Generate a PDF report that includes answers to the following questions, as well as analysis and accompanying figures. It may be in the form of a notebook (e.g. jupyter) or simply a document with a commented and formatted code and figures within text.

## Intro questions

1. Explain the advantages and disadvantages of 16S sequencing and Whole Genome Shotgun metagenomic sequencing.
2. What is the difference between Operational Taxonomic Units (OTUs) and Amplicon Sequence Variants (ASVs)?
3. Briefly describe how would one go about doing functional analysis based on the 16S microbiome data and what would be the main limitations of such analysis?

## General Microbiome Analysis

In this example, the goal is to identify microbes associated with some phenotypes. We provided two tables:

1. test\_microbial\_abundance.txt: microbial count data provided from a pipeline generating ASV counts for each sample (across columns region, sampleid, asv\_id and asv\_count)
2. test\_phenotypes.txt: basic phenotype data for each sample: region, sampleid, age, gender and bmi (body mass index)  
     
   For some samples, the data for one or more phenotyping variables is missing. Explain why/how you dealt with this missing data in the subsequent analyses.

The microbiome data pipeline identified a large number of ASVs in the data. In order to visualize the similarity between samples in a dataset, it is common to employ some form of multidimensional scaling, e.g. Principal Component Analysis (PCA) or Principal Coordinate Analysis (PCoA), during exploratory analysis.

1. What is the difference between PCA and PCoA?

1. For the purpose of this test:
   1. Normalize the microbial counts and run PCA on the microbiome data.
   2. Generate a 2D PCA plot with the first two principal components, color points by the geographical location (region column).
   3. What are the top 5 microbes (in terms of ASV IDs) that contribute the most to the first two principal components?
2. Find ASV IDs associated with BMI, adjusted for age and gender, using a method of your choice (briefly explain why you used that method) and show a list of the top 5 most significantly associated ASV IDs.

## Microbial-metabolite associations

One goal of a project might be to identify statistical associations between the abundance of a microbes and levels of blood metabolites. In one such project we measured the levels of blood metabolites using mass spectrometry, and relative abundance of microbes using 16S sequencing. The data is provided in tables mm\_microbial\_abundances.txt and mm\_metabolite\_values.txt,with metabolite IDs (mtb\_id) and already normalized and transformed metabolite values mtb\_value. Keep in mind, that unlike 16S sequencing data, mass spectrometry metabolomics data is not compositional – all metabolite levels in a sample do not add up to “100%”.

As a first pass to start identifying bacterially produced metabolites, one may want to try a simple linear regression between microbial abundance levels (or its transformed values) and metabolite levels.

1. What are the potential problems of using linear regression to microbe-metabolite associations?
2. What could be some of the approaches alternative to linear regression to find metabolites potentially produced by microbes? (Do not need to run them, just describe.)
3. In a typical dataset, there may be many microbes and many metabolites so the number of all possible combinations might be too large to run in a reasonable time. Typically, not every combination needs to be tested either.  
      
   Use some meaningful way of selecting a subset metabolites and microbes to run a reduced number of regressions.
4. Write a code to run these regressions on your laptop or desktop computer and report the top 10 microbial-metabolite hits (in terms of ASV IDs and MTB IDs), including the p-values and confidence intervals for the effect size.
5. Briefly describe few ways to how to adjust p-values when we test large number of hypotheses and the advantages or disadvantages of adjusting p-values in such manner.