**Metagenomics Analysis of stool 16s rRNA Amplicon Sequences of Gut Microbiome of Healthy and Diseased Patients using R Programming**

**Introduction:**

Chronic immune-mediated gastrointestinal ilnesses including Crohn’s disease and ulcerative colitis are important causes of the generalised inflammatory bowel diseases. It is becoming more and more clear that combinations among genetic and environmental factors, which might skew immune responses against healthy gut flora, are what cause the pathological processes of IBDs. The purpose of this study is to evaluate and analyse the relative abundaces of the different species in UC and CD in the Saudi population.

**Steps to be followed for metagenomic analysis :**

1. **Installed and loading R packages:**

Firstly installed R and R studio and installed these packages into R Studio console:

1. BiocManager
2. Dada2
3. Rcpp
4. Phyloseq
5. Ggplot2
6. Biostrings
7. Decipher
8. Phangorn

After installing these packages load these packages with the help of code:

CODE:



1. **Raw reads from metagenomic samples:**

4 Experimental data were retrieved of diseased patient from (<https://www.ebi.ac.uk/ena/browser/view/PRJEB57347>) and healthy patient from (<https://www.ebi.ac.uk/ena/browser/view/PRJNA872871>).

Diseased : Illiumina NovaSeq sequencing generated Fastq files for paired reads were downloaded from ENA run accession ID **ERR10482484** and **ERR10482516.**

Healthy : Illiumina MiSeq sequencing generated Fastq files for paired reads were downloaded from ENA run accession ID **SRR21460907** and **SRR21460910.**

Set the path of all these sample.

CODE:

#specify path where the FASTQ files are located

PathD <- “-/disease”

PathH <- “-/healthy”

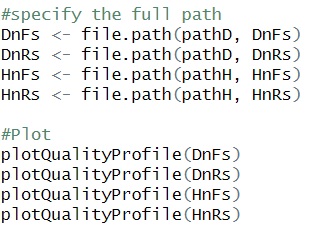
list.files(pathD)

list.files(pathH)

1. **Plot Quality Profile**

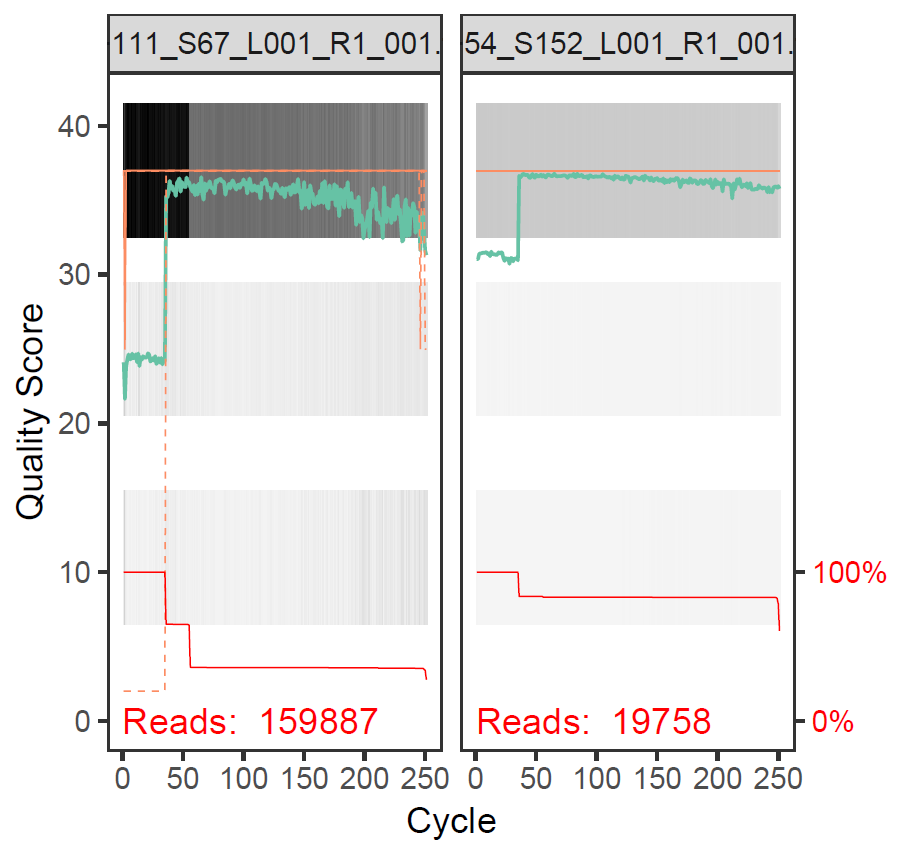
A grey-scale heat map depicts the distribution of quality scores at each point, with darker colours correlating to higher frequency. The shown lines represent positional summary statistics: green represents the mean and orange represents the median.If the length of the sequences varies, a red line will be displayed to illustrate the percentage of reads that extend to at least that point. In general, the scores do not fall below 30, and to be on the safe side, we will subtract bp from the end and beginning of those scores with quality less than 30.

