



## Student Handout for DNA Subway Purple Line (alpha testing documentation)

*You can also watch a video tutorial of the steps outlined here on our [YouTube Channel](#)*

### What is DNA Subway?

Metabarcoding uses high-throughput sequencing to analyze hundreds of thousands of DNA barcodes from complex mixtures of DNA. In a typical experiment, DNA is isolated from sterile swabs or material taken from different environmental locations or conditions. PCR is then used to amplify a variable region of the DNA, such as the 16S ribosomal RNA gene. The amplified DNA is then sequenced via high-throughput methods that use metabarcodes to differentiate between samples. Finally, specialized software processes the sequence data and identifies the different taxa (species or higher level taxa) present as well as their abundances in the different samples.

DNA Subway is an online platform with the capability to analyze different types of genetic data. The “Purple Line” of DNA Subway implements a simplified version of the QIIME 2 (pronounced “chime two”) workflow to analyze data microbiome data (e.g. 16S ribosomal RNA gene sequences). Using the Purple Line, you can analyze high throughput sequencing reads to identify taxa in microbial or environmental DNA samples.

The QIIME2 analysis steps performed by the Purple Line are outlined in the diagram below. Once the sequences are analyzed, the results can be visualized to allow comparisons between samples and different conditions summarized in the metadata.

## Goals and Learning Outcomes

In this tutorial, you will use DNA Subway to learn how to:

1. Process your raw sequence data to obtain high-quality sequences.
2. Obtain a table of taxonomic and abundance information for the different types of bacteria observed in each bean beetle sample.
3. Interpret graphs that describe various aspects of sequence analysis



### Before Getting Started, make sure you have the following:

1. Have a valid CyVerse account.
2. Provided your CyVerse username to your instructor so they may share the **fastq.gz sequence files** from your experiment with you in CyVerse.

## Important note on sample data shown in this tutorial

This tutorial uses a sample dataset that only contains four samples. These sample data come from DNA extracted from bean beetles fed on two different kinds of diets: Blackeyed Peas (BEP) and Adzuki Bean (ADZU). There are two beetles from each bean host, diet, in this Demo-Data file.

Since the values and settings shown in each step in this tutorial will be specific to this sample dataset, you will need to input different values when working with your own dataset that you generated from your own experiments.

In this tutorial we will walk you through how to determine the appropriate settings for your dataset.

## Important note about time required to complete analysis


For each step in the analysis, you will see a note describing the expected run time for that step. However, this is only an estimate. The actual time required to complete the DNA Subway analysis depends on several factors, including, but not limited to:

- Number of samples being processed
- File size
- Number of users logged into the CyVerse servers at any given time
- Cyverse or DNA Subway website maintenance

Therefore, it is essential that you give yourself ample time to complete your analysis. This may be anywhere from 1 day to 1 week. It is unlikely that analysis would take longer than a week, but if you experience difficulties, please contact the DNA Subway website hosts. Their contact information can be found on the bottom of the DNA Subway homepage.

## Starting an Analysis

### *Create a Microbiome Analysis Project*

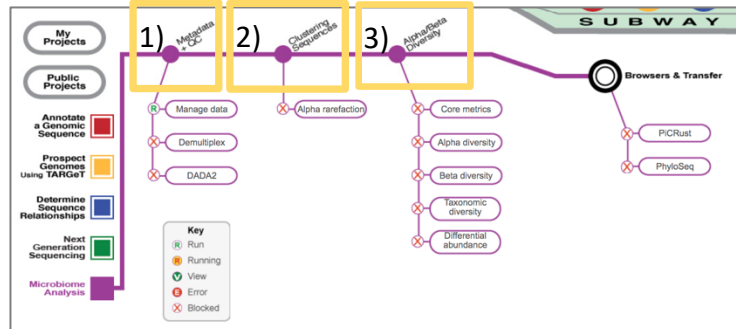
- Log-in to DNA Subway at **dnasubway.cyverse.org** using your CyVerse username and password.
  - (Note: DNA Subway works best with Chrome or Firefox browsers) 
- Click the purple square titled **Metabarcoding Analysis**, to begin a new project.
- You will be taken to a page where you need to select a few options:
  - **Select Project Type:**
    - check **Paired End Reads**
  - **Select File Format:**
    - check **Illumina Casava 1.8**
  - **Name your Project:**
    - **Project Title:**
      - Type in a title for your project, for example **"My First Microbiome Analysis"**
  - **Description:**
    - Type in a brief description of your analysis. You can type in your groups name, number of samples to be analyzed, and the experimental treatments. For example:
      - **"This analysis compares the microbiomes of bean beetles fed on adzuki bean or blackeye pea diets."**
- Once you are finished setting up your project, click on the **Continue** button

