecology analyses. Two microbiome community datasets are available, one based on colony phenotypes of cultured bacteria, and a second based on 16S rRNA gene sequencing of cultured bacterial colonies. We provide step-by-step instructions with computer screen shots that guide students through large dataset manipulations using MS Excel and data analyses using Excel or RStudio software. Basic community ecology variables and indices may be readily calculated to describe and compare microbiomes from this model organism.

Keywords: bean beetles, microbiome, community ecology, course-based undergraduate research experience

Introduction

Studies on microbiomes, the communities of microbes (bacteria, viruses, fungi, and archaea) living symbiotically with all metazoans, are becoming feasible for undergraduate laboratory courses. Continued improvements in DNA sequencing technologies and the simultaneous decreases in the cost for whole community sequencing of 16S rRNA genes have made it possible for students to conduct original experiments and collect data on microbial communities that would not have been accessible only a few years ago. Microbial community diversity can be based 16S rRNA sequence data or on more traditional colony phenotype analyses. In this study, we show how data our students collect may be used to perform ecological community analyses on microbiome data.

Microbiome Biology and Bean Beetles

In the past decade, both interest in and research on microbiomes, including their implications for human health (Christian, Whitaker and Clay 2015, Costello *et al.* 2012, McFall-Ngai *et al*. 2013, The Human Microbiome Consortium 2012, Young 2016), have increased dramatically. The fact that this is a very new field of study and the potential connections to individual human health and public health make microbiomes a compelling area in which students may conduct research.

An added feature of microbiomes that makes their study suitable for undergraduate students is the existence of these communities in (and on) all metazoans, including insects (Engel and Moran 2013). Furthermore, model insect species are already being used in course-based research with undergraduate students and are suitable for microbial community studies. Insect-microbe interactions have broad implications for human affairs and many insect species are every easily maintained in laboratory environments. The bean beetle, *Callosobruchus maculatus*, is a model species that is particularly suitable for course-based undergraduate research experiences (CUREs) (Beck and Blumer 2007) and very little is known about the microbiomes in this species.

The data that are collected in any microbiome study consist 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.

In this study, microbiome data from bean beetles 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 for which we have

posted datasets at www.beanbeetles.org. The colony phenotype data consist of tabulations of taxonomic units defined by unique (but limited) combinations of easily observable traits in bacteria cultured from individual microbiome extracts from bean beetles. Students picked one colony from the cultured bacteria they observed, a portion of the 16S rRNA gene of that colony was amplified using PCR and subsequently sequenced. The taxonomic units in this colony-based sequence dataset were identified by BLAST analysis of the 16S sequence. The lowest reliable taxonomic unit is at the level of genus. These 16S data are not truly community data since only one or two taxa were identified per individual beetle, but the collection of taxa from a given set of conditions constitutes an approximation of the microbiome community under those conditions.

Course Context: CUREs

The study of insect microbiome communities is an ideal field in which course-based undergraduate research may be conducted (Cole *et al*. 2018). An instructor may choose to have students collect their own microbiome data or download and analyze those data freely available at www.beanbeetles.org. Similarly, an instructor may choose to have students evaluate one or both datasets (i.e., colony phenotypes and colony-based 16S sequences). The datasets and their analyses are not contingent on each other. In Notes for the Instructor, we describe how the data analyses may be streamlined to best meet the needs of your course and your student learning goals.

Student Outline

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.

Microbial Community Analysis Using Colony Phenotype Database

Questions

Using data from the colony phenotype database and the analyses described below, answer the following 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 species (taxon) diversity that you use?
3. Is there a relationship between number of samples and taxonomic diversity? If so, what might explain this?
4. Which communities are most similar (different)?

Database description

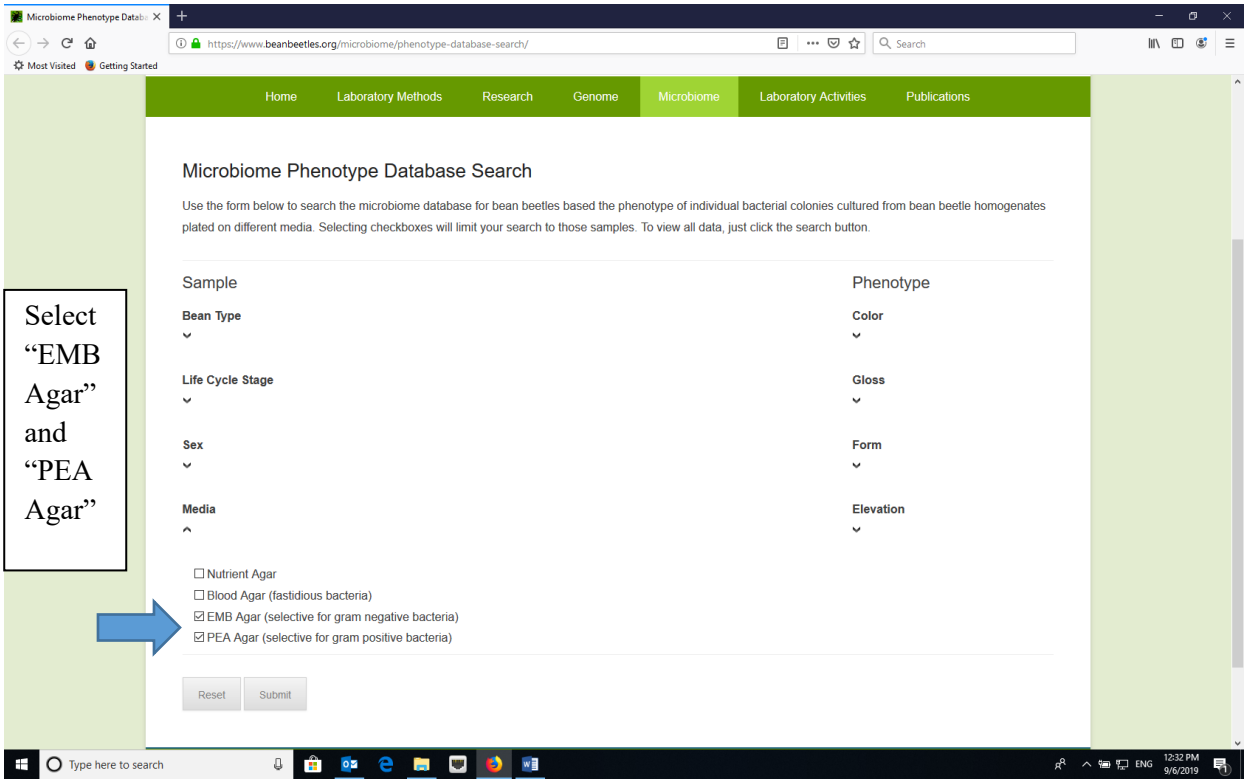
This database contains data for the microbial community of bean beetles based on colony phenotypes of individual bacterial colonies cultured from bean beetle homogenates plated on different media. The bacteria cultured from each beetle are represented by multiple rows of data with each row representing a different combination of phenotypic characters. A unique combination of phenotypic characters is taken to represent a unique bacterial taxon.

Access the database at https://www.beanbeetles.org/microbiome/phenotype-database-search/.

The database allows you to limit your search by bean host species, sex, life cycle stage, media on which bacteria were grown, and colony phenotype. EMB and PEA plates select for gram negative and gram positive bacteria, respectively.

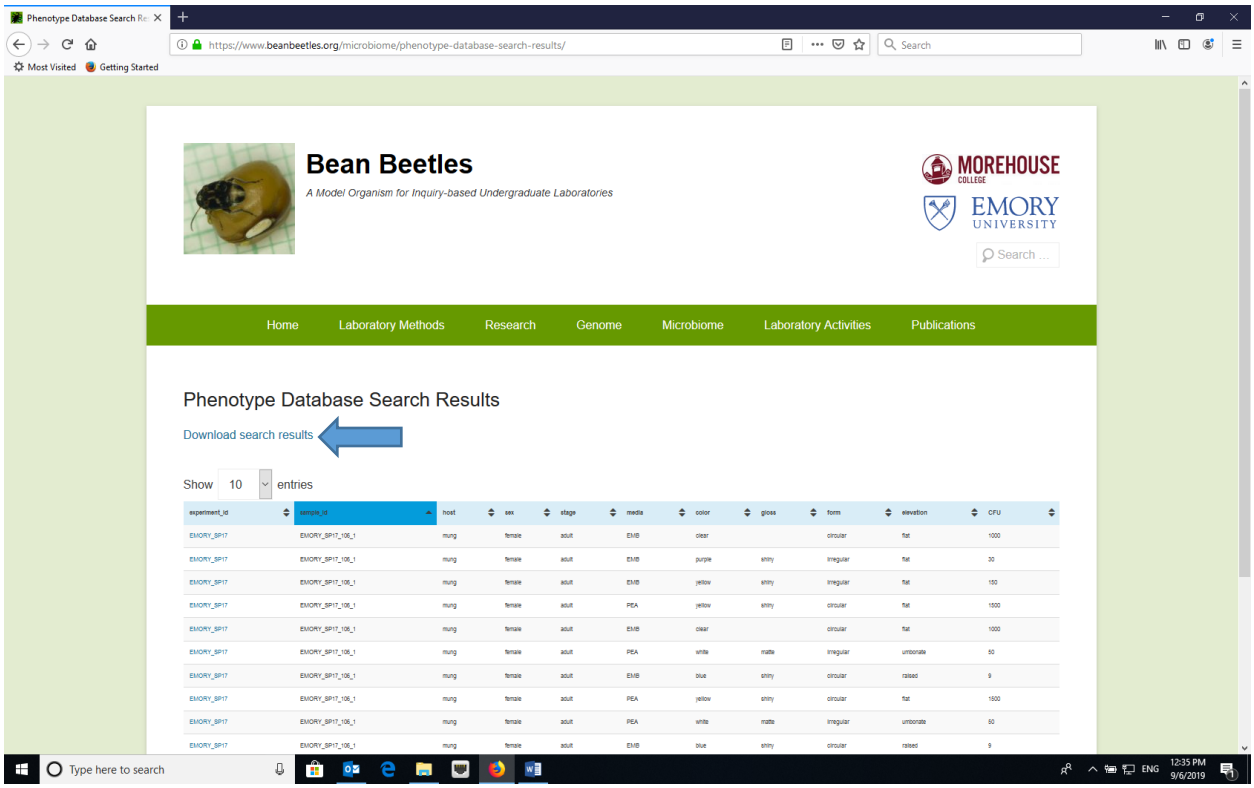
So, a colony with a particular combination of phenotypic characters on an EMB plate is a different bacterial taxon than a colony with the same combination of characters on a PEA plate. The same is not true if we include blood agar or nutrient agar plates in our sample. Those plates could be consider independently.

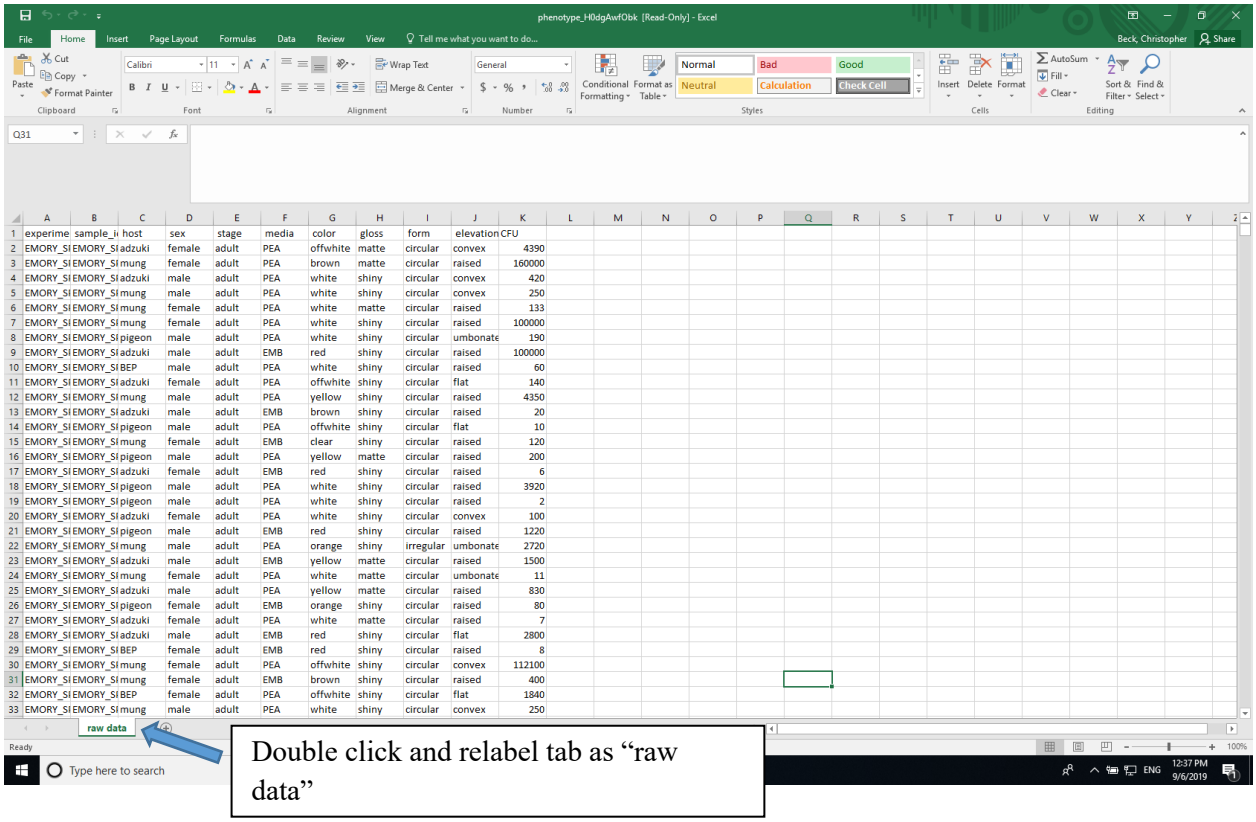
We want to limit our search to PEA and EMB plates (or Nutrient Agar only). After limiting by media type, click "Submit."



Downloading Data

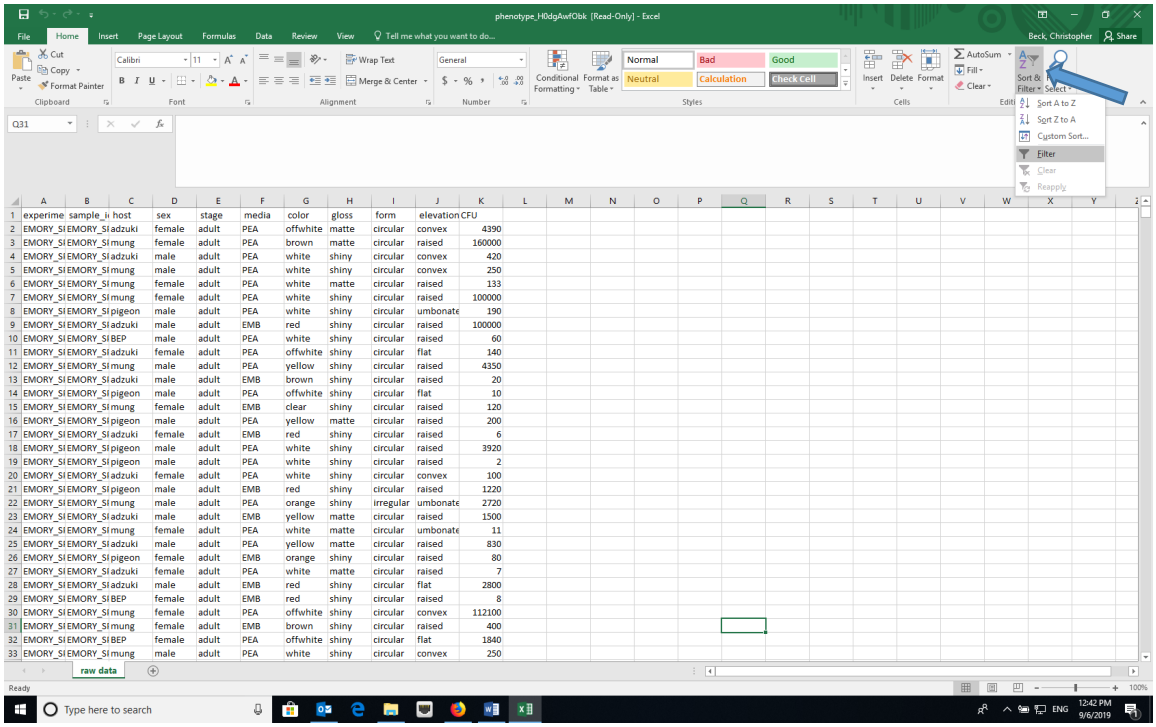
While we can view the data on the website, we want to download the data to manipulate. Click the download link to download a csv file with the data. Then, re-save the file as an Excel file and rename the tab "raw data."