CODE:

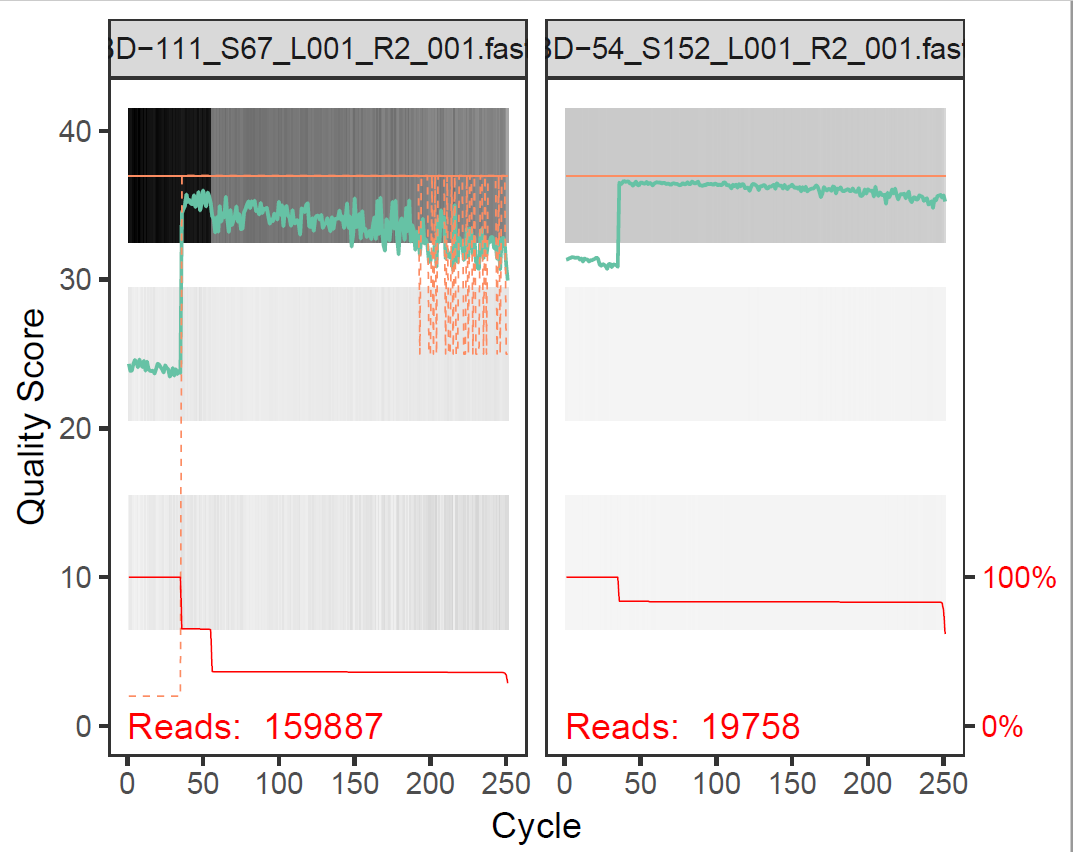


**Result:**

Forward reads – The quality score is less than 30 upto 45 bp and hence trimmed.



Reverse reads-The quality score is less than 30 upto 40 bp and hence trimmed.

