GROUP:

NAMES:

**Pipeline for analyzing 16S microbiome data**

**Project-specific starting Files**

The names of these files will change depending on the project that you choose to work on. For all of the data sets to choose from, you will be provided with (1) a single file containing quality-filtered paired-end reads in fasta format for all project samples, (2) a mapping file containing formatted for QIIME that contains all of the metadata associated with your samples, and (3) a metadata file that is identical to the mapping file except that the leading pound symbol (#) has been removed. These are formatted the same way as the files we used in class – the corresponding file being listed in parentheses below.

*All of these files will be stored in the /home/biol3004/shared directory.*

**(1) sequences.fna (example: HMP\_5BS\_combined\_seqs.fna)**

**(2) mapping.txt (example: HMP\_5BS\_mapping.txt)**

**(3) metadata.txt (example: HMP\_5BS\_metadata.txt)**

**QIIME Pipeline**

Based on your notes and the class assignments/tutorials, reconstruct the QIIME pipeline using the files with generic IDs above. For each step, include the following:

**(1)** The name of the script (e.g. pick\_closed\_reference\_otus.py)

**(a)** The purpose of the script and a brief description of the required input files and the output files that are generated.

**(b)** The command required to run the script (noting whether this should be submitted as a PBS script or can be run in an interactive session). Include the complete paths to the required files.

*You ONLY need to include the QIIME commands here (not logging on to MSI, the lab computer, module load qiime, etc.).*

**EXAMPLE:**

(1) pick\_closed\_reference\_otus.py

(a) This script takes input 16S rRNA gene sequence reads in fasta format (sequences.fna) and maps them to a reference set of 16S rRNA gene sequences that have been clustered based on 97% identity (Greengenes; 97\_otus.fasta). Taxonomy is then assigned (Greengenes; 97\_otu\_taxonomy.txt). The parameters for OTU picking are defined in the parameters file (usearch\_ref\_params.txt). The output for this script is a directory containing an OTU table in biom format (otu\_table.biom) with the number of reads in each sample that map to each OTU and the taxonomic identity of each OTU. It also produces a new directory called ‘usearch61\_ref\_picked\_otus’ with files describing the abundances of each OTU, reads that failed to map, and the clustered reference sequences.

(b) pick\_closed\_reference\_otus.py -i /home/biol3004/shared/sequences.fna –r /home/biol3004/shared/97\_otus.fasta -o /home/biol3004/*X500*/CR\_otu\_picking -t /home/biol3004/shared/97\_otu\_taxonomy.txt -p /home/biol3004/shared/usearch\_ref\_params.txt

This job takes about 25 minutes and requires that a PBS script is submitted to MSI