## DNA Subway Home Page

After clicking continue you will be taken to the DNA Subway Home Page. This is where you will navigate through the steps of QIIME2. It is a good idea to familiarize yourself with a few things before getting started:

The analysis is done in 3 main steps:

- (1) **Metadata + QC**
- (2) **Clustering Sequences**
- (3) **Alpha/Beta Diversity**



Each main step has several processes below them that need to be performed. You can take a look at the diagram at the end of this tutorial (**Conceptual Overview of QIIME2 and DNA Subway**) to get a more detailed overview of the main QIIME2 processes being performed during each of these steps.

Also notice that each process on the DNA Subway home page has a symbol next to it. Each symbol represents the current status of a given process.

- The red X in a white circle means that you are not able to start this process yet.
- The green R in a white circle indicates a process is ready to be started.
- The red R in a yellow circle indicates a process that is currently running.
- The white V in a green circle indicates that the process has finished, and there are graphs or data plots that are ready to be viewed.
- The error symbol indicates that a major problem has occurred, and the process has stopped running and did not complete successfully. If you receive this error, you will likely not be able to proceed to subsequent steps in the analysis. In such a case you may need to contact the developers to help troubleshoot the error. You can contact them by navigating to the **Contact Us** link on the bottom of the home page.

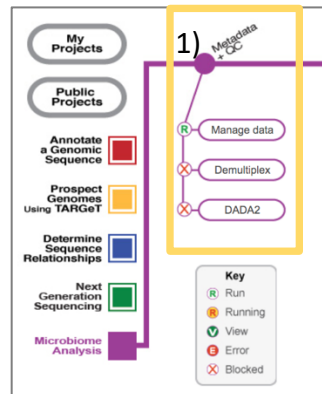


## Step 1: Metadata + QC

There are three sub-processes in this step where we will A) upload our files (manage data), B) Demultiplex the reads, and C) use DADA2 to perform initial quality checks (QC) on the data.

### **Manage data**

- Click on **Manage Data**
- A new window will appear where you will need to upload your **fastq.gz** files. You will also be able to create a **metadata.tsv** file.
- To upload **FASTQ** files
  - *click on the folder titled **Shared with me**.*
  - *Type the CyVerse username of the person who shared the datafiles with you in the search bar (your instructor's username) and click search. Click on the full name that appears below the search bar and the CyVerse directory of the person who shared files with you will appear.*
  - *Go to the folder containing the dataset shared with you and select the **fastq.gz** files in that folder. There will be two fastq.gz files for each sample in the dataset. The Demo-Data shown in this tutorial has 4 samples so there would be 8 fastq.gz files in that dataset.*
  - *After your files have been uploaded, it should return you to the Manage data window so you can create your metadata file.*
- To create **Metadata** files
  - Click on **Create new**
  - Wait a few seconds and the metadata file will be automatically created for you. Check that all the data files are correct, then click **Save**.
  - After saving the metadata file, click **Back**.
  - You will now see a metadata.tsv file. In this section you will also see three options: **Edit**, **Validate**, and **Rename**. Click on **Validate**
  - If anything is wrong with your metadata file, you may see a warning message (in yellow text) or an error message (in red text). The messages will let you know that there is something you need to fix in your metadata file.
  - If you get a warning or error message, consult with your instructor.
- Once the files have been validated, and you see no errors or warnings, click on **Run**.
- After clicking **Run**, you should be returned to the DNA Subway Homepage



- You are now ready to proceed to the next step: **Demultiplex Reads**

### ***Demultiplex Reads***

During sequence analysis, **QIIME2** takes the separate **fastq** files, representing sequences from our different samples, and analyzes them all together at the same time.

**Demultiplexing** refers to the step where the **QIIME2** software sorts through all the sequences and labels them with the appropriate information to keep track of which sequences came from which samples.

- Click **Demultiplex reads**
- **Random sequences to sample for QC**
  - This step will check the quality of 10,000 sequences and then calculate an average quality score for each nucleotide position across all sequences.
  - You can choose to increase or decrease the number of sequences that will be quality checked. For the purposes of this tutorial, you can leave the default value of 10,000.
- Then click **demux reads**

*Run time:* Depending on the size of your dataset, this step will likely take between 20 minutes up to a couple of hours. The process will continue running even if you close your browser, so it is okay to close it and work on other things while you wait. Come back to the periodically to check the status of this step.