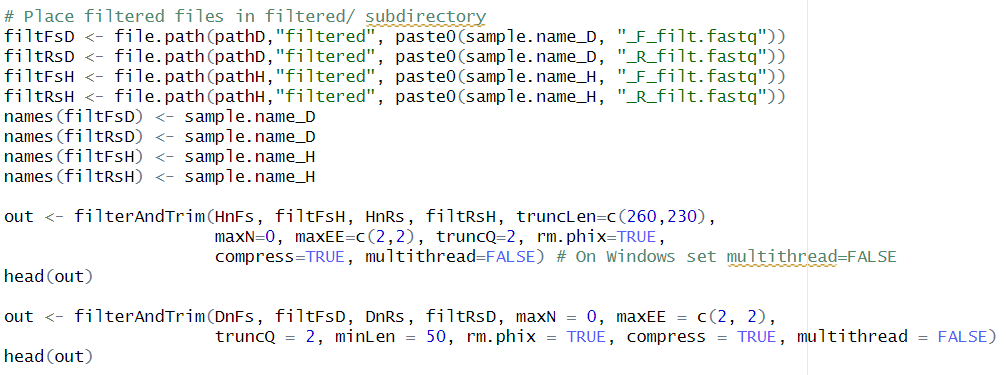


1. **Trim off low quality reads and filtered:**

To get rid off the low quality profile score reads, trim the basepairs having score<30.

truncLen function assign the values as C(FWD, REV), maxN is the maximum number of ambiguous bases allowed,maxEE = c(FWD,REV) the maximum number of errors allowed for an individual read,truncQ truncates reads at the first instance of Q score less than or equal to the value specified (a Q score of 2 is very bad i.e. ~63% chance of a base call being incorrect)

CODE:



1. **Learn error rates**

Error rate function tell us about identifying the correct base during sequencing. Error rate plot for every possible 16 subsitutions of bases. Black line tells estimated error rate and the red line shows expected error rate of each substitution.In general, frequency of errors decreases as quality score increases.

Here the black line are good fit to the observed rates(points).

CODE:

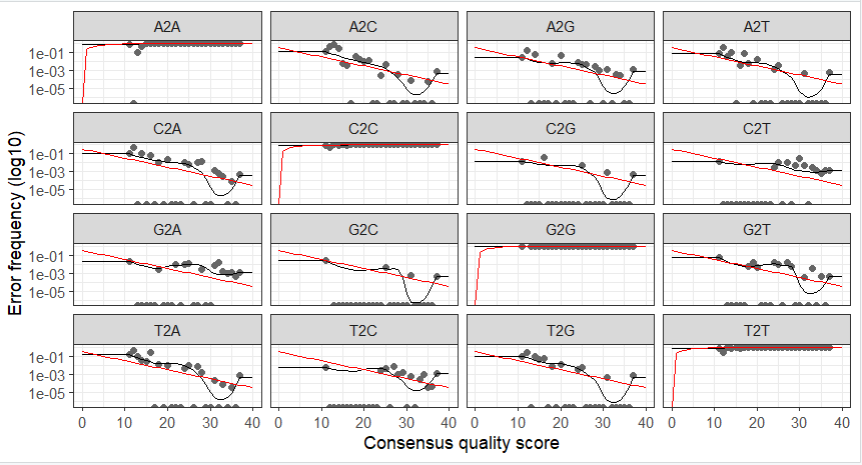
#learn error rates

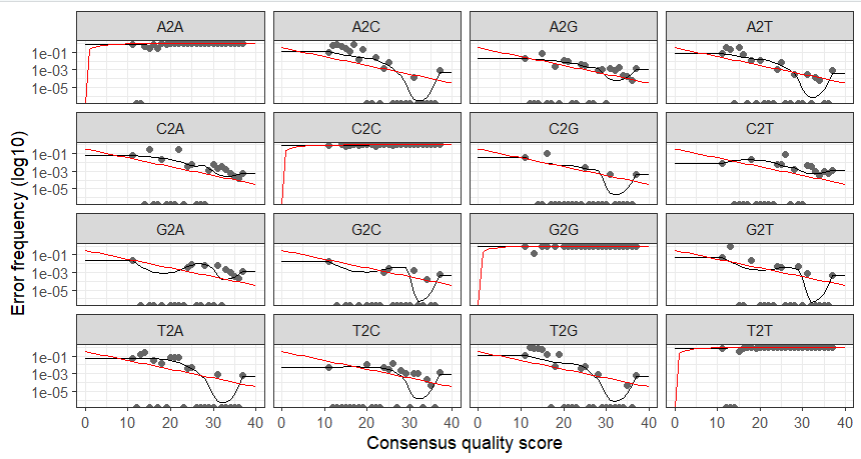
errF <- learnErrors(filtFs, multithread=FALSE) # for forward reads

plotErrors(errF, nominalQ=TRUE)