Bean Beetles
A Model Organism for Inquiry-based Undergraduate Laboratories
MOREHOUSE COLLEGE
EMORY UNIVERSITY
Home
Laboratory Methods
Research
Genome
Microbiome
Laboratory Activities
Publications
Phenotype Database Search Results
Download search results
Show 10 entries

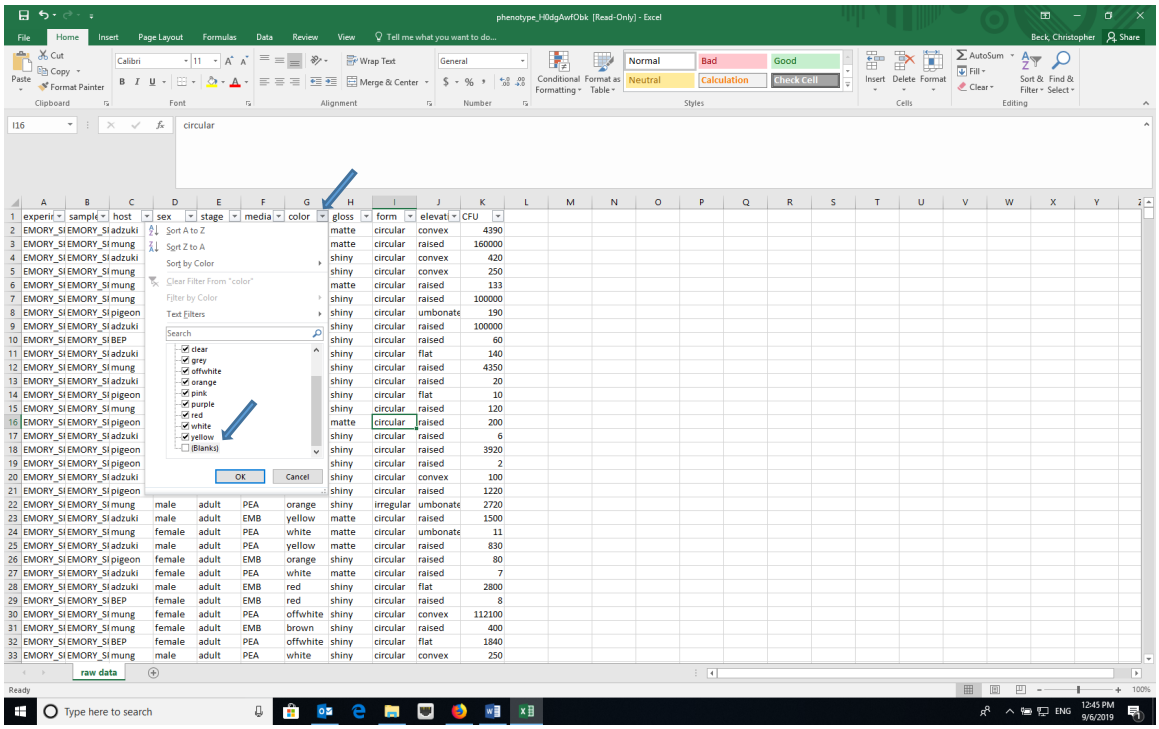

raw data
Double click and relabel tab as “raw data”

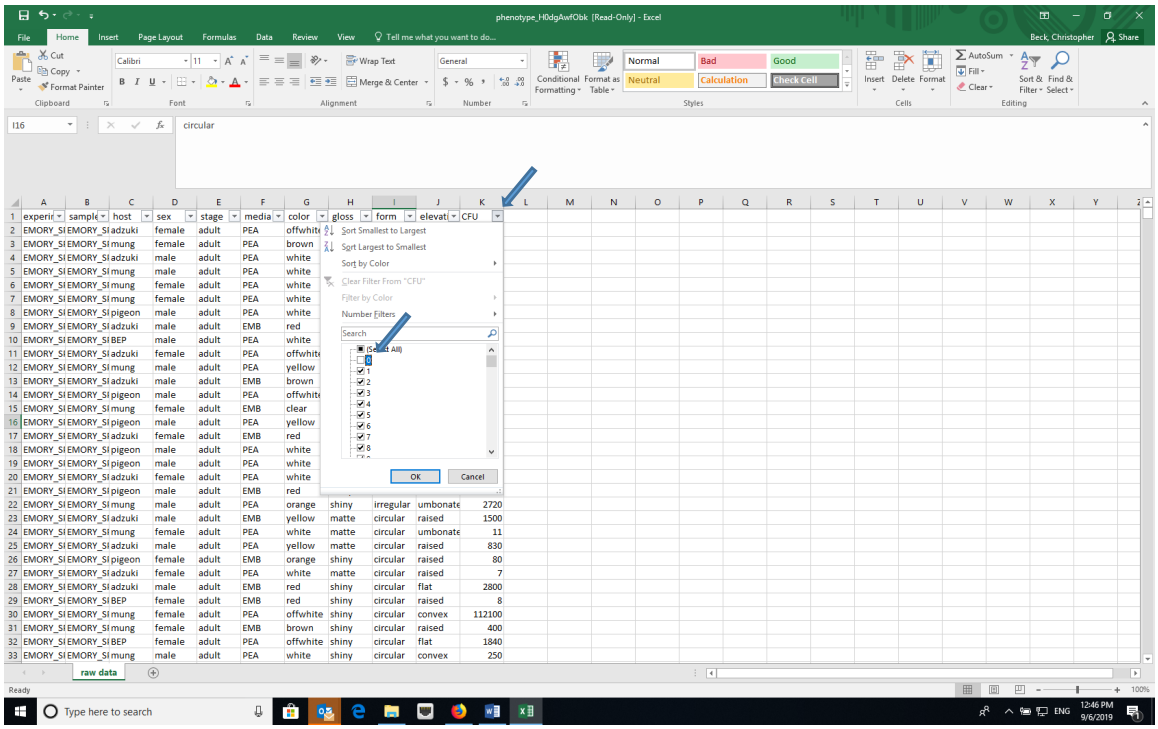
Data manipulation

1. The raw data have some missing values for phenotypic characters or values of zero for CFU (colony forming units) (which represents missing data for CFU). These rows need to be deleted. The easiest way to do this is with the Filter function in Excel. In the "raw data" worksheet, turn on filtering by selecting Filter under Sort and Filter.



Then, for each of the phenotypic characters, click the down arrow in the column heading and unselect "Blank" at the bottom of the list and unselect "0" in the CFU column.

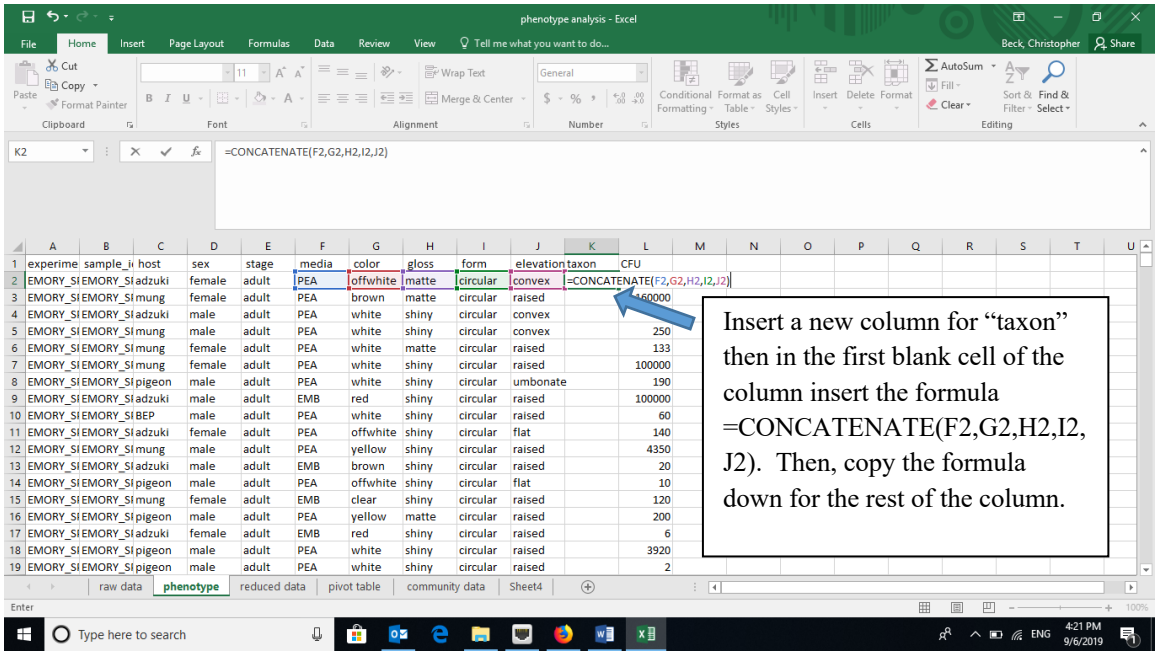




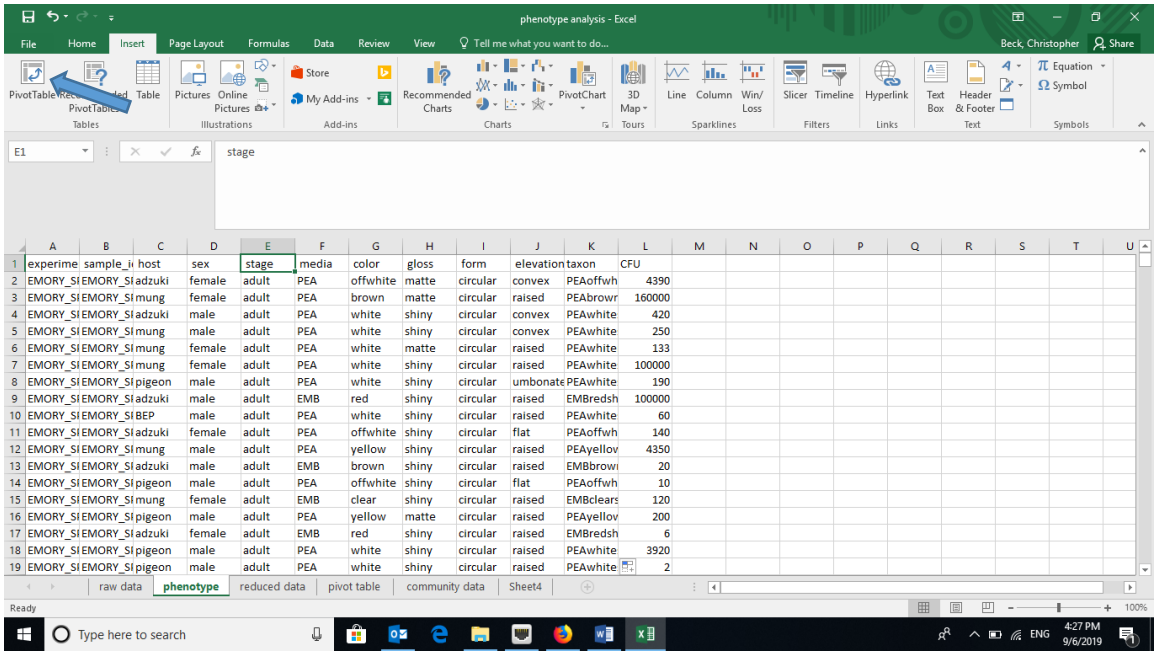
Next, select all (CRTL-A or CMD-A) and copy and paste into a new worksheet and name that sheet "phenotype". The "phenotype" worksheet represents the cleaned raw data. If you will be analyzing the data in R later, copy and paste a second time into another blank sheet and name that sheet "community".

2. With our new "phenotype" dataset, we need to define a bacterial "taxon" based on the combination of media and the four phenotypic characters (color, gloss, form and elevation). One way to do this is to create a "taxon" name by

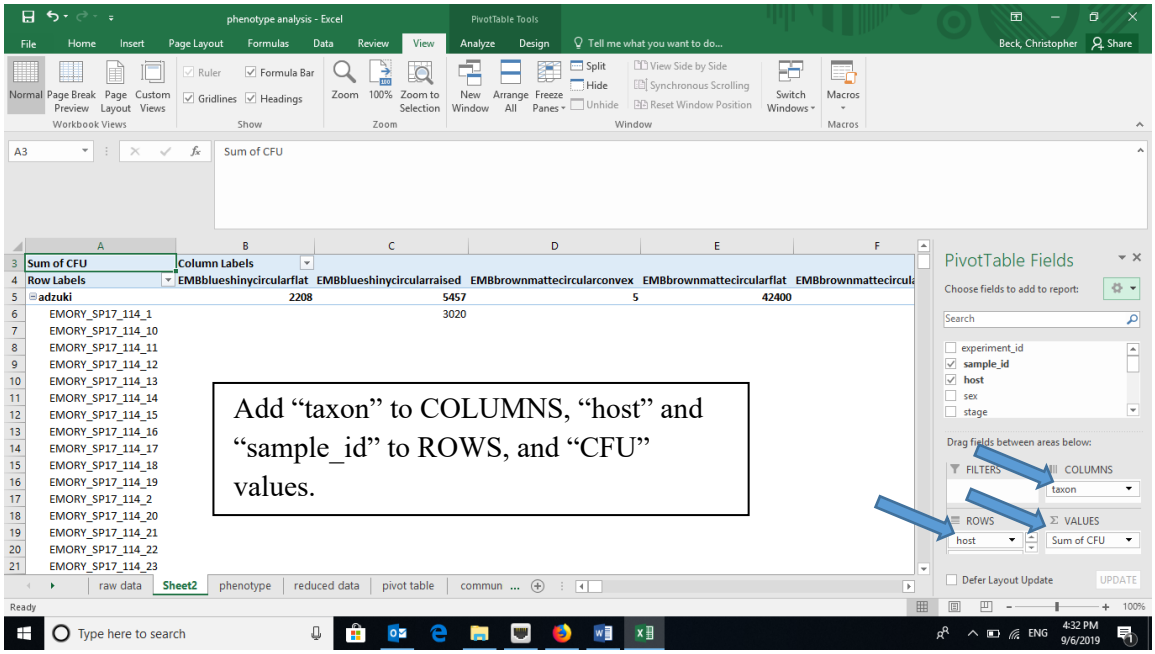
concatenating the media and the four different phenotypic traits. You can do this using the CONCATENATE function in Excel (=CONCATENATE(F2,G2,H2,I2,J2). After you create the "taxa" names, you might want to select the column and then re-paste it in the same column by pasting values (using paste special) to remove the formula.



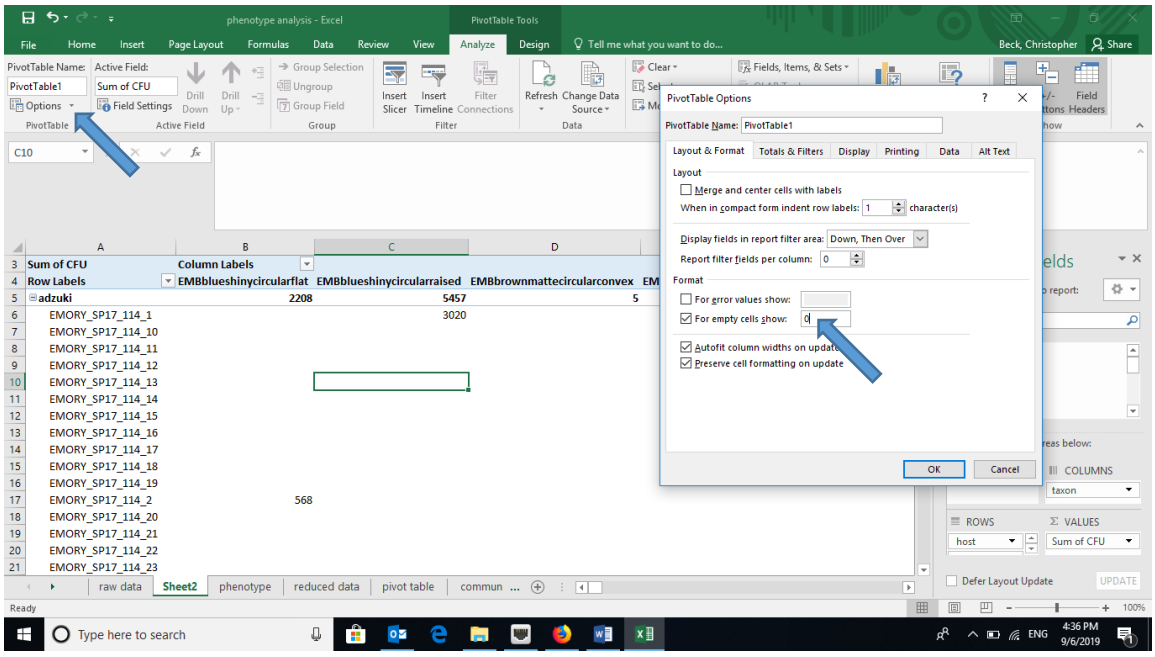
3. Now, we need to consolidate the data for each host species, each sex, or the combination of the two by the bacterial taxa. The easiest way to do this is with the Pivot Table function in Excel.
4. When clicked on a cell within the data, create a Pivot Table (Insert -> Pivot Table OR Data -> Summarize with Pivot Table, depending on your version of Excel). Make sure that the data source includes the top row, which has the column headings.



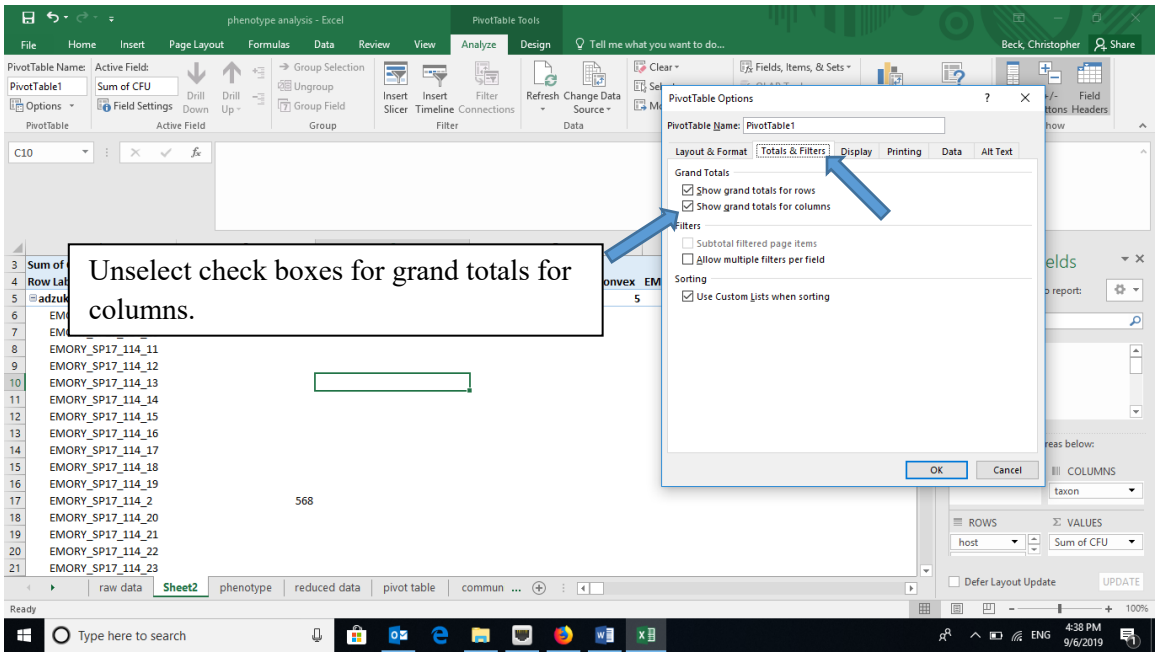
Set the host and sample_id as the rows as these represent the treatment and individual communities, respectively. The new "taxon" column should be the columns in the pivot table. The Values should be a SUM of the CFU (colony-forming units, a measure of density), which will be shown as "SUM of CFU" using the Options menu.



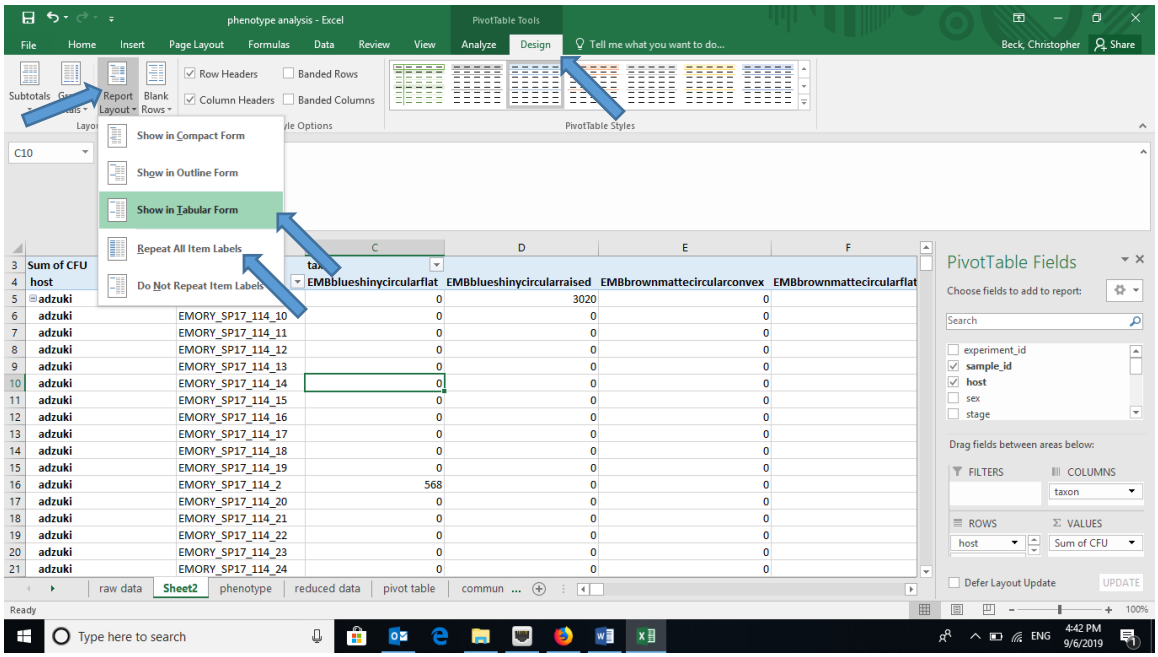
5. You can add zeros to all of the empty cells in the Pivot Table using the Options menu.