- When demultiplexing is complete, click on the link titled **Demultiplexing Summary**
- A new window will appear that says **QIIME2view** in the upper corner. Expand this window to view it on the full screen. You can also scroll up and down to view the entire contents of the page.
- You will notice that the window has two tabs. The first tab is the **Overview** tab, which provides summaries of the number of sequences and samples processed during the demultiplex step.

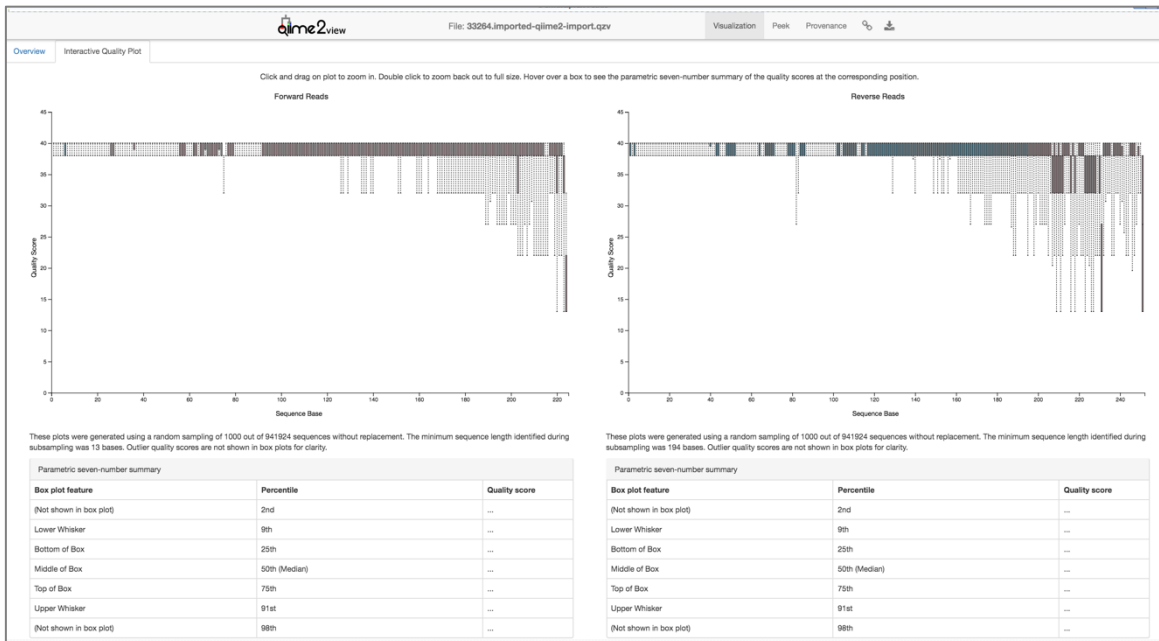
### Description of the **Overview** tab

- The overview tab shows you the number of sequences contained within each sample. It shows you this information in three ways:
  - **Demultiplexed sequence counts summary** table
  - **Histogram**
  - **Per-sample sequence**

- The **Demultiplexed sequence counts summary** is a summary table that provides information regarding the total number of sequences that were demultiplexed at this step.
  - Based on the sample **Beetle\_Diet\_Microbiome** dataset, there were a total of 637,927 sequences (total from 4 samples) that were sorted and grouped during this step.
  - The *Minimum* value represents the sample with the least number of reads (149,943).
  - QIIME 2 also calculated the *Median* and *Mean* number of sequences across all 6 samples as well as the sample with the most sequences (169,742), which represents the *Maximum*.
- The **Histogram** is a type of bar chart that summarizes the distribution of the sequence data. It does this by grouping the numbers of sequences per sample into ranges (x-axis). The height of each bar shows how many samples fall into each range of sequence counts. Ideally, we would want to have the same number of sequences in all of our samples (although this rarely happens). The reason we want this will be explained in subsequent steps.
- Finally, **Per-sample sequence counts** table, found below the histogram, shows the exact number of sequences observed in each sample.

