Colony-based Sequencing Analysis

Introduction

Your group or class has collected data on the microbial community of bean beetles based on 16S rRNA sequencing of individual bacterial colonies cultured from bean beetle homogenates plated on different media. Since only a small number of colonies 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 host bean species, sex, and other sample or experimental variables.

Data manipulation

1. We need to consolidate the data for each host type, 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 -> Pivot Table). Make sure that the data source includes the top row, which has the column headings. Set the treatment(s) that you are interested in as the rows and the taxonomic level you are interested in as the columns. The Values should be a COUNT of the sample_id, as each row in the dataset represents a single sample.
3. You can add zeros to all of the empty cells in the Pivot Table and remove the Grand totals for Columns using the Options menu. (You want to keep the Grand totals for rows to calculate diversity indices.)
4. You can remove the "blanks" column using the Column labels dropdown and unselecting "blank".
5. Copy and paste (as values) the pivot table to a new worksheet and remove any extra rows at the top. The top row should have the taxa names.

Calculating diversity indices

1. Species richness – the number of unique species in a sample
    a. Although you could manually count the number of cells with values greater than zero for each treatment, using the COUNTIF formula in Excel is easier (e.g., =COUNTIF(range,">0")).
2. Simpson Index – the Simpson Index incorporates both species richness and species evenness.
    a. $D = \Sigma(n/N)^2$, where n=number of individuals of a particular species and N=total number of individuals in a sample. D increases as diversity decreases, which is counterintuitive.
    b. Reciprocal Simpson – 1/D
    c. Inverse Simpson – 1-D
    d. Using the grand totals for each treatment, calculate the proportions squared. Using the Excel trick that $ before a column or row prevents Excel from iterating when copying a formula makes this easy.
    e. Calculate the sum of the proportions squared (=SUM in Excel) to calculate the Simpson Index.
    f. Calculate the reciprocal and inverse Simpson using formulas in Excel.
3. Shannon-Weaver (Shannon-Weiner) Index – also incorporates species richness and species evenness

a. H=-Σ*p*ln*p*, where p is the proportion of individuals of each species in a community (i.e., n/N).
b. Using the grand totals for each treatment, calculate the proportions. Using the Excel trick that $ before a column or row prevents Excel from iterating when copying a formula makes this easy.
c. Note that ln*p* is undefined if *p*=0, so you can use an "IF" statement in Excel to prevent the calculation of undefined values. For example, =IF(C2>0,(C2/$EX2)*LN((C2/$EX2)),"")

Calculating community similarity (distance)

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

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

Where:

- i & j are the two sites,
- S_i is the total number of specimens counted on site i,
- S_j is the total number of specimens counted on site j,
- C_{ij} is the sum of only the lesser counts for each species found in both sites.

In Excel, S_i and S_j are just the grand totals for a particular community. To calculate C_{ij}, we need to find the taxa that are present in both samples and then find the minimum. We can use the following formula for a particular taxa:

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

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

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 diversity that you use?
3. Which communities are most similar (different)?
4. Do your answers to the questions above depend on the taxanomic level of analysis?

Data Manipulation in R

1. Import the reduced dataset into RStudio.
2. Attach the imported dataset to the dataframe (*attach(dataset)*)
3. Create a community matrix for a particular treatment

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

4. If you want to look at two factors at the same time, creating the community matrix is a little more complicated.

   First, create subsets of the data using the subset command.

```
> male<-subset(colony,colony$sex=='male')
> female<-subset(colony,colony$sex=='female')
```

   Then, create a community matrix for each subset.

```
> male_table<-table(host,genus)
> female_table<-table(host,genus)
```

   Then, put them back together.

```
>community_2<-rbind(male_table,female_table)
```

   Finally, rename the rows so that they have unique names.

```
>rownames(comm2)<-c('adzuki_male','BEP_male','mung_male','pigeon_male'
,'adzuki_female','BEP_female','mung_female','pigeon_female')
```

Calculating diversity indices

1. Load the BiodiversityR package.
2. Species Richness

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

3. 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)
```

4. Shannon

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

Calculating community similarity (distance)

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

This gives a matrix of all of the pair-wise distances between samples using the Bray Curtis index of distance.