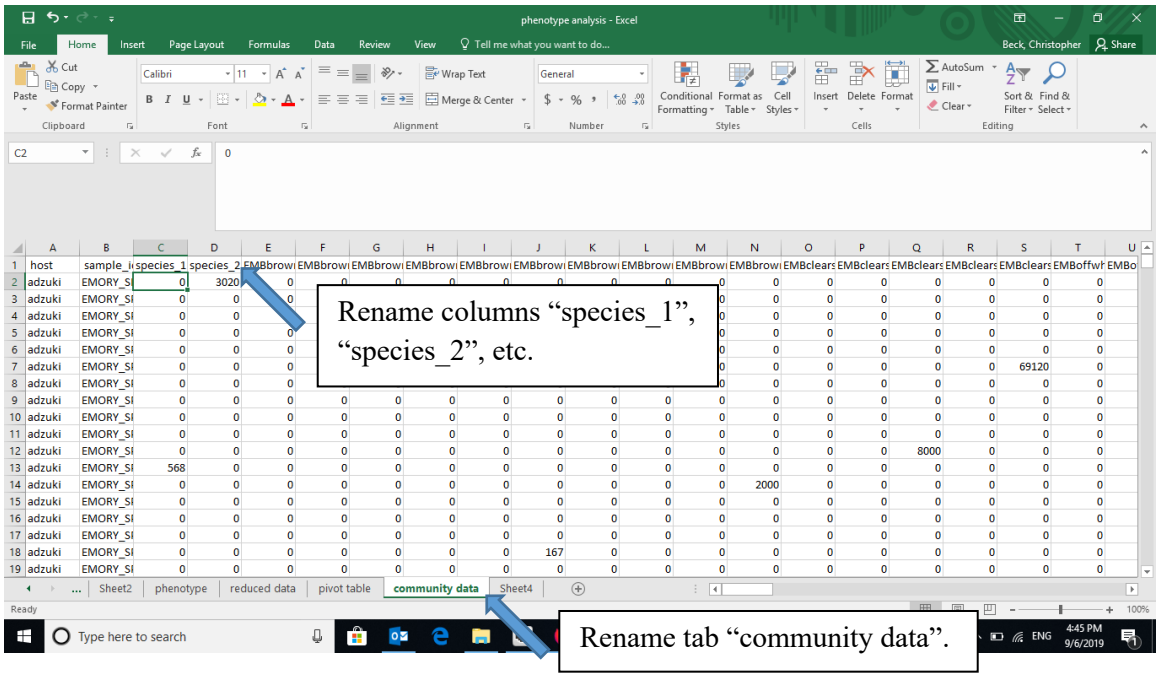
6. Remove the Grand totals for Columns using the Options menu or the Design tab depending of your version of Excel.



7. To 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”.

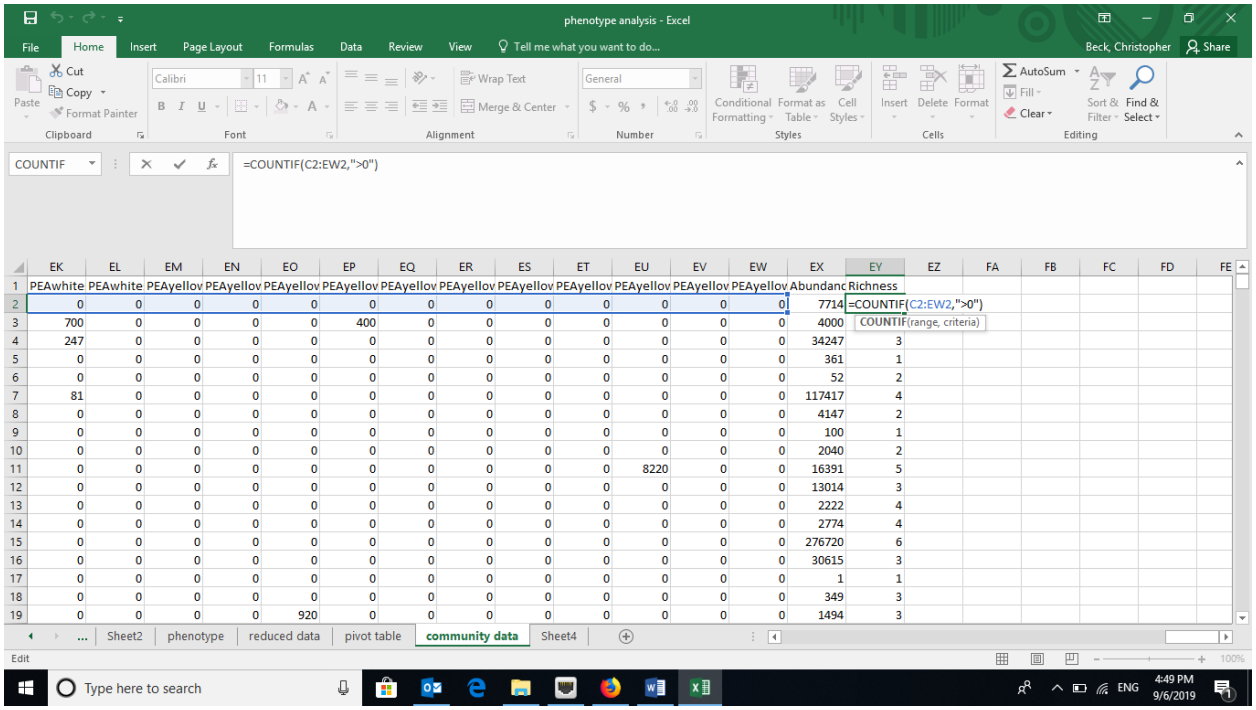


8. Copy and paste (as values) the pivot table to a new worksheet and remove any extra rows at the top. Each of the columns in this new worksheet represents a unique bacterial taxon. The exact phenotype doesn’t matter, so we are going to rename them species_1, species_2, …. Title this worksheet tab “community data”.

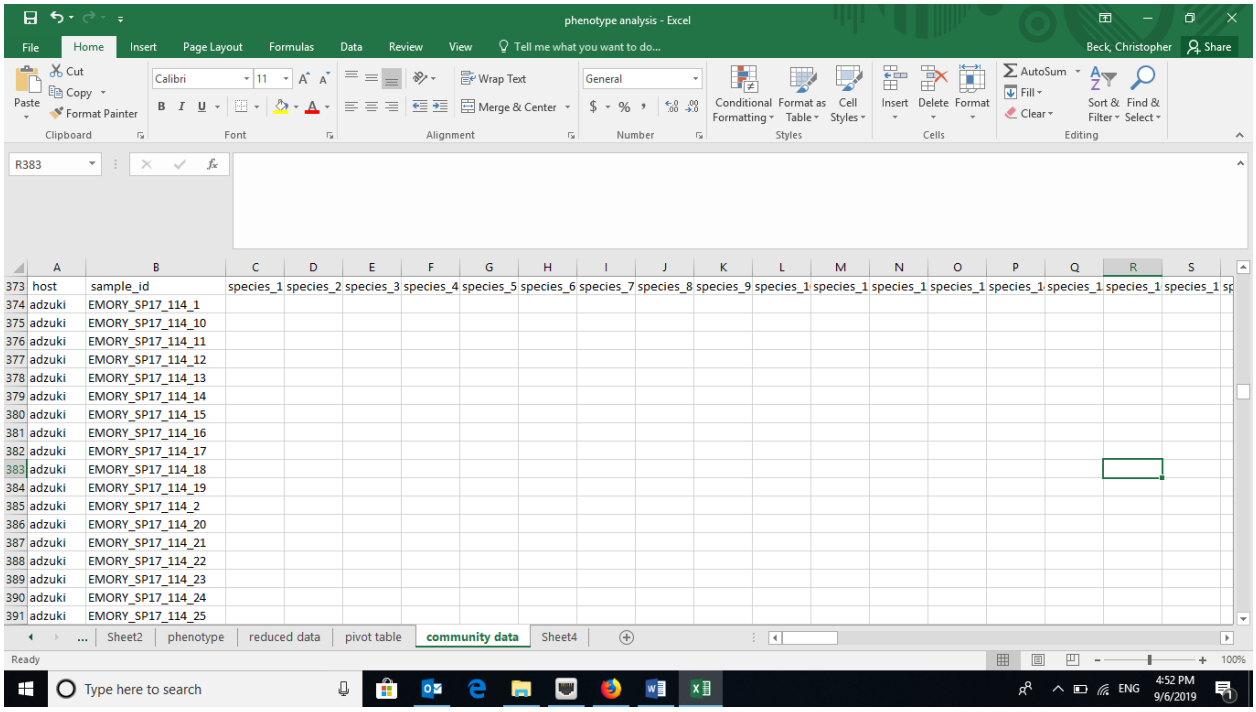


Calculating diversity indices

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” the is range of cells in the spreadsheet containing the data, such as “C2:EW2”).

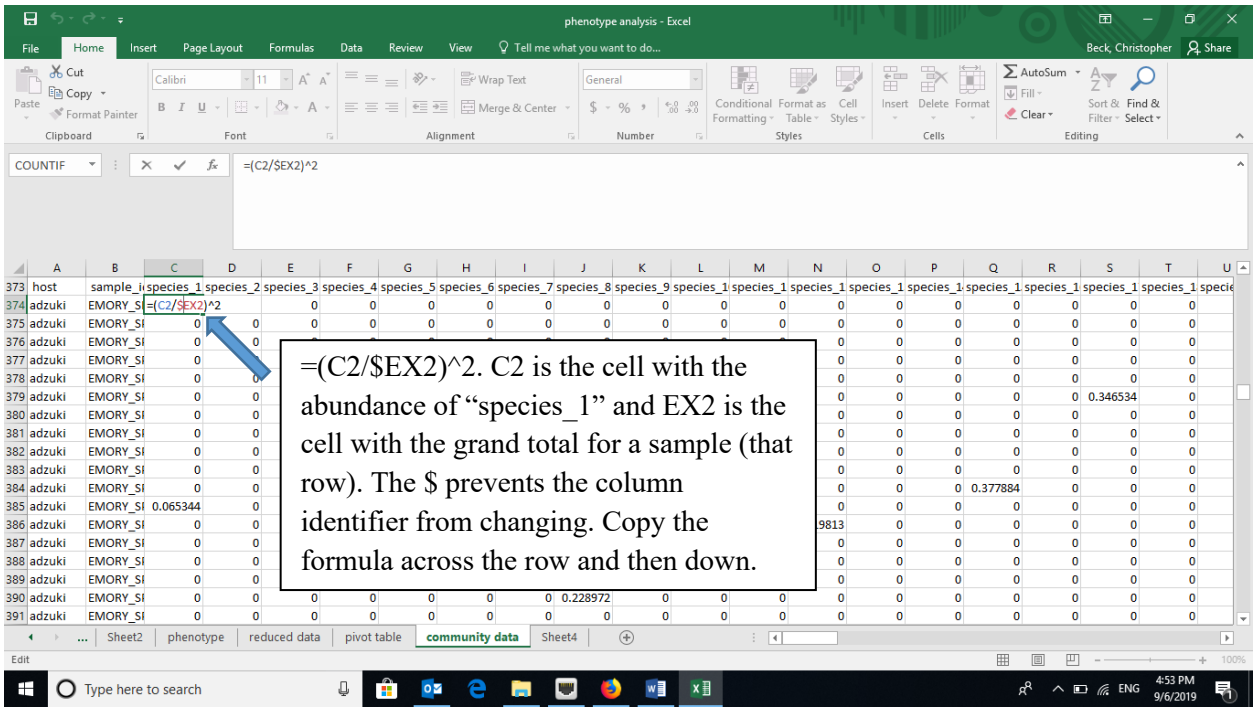


2. Simpson Index – the Simpson Index incorporates both species (taxon) richness and species (taxon) evenness.
    a. D = Σ(n/N)^2, where n=number of individuals of a particular species (taxon) and N=total number of individuals in a sample. D increases as diversity decreases, which is counterintuitive.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

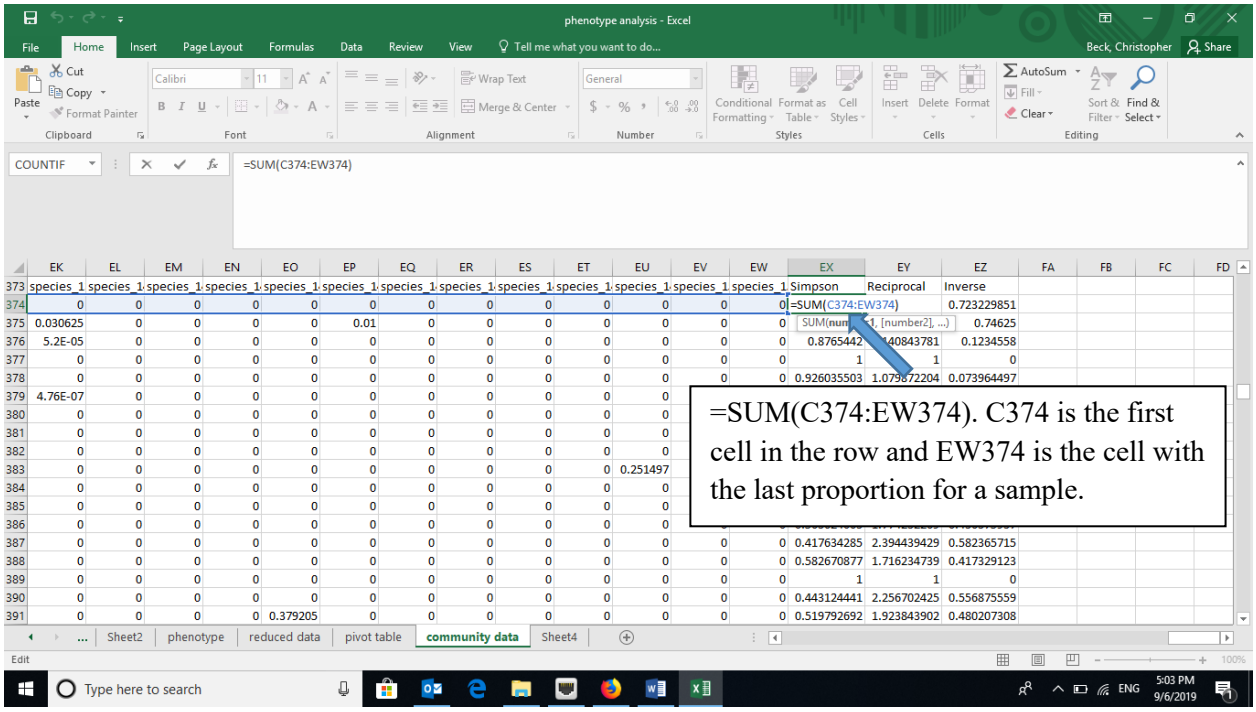


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
373	host	sample_id	species_1	species_2	species_3	species_4	species_5	species_6	species_7	species_8	species_9	species_1	species_1	species_1	species_1	species_1	species_1	species_1	species_1
374	adzuki	EMORY_SP17_114_1																	
375	adzuki	EMORY_SP17_114_10																	
376	adzuki	EMORY_SP17_114_11																	
377	adzuki	EMORY_SP17_114_12																	
378	adzuki	EMORY_SP17_114_13																	
379	adzuki	EMORY_SP17_114_14																	
380	adzuki	EMORY_SP17_114_15																	
381	adzuki	EMORY_SP17_114_16																	
382	adzuki	EMORY_SP17_114_17																	
383	adzuki	EMORY_SP17_114_18																	
384	adzuki	EMORY_SP17_114_19																	
385	adzuki	EMORY_SP17_114_2																	
386	adzuki	EMORY_SP17_114_20																	
387	adzuki	EMORY_SP17_114_21																	
388	adzuki	EMORY_SP17_114_22																	
389	adzuki	EMORY_SP17_114_23																	
390	adzuki	EMORY_SP17_114_24																	
391	adzuki	EMORY_SP17_114_25																	

    e. To calculate the proportion squared for each taxon, 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/$EX2)^2. Copy the formula across the row and then down.

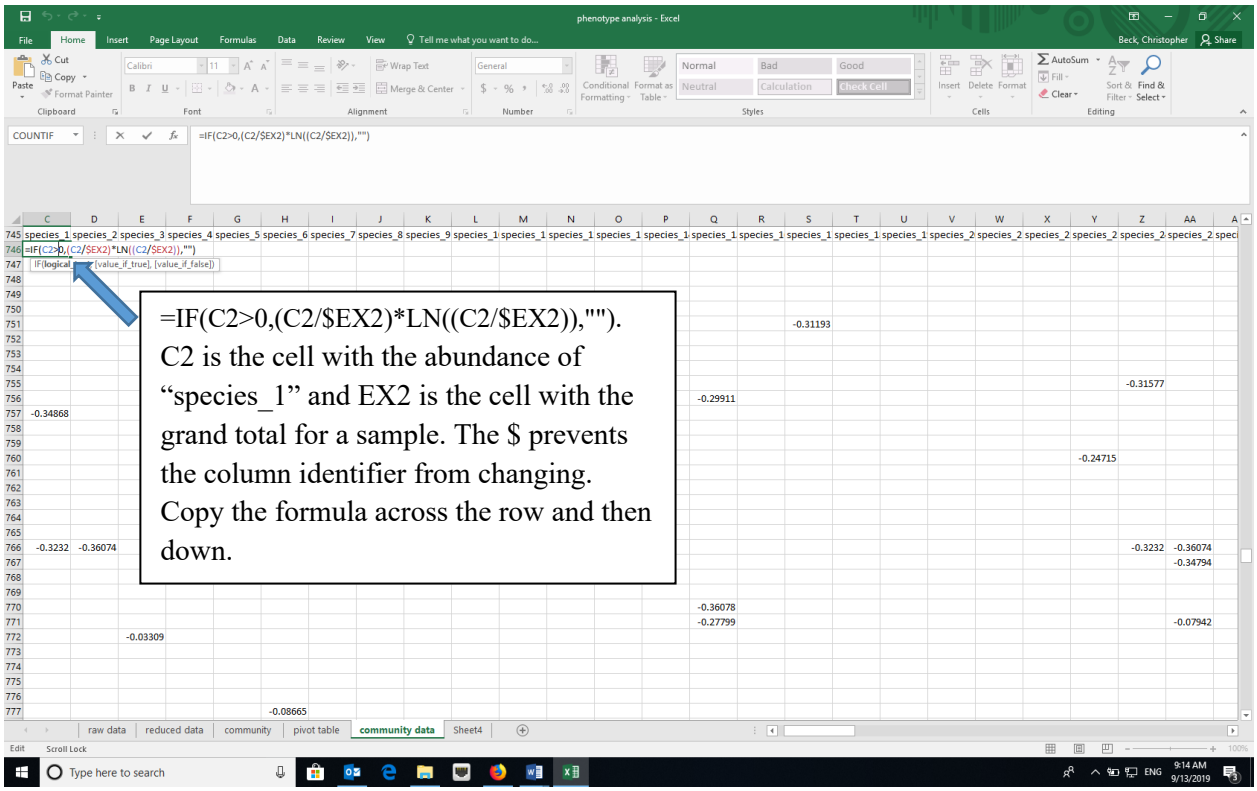


f. Calculate the sum of the proportions squared (=SUM in Excel for each row, a different microbial community) to calculate the Simpson Index.