#### Description of the **Interactive quality plot** tab

- This tab provides an interactive plot that allows us to assess the overall quality of the sequences that were demultiplexed
- There is a lot going on in these plots, so let's walk through them slowly.
- You will see box-plots for both the forward and reverse sequences titled Forward Reads and Reverse Reads. Another name for sequence is "read," which is the term being used in these plots.
- The axes of the box-plots are labeled **Sequence Base** (x-axis) and **Quality Score** (y-axis). Essentially, what is being shown here are box-plots for the *average* quality score for each nucleotide-base position along a sequence calculated from 10,000 randomly chosen sequences from our dataset.
- The x-axis ranges from 0 to about 250 because 250 bases was the maximum sequence length observed in the 10,000 randomly selected sequences that were used to calculate the average quality score for each base.



- The y-axis shows ranges from 0 to 45, representing the range of possible quality scores for each nucleotide base.
- In general, a quality score above 30 is considered high quality, meaning that you can have high confidence that the nucleotide base was called correctly during sequencing. Scores above 25 are considered good/acceptable. Any score below 25 is considered low quality, meaning you cannot be confident that the nucleotide called during sequencing was correct.
- Let's take a closer look at our Forward Reads.
- For our forward reads, the plot shows that on average, the Quality Score for our base-calls are above 30, indicating that for the most part, they are high quality reads. You may notice that there are a few outliers, the first one showing up around Sequence Base 79. The further down the read (along the x-axis) we travel, the more outliers appear. Once we reach Sequence Base ~190, the Quality Score of the outliers fall below 30. You will also notice that the inter-quartile spread for the box-plots starts to get wider after Sequence Base 200.
- What we are trying to assess for our data is the point along the reads where the majority of the base calls fall below a Quality Score of 30. In order to do this, we need to zoom into our data and examine the **Parametric Seven-Number Summary** below the box-plots.
- Click and drag on the plot (towards the later points of the X-axis) to select a group of points to examine more closely.
- Once you have zoomed into the plots, you can better examine the quality scores of each Sequence Base.

- The **Parametric seven-number summary** shows you the inter-quartile values for the box-plots of a specific sequence base. Click on any base you would like to view summaries for in the table below the plot and examine the quality scores at the 25<sup>th</sup>, 50<sup>th</sup>, and 75<sup>th</sup> inter-quartile ranges.
- So why are we looking at the quality of base calls so closely? Well, later in the analysis we will remove, or trim-off, the low-quality portions of our reads so that our analyses are done on only the high quality portions (the portions of our sequences that we can be confident were correctly sequenced). At this step, we are determining the specific point along our reads that we will trim-off later.
- Choosing the exact location to trim-off is tricky and subjective. In general, trimming at locations where **the bottom 25<sup>th</sup> percentile of quality scores** fall below a value of 30 while also keeping as many of the bases as possible is recommended.
- For your data, you may need to examine several bases before you identify position along the reads where the bottom 25<sup>th</sup> percentile of quality scores fall below a value of 30 that does not result in a sequence that is too short (e.g. below 150bp).
- For our sample dataset, we will trim our forward reads to **224bp**, which is the point along our reads where the bottom 25<sup>th</sup> percentile of quality scores falls below 30.
- Once we have chosen the exact location that will be selected for the trim step, we need to do the same for the Reverse Reads.
- After examining several Sequence Base positions along the X-axis, it seems that position **231** is the earliest point at which the Bottom 25<sup>th</sup> percentile falls below Q30. So, that is the trim value we will select for the reverse reads in the next step.
- Once you are done, close out of the QIIME2 Viewer window and return to the DNA Subway home page.

## **DADA2**

The next step in the DNA Subway pipeline is performed using **DADA2** (Divisive Amplicon Denoising Algorithm). **DADA2** “filters” out poor-quality sequences by trimming off the portions that have low quality base-calls. It also identifies and removes reads that were likely created by amplification or sequencing errors (such as chimeric sequences). Finally, it pairs (or “joins”) the forward and reverse reads.

The end result of all of this filtering is a set of high-quality representative sequences that we will use for further analysis.

- Click on **DADA2** .
- You will see 4 boxes where you can input numerical data
- Enter the following:
  - **Forward:**
    - `trimLeft=0`
    - `truncLen=224`
  - **Reverse:**
    - `trimLeft=0`
    - `truncLen=231`
- Then click on **Trim reads**

Forward trimLeft: 0	truncLen: 224
Reverse trimLeft: 0	truncLen: 231

*Run time: Depending on the size of your dataset, this step will likely take the longest time to complete. For this sample dataset with just 6 samples, it should take ~15 minutes as long as the server is not too busy. Refresh the page periodically until you see the view symbol.*

Once you see the view symbol (white V in a green circle) next to **DADA2** on the home page, you know that the process is complete.

