

## Excel Tutorial Part 1: Accessing 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 open raw microbiome data files, and prepare your data for down stream analyses.

In order to follow along with us today, you have already collected samples in your environment, sequenced the microbiome for those samples using high through-put next generation sequencing technology, and raw data has been generated for you.

Within the window on your screen, you can see a list of files. This is the output from the qiime open-sourced bioinformatics pipeline for 16S microbiome data analysis. Sequencing centers can provide these files to you, or you can generate them yourself with the qiime pipeline using a server or a personal computer with adequate computing power. The qiime analysis pipeline will produce html charts and raw data files including several *taxa summaries*. In the next few minutes we will demonstrate how to identify and open the raw data text files, so that in the future you may perform downstream analyses and produce figures to complete your research project.

The first file in this window is the log file, which lists the process and performance of the qiime bioinformatics pipeline. It does not contain any microbiome data. The second file with the extension “.biom” is a specialized file for advanced bioinformatics analyses, and will not be useful to us today. We will open up the third file with the “.txt” extension to access your raw microbiome data.

These .txt files are tab delimited text files that can be opened up in Microsoft Excel. The file name describes the type of raw data you will find when you open the file. First, the acronym OTU stands for Operational Taxonomic Unit, which is the scientific system used to identify microbial taxa.

Further in the name you'll find shorthand for the removal of chloroplast, mitochondrial and unidentified sequences, which will remove eukaryotic contamination, artifacts and incomplete sequences from your data set.

Finally, you can see the L2 designation for your level of taxonomy. For example L2 or level 2 represents the Phylum level of identification for the microbes found in your environmental samples. L3 or level 3 represents the microbes identified to their Class. L4 represents your data identified to the Order, L5 to the Family and L6 to the Genus level.

The “taxa\_summary\_plots” folder at the bottom of this window contains html documents describing your dataset output.

For our work today, we will focus on the Genus level L6 .txt file. This is a tab delimited text file that can be opened in a number of different applications. To open this file in Microsoft Excel, right click on the file name with your mouse cursor. Select the second command option from the list called Open With, and select Microsoft Excel from the selection window that appears.

In most versions of Excel, the tab-delimited feature of these text files will automatically parse your data into separate columns. If your version of Excel does not immediately recognize the tab-delimited format, all of your data will display in the first column. You can remedy this by selecting your data column, clicking on the Data tab at the top of the Excel window, and selecting the Text to Columns tool from the menu bar.

In the window that pops up, choose the Delimited selection and click Next. In the second window, select the Tab option, and click Finish. Your data should now be separated into several columns.

Please now look to cell A1 of your table. You will see an informational note from the qiime software pipeline that this text file was constructed from the biom bioinformatics file. In cell A2, you see the column header #OTU ID, which stands for Operational Taxonomic Unit Identification. Widen column A by double clicking on the letter A at the top of the column or by using your mouse cursor to drag the right edge further to the right.

Taxonomy for each OTU in your dataset is reported for every level—k for kingdom, p for phylum, c for class, o for order, f for family, and g for genus—separated by semi-colons. The qiime bioinformatics pipeline relies on extensive microbial databases to report taxonomy for each OTU, but some of the OTUs in your dataset may not be defined yet. This column will reflect the most recent microbial identification available to date.

Please now look to cell (pick one) for an example of OTU identification. This OTU is assigned to the phylum (???) and class (???). You can now perform a literature search on this phylum and class assignment to learn more about the organisms from this OTU within your sample.

In order to successfully start using commands in Excel for this data table, you must select row 1 with your mouse cursor. Right click on the selected row, and select Delete from the options menu. The header of each column in your data table should now occupy row 1.

To summarize, you now have taxonomy information listed for each OTU in column A, and relative abundance data is displayed for each of your samples in the subsequent columns of your spreadsheet.

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