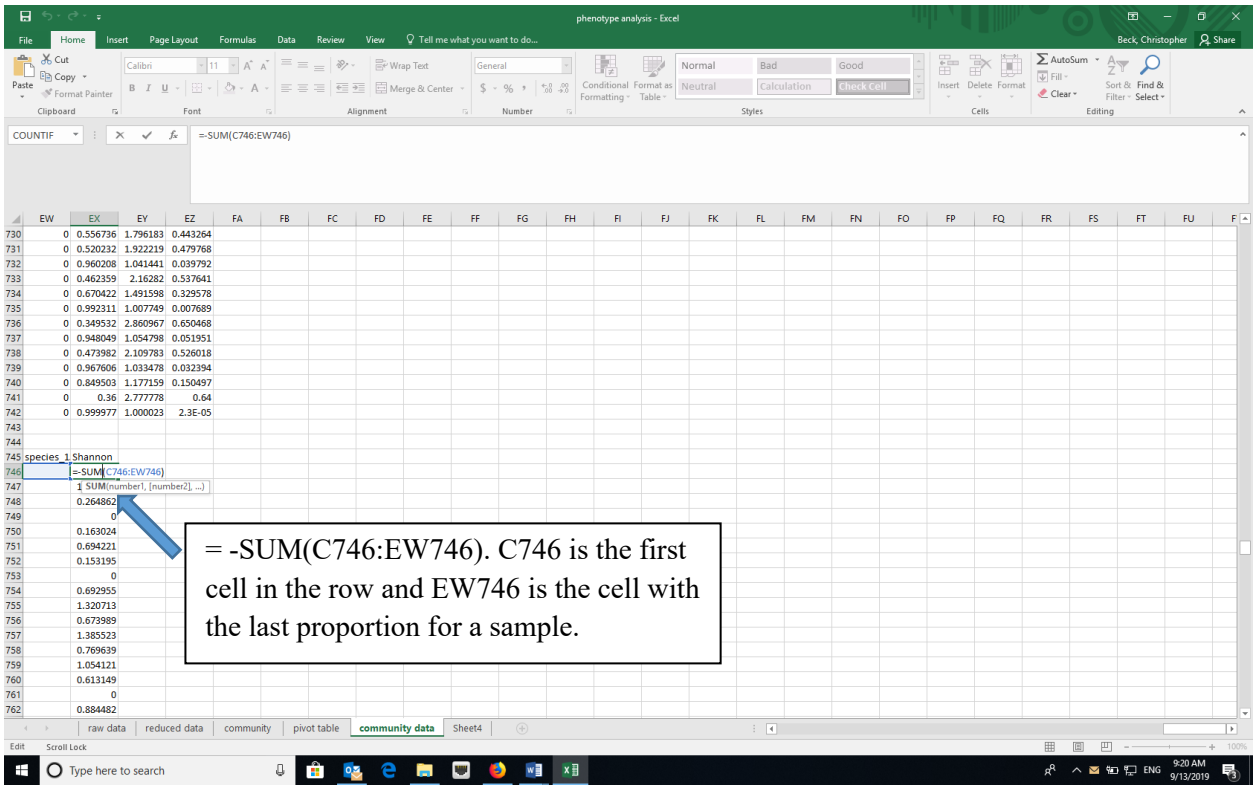


g. Calculate the reciprocal (e.g., =1/EX374) and inverse Simpson (e.g., =1-EX374) using formulas in Excel.

3. Shannon-Weaver (Shannon-Weiner) Index – also incorporates species (taxon) richness and species (taxon) evenness
    a. H=-Σ*p*ln*p*, where p is the proportion of individuals of each species (taxon) in a community (i.e., n/N).
    b. Create a new data array below the original using the same row labels (treatment variables) and the same column labels (species).
    c. Using the grand totals for each community, 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.
    d. Note that ln*p* *is* undefined if *p*=0, so you can use an "IF" statement in Excel. For example, =IF(C2>0,(C2/$EX2)*LN((C2/$EX2)),"")

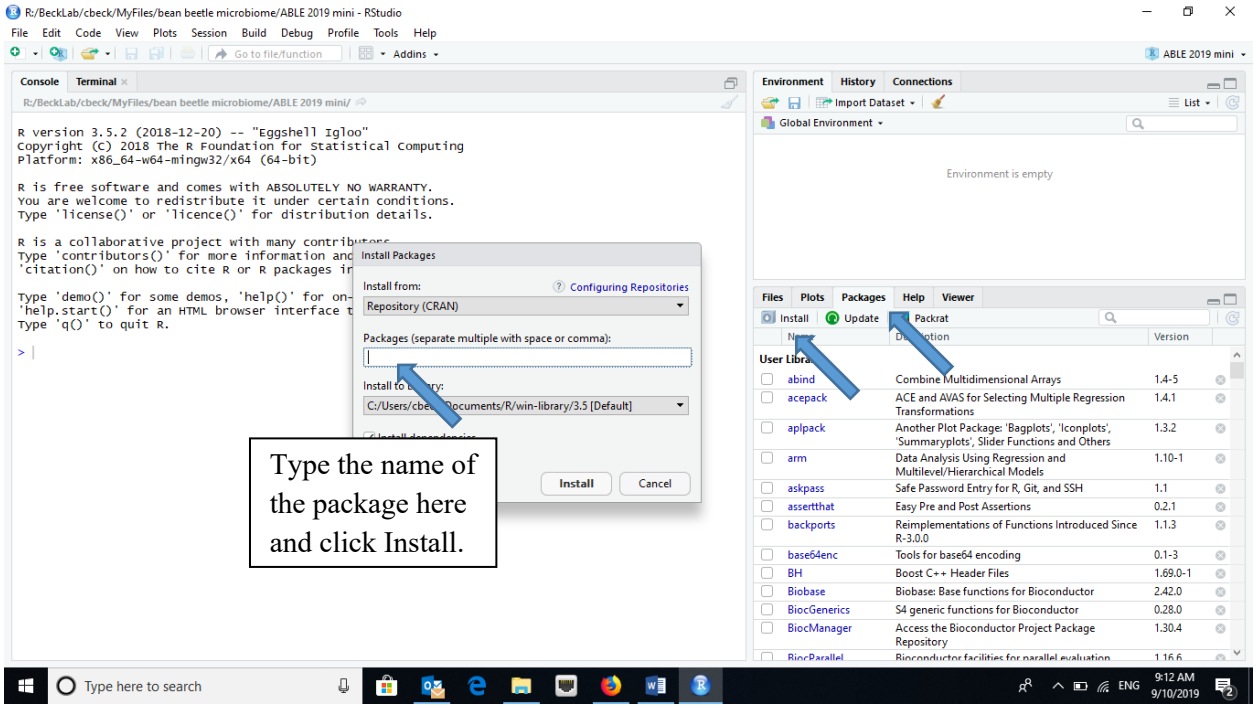


    e. Calculate the negative sum of the proportions (*p*ln*p*) (=-SUM in Excel for each row, a different microbial community) to calculate the Shannon-Weaver Index.

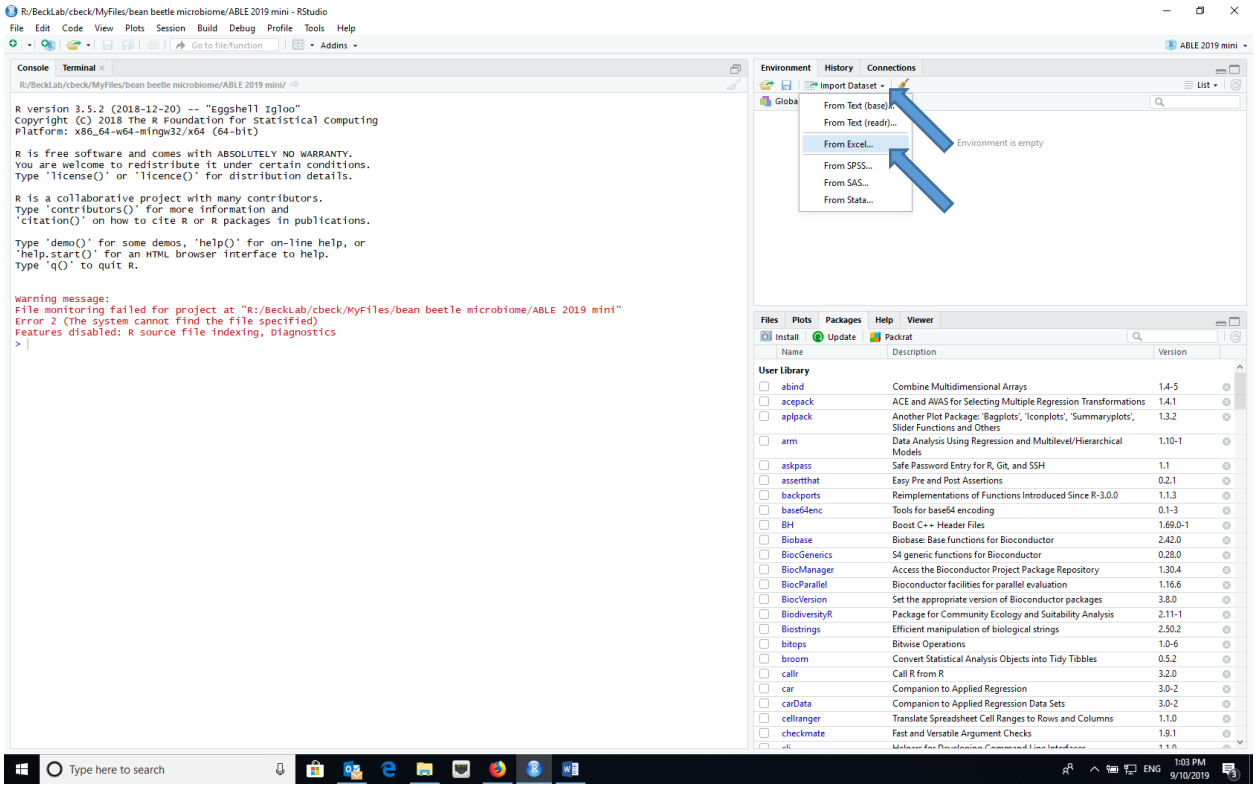


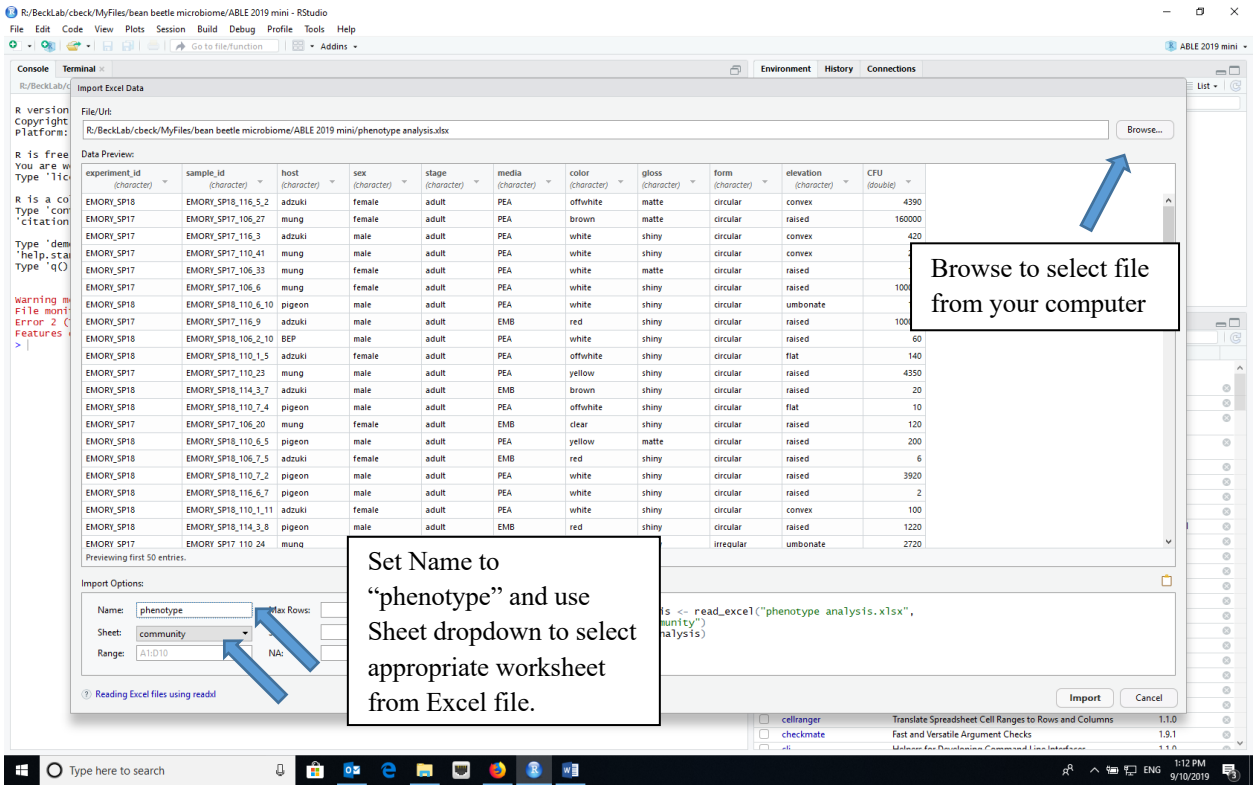
Data Manipulation in R

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 ("community" that you created above in the Excel section) into RStudio.





5. Attach the imported dataset to the dataframe using the attach command in the console (`attach(phenotype)`)

6. Create a community matrix for each sample. Since each "taxon" is defined by the combination of media type and the four phenotypic characters, we can use the `dcast` function in the `reshape2` library, using the following command.

```
> comm_pheno=dcast(phenotype,sample_id~media+color+gloss+form+elevation,valu
e.var = "CFU",fun.aggregate = sum)
```

7. The first column in the resulting data table is the sample_id. The sample_id needs to become the row name and then deleted, using the following commands.

```
> row.names(comm_pheno)<-(comm_pheno$sample_id)
> comm_pheno<-comm_pheno[,-1]
```

8. Now, we need to create an environment matrix with the sample metadata. First, we create a data table with the sample names and sample metadata, which are in the second through fourth columns of the "phenotype" dataframe. Second, we remove all of the duplicate values for the sample_id using the `distinct` function in the `dplyr` package. Third, we need the sample_id to become the row name. (Note that this causes an error message and for some reason if you delete the column with the row name, the row names disappear. In this case, we do not need to delete the sample_id column because we can specify the factor of interest.) Last, we need to set host and sex as factors.

```
> pheno_meta<-phenotype[,2:4]
> pheno_meta<-pheno_meta %>% distinct(sample_id, .keep_all = TRUE)
> row.names(pheno_meta)<-(pheno_meta$sample_id)
> pheno_meta$host<-as.factor(pheno_meta$host)
> pheno_meta$sex<-as.factor(pheno_meta$sex)
```

You can ignore the error message about setting row names on a tibble being deprecated.

Species accumulation curves

Species accumulation curves are often used to visualize whether all of the taxa in a community have been sampled. As the number of samples increases, the total number of unique species sampled should increase. However, the relationship between the number of samples and the number of unique species should asymptote. If so, we can say that we have sampled the majority of species in the community. However, if the slope of the relationship is steep, this suggests that the community is incompletely sampled.

For each treatment separately, create the data for the species accumulation curve using:

```
> Accum.label1<-specaccum(comm_pheno,method='exact',permutations = 100,
conditioned=TRUE,gamma='jack1',w=NULL,
subset=pheno_meta$factor_variable=="factor")
```

`factor_variable=="factor"` refers to the factor being evaluated such as host in our dataset, and the "factor" is one state of that variable. For example, `host=="mung"` would do a species accumulation curve for microbiome communities of beetles that emerged from mung beans. Note that two equals signs are necessary.

You need to run the command above for each treatment group separately. Change the label and factor terms appropriately for additional analyses.

a. Plot the first species accumulation curve using:

```
> plot(Accum.label1,col="red")
```

b. Plot each additional species accumulation curve using:

```
> plot(Accum.label2, add=TRUE, col="blue")
```

If the second curve extends beyond the y-axis, replot the curves in the opposite order (i.e., plot curve 2 first and then curve 1).

Calculating Diversity Indices

You can calculate the diversity indices described above in the Excel exercise using functions in the `BiodiversityR` package.

1. Species Richness

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

2. Simpson

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

This calculates the inverse Simpson described above.

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

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

3. Shannon

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

4. To calculate all of the biodiversity indices and merge them with the metadata for future analysis. You can then use this dataset for analysis of differences between treatments with t-tests, ANOVAs, or their non-parametric equivalents.

```
> pheno_diversity<-diversityvariables(comm_pheno,pheno_meta)
```

Because the `diversityvariables` function calculates a range of diversity indices, sometimes an error occurs because a particular index cannot be calculated with the dataset (e.g., requires taking log of zero). If this is the case, you can use the `diversityresult` function above and place the results in a dataframe. Then, you can combine the dataframes with the metadata using the `cbind` function.

```
> Simpson<-diversityresult(comm_pheno,index="Simpson",method="each
site")
> Shannon<-diversityresult(comm_pheno,index="Shannon",method="each
site")
> pheno_diversity<-cbind(pheno_meta,Simpson,Shannon)
```

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.

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

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(comm_pheno, 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(comm_pheno, method="horn", binary=FALSE, diag=FALSE, upper=FALSE)
```

How different (similar) are the communities?

Adonis is an approach to testing whether communities differ based on a treatment. It is the equivalent of an analysis of variance, but comparing distance matrices. "community_adonis" stores the results of the analysis, "comm_pheno" is the community matrix, "factor_variable" is the treatment (e.g., host), and "pheno_meta" is the name of the dataset that has the treatment data for each community. In the code below, we use Morista-Horn to estimate distance between communities.