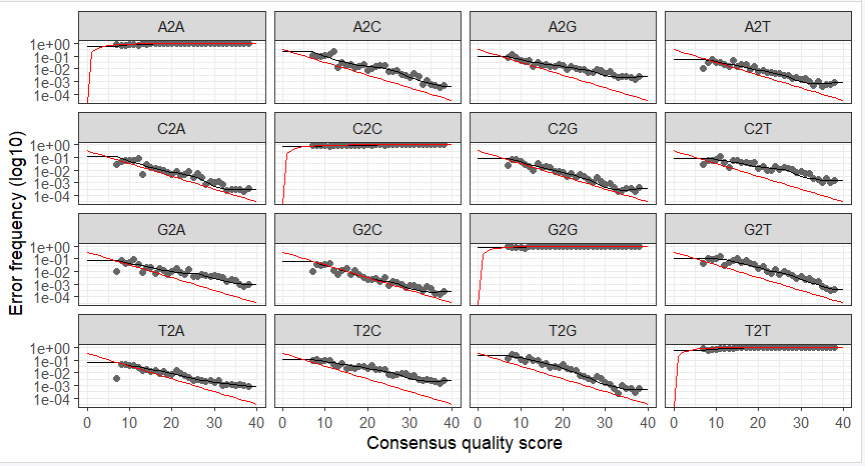
**Result:**

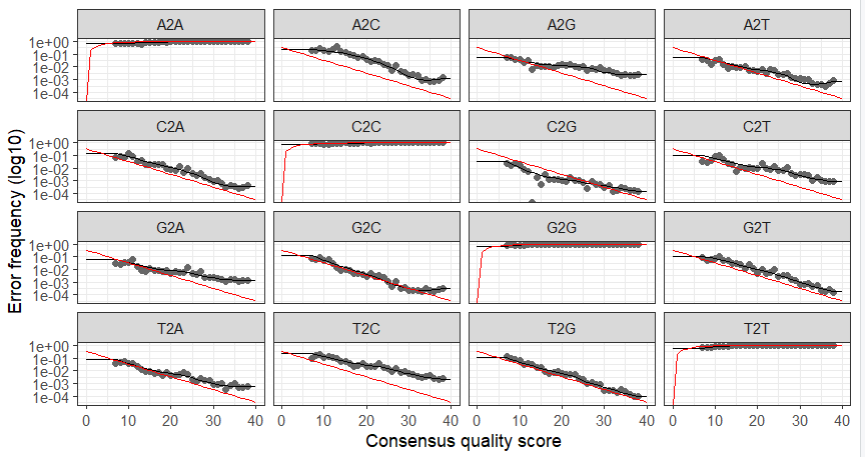
Diseased: Forward and Reverse strand





Healthy: Forward and Reverse reads

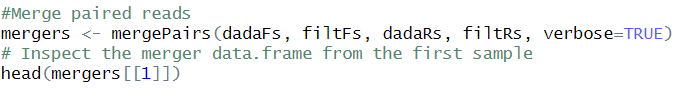




1. **Metagenomics assembly**

Merge the paired reads only if they exactly overlap (around 20 nucleotide of overlap) this can be much more particularly for these longer hypervariable region This is because both forward and reverse reads have been denoised and should be error free.

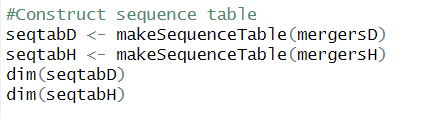
CODE:



1. **Sequence table**

The dada2 sequence table is analogous to traditional OUT table.The sequence table is a matrix in which coloumn corresponds to sequence variants, and rows correspond to samples.