- Click on **DADA2** to view the output of this step.
- You should see three links under the results section:
  - **Trim table**
  - **Stats**
  - **Representative sequences**
- The most important information we need to complete the analysis is found under **Stats**, so let's first take a there.

- After you click on **Stats** , you will see the following table:

The screenshot shows the qiime2view interface for file 51207.stats.qzv. It includes a 'Back' button, a 'Download metadata TSV file' button, and a note that the table below may not reflect dynamic sorting or filtering. The table displays the following data:

sample-id	input	filtered	percentage of input passed filter	denoised	merged	percentage of input merged	non-chimeric	percentage of input non-chimeric
ADZU.1	157687	146639	92.99	146156	102924	65.27	96210	61.01
ADZU.2	169742	157302	92.67	157025	110818	65.29	103529	60.99
BEP.1	149943	140106	93.44	139729	126209	84.17	118359	78.94
BEP.2	160555	150481	93.73	148750	141970	88.42	137063	85.37

Showing 1 to 4 of 4 entries

- This table shows us how many sequences were removed, or filtered out, either because they contained too many incorrectly called bases, were chimeric sequences, could not be properly joined during the merge step, or were otherwise low-quality reads.
- The final column, titled “**non-chimeric**”, contains the final sequence counts for our samples, which represent the high-quality sequences that survived the filtering process.
- Although we have lost many sequences during the filtering process, it is okay because we do not want any low-quality or noisy data messing up our future analyses.
- When analyzing your own data, you want to pay attention to how many sequences were lost during the analysis. In some cases, some samples may have been of such poor quality that most (sometimes all!) of your sequences are gone. For example, if the majority of your sequences were removed in all of your samples, that might be an indication of a poor sequencing run.
- You will also want to record the number associated with the sample that has the lowest number of sequences. For our dataset, ADZU.1 has the lowest total sequence counts with **96,210**. The lowest total sequence count only should consider actual samples. If your dataset included a negative control do not use the negative control sequence count.
- Find the sample with the fewest sequences in your dataset, and record the number of sequences in your lab notebook.
- You can also download the table with all the information by clicking on the **Download metadata TSV file** button.

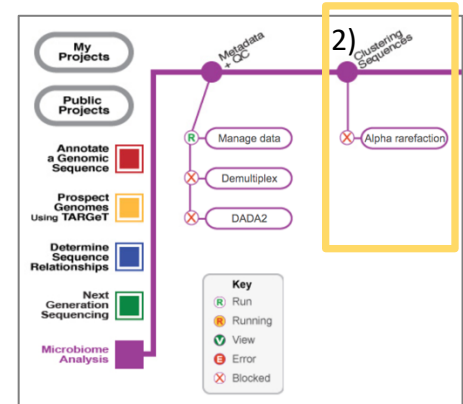
- Once you are done viewing the contents of this window, close out the window, which should return you to the home-page.
- Then, proceed to **Step 2: Clustering Sequences**

## Step 2: Clustering Sequences

At this step, you will begin make initial assessments about the **diversity** of the samples within your dataset.

### Alpha rarefaction

Alpha rarefaction introduces the concept of **sequencing depth**.



Sequencing depth refers to the number of sequences that were obtained in each of your samples. Take a closer look at the **stats** table that resulted from the **DADA2** step. You will notice that some samples have many more sequences than others. There are many different reasons why some samples may have more sequences than another sample, such as differences in yields of PCR amplification and DNA extraction between samples. While it is normal for different samples to have different number of total sequences, it does make things tricky when trying to interpret **diversity** of microbiome data.

It is important to keep in mind that the total number of sequences in a sample, or **sequencing depth**, will greatly affect the number and types of bacteria we observe. It will also affect the conclusions we make about the “diversity” of our samples. Samples with greater depth (samples with many sequences) tend to show higher diversity, while samples with lower depth (samples with fewer sequences) have lower diversity. However, that may not be representative of the true diversity observed in the samples, but instead may simply be a consequence of the differences in total sequences between samples.

Historically, the approach that has been used to adjust for differences due to sequencing depth is to perform **rarefaction** on your samples. Rarefaction involves random subsampling of a defined number of sequences from each sample. The result is that each sample has the same number of sequences. The resulting diversity that is observed based on those randomly subsampled sequences can be more easily compared.



We will revisit the concepts of sampling depth and rarefaction again during the **Ranacapa Tutorial**. For now, we will proceed to the **Alpha Rarefaction** step in DNA Subway, which will perform the random subsampling based on a value we specify.