```
> community_adonis<-adonis2(comm_pheno ~ factor_variable, data = pheno_meta,
method="horn")
> community_adonis
```

Microbial Community Analysis Using Colony-based Sequencing Database

Questions

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

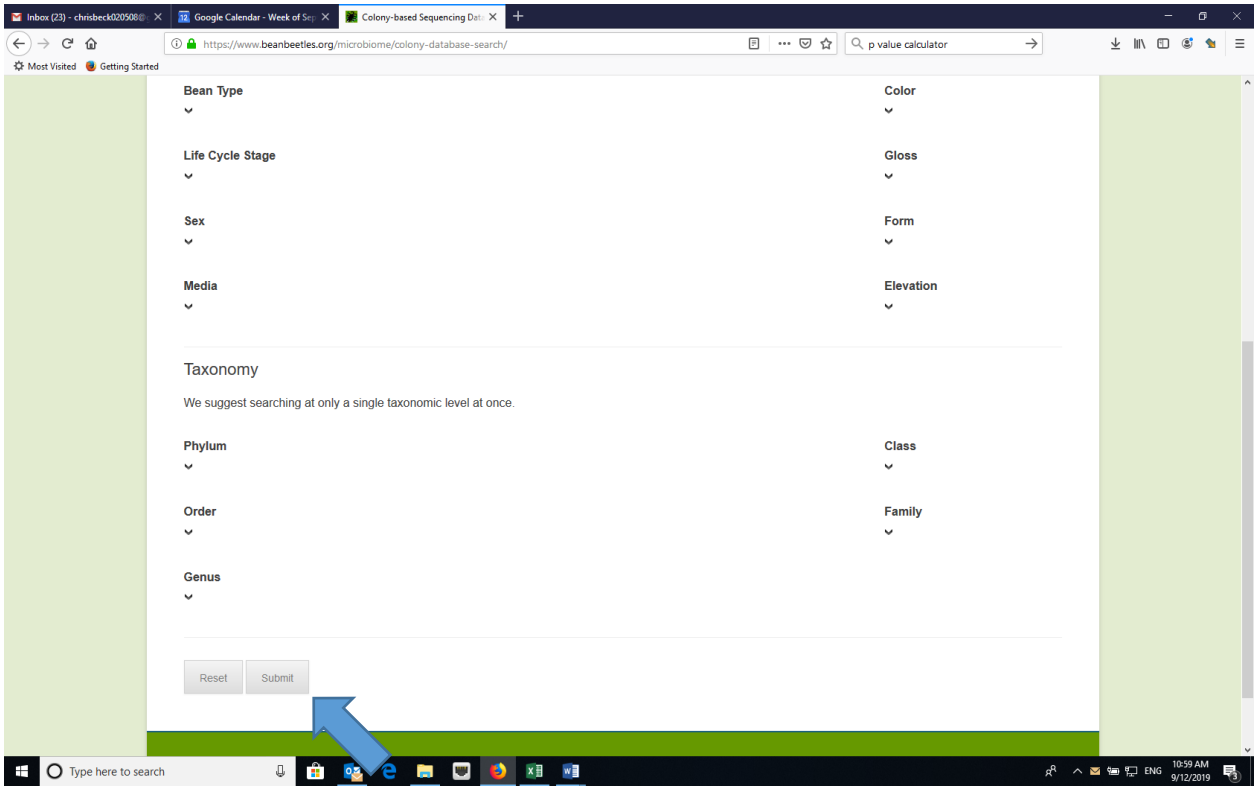
1. Which taxa are most prevalent in the bacterial communities in bean beetles?
2. Do the most prevalent taxa vary based on host bean type?
3. Based on the diversity indices that you calculated, which treatment had the highest (lowest) diversity?
4. Does the answer depend on the measure of species (taxon) 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.

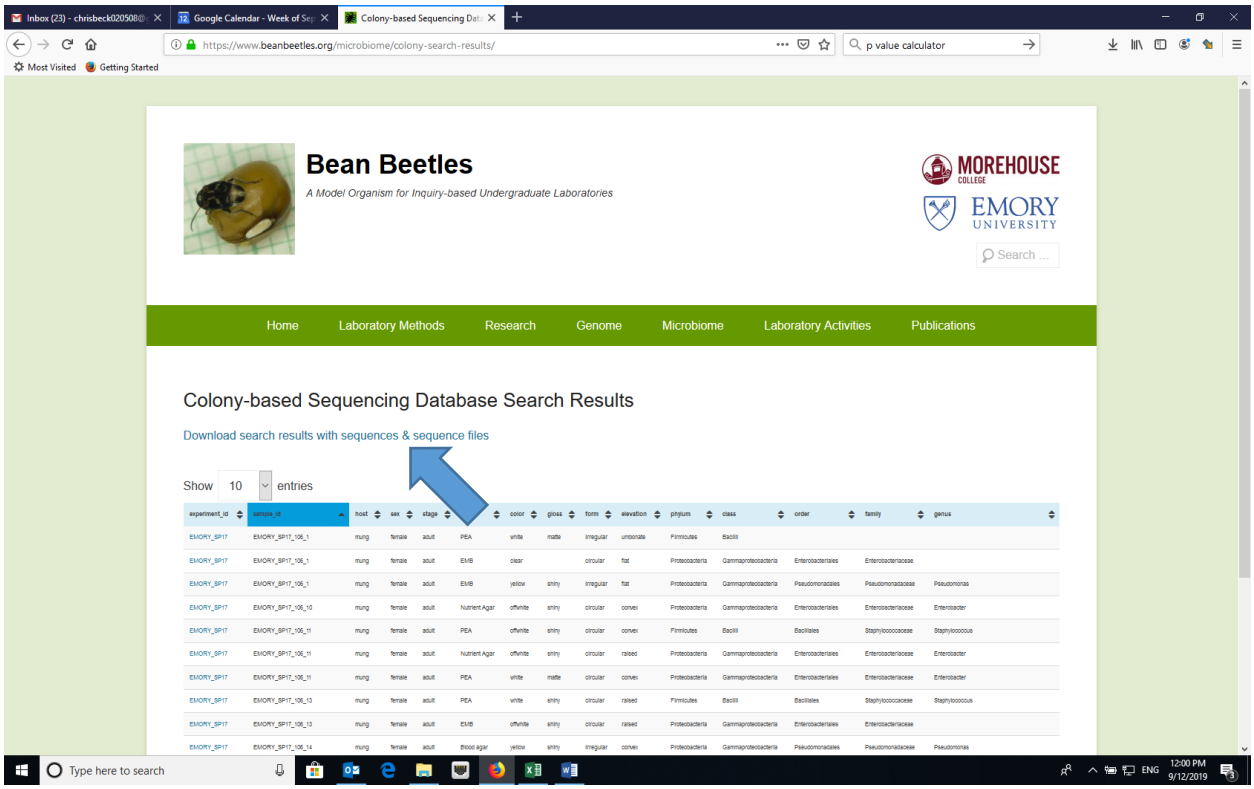
Access the database at https://www.beanbeetles.org/microbiome/colony-database-search/.

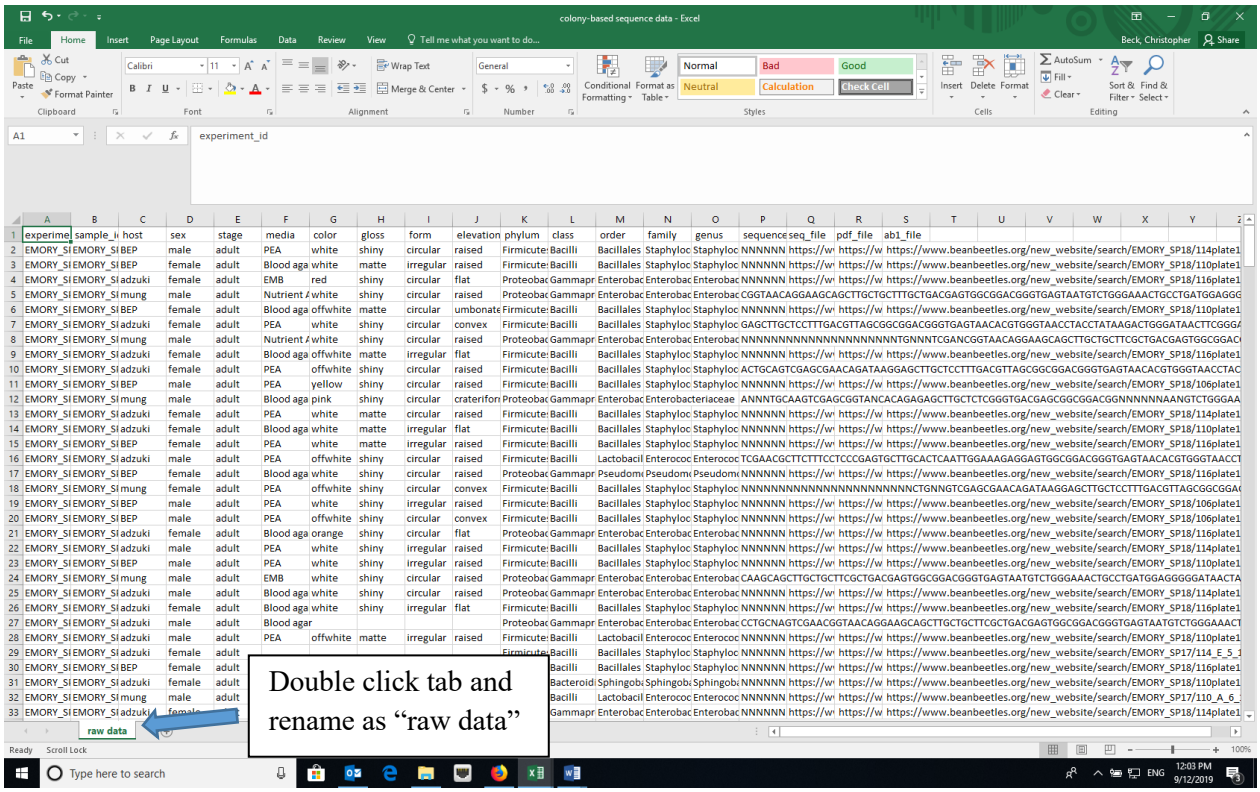
The database allows you to limit your search by bean host type, sex, life cycle stage, media on which bacteria were grown, colony phenotype, and bacterial taxonomy. Since we are interested in making comparisons between bacterial communities based on host species and sex, we want to download the entire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



Downloading Data

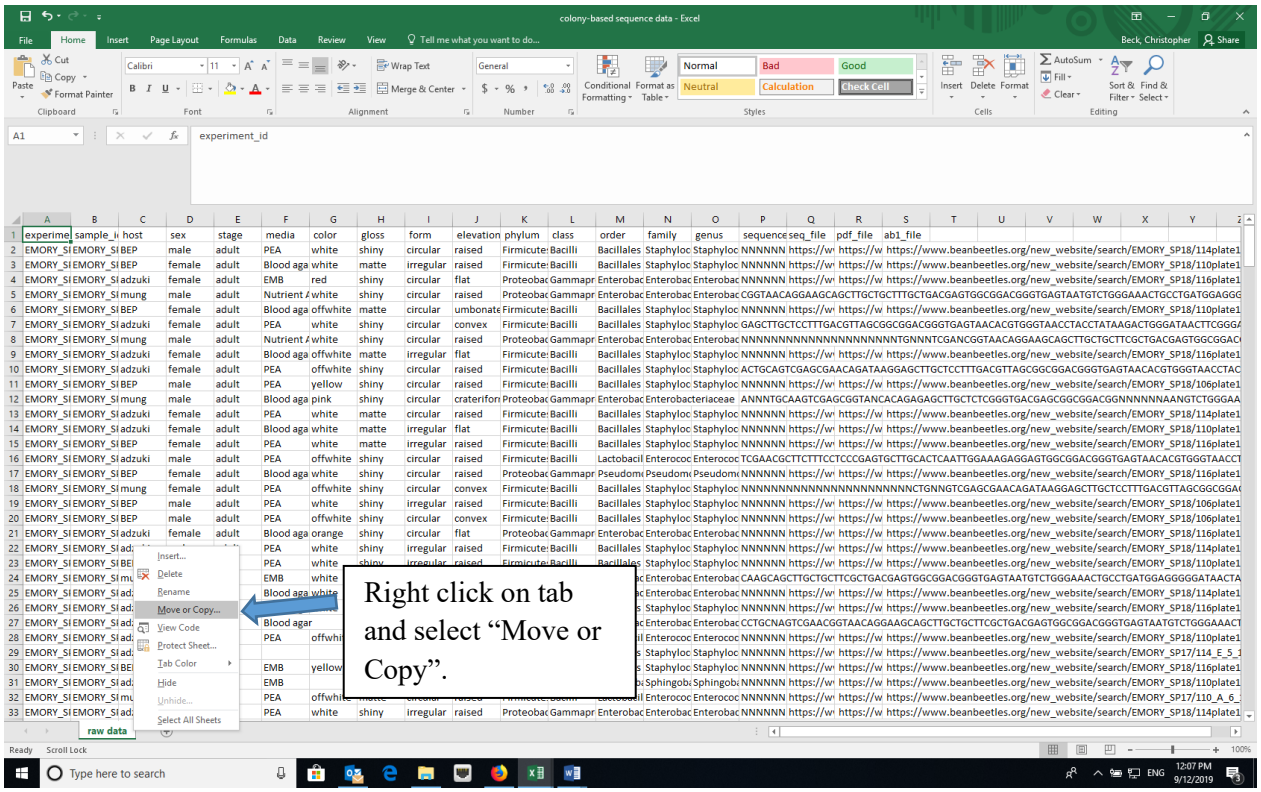
While we can view the data on the website, we want to download the data to manipulate. Click the download link to download a csv file with the data. Then, save the file as an Excel file (name the file "colony-based sequence data") and rename the tab "raw data."

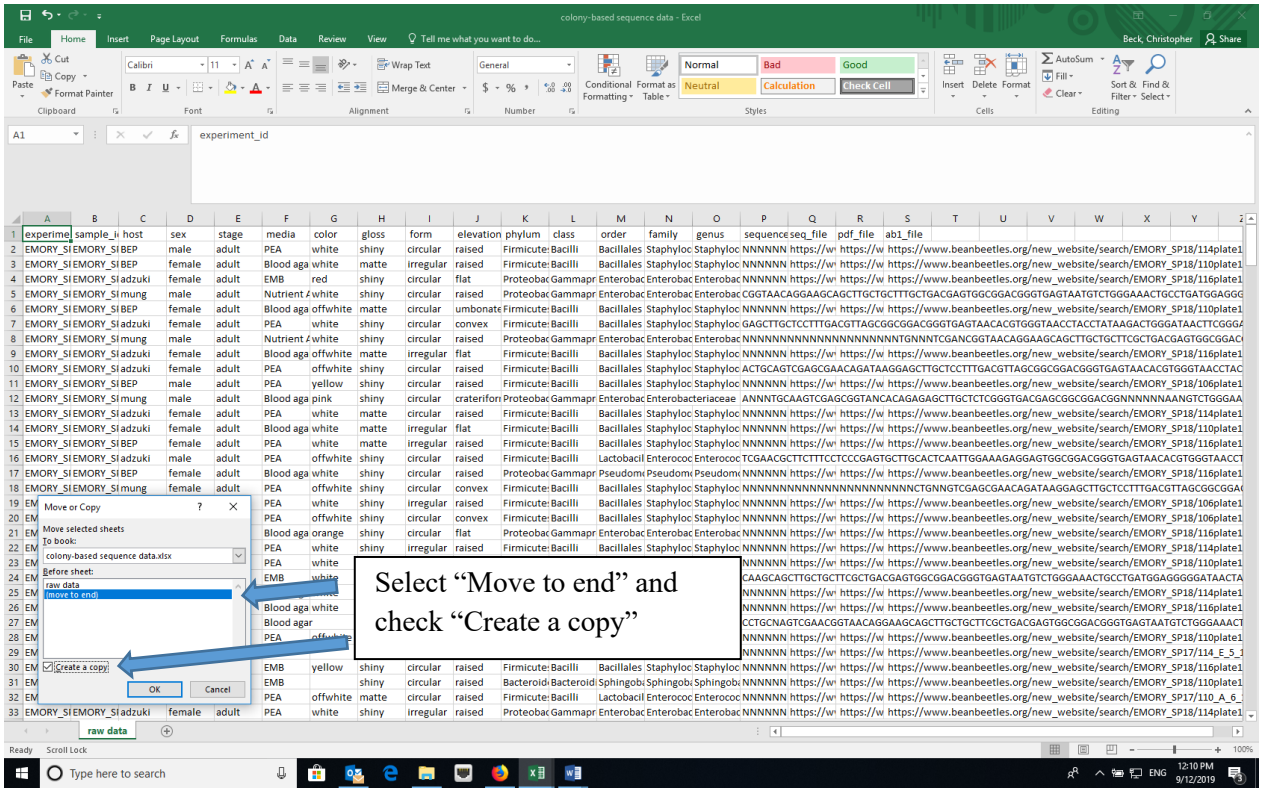




Data Reduction

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”).

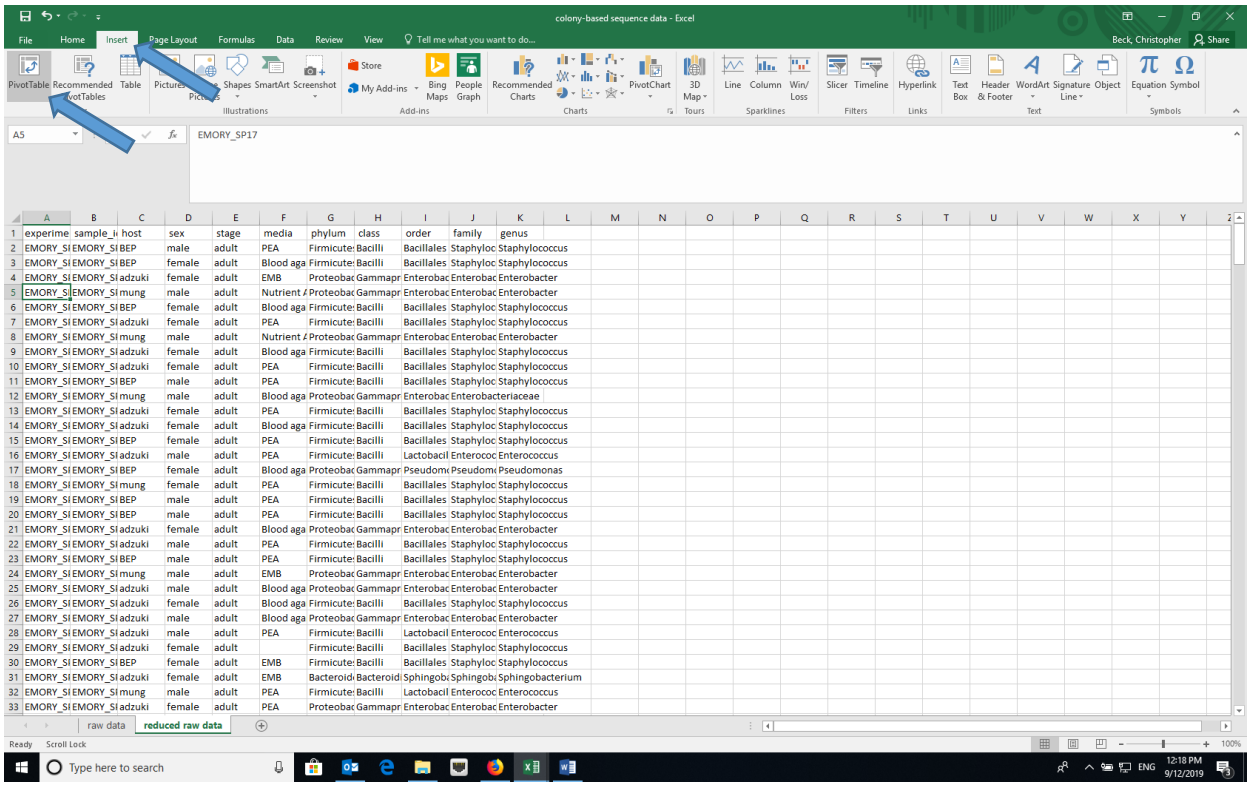

Right click on tab and select “Move or Copy”.

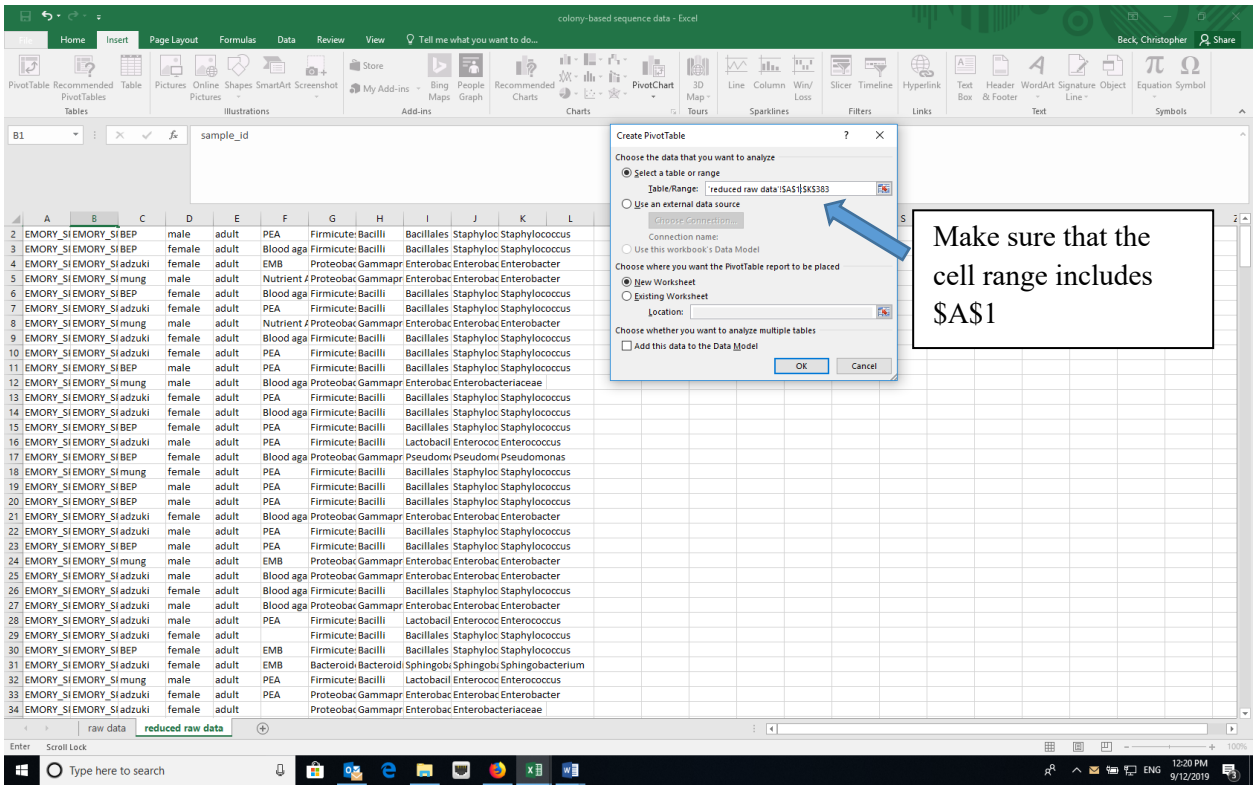

Select “Move to end” and check “Create a copy”

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.

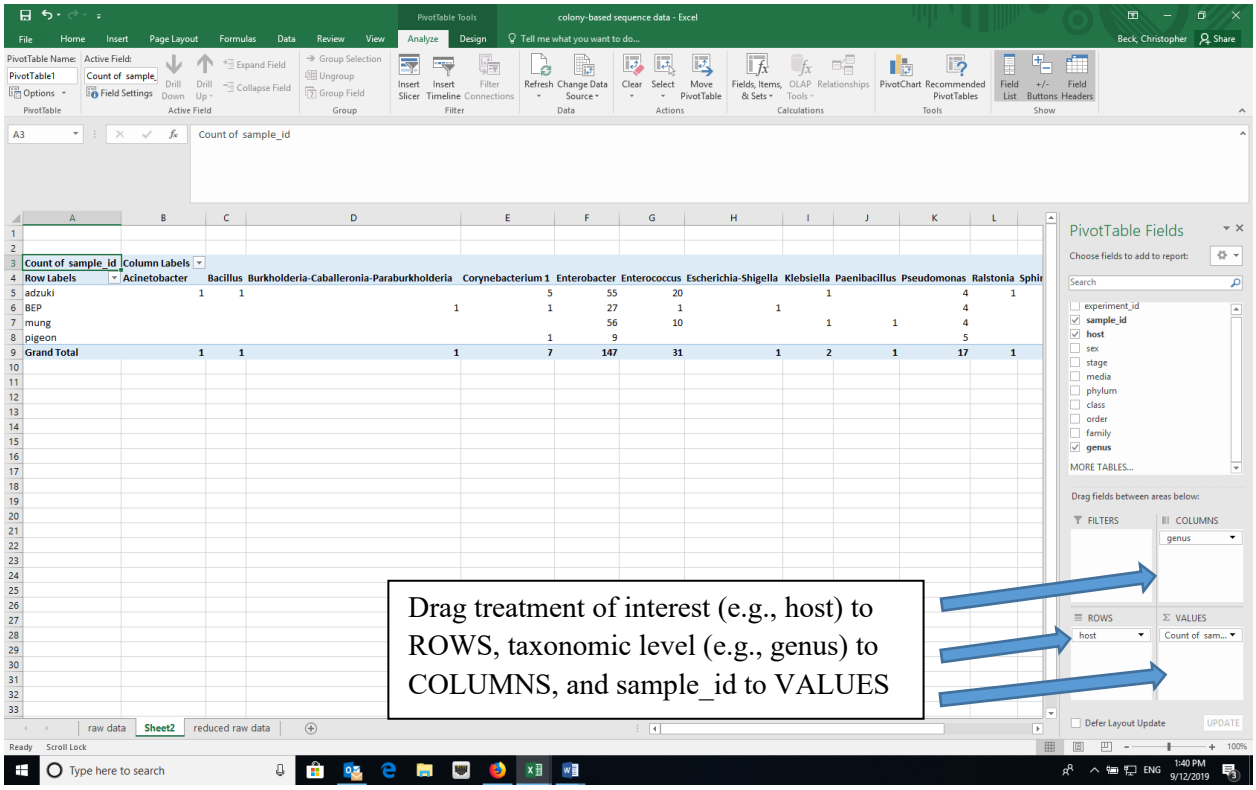
Data Manipulation

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 “A1”), which has the column headings. Click OK to create the pivot table in a new worksheet and label the tab “pivot table”.

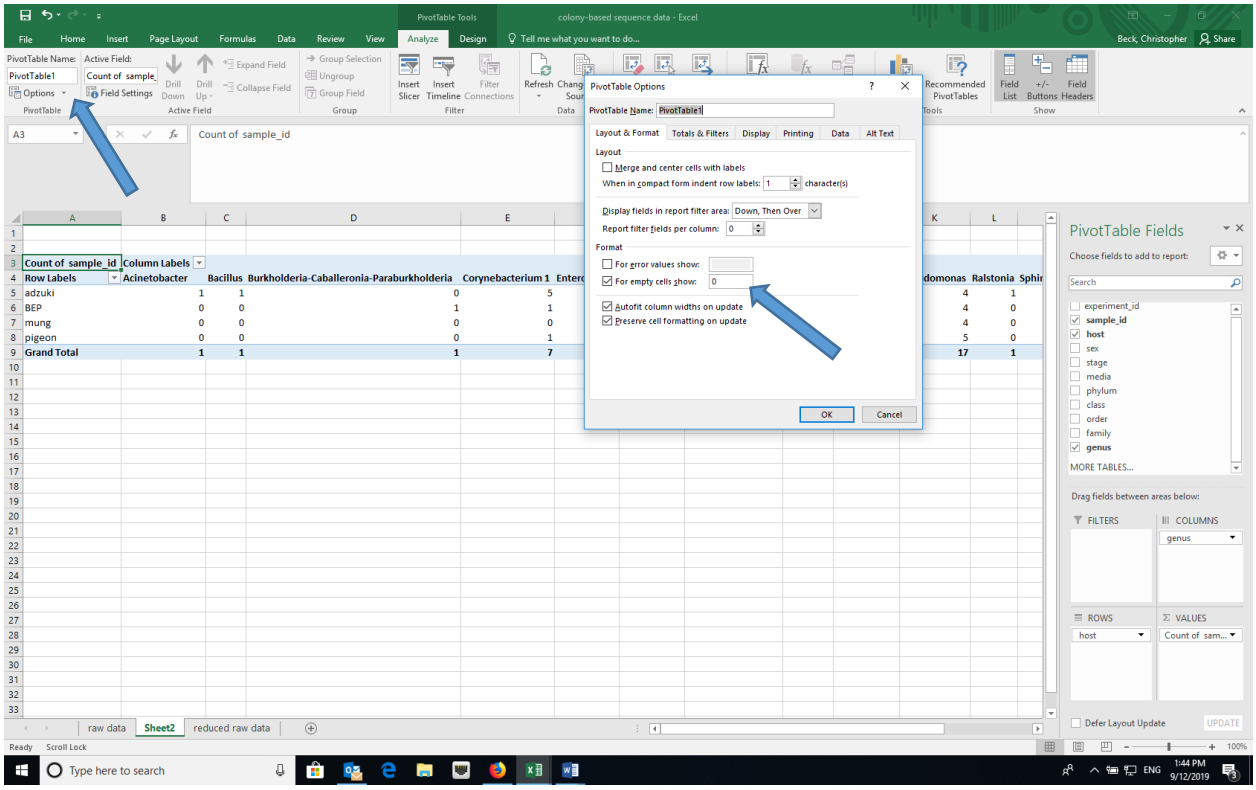


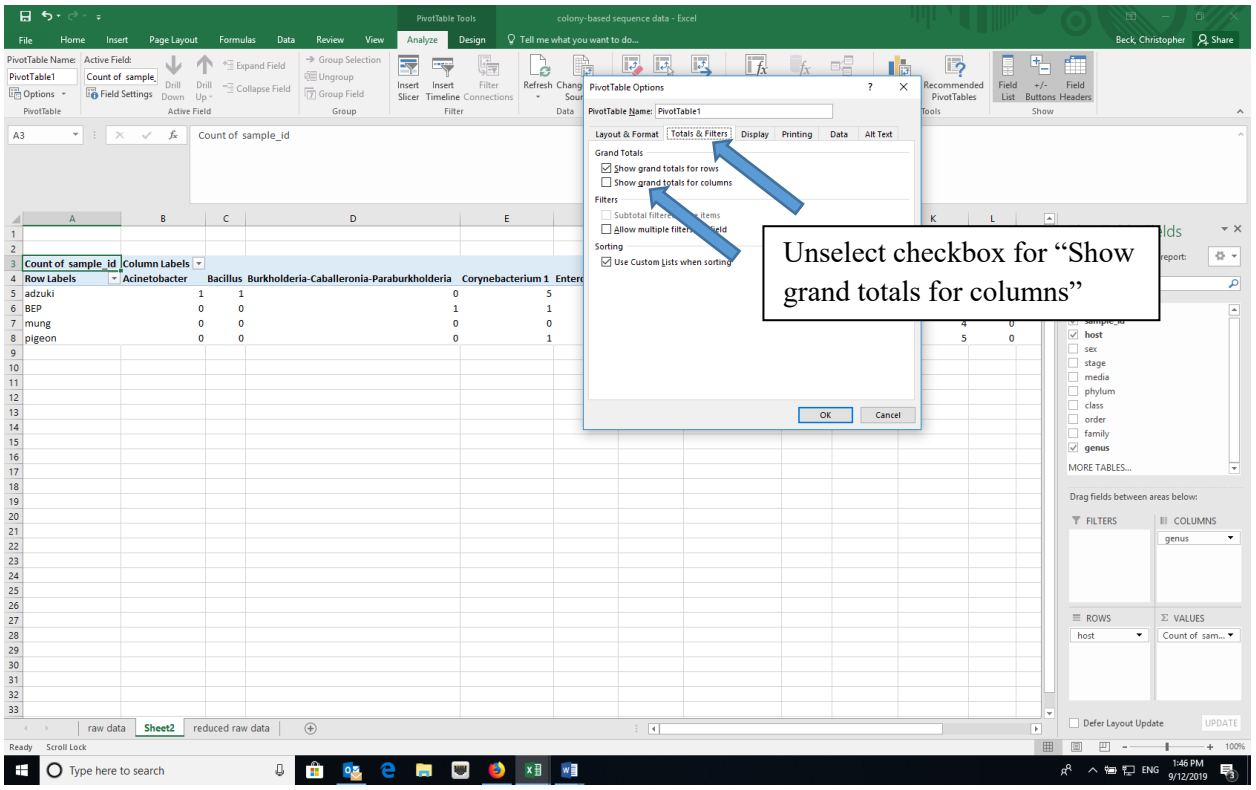

Make sure that the cell range includes A1

3. 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.

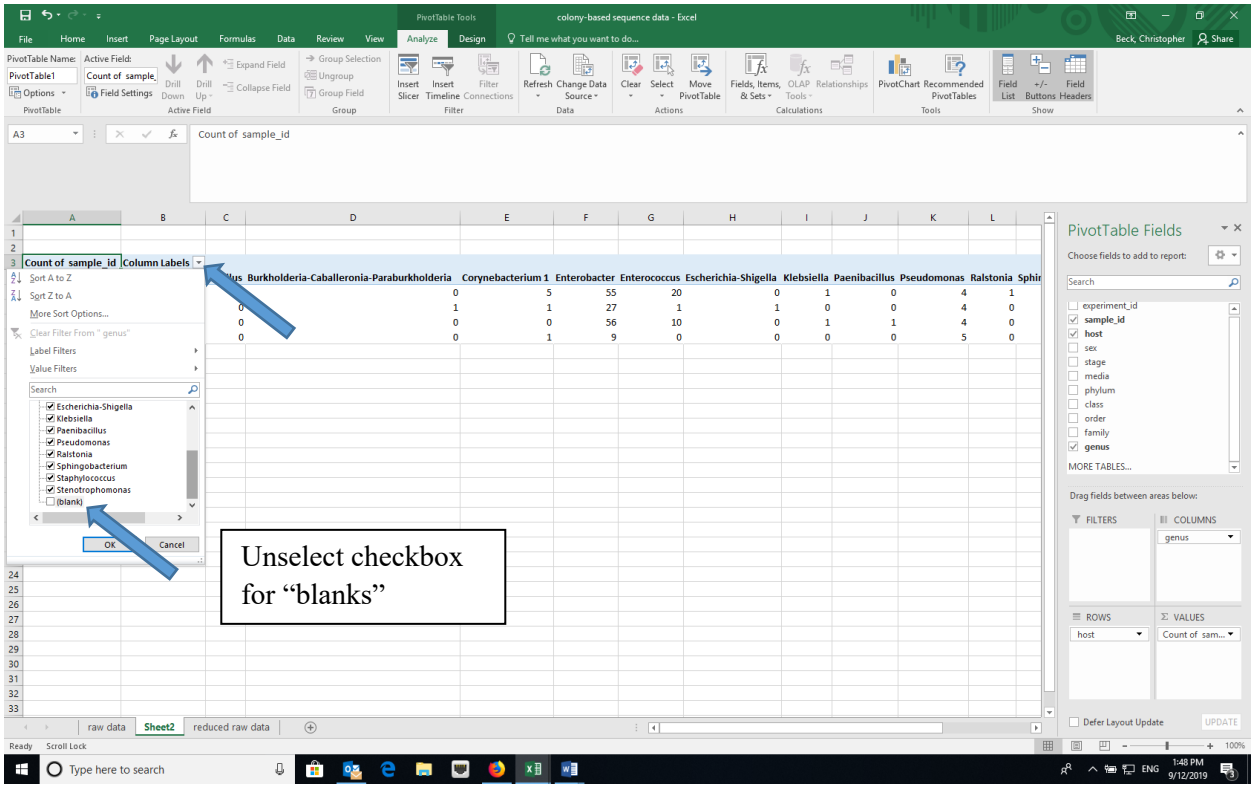


4. You can add zeros to all of the empty cells in the Pivot Table using the Options menu and remove the Grand totals for Columns using the Options menu or the Design tab depending of your version of Excel. (You want to keep the Grand totals for rows to calculate diversity indices.)

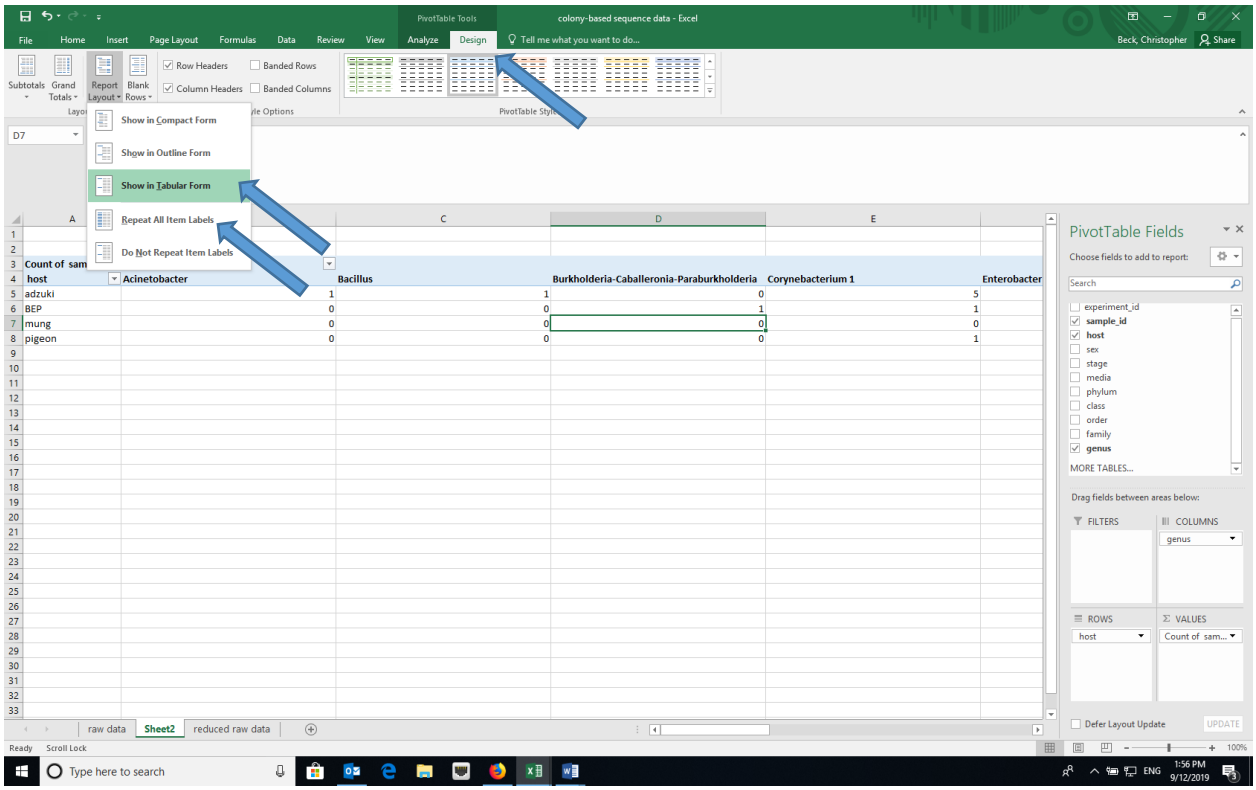




5. You can remove the “blanks” column using the Column labels dropdown (located at upper left of the sheet) and unselecting “blank”.

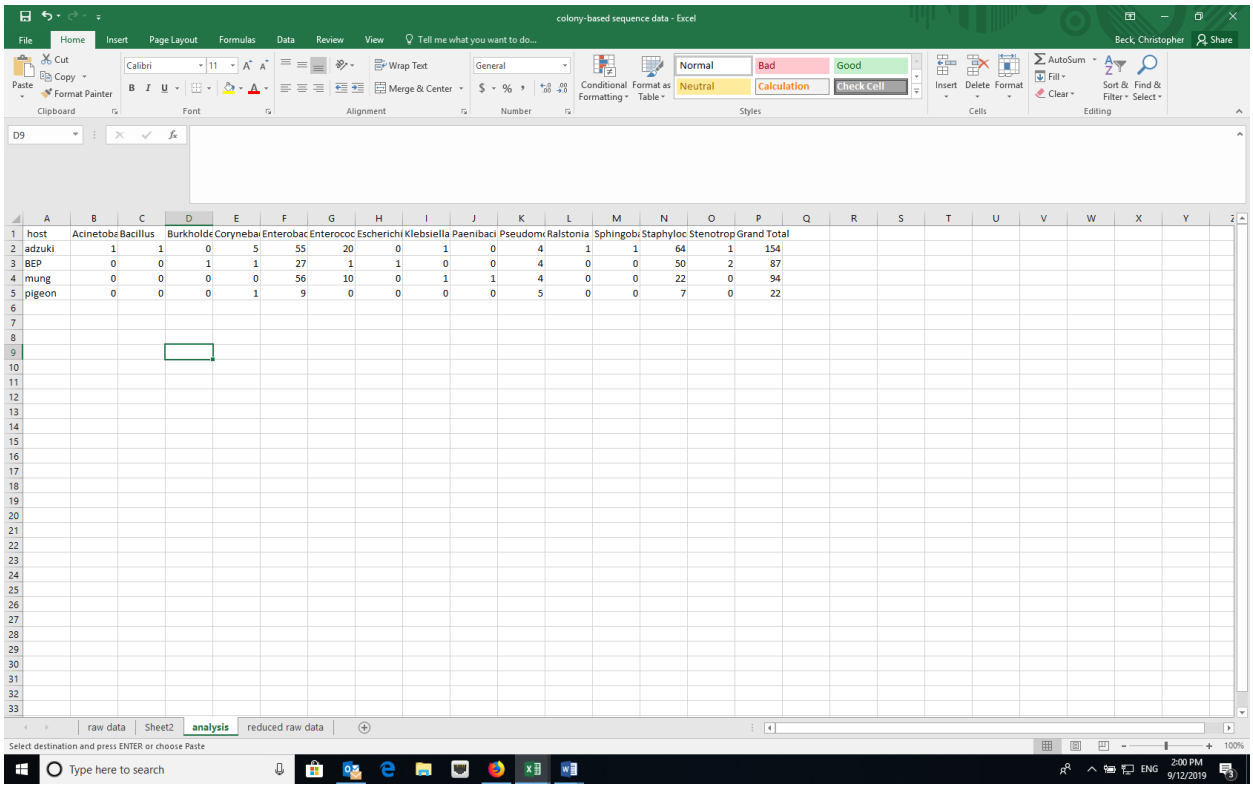


6. If you selected more than one treatment for the rows, you can get the treatment data to repeat for each sample. In the Design tab, select “Report Layout” and choose “Show in Tabular form” and “Repeat All Item Labels”.



Count of sam						
host	Acinetobacter	Bacillus	Burkholderia-Caballeronia-Paraburkholderia	Corynebacterium 1	Enterobacter