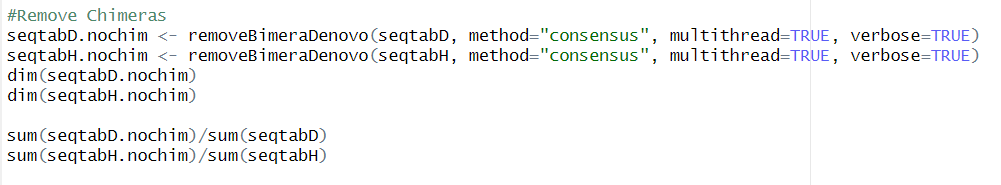
CODE:



1. **Remove Chimeras**

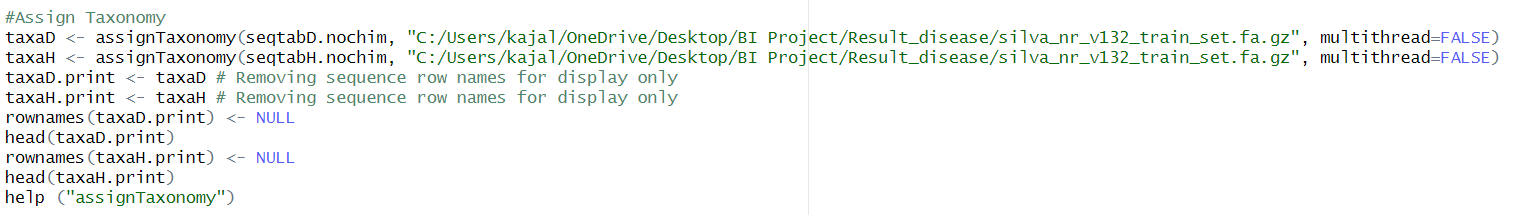
Chimeras are basically the PCR artifact so basically what code does is that it takes sequence table and it perform MSA and do sequence alignments with all possible combinations to see if there’s any overlap.

CODE:



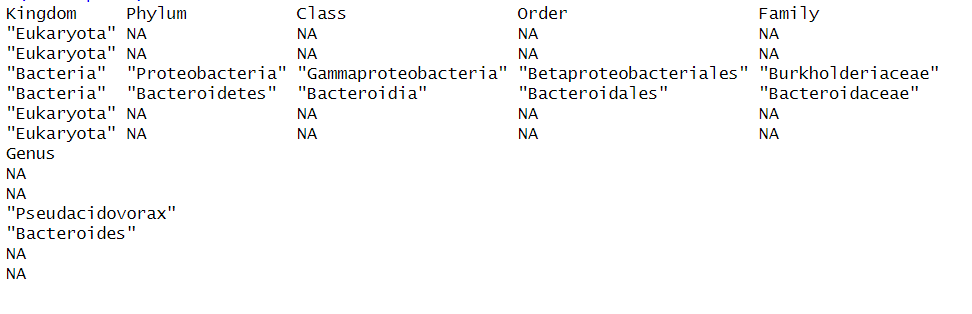
1. **Assign Taxonomy**

To assign a Taxonomy DADA2 package provides naïve Bayesian classifier method for this purpose. A character matrix of assigned taxonomies to the sequence variants .Rows correspond to the provided sequences,columns to the taxonomic levels such as kingdom, phylum, class, order, family , genus.Taxonomy can be assigned using sequence table as an input and get the levels with the help of SILVA database in assignTaxonomy function.

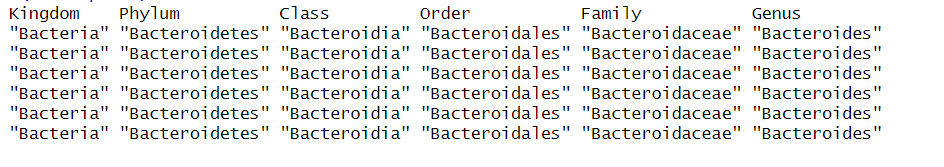
CODE:  


**Result :**

Diseased – Top 6 taxonomy



Healthy – Top 6 taxonomy



Top family of microbiota has been showed .From this we can infer that Healthy patients contain more **‘Bacteroidaceae’** community campare to diseased patient and also diseased one contain more **‘Burkholderiaceae’** family compare to healthy one.

