# Tutorial Descriptions

Each tutorial is expected to be at least 30 minutes. Tutorials requiring more work will be split up into teams rather than single people (this is TO-BE-DETERMINED).

## Tutorials 1 and 2: Phylogenetics through High Performance Alignment and Tree Construction

**Tutorial 1: Alignments:** Students will align H. Influenzae 16S genes using Cipres (http://www.phylo.org/). Students will compare **Muscle** and **MAFFT** alignments, comparing time and using SeaView (or another visualization tool) to examine the alignment.

**Tutorial 2: Tree Inference:** Then students will build the trees using **RaxML** and **FastTree**. The students will comment on differences in the trees due to the different methods and biology.

In these tutorials, the students will be expected to dedicate a large portion of the class to how the underlying algorithms work for both alignment and tree inference. An analysis of the pros and cons of each of the 4 methods mentioned above are expected.

*Instructions for the Datasets:*

*16S rRNA gene for Haemophilus influenzae:*
*/mnt/HA/groups/nsftuesGrp/data/Haemophilus_influenzae_16S.fasta*

*GlnS for Haemophilus influenzae:*
*/mnt/HA/groups/nsftuesGrp/data/Haemophilus_influenzae_GlnS.fasta*

## Tutorials 3 and 4: Comparative Taxonomy using 16S rRNA

(Suggested Reading: Numerical Ecology by Legendre and Legendre (to be posted on Bblearn, email me if you cannot find it)
**Tutorial 3: Diversity, Scaling, and Visualization**: Students will review alpha/beta/gamma diversity metrics for the class (please make sure that you also cover Simpson, Chao, Unifrac and Bray-Curtis in this introduction). Students will be expected to use the VEGAN package in R to implement these diversity measures on either the Guerrero Mat samples or the Human Microbiome Project (students can process use any OTU clustering/classification method or even use pre-processed methods) before importing into VEGAN. Then, please do metric (and non-metric) multidimensional scaling (http://www.r-bloggers.com/7-functions-to-do-metric-multidimensional-scaling-in-r/ and

http://strata.uga.edu/software/pdf/mdsTutorial.pdf ). Note that there is an HMP metadata file with mapping to body site that may be interesting to note on your ordination. Students will verify ordination using Unifrac distance matrix to a Bray-Curtis distance matrix using Procrustes analysis and will get the confidence of the distance matrix using ANOSIM (don't have to review ANOSIM).

**Tutorial 4: Canonical Analysis**: Students will review Canonical Correspondence Analysis, Redundancy Analysis, and distance-based RDA. Students will show examples of how these methods can be applied to the Guerrero Microbial Mat, since this is a stratified sample with approximately continuous explanatory variables. In addition, students will review statistical hypothesis testing including PERMANOVA and ANOSIM.

*Instructions for the Datasets:*
*HMP Data:*

*/mnt/HA/groups/nsftuesGrp/data/HMP*

*Guerrero Negro Microbial Mats:*
*http://www.ncbi.nlm.nih.gov/bioproject/29795*http://www.ncbi.nlm.nih.gov/bioproject/29795
*Click on List all 10 'Metagenome' projects...*

*Let's download data for Microbial Mat 01 (you will have to repeat this procedure for the other nine):*
*Click on PRJNA29605*http://www.ncbi.nlm.nih.gov/bioproject/29605
*Click on 10530 in Genomic DNA row*
*Click on "Display Settings:" (upper left)*
*Under Format, Select "FASTA", hit "Apply"*
*Click on "Send:" (upper right)*
*Under "Choose Destination", click "File"*
*it will get downloaded as sequence.fasta which you should rename.. I guess mat01.fa*

## Tutorials 5 and 6: Assembly and Binning

**Tutorial 5:** D**e novo assembly:** Students will review the difference between read mapping and *de novo* assembly. Use at least three samples from an infant gut microbiome dataset to demonstrate a showcase of assembly with IDBA package.
***read mapping to assembled contigs*:** student will cover this as an essential step for GroopM binning.

- map reads from each sample to the assembled contigs using Bowtie2

**Tutorial 6: Binning Contigs: S**tudents will discuss the underlying assumptions for binning contigs (why do we perform binning process and based on what traits do we group
the contigs). Students will use Bowtie2 to map reads to assembled contigs using Bowtie2 (both tutorial students can collaborate on this step). demonstrate binning with GroopM. Compare the binning result in comparison to the original paper which used Bowtie for read mapping and ESOM for binning (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3530670/table/T1/) and to the GroopM paper which uses BWA for read-mapping and GroopM for binning (https://peerj.com/preprints/409/). The way to do this comparison is by:

- Getting the binning of contigs using GroopM
- Extracting bins with high completeness as reported by GroopM (based on marker genes presences)
- Annotating the taxonomy for these more complete bins (BLAST) and see if they match the identity of complete genomes recovered in the original paper.

