Tutorial Descriptions

Each tutorial is expected to be at least 30 minutes. Tutorials will consist of two people. Undergrads will be required to do 1 and Graduate students will be required to do 2. Every week all students will come up with questions about the tutorials and the next week, everyone will be quizzed on the previous week's tutorials. Making quiz questions will be recorded as class participation, while the quizzes will be graded.

Instructions for the Datasets:
HMP Data:
/mnt/HA/groups/nsftuesGrp/data/HMP

Tutorials 1 and 2: "Meta-Packages" and their features

Students will review a) a website that specializes in whole-genome and 16S metagenomics analysis and b) a web application that requires in-house and/or external computing resources.

Tutorial 1: MG-RAST: Students will sign up for http://metagenomics.anl.gov and input the Urban Transit dataset (http://msystems.asm.org/content/1/3/e00018-16, https://www.ncbi.nlm.nih.gov/bioproject/PRJNA301589/), a WGS dataset, and the Guerrero Microbial Mat, a 16S data, to it for analysis. Students will demonstrate all the functions that MG-RAST can perform.

Tutorial 2: MEGAN: Students will download http:// ab.inf.uni-tuebingen.de/software/megan/ (Dr. Rosen will provide license) and input the Urban Transit dataset (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA301589/), a WGS dataset, and the Guerrero Microbial Mat, a 16S data, to it for analysis. Students will demonstrate all the functions that MEGAN can perform.

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 3 and 4: Phylogenetics through High Performance Alignment and Tree Construction

Tutorial 3: Alignments: Students will align H. Influenzae 16S genes using Cipres (http://www.phylo.org/). Students will compare **Muscle** and **MAFFT** alignments, comparing time that it takes.

Tutorial 4: 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 alignment and tree methodology and the biological differences.

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 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

Tutorial 5: Metagenome Assembly

Students will review how to use a metagenomic assembler and the metrics involved in assessing its performance.

De novo assembly: Use samples provide to demonstrate assembly of contigs by comparing the IDBA-UD with the IDBA-hybrid packages. Students will be expected to overview and compare metrics of the two assemblers, such as N50, average contigs' size, etc using pbsuite quality assessment tool (https://sourceforge.net/p/pb-jelly/wiki/Home/). Write Saeed (sdkeshani@gmail.com) for more information and to guide you.

Tutorials 6 and 7: Taxonomic Classification of whole-genome/transcriptomic data

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 classification of sequences vs. abundance estimate of samples. 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.

Tutorial 6: Examples of Read-by-Read classification:

- **Diamond+MEGAN:** Students can discuss how alignment with last common ancestor algorithm works.
- **CLARK:** Students can discuss how k-mer based composition algorithms work.

Tutorial 7: Example of Abundance Estimation:

- **WGSQuikr**: Students can 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.
- **MetaphlAn:** Students will discuss the algorithm behind MetaphlAn. Use the latest version of MetaphlAn on all available samples in the infant gut dataset.

Instructions for the Datasets and Choosing methods:
<see instructors and make sure to correspond with instructors>

Tutorials 8 and 9: Functional Annotation of Microbial Surveys

Tutorial 8: Picrust: Students will explore the predicted functional potential of gorilla gut microbiome, generated via 16S sequencing:

1. Download **amplicon** metagenomes from http://metagenomics.anl.gov/mgmain.html?mgpage=search&search=6321 ; Please begin with the fna files that have already been dereplicated.
2. Generate an OTU table via QIIME using **closed** reference OTU picking (http://nbviewer.jupyter.org/github/biocore/qiime/blob/1.9.1/examples/ipynb/illumina_overview_tutorial.ipynb) using **GreenGenes 13.8 release clustered at 97% identity**.
3. Copy number normalize OTU table and predict functional content (**KEGG Orthologs**) via PICRUSt: https://picrust.github.io/picrust/tutorials/metagenome_prediction.html

Tutorial 9: Tax4fun: Students will explore the predicted functional potential of gorilla gut microbiome, generated via 16S sequencing

1. Download **amplicon** metagenomes from http://metagenomics.anl.gov/mgmain.html?mgpage=search&search=6321 ; Please begin with the fna files that have already been dereplicated

2. Generate an OTU table via QIIME using **closed** reference OTU picking (http://nbviewer.jupyter.org/github/biocore/qiime/blob/1.9.1/examples/ipynb/illumina_overview_tutorial.ipynb) using **SILVA**.
3. Predict functional content (**KEGG Orthologs**) via Tax4Fun: http://tax4fun.gobics.de/RPackage/Readme_Tax4Fun.pdf

Tutorials 10 and 11: Functional Annotation of Metagenomes/Transcriptomes: Normalization and Differential Abundance of Metagenomes/Transcriptomes

Tutorial 10: 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. Using an oral metatranscriptome dataset (http://mbio.asm.org/content/5/2/e01012-14.abstract, SRP033605), compare the differential expression results obtained from DESeq2, EdgeR, and Limma (http://www.bioconductor.org/help/workflows/rnaseqGene/).

Tutorial 11:
Using the following IBD WGS dataset (http://www.nature.com/nature/journal/v464/n7285/full/nature08821.html): http://www.ebi.ac.uk/ena/data/view/ERA000116&display=html, compare MUSiCC normalization (https://genomebiology.biomedcentral.com/articles/10.1186/s13059-015-0610-8) to MicrobeCensus normalization (http://genomebiology.biomedcentral.com/articles/10.1186/s13059-015-0611-7). Annotate reads using a top gene approach after aligning each read to a peptide database via mBLASTx (see https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4778032/). Apply normalization at the pathway level and downstream functional analysis should be compared to demonstrate how the normalizations affect downstream performance. Students must discuss the importance and difference between methods of count normalization and differential expression analysis.

Tutorials 12: Functional Annotation of Metagenomes/Transcriptomes: Metabolic Analysis

Tutorial 12:

- 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). Students will also determine the abundance of these pathways and present on how this software aims to summarize functional capacity of metagenomes (and actual capacity of transcriptomes).

Instructions for the Datasets:
The IBD dataset from Instructors.