Final report for assessment:

============================

You should write a report ~ 20-30 pages with two parts

a) Overview of the metagenomics course

You will give an overview of metagenomics (both Amplicons and Whole-Genome ShotGun sequencing).

This may include information on (but not limited to):

-Sequencing platforms and inherent problems associated with them

-Different strategies for processing Amplicons (De novo and reference based) Merits/Demerits

-Different strategies for processing Whole-Genome Shotgun sequencing data including their merits/demerits

-Assembling genomes (single vs metagenomes), principles and techniques

-What sort of multivariate statistical analysis one normally performs on the tables generated from both technologies with their merits/demerits

-Information on databases and software one normally uses for processing metagenomics data

- Information on annotation and functional assignments (KEGGs, SEEDs subsystems, Pfams, HMMs)

- You can read through numerous papers (uploaded on moodle) to write your narrative. Please also look up terms/papers that are mentioned in the course slides.

b) Crohn's dataset

This should include some of your results from the tutorials but not necessarily all of them. I am looking for something coherent, nicely presented with proper references, and that indicates that you have understood key concepts. You can use the plots/results generated from QIIME, SEQenv, Metaphlan, HumaNN, LEFSE, and UPARSE.

Then you can add extra content for extra credit, further statistics in R or Biopython scripts to address particular questions. Ideas for extra analysis:

-Use R with Kruskal-Wallis, t-tests and Benjamini-Hochberg corrections to search for significantly different taxa between CrohnÊ¼s and Healthy individuals. You may take hints from KW.R and NB.R scripts on

http://userweb.eng.gla.ac.uk/umer.ijaz/bioinformatics/ecological.html

Your multivariate statistical tutorial also contain information on Benjamini-Hochberg. In particular you can google up how R's p.adjust() method is used. You can also get some hints from how I have used it on my ecological.html page

-Use R to explore effect of generalized Unifrac metric on PCoA/NMDS plots (take hints from phyloseq.R on my ecological.html page). You have OTUs sequences which you can align (mafft-ginsi) and then use FastTree to generate the tree. Essentially, I am looking for differences between phylogenetic-aware measures (Unifrac, Generalised Unifrac, Comdist) versus Bray-Curtis (or any that don't require a phylogenetic tree) in the context of Crohn's dataset to see if the tree gives better discriminatory power.

You can also do the same with diversity information (pd() that takes a phylogenetic tree) versus R's diversity(). In this regard, it would also be useful for you to go through Vegan tutorial:

http://cc.oulu.fi/~jarioksa/opetus/metodi/vegantutor.pdf

-In addition to the abundance tables (phylum.csv, class.csv, order.csv, genus.csv, family.csv, otu\_table.csv), you can also include the ENVO table generated from SEQenv. Use the following command to copy it to your ITutorial\_YOURNAME folder:

cp /home/opt/tutorials/ITutorial\_Umer/SE/otus\_N1000\_blast\_F\_ENVO\_samples\_labels.csv ENVO\_table.csv

Marks will follow the following rough guide:

A->good overview, good understanding and additional original analysis

B->substantial overview, tutorials results with good understanding and/or additional original analysis

C->satisfactory overview + satisfactory results plus satisfactory explanations