1. **Construct phylogenetic tree**

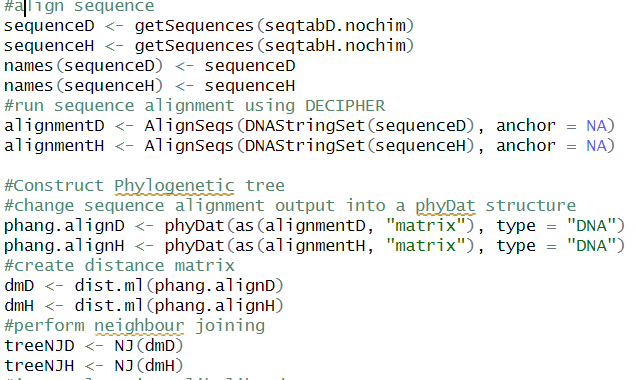
A tree was needed to get the information about possible phylogenetic realtionships between them, sequence alignment was needed before tree construction. Several methods are used to construct tree ,in this case we use neighbour joining method to construct phyogentic tree.

Distance matrix was constructed .

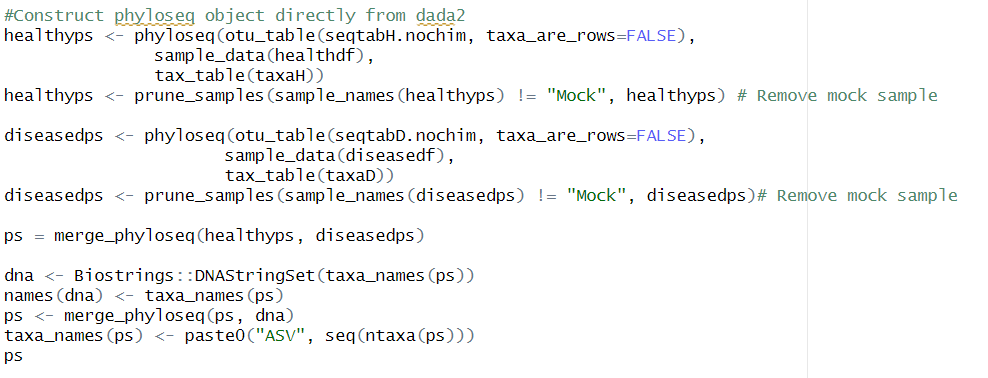
**The computational power needed was much higher than available and hence not shown!**

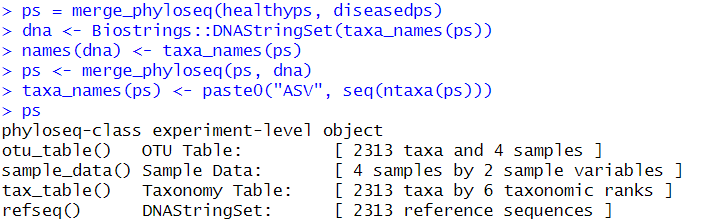
We can use Phangorn package of DADA2 to construct phylogenetic tree.

CODE:



1. **Construct and merge phyloseq**

CODE:  




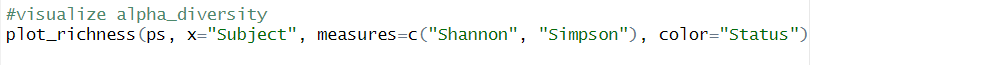
1. **Visulalise alpha diversity**

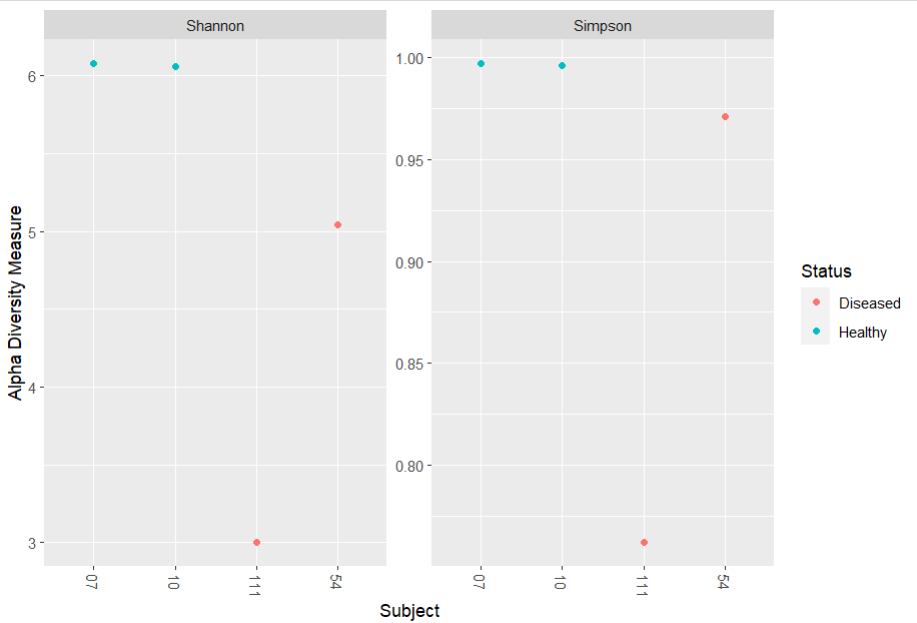
Alpha diversity gives the measure of richness in ecological community, by usually taking into account the number of different species observed.