adzuki	1	1	0	5	
BEP	0	0	1	1	
mung	0	0	0	0	
pigeon	0	0	0	1	

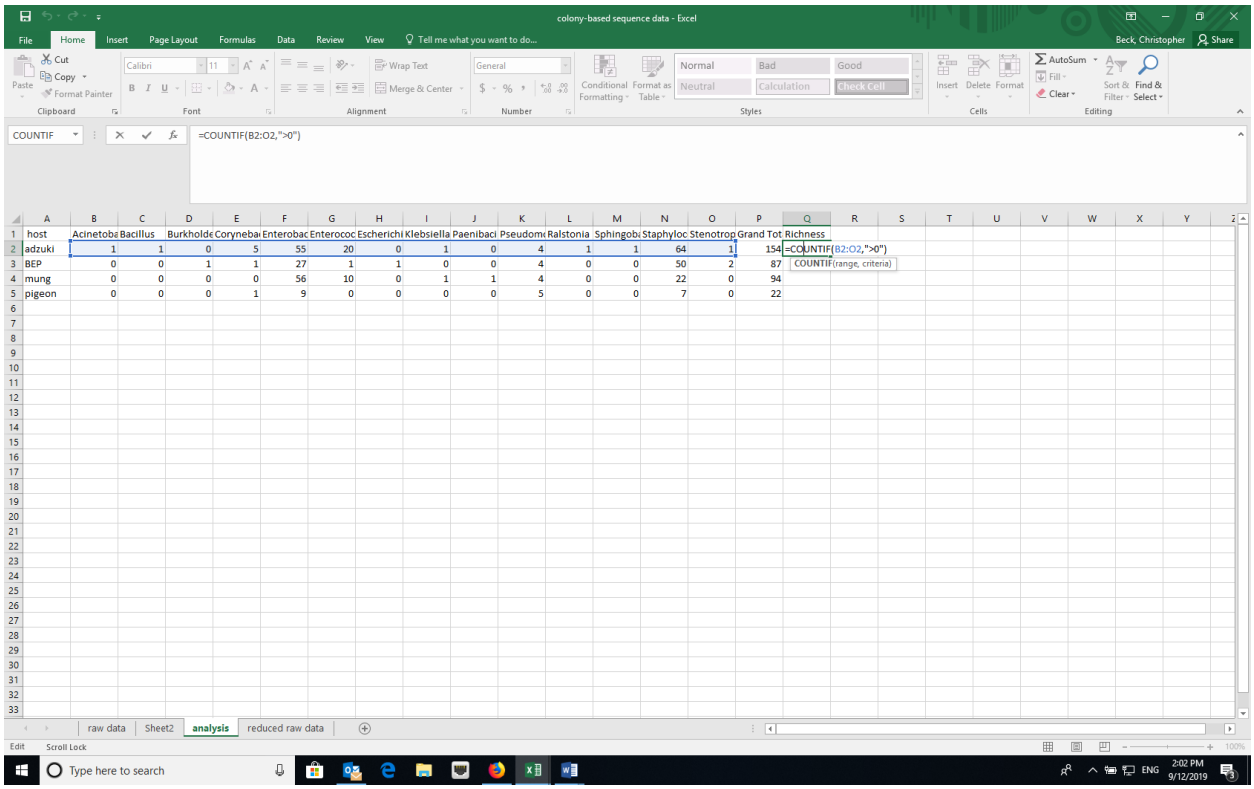
7. Copy and paste (as values) the pivot table to a new worksheet and remove any extra rows at the top. The top row should now have the taxa names. Name this tab “analysis”. Conduct the community ecology analyses that follow in Excel on the “analysis” sheet.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	host	Acinetoba	Bacillus	Burkholde	Corynebac	Enterobac	Enterococ	Escherichi	Klebsiella	Paenibaci	Pseudom	Ralstonia	Sphingobi	Staphyloc	Stenotrop	Grand Total
2	adzuki	1	1	0	5	55	20	0	1	0	4	1	1	64	1	154
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

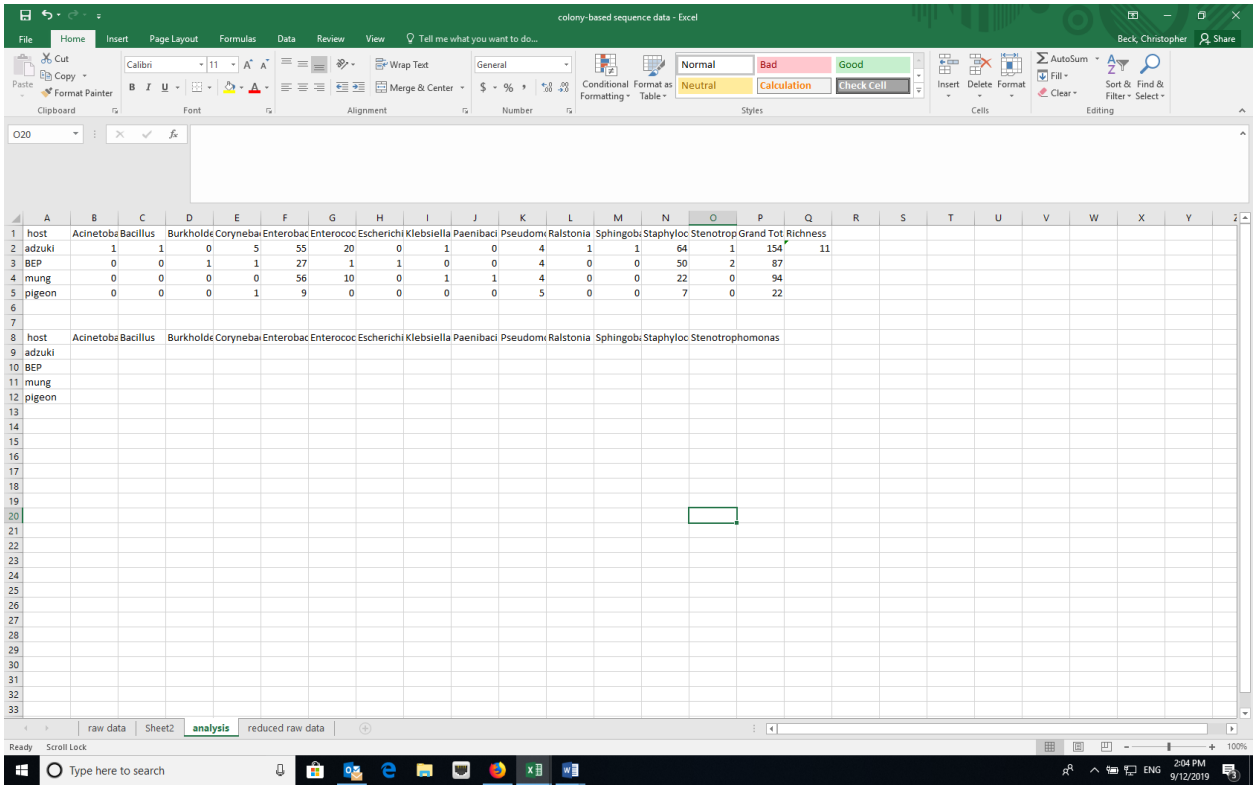
Calculating Diversity Indices

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.

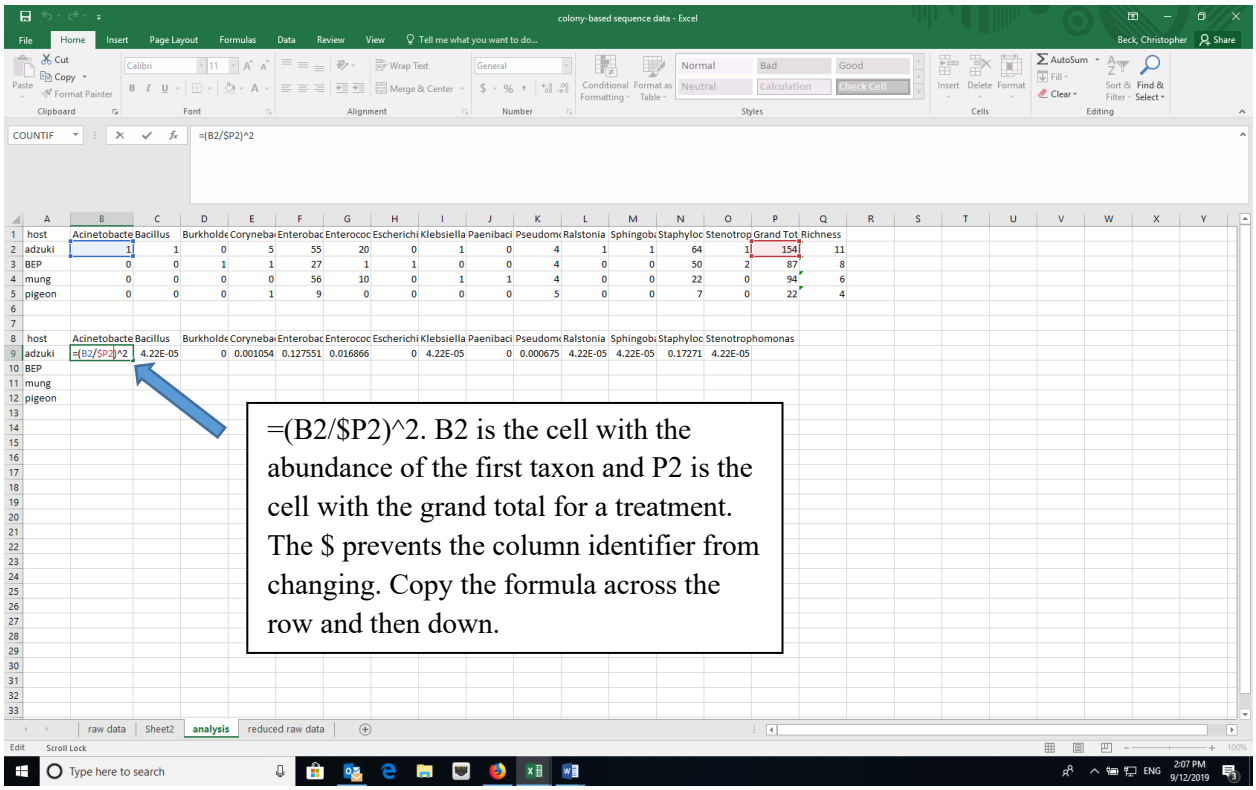


2. Simpson Index – the Simpson Index incorporates both species (taxon) richness and species (taxon) evenness.

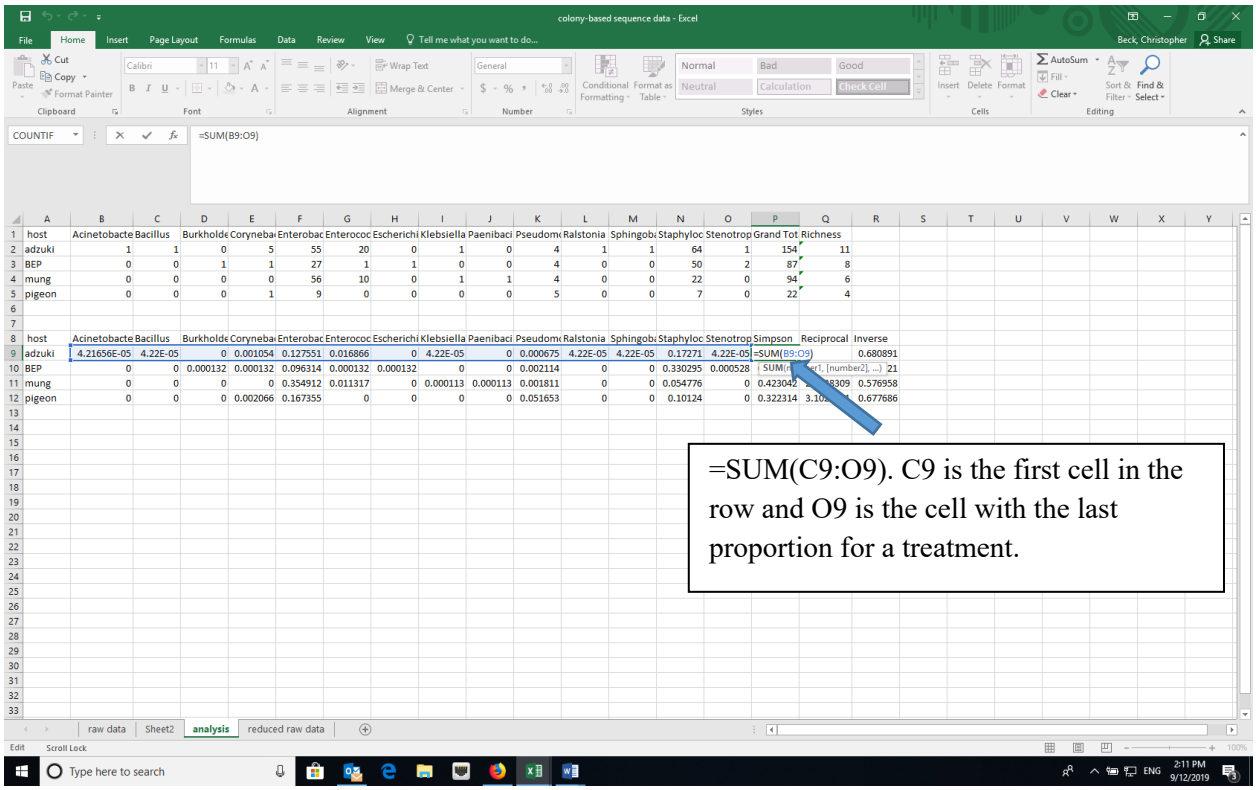
    a. $D = \Sigma(n/N)^2$, where n=number of individuals of a particular species (taxon) and N=total number of individuals in a sample. D increases as diversity decreases, which is counterintuitive. A reciprocal or inverse index would be more intuitive and are easily calculated.
    b. Reciprocal Simpson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



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.

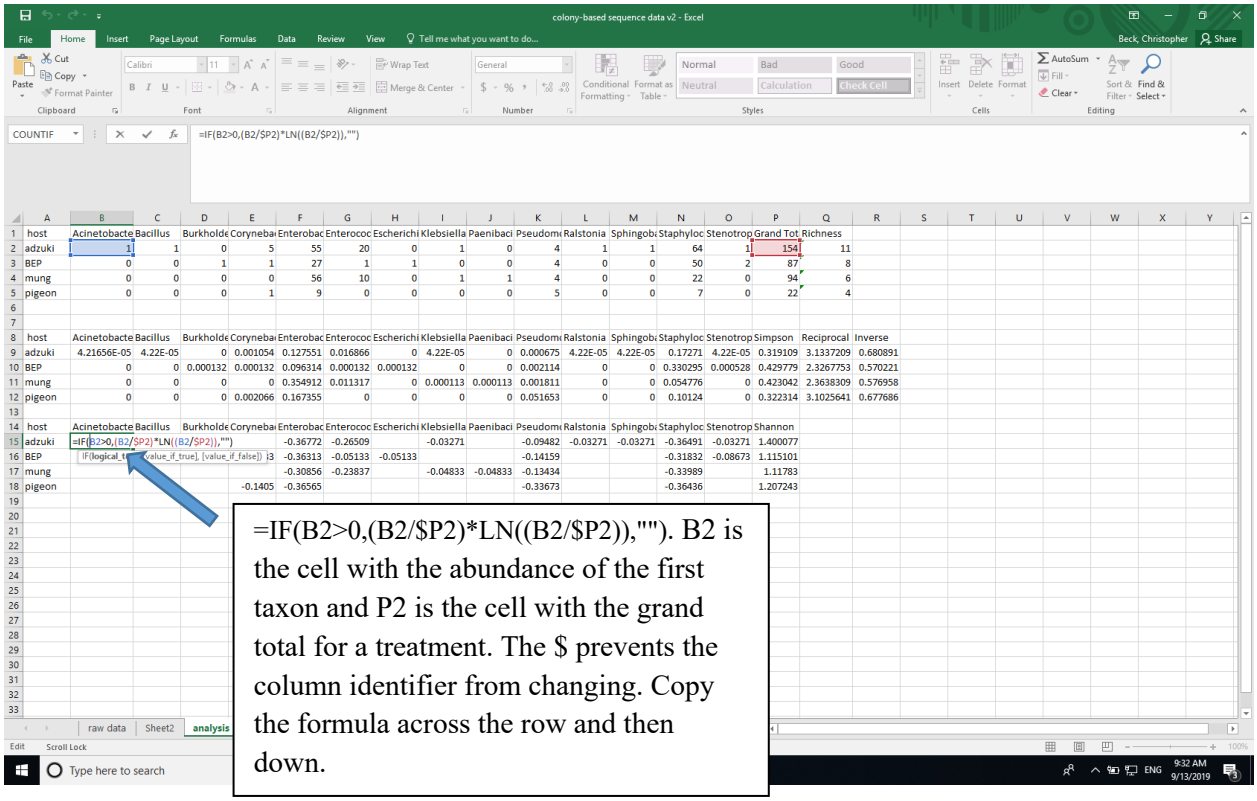


f. Calculate the sum of the proportions squared (=SUM in Excel) to calculate the Simpson Index.

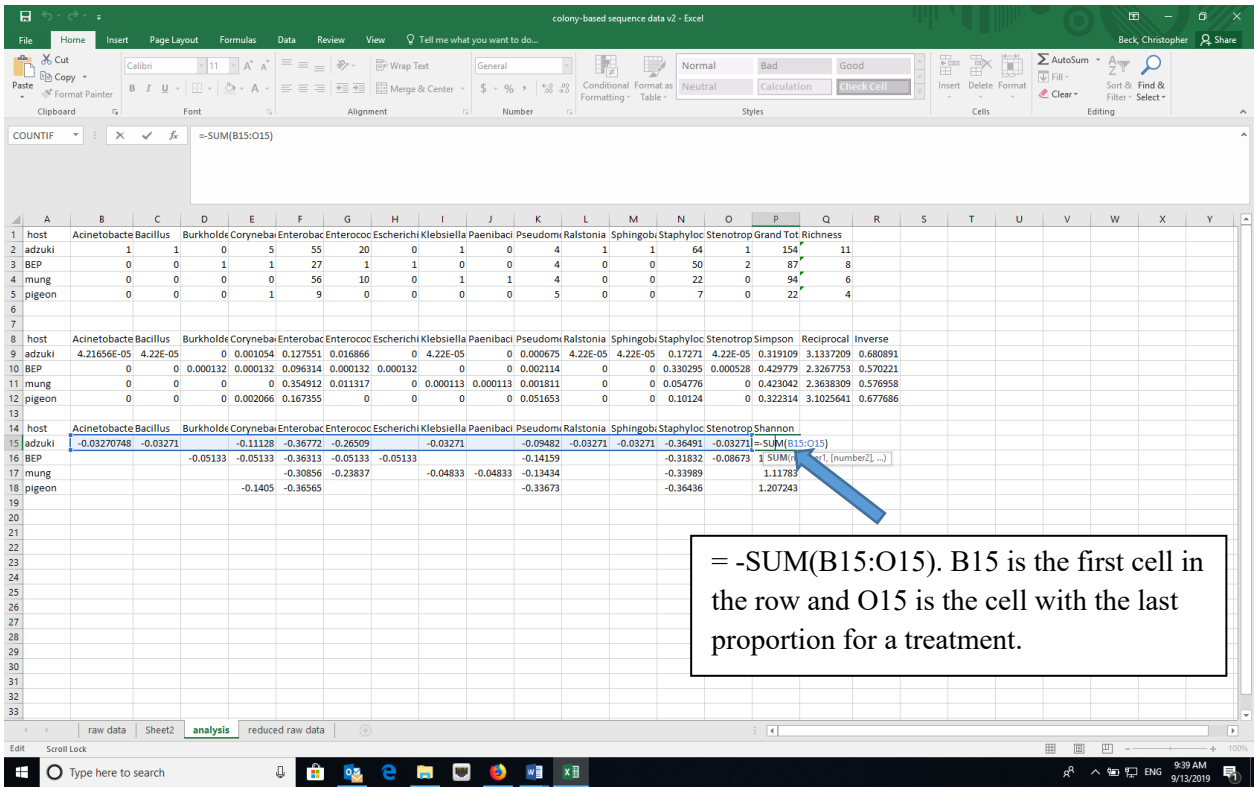


g. Calculate the reciprocal (e.g., =1/P9) and inverse Simpson (e.g., =1-P9) using formulas in Excel.

3. Shannon-Weaver (Shannon-Weiner) Index – also incorporates species (taxon) richness and species (taxon) evenness
    a. H=-Σ*p*ln*p*, where p is the proportion of individuals of each bacterial taxon in a community (i.e., n/N).
    b. Create a new data array below the original using the same row labels (treatment variables) and the same column labels (species).
    c. 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.
    d. 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)),"")

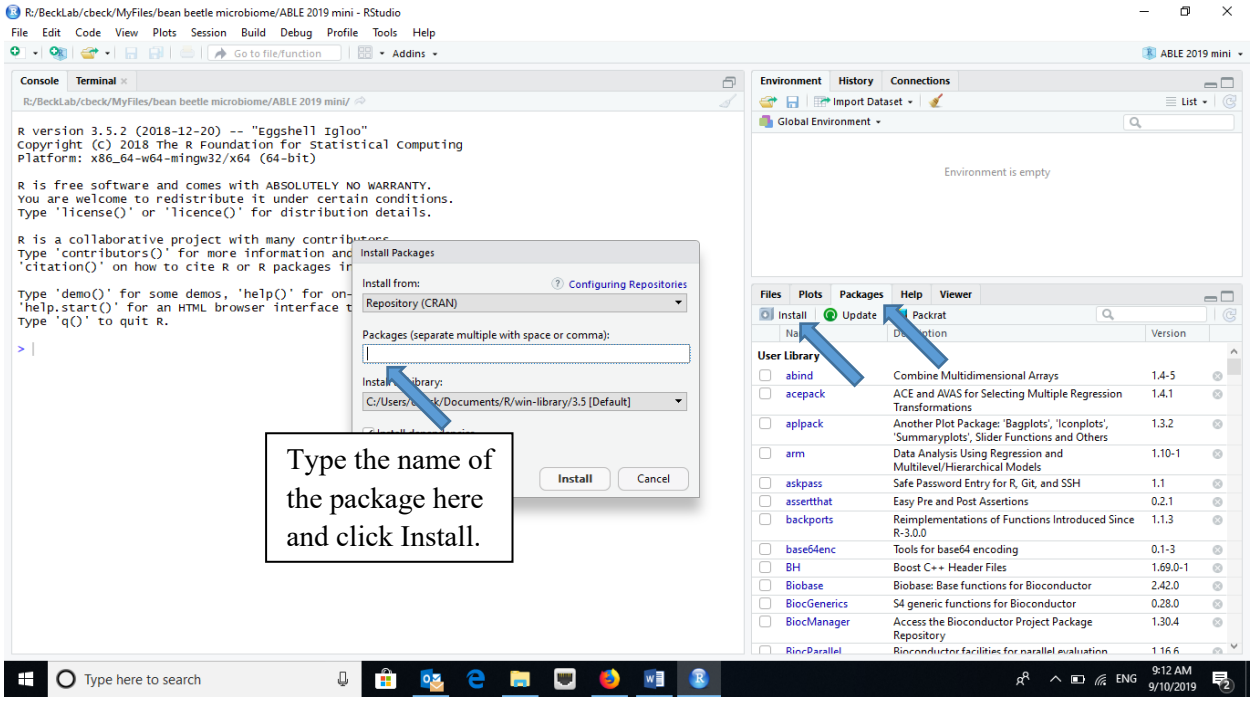


e. Calculate the negative sum of the proportions ($p \ln p$) (=-SUM in Excel for each row, a different microbial community) to calculate the Shannon-Weaver Index.

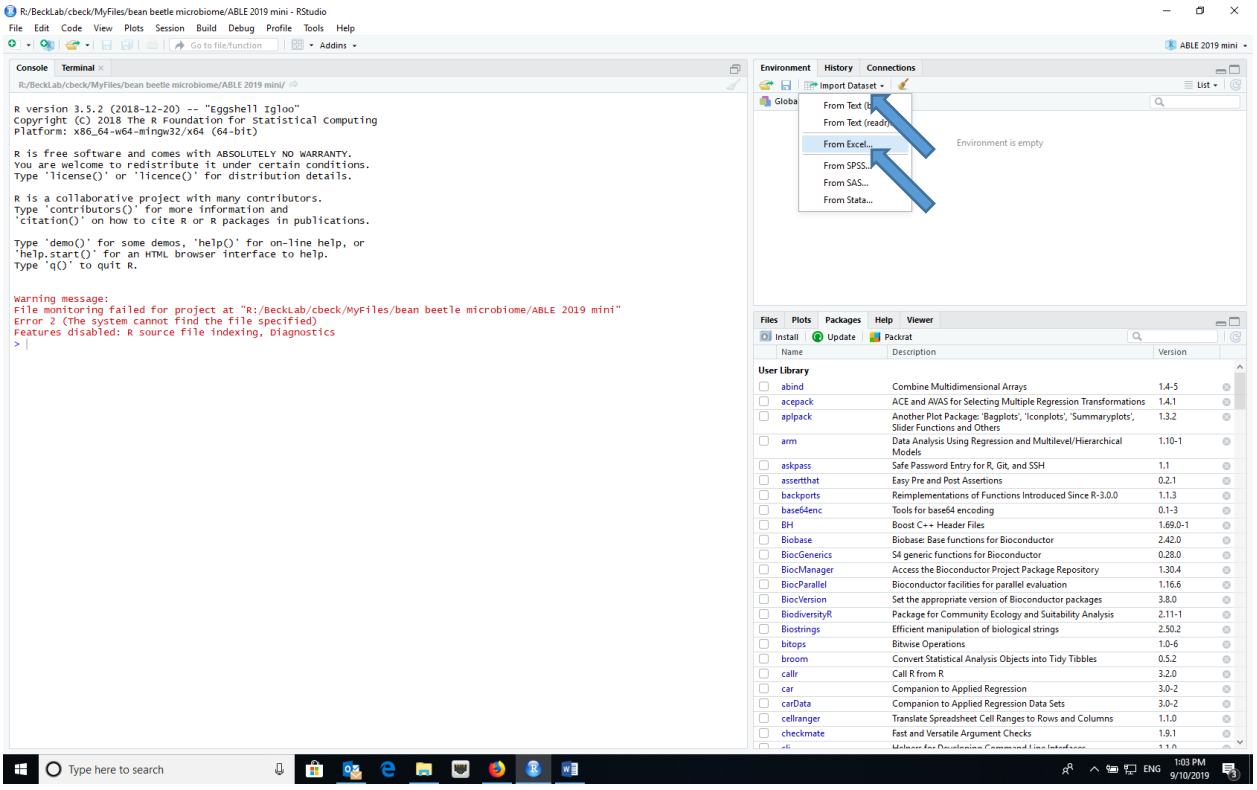


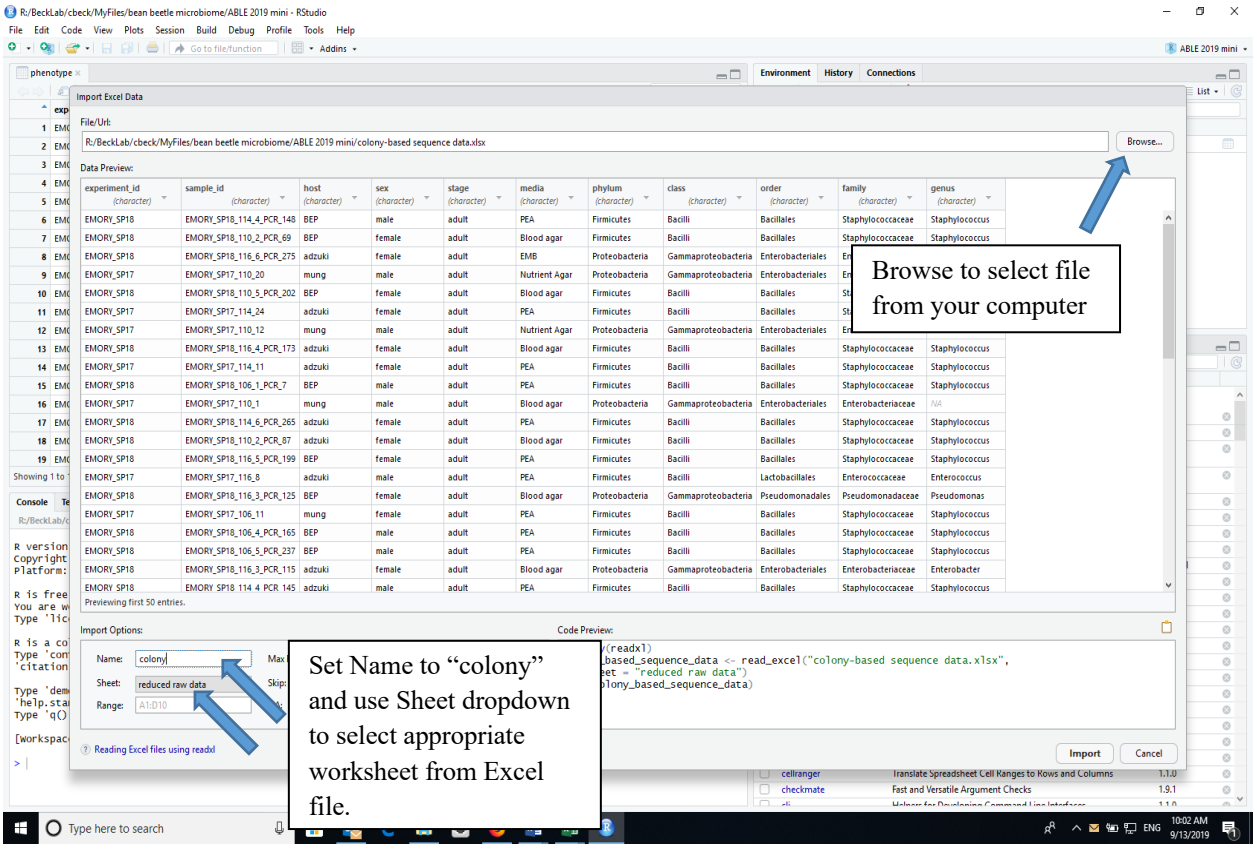
Data Manipulation in R

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.


R:/BeckLab/cbeck/MyFiles/bean beetle microbiome/ABLE 2019 mini - RStudio
Import Dataset
From Excel...
From SPSS...
From SAS...
From Stata...
Environment is empty


R:/BeckLab/cbeck/MyFiles/bean beetle microbiome/ABLE 2019 mini - RStudio
Import Excel Data
File/Url:
R:/BeckLab/cbeck/MyFiles/bean beetle microbiome/ABLE 2019 mini/colony-based sequence data.xlsx
Browse...
Data Preview:
Import Options:
Name: colony
Sheet: reduced raw data
Range:
Code Preview:
Import
Cancel
Browse to select file from your computer
