Excel Tutorial Part 4: Focusing on Taxonomy
By Jimiane Ashe

Welcome!

This tutorial is brought to you by the Research Experiences in Microbiomes Network (or REMNet) and the City University of New York.

In this video you will learn how to isolate and analyze a subset of your project's OTUs (or Operational Taxonomic Units) at a specific taxonomic level using the Text to Columns, Remove duplicates and Sort tools and the SUM IF and COUNT IF functions in Microsoft Excel. To do this, you will also learn how to convert your relative abundance values into their number of sequences.

In order to follow along with us today, you have received taxa summary files and a DataSummary document for your analyzed microbiome samples, and you have opened them in Excel. If you need a refresher, please review our tutorials called Accessing your Microbiome Data, Understanding your Microbiome "DataSummary" Document and Organizing your Microbiome Data.

As you work with us today, you are welcome to use keystrokes or your mouse to utilize these commands and functions or to select necessary cells.

To orient you to this table, column A contains the hierarchy of taxonomy for each OTU identified in the project. Columns B through E report the relative abundance for each sample in this project.

Before we get started, we will save our tab-delimited .txt file as an Excel .xlsx file. Many of the tools and functions are not compatible with the tab-delimited file type.

#File > Save As… > Format: Excel Workbook (.xlsx) > click SAVE

In this example, we will isolate the OTUs at the Order taxonomic level. We will use the Text To Columns Tool to separate the taxonomic levels in Column A. You will need 5 empty columns to your right to complete the next task.

#Eg1: select sample columns and slide to right 3 columns
#Eg2: select 2 empty columns, click Insert, select Columns

#Select column A
#Data > Text to Columns… > Delimited > click Next > select semicolon > click Finish

Now your table has the full taxonomy separated into 6 columns for each OTU.
#(Label columns)

Before we can proceed with any functions or formulas, we must convert the relative abundance values to the number of sequences for each OTU. Insert a row below your header row so we can paste in this information.

#Select Row 2 > Right click > Select Insert –OR–
#Select Row 2 > Select Insert at the top of your screen > Select Rows from the menu

Next go to the FinalSeqStats tab in your "DataSummary" document for your dataset.

Relative abundance for this dataset was calculated after unidentified sequences, incomplete sequences or "bad reads", and eukaryotic contamination or "non-target" sequences were filtered out. In the final column, you will find the final number of sequences remaining for each sample after these filtering steps.

#Select cells > COPY > and PASTE SPECIAL > Transpose under our sample names in the data sheet.

Relative abundances for the OTUs in each sample are a proportion of this total number of filtered sequences. We simply need to multiply these two values.

We will make new headers for each column.

We can create a multiplication formula by typing the = sign in cell G3.

= > click on the first relative abundance value in your first sample > * > select the cell with the number of sequences in that sample B2.

We now need to modify our formula to hold the number of sequences constant.

In your formula bar, type a $ sign between B and 2, and hit enter.

Apply this formula to all of the orders in your new column:

#Grab blue corner square with black cross and slide mouse down
#Double click on bottom corner
Edit > Fill > Down

Click on a few cells to verify that each relative abundance has been multiplied by cell B2. Next apply this to your other samples.

To double-check our formula, we can calculate the sum of all of our OTUs. If our formula is applied correctly, the sum of the column will equal the correct number of sequences originally obtained for that sample.

All formulas can be accessed by typing the = sign in a cell.
You can type in the SUM function or you can choose SUM from a list of functions.

In the formula bar, you will see your cursor blinking between a closed pair of parentheses

Select the values in that column, and hit enter. Apply this formula to your other samples. Format your cells to whole integers.

#Select your columns > Format > Cells… > Select Number > Select 0 Decimal Places > Click OK

We can see that our calculated number of sequences matches the original value.

Next we will remove all duplicate orders.

#Select Orders column > copy and paste to new column > Data > Remove Duplicates

In the Remove Duplicate window, Excel will report how many duplicates were found, and how many will remain after we use this tool. Click the Remove Duplicates button.

Sort your remaining orders alphabetically

#Data > Sort > click “My List has headers” > under Order: A to Z > click OK

We will now demonstrate how to use the SUM IF function to quantify the number of OTU sequences for each order. The SUM IF function selectively sums a specific subset of values from a larger list.

#Type SUMIF header

Type an = sign to access the functions.
You can type in the SUMIF function or you can choose SUMIF from a list of functions.
In the formula bar, you will see your cursor blinking between a closed pair of parentheses where you will need to select the range, criteria and the sum range.

The range will be the full-unmodified list of orders from column D. Select column D, and type a comma.

The criteria will be one specific order. Select the order from cell H2, and type a comma.

The sum range will be the relative abundance data from our samples.
Select those columns, and hit Enter.

#Click on formula bar to review

Apply this formula to all of the orders in your new column using your favorite method:

#Double click on bottom corner

We will now demonstrate how to use the COUNT IF function to count how many OTUs from your samples are in each order. This will allow you to see which orders are most abundant in your dataset.

#Type COUNTIF header
= > COUNTIF > range: order column > type comma > criteria: new order cell > hit Enter
Click on formula bar to review
#Fill down

To summarize, you learned how to use the Text to Columns, Remove Duplicates and Sort tools. You converted your relative abundance values to their number of sequences. You also learned the SUM IF and COUNT IF functions and formulas, and how to apply it to all cells in a column.

Thank your for participating. If you found this helpful, we invite you to view our other tutorials guiding you through your microbiome data set.