Various matrices used to measure alpha diversity are:

**Shannon diversity** an estimator of both species richness and evenness,but with weight on the richness.

**Simpson diversity index** based on the probability that two entities taken from sample at random are of different types.It ranges from 0 to 1.

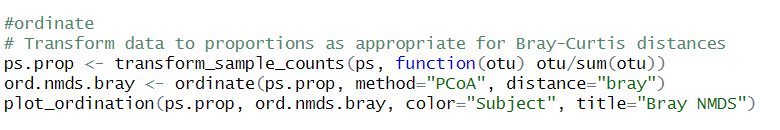
CODE:  


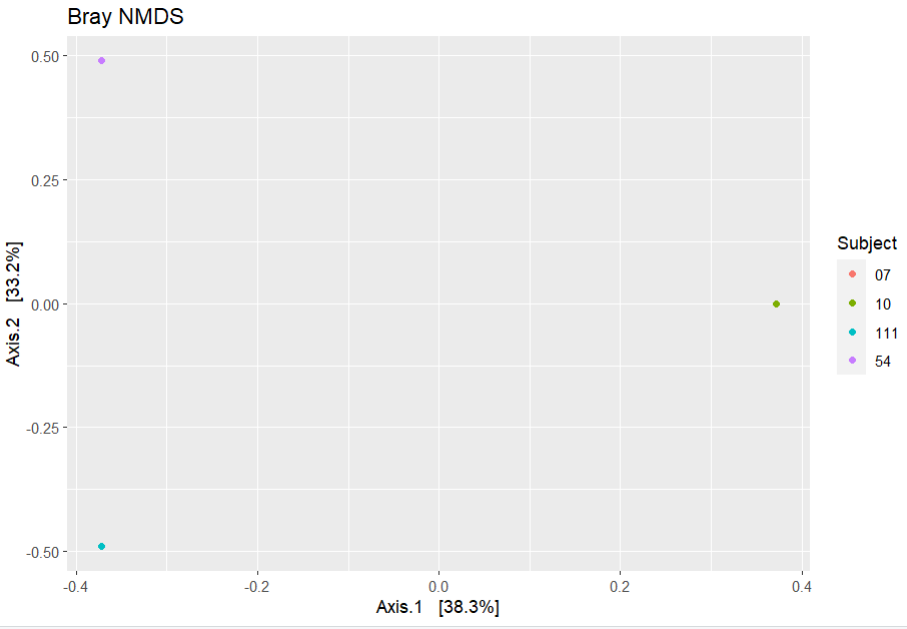


1. **Visulalise beta diversity**

It express the difference between two groups or we can say comparison between two groups of species such as treated and control groups.

The graph was plotted using **Bray-Curtis Dissimilarilty range(0 to 1)** where the examination of abundance of microbes which are shared between two samples.”dissimalirty is zero provided both the samples share maximum microbes else 1.

CODE:  
 

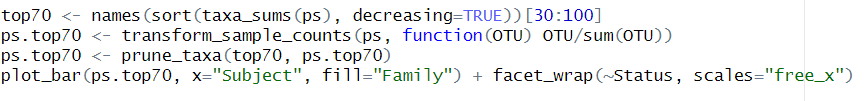


1. **Abundance Graph**

A abundance plot  helps us to visualize species abundance distributions.Here, sorting of no of individual species is done in descending order, and the proportion of the total number of individuals for each species is then plotted on the log scale against the species rank.

