Project Descriptions

Students may have options of A) doing microbiome analyses, B) benchmarking metagenomic tools or doing primer design, or C) analyzing a time-series metagenomic dataset.

A) For each project, a student will pick a dataset and a particular metagenomic software package. Not only will students have to analyze the data, but they will have to do a review of a particular metagenomics software package and discuss its advantages and disadvantages. The students will be graded on how much analyses they were able to complete as well as the package review.

Each student will focus on a particular package, but if that package could not complete a particular analysis, it is expected that the student install and/or use other code to complete the analysis. In the final presentation, students will review what their particular package was able to complete and what it could not.

For 16S data, at the MINIMUM: the taxonomic composition (relative abundance of different taxa in the sample) must be assessed. Diversity and multidimensional scaling must be performed for visualization. Statistical tests to determine the significance of the differences must be performed. Also, the potential function of the samples must be assessed using Picrust. Feature selection should also be applied to determine discriminative species between different types of samples (e.g. healthy vs. disease).

For metagenomic data, at the MINIMUM: the taxonomic composition (relative abundance of different taxa in the sample) and function (abundance of OGs, protein families, and metabolic pathways) must be assessed. Diversity metrics and multidimensional scaling, differential metabolic pathways from the data. Assemblies of the highest abundant taxa (when desired). Statistical tests showing significance. Extra analysis is encouraged for any futher analysis that the package offers. Feature selection should be used as an approach to determine any discriminative species/genes/protein families/metabolic pathways between different types of samples (e.g. healthy vs. disease).

For metatranscriptomic data, at the MINIMUM: the transcripts/genes must be assembled via a de novo assembly approach. Taxonomic and functional (protein families, orthologous groups or pathways) composition of the assembled genes must be assessed. Diversity metrics of gene content must be evaluated. Identify genes that are differentially expressed between different types of samples (e.g. control or disease). Evaluate the functional implication as revealed by the functions/pathways making up of the differentially expressed genes.There will be X datasets to analyze:

1. One dataset will be 16S rRNA data from mice that were given a high-fat diet (HFD) vs. LFD.

For metagenomic datasets, students will pick from the following software packages:
- Biobakery (https://bitbucket.org/biobakery/biobakery/wiki/Home)
- MG-RAST (https://metagenomics.anl.gov/)
- MEGAN (http://ab.inf.uni-tuebingen.de/software/megan5/)
- KBase.us
- STAMP (http://kiwi.cs.dal.ca/Software/STAMP)
- smashcommunity (http://www.bork.embl.de/software/smash/)
- PhyloSeq/dada2 (if it's 16S)

2. Metagenomics of an Urban microbiome.
3. Metagenomics of Leaf Microbiome.
4. Perhaps an algal dataset?

For metatranscriptomic datasets, Oases/Cufflinks and then some of the packages above.

5. A metagenomic and/or RNA dataset from ant guts

B) Benchmark Metagenomic Tools or Primer Design

i) Design primers specific for different marker genes and ant-specific gut microbiota from different taxa
ii) Benchmark PhylopythiaS+ and MetaPhyler on datasets. Use results from many more algorithms to see if you can design a ensemble of classifiers to do better than the original. (Success in this project will result in your name on a journal publication).
iii) See how database size affects different taxonomic classifiers.

C) Time-series data

We will be analyzing the time-series data from http://genomebiology.com/2014/15/7/r89 , which is a publication about 2 subjects who took samples of their gut and saliva every day. I have put the data HERE. Using the data, can you uncover periodicities of the data? Can you model trends in the data? Can you find correlations of species with the metadata? Can you use autoregressive moving average (ARMA) techniques to characterize weeks or seasons? Are the weekdays different from the weekends? with Which species seem to co-vary? Is there correlation between saliva and gut

or is there correlation between samples and metadata? Etc. etc. Find the data in /mnt/HA/groups/nsftuesGrp/project_data/David-timeseries