## Excel Tutorial Part 3: Organizing your Microbiome Data

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 organize your microbiome taxa summary data from a file called the otu_table. We will go through the Average and Count functions, the Sort and Conditional Formatting tools, and formulas for the SUM function to identify important trends among the samples in your project.

In order to follow along with us today, you have received taxa summary files produced by high throughput next generation sequencing and the qiime bioinformatics pipeline, and you have opened this file in Microsoft Excel. If you need a refresher on how to open these files in Excel, please review our tutorial called Accessing your Microbiome Data.

To orient you to this table, column A contains the hierarchy of taxonomy for each OTU (or operational taxonomic unit) identified in the project. Columns B through AH 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

First we will use functions in Excel to count the number of samples in this project. Select the first row of your sample names from column B to AH, and look to the status bar at the bottom of your Excel window. Depending on your version of Excel, a left or right mouse click on the statistics segment will open the menu options. The Average, Count and Sum options will be used today. Excel has counted 10 samples in this project.

Now please look to cell E2. The numerical values reported for each sample are the relative abundances, which can be defined as the relative proportion of a specific OTU within each sample. Therefore the sum of all relative abundance values for each sample will equal 1.

To verify this, we will select the sample information in column B. The status bar reports that the sum of this column is indeed equal to 1. The count here is reporting the number of OTUs plus one cell for your header row. This project contains 1275 OTUs per sample.

Each of the tools and functions we will demonstrate in the remainder of this segment involves selecting specific columns and rows. You are welcome to use keystrokes or your mouse to select the necessary cells.

Next we will sort the sample columns horizontally. Select all of the sample columns in your project. Do not include the OTU ID column in this selection. Click on the Data tab in Excel, and select the Sort tool. The default option is to sort by rows. To sort by columns click on Options…, select “Sort left to right”, and click OK.

Select Row 1 to sort your samples by the column headers. The order of sort for this example will be alphabetic from A to Z. Choose the appropriate “sort on” and “order” selections for your project, and click OK. To view the full names for each sample, you can use your mouse cursor to widen each column.

We will now explore a very useful tool in Excel called Conditional Formatting. There are many features within this tool that will allow you to identify trends within each sample and within your entire project. There are also features that will allow you to compare values between samples across a subset of your entire project.

Go to the Home tab at the top of your Excel Window, and select the Conditional Formatting tool. We will now demonstrate the Top/Bottom Rules from the pull down menu, and select the top 10% from the subsequent menu. In the New Formatting Rule window that appears, you may tailor several options specifically for your project. We will proceed with the Classic style, top 10% and red highlighting. Click OK to view your selections applied to the data.

In column B, any cell highlighted in light red with dark red text indicates an OTU that has a relative abundance within the largest 10% of abundances for the sample in Column B.

You learned earlier about the sum option in the statistics bar, but now we will scroll to the right of the sample columns to demonstrate the SUM function. This method will report sums for each OTU in your project in a new column.

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. Use your mouse or keystrokes to select the relative abundance values in that row. You may also manually type the first and last column letters separated by a colon to select these cells. Hit enter to complete the formula.

To apply this formula to all of the OTUs in your project, you can execute 1 of 3 options. In the bottom right corner of your formula cell (in our example cell AJ2), when the mouse cursor forms a black cross, you can 1) double click, 2) click and drag that corner to the last row of your OTUs. Or 3) Select all cells from your first formula to the bottom and select Edit > then Fill > then Down from the pull down menus at the top of Excel. We are demonstrating option 2 now.

Get some additional practice using the Conditional Formatting tool to explore this new column of Sums. To review the process again, go to the home tab, select your Sum column, select the Conditional Formatting tool, select Top/Bottom Rules, select Top 10%, make any desired modifications the New Formatting Rule window, and click OK to apply. The 10% of samples with the highest summed relative abundance are now highlighted in red.

Remember that these summed values can now identify which OTUs are most abundant across your entire project. We can explore a second feature of the Sort tool to identify these OTUs.

Select all cells in your sheet by clicking on the small square with a shaded half triangle at the top left of your table. This square is above row 1 and to the left of column A. Click on the Data tab at the top of your Excel window. Click the sort tool. Select the Options button in the Sort window. Select Sort top to bottom. Click OK.

Our table has a header row, so we must click the “My list has headers” option in the top right corner of the window. Under “column” select your new sum column. Under Order select the “largest to smallest” option to bring your most abundant OTUs to the top of your table, and click OK to apply your selections.

All of your OTU IDs are now sorted by the largest relative abundance sums. We will scroll to the right to view the sum column. We have identified that OTUs from row 2 to row 128 are the most abundant across our entire project, and we are displaying these cells in bold for you now.

To summarize, you learned how to find averages and counts using the statistics bar at the bottom of your Excel window, you practiced using the Sort and Conditional Formatting tools, and you learned how to apply the SUM function.

You can now use these tools to identify trends across your entire project or within a subset of your project. For example, you can focus on a specific environment or sample type, or compare specific experimental samples to control samples.

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