Note: this tutorial was designed for a 454 dataset, but the scripts and arguments should be identical. Connect with Ashley if you have issues executing the script. 15 aug 2014.

4. Exploratory rarefaction (to determine an appropriate even depth of subsampling)

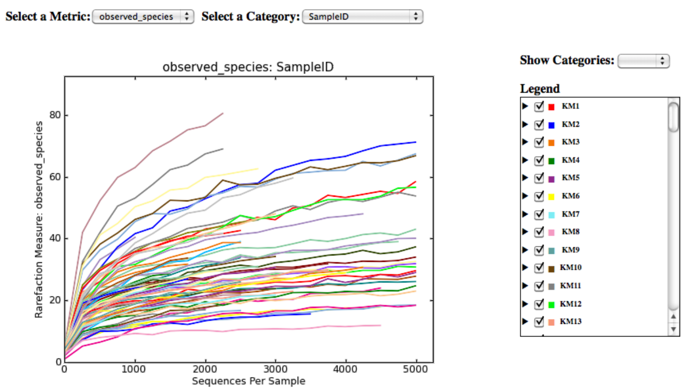
Before we start any ecological analyses, we want to evenly sample (rarefy) all the samples to an even number of sequences so that they can be directly compared to one another. But, before we rarefy the OTU table, we need do make a decision about what the best number of sequences will be for rarefaction. As a rule, we must rarefy to the minimum number of sequences per sample. But, looking at the dataset, there is at least one outlier sample (KM43) that has far fewer sequences than the others - only 148 when compared with the average of 5188. Though there may be a wide range in sequencing success among samples, the very low number of sequences in KM43 in likely due to an unknown error in the method, and not for a biological reason. We could remove this sample and rarefy to an even depth of 861 sequences rather than 148. But, it could be that the caterpillar gut is so simple, that 148 sequences is quite sufficient to well-sample all of the diversity, and so it wouldn't actually compromise the amount of information. The only way to check is to make some rarefaction curves to and look for the asymptote.

For this exploration, we set the -e argument (which sets the upper limit for rarefaction depth) near the mean that was observed in the dataset (from the previously executed `per_library_stats.py` command). This should provide a comprehensive range over which to compare all samples, and to figure out to what depth we want to evenly sample the final OTU table. This will take a little bit of time to complete. We also set the -n argument (which sets the number of rarefaction steps evenly spaced between the minimum (default = 10) and the maximum (we set it to 5000). Thus, 20 subsamplings between 100 and 5000 sequences will be conducted for every sample

```
$ alpha_rarefaction.py -i Manduca_otu_table.biom -o alpha_rarefaction_max5000/
-t Manduca_rep_phylo.tre -m Manduca_map.txt -e 5000 -n 20
```

Through your computer's file browser folders, navigate into the new alpha_rarefaction_max5000 directory, then the rarefaction_plots directory, and then and double click on rarefaction_plots.html. Your web browser will open and there will be to drop-down menus at the top of the page. Select "observed_species" (no. OTUs = richness) from the left dropdown menu, and "Sample" from the right. Notice that the right drop-down menu includes all of the categories that were specified in the mapping file.

Inspect the plot:


Select a Metric: observed_species Select a Category: SampleID

Show Categories:

To what depth should we rarefy our samples in this dataset? Remember - we want to maximize the number of samples we use (if a sample has fewer sequences than our chosen rarefaction depth, we will omit it from our analyses, thereby forfeiting any information that it offers), AND maximize the number of sequences that describe the communities as comprehensively as possible.

Toggle different options for each menu, select and de-deselect samples/barcodes, etc. on the legend of the right hand side. Take a few minutes to explore the data. Then, I've prepared some alpha_rarefaction at different a depth near the maximum depth (30,000) for in the analyzed data folder - also explore these plots one if you like.

Explore the new alpha_rarefaction directory and its subdirectories. The subdirectory "rarefaction" contains individual .biom OTU tables for each of ten replicates at every subsampling depth. The number of sequences is given first in the title, and then the replicate. For example, the file called "rarefaction_259_8.biom" indicates an even subsample of 259 randomly selected sequences from each sample, replicate 8 of this depth. The subdirectory alpha_div_collated gives the final calculations of observed_species (richness), chao (chao estimator), and PD (Faith's phylogenetic diversity) for each depth. These files are useful for plotting rarefaction curves outside of QIIME. The subdirectory alpha_rarefaction gives files of calculations of observed_species, chao, and PD for each depth, each rep - essentially the information that is summarized in the collated folder.

Note: In the alpha_rarefaction subdirectory, these files aren't really "biom" files though they have the .biom extension - they are not OTU tables, and the extension is confusing.

5. Rarefaction to one depth for the final OTU table

***

Important decision: What is the appropriate rarefaction depth?

***

Because the samples seem to be approaching sequencing saturation at low sequencing depths, we will omit only the very low sample outlier, KM43, and rarefy to the new minimum (KM23), which is 861 sequences per sample.

```
$ single_rarefaction.py -i Manduca_otu_table.biom -o
Manduca_otu_table_even861.biom -d 861
```

Summarize the *final* OTU table:

```
$ per_library_stats.py -i Manduca_otu_table_even861.biom

py -i Manduca_otu_table_even861.biom
Num samples: 76
Num otus: 389
Num observations (sequences): 65436.0

Seqs/sample summary:
 Min: 861.0
 Max: 861.0
 Median: 861.0
 Mean: 861.0
 Std. dev.: 0.0
 Median Absolute Deviation: 0.0
 Default even sampling depth in
  core_qiime_analyses.py (just a suggestion): 861.0

Seqs/sample detail:
 KM1: 861.0
 KM10: 861.0
 KM11: 861.0
 KM12: 861.0
 KM13: 861.0
 KM14: 861.0
 KM15: 861.0
 KM16: 861.0
```

Notice that now we are down to 76 samples total - if you scroll down and look for our previous low-sequence outlier (sample KM43 with 148 sequences) you will not find it. Because we rarefied to 861 sequences, and samples that don't have minimum that number of sequences are automatically omitted.
Note: there are also other rarefaction scripts available through QIIME, like multiple_rarefactions_even_depth.py. Explore these and check out the other options available.