Set Name to “colony” and use Sheet dropdown to select appropriate worksheet from Excel file.

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.

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

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 fuction. 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 Grander View of Life. New York: HarperCollins Publishers.

Materials

This is a computer-based activity so students will need individual computers with internet access. This study could be conducted by students working in pairs at one computer, but there is more to be gained by having students work individually while collaborating. Basic data manipulation requires Microsoft Excel. Data analyses may be conducted using Excel or the freely available RStudio program. Configuring computers prior to conducting the data manipulation and analyses is strongly suggested since it will minimize student frustration.

RStudio is an open access statistical analysis interface that requires the installation of the open access R program. R and RStudio may be downloaded from:

https://cran.r-project.org/
https://www.rstudio.com/products/rstudio/download/

Notes for the Instructor

We advise instructors to conduct the entire data manipulation and analysis that you want students to conduct prior to meeting with a class. That will ensure you are familiar with potential problems students may have when conducting this study. The screen shots included in the Student Handout are from a WindowsOS computer so the screens that you and your students see will depend on the computer operating system and the version of Excel that you use. Microsoft Excel (or a similar spreadsheet program) is required to manipulate data prior to conducting the community ecology analyses. Once data are configured and organized, the data may be analyzed using either Excel or RStudio. RStudio will permit a more comprehensive analysis, but Excel will permit students to calculate the basic variables and indices necessary for community comparisons, as well as understand the underlying formulas used for calculating the indices.

The data manipulation required prior to data analysis could be performed by the instructor as a way to save time and have students just focus on the community ecology comparisons. This streamlining of the data manipulation may be appropriate depending on the specific learning goals for your course.

The basic community ecology variables: species (taxon) richness, Simpson Index (as well as the reciprocal and inverse Simpson), and Shannon-Weaver Index may be calculated using either Excel or RStudio. Calculating an index of community distance (or dissimilarity), such as the Bray Curtis Dissimilarity Index or the Morista-Horn Index of Dissimilarity, as well as plotting species accumulation curves, will be much more tractable using RStudio.