### ***Finding a good value for rarefaction***

Finding a good value for the rarefaction step can be tricky, so tricky in fact, and there is considerable debate in the literature about if and how rarefaction should be performed. However, for our purposes, we will go ahead and perform rarefaction so we can move through the analysis. In the end, the final data table we need will not be affected by this rarefaction step.

For our sample dataset, we identified a value of **96,210** sequences, which was associated with our smallest sample. That is a decently high value, so, we will input **96210** for the sampling depth. If your dataset included a negative control, do not use the negative control sequence count for rarefaction.

---

**Parameters**

<b>Trim</b>
TrimLeftF: 0
TruncLenF: 224
TrimLeftR: 0
TruncLenR: 231
Metadata: metadata.tsv

Min. rarefaction depth:

Max. rarefaction depth:

- **For your data, you will want to input the value associated with your smallest sample not including your negative control.**
- Then, click on **Submit job**

*Note: Depending on the size of your data, this step could take a few to several minutes. It may take longer than the “demultiplex” step, but should be shorter than “Trim reads” step. Refresh your page periodically until you see the green “V” symbol next to this step.*

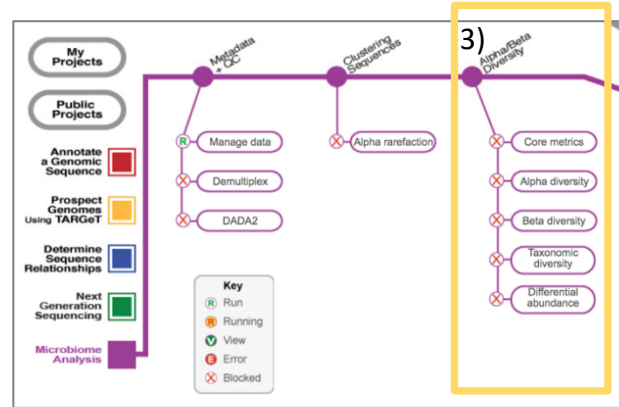
When the step is done, proceed to the **Step 3: Alpha and Beta Diversity**

## Step 3: Alpha and Beta Diversity

This step is meant to examine *Alpha Diversity* (the diversity of species/taxa present within a single sample) and *Beta Diversity* (a comparison of species/taxa diversity between two or more samples).

Alpha diversity answers the question - “how many different species are in a sample?”;

while beta diversity answers the question - “how similar/different are species between samples?”.



- Click on **Core metrics**
- For sampling depth, select the x-value where rarefaction plots begin to plateau. If that value is greater than 20,000, then type in **20,000**, as that is the maximum allowed value for the Core metrics step.
- Select **Greengenes (515F/806R)** as the classifier from the drop-down menu.
- The click on **Submit job**.

This step may take several minutes to complete. Refresh your page until you see the view symbol.

When this process is finished, you should see view symbols on the **Core metrics** and the three subsequent steps (**Alpha diversity, Beta diversity, and Taxonomic diversity**)

Although this step calculates Alpha and Beta diversity, we will not look at these results in DNA Subway. Instead, we will cover Alpha and Beta diversity in more detail using different software (**RShiny App** described in the RShiny Bean Beetle Microbiome handout).

So, for now, let's skip directly to the **Taxonomic Diversity** tab.

## ***Taxonomic Diversity***

- Click on **Taxonomic diversity**
- Click on **Bar Plots**
- The first plot you will see is a taxonomy stacked bar plot.
- Go to the **Taxonomic level** category and select **Level-5**. This level will show us the taxonomy of our sequences at all levels down to Family. We use level 5 because sequencing just one of the variable regions of the 16s gene (v4, in our case) does not provide sufficient data to be able to identify many of the bacteria to the genus or species level.
- Then, go to the **Download** section, and download the data as a .csv file
- This file will be saved to your computer.

## ***Final Step***

This level-5.csv file is your **taxonomy file**, or the data table that you will need for all future community analysis. Make sure you save this file somewhere where you can access it for your future.

Although there are other steps you can perform on DNA Subway, for now we will stop here since this step provides us with the taxonomy file we need for further analysis.

Now that you have downloaded your **level-5.csv**, you can proceed to the handout:

**RShiny Bean Beetle Microbiome.**

*This tutorial was developed by Anna Zelaya, PhD for the Bean Beetle Microbiome Project, supported by National Science Foundation Grants, DUE-1821533, and DUE-1821184 to Morehouse College and Emory University, Atlanta, Georgia, USA. Any 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. Revised by L. Blumer November 2023.*