Instructions for the Datasets:

- Acquire microbiome reads from NCBI site (in SRA format) ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByExp/sra/SRX/SRX144/SRX144807/
- Use SRA toolkit to convert sra files to fasta/fastq (will be covered in class on week 3)
- Use IDBA-UD for assembly (at least 3 samples altogether)

## Tutorials 7 and 8: Taxonomic Classification and de novo clustering and phylogeny

**Tutorial 7: Taxonomic classification**: Students will review recent methods in bioinformatics for predicting the taxonomy collected from whole genome shotgun (WGS) sequencing runs, and the students should highlight the differences between clustering, alignment, and classification of sequences. While older methods use these techniques to label every sequence, newer techniques presented in this section use “tricks” to circumvent labeling everything. . It is expected that the students compare and contrast the differences between the approaches, WGSQuikr, MetaPhlAn and Phylosift.

- **WGSQuikr**: Students will discuss how WGSQuikr how it relates to compressive sensing. Use WGSQuikr (using the Greengenes 94 database) on all available samples in the infant gut dataset (way to get this data is provided for tutorials 5 and 6)..

- MetaphlAn: Students will discuss the algorithm behind MetaphlAn. Use the latest version of MetaphlAn on all available samples in the infant gut dataset (way to get this data is provided for tutorials 5 and 6).
- Phylosift: Students will discuss the algorithm behind Phylosift. Use the latest version of MetaphlAn on all available samples in the infant gut dataset (way to get this data is provided for tutorials 5 and 6).

Instructions for the Datasets:
See Tutorials 5 and 6.

**Tutorial 8:** DNA Clustering and Alignment: Tutorial 7 covers the topic of identifying the taxonomic composition of samples. However, sometimes, one may wish to just cluster similar fragments together, which may commonly be orthologs between many species. One such conserved gene is the 16S rRNA gene. Here, students will learn they can align over 7000+ sequences by clustering them into groups first and then one representative per cluster. Students will compare how many groupings they get from clustering at 97% identity compared to 90% identity, and how this affects the time to align the sequences.

DNA Clustering: Students will compare the performance of CD-HIT, Uclust and Uclust_ref to Usearch and Usearch_ref (using QIIME) to group clusters into 90% and 97% identity. Students are expected to spend a good portion of the class comparing and contrasting the algorithms of CD-HIT, Uclust, USearch, and UParse.

Alignment: Students will align the representative sequences of each cluster using 90% and 97% identity from the previous step. They will do this in QIIME using PyNAST, MUSCLE, INFERNAL. The PyNAST, MUSCLE, and INFERNAL algorithms will be compared and contrasted by then building their phylogenetic tree relationships.

Instructions for the Datasets:
- /mnt/HA/groups/rosenGrp/data/yemin/mytemp/All_16S.fasta

## Tutorials 9 and 10: Functional Annotation of metagenomes

**Tutorial 9:** To functionally annotate genomes, one must first annotate the start and stop of genes (this is traditionally done with HMMs which usually assume the entire genome is being scanned). However, with metagenomics, only short reads exist, so this process has some optimizations and a good software to do this is FragGeneScan. However, to annotate what protein family a gene is from, this is still done using HMMs, and HMMer is a popular tool that is accurate but

at the sacrifice of speed. Finally, knowing the gene and protein family can provide much information, but sometimes one wants to go further and investigate what metabolic pathways these genes are involved in and what could be upregulated.

- FragGeneScan - gene discovery from contig(s), Students will be expected to annotate genes within the infant gut microbiomes and give statistics on how many genes were found and the coverage of these genes
- HMMer - functional annotation to protein families / Pfams, Students will use HMMer to annotate these genes with different protein family domains

**Tutorial 10:**

- HuManN: Students will use HuManN to identify the abundance of orthologous gene families (students should explain differences between COG, NOG, eggNOG, etc.), the presence/absence of each pathway in a community (students should review what the term "coverage" means). If possible, it would be nice to modify the code to use Metacyc instead of KEGG. Finally, students will also determine the abundance of these pathways.

*Instructions for the Datasets:*
*The infant gut microbiome samples from Tutorial 6.*

**Tutorial 11: Metatranscriptomic analyses**

16S and metagenome datasets reveal what potential is in samples but not necessarily what is being expressed in any given moment. RNA-seq can reveal this. A common process is to de novo assemble the RNA-seq reads into genes and then get differential expression of the expressed genes between samples. Students will do this through **IDBA_MT** and **Cufflinks**. Students will also **BLAST** the highly differentially expressed genes to the databases to see which genes are differentially expressed between samples and how. Concepts like FPKM/RPKM must be discussed and full discussion on how IDBA_MT and Cufflinks works is expected.

*Instructions for the Datasets:*

- *Acquire metatranscrimptomic reads from this study (http://www.pnas.org/content/111/22/E2329.full.pdf), which were stored in NCBI SRA (http://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP019038)*