Choosing the dataset(s) that students evaluate, colony phenotype or colony-based sequence is a matter of preference and the time you want students to spend on these analyses. The two datasets are not simply different methods of defining taxa in a community analysis but also represent different ways of defining a community. Analyzing both types of data to address similar questions could lead to a very productive discussion with students on the limitations of each database and the analysis of the data within them. Each database could be used independently of the other and there is no necessity for one database analysis to precede the other. If students worked on one database, the approximate time required to perform the data manipulation and the analysis using either Excel or RStudio is one and one-half hours. Completing all the activities presented here would require two 3-hour laboratory periods.

Cited References

Beck CW, Blumer LS. 2007. Bean beetles, *Callosobruchus maculatus*, a model system for inquiry-based undergraduate laboratories. Pages 274-283 In O'Donnell MA, editor. Tested studies for laboratory teaching. Volume 28. Proceedings of the 28th Workshop/Conference of the Association for Biology Laboratory Education (ABLE).

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 fuction. 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 Grander View of Life. New York: HarperCollins Publishers.

Acknowledgments

This work was supported in part by the National Science Foundation awards DUE-1821533 and DUE-1821184 to Emory University and Morehouse College. Opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation, Emory University, or Morehouse College.

About the Authors

Larry Blumer is Professor of Biology and Director of Environmental Studies at Morehouse College where he teaches ecology, environmental studies, and introductory biology. Blumer also teaches in the SEA PHAGES program (Freshmen research immersion) at Morehouse.

Chris Beck is a Professor of Pedagogy in the Department of Biology at Emory University, where he teaches an upper-level ecology lab. Along with Larry Blumer, he developed the bean beetle as a model system for inquiry-based teaching in undergraduate biology laboratory courses.

Mission, Review Process & Disclaimer

The Association for Biology Laboratory Education (ABLE) was founded in 1979 to promote information exchange among university and college educators actively concerned with teaching biology in a laboratory setting. The focus of ABLE is to improve the undergraduate biology laboratory experience by promoting the development and dissemination of interesting, innovative, and reliable laboratory exercises. For more information about ABLE, please visit **http://www.ableweb.org/.**

Advances in Biology Laboratory Education is the peer-reviewed publication of the conference of the Association for Biology Laboratory Education. Published articles and extended abstracts are evaluated and selected by a committee prior to presentation at the conference, peer-reviewed by participants at the conference, and edited by members of the ABLE Editorial Board. Published abstracts are evaluated and selected by a committee prior to presentation at the conference.

Citing This Article

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 41. Publication of the 41st Conference of the Association for Biology Laboratory Education (ABLE) https://doi.org/10.37590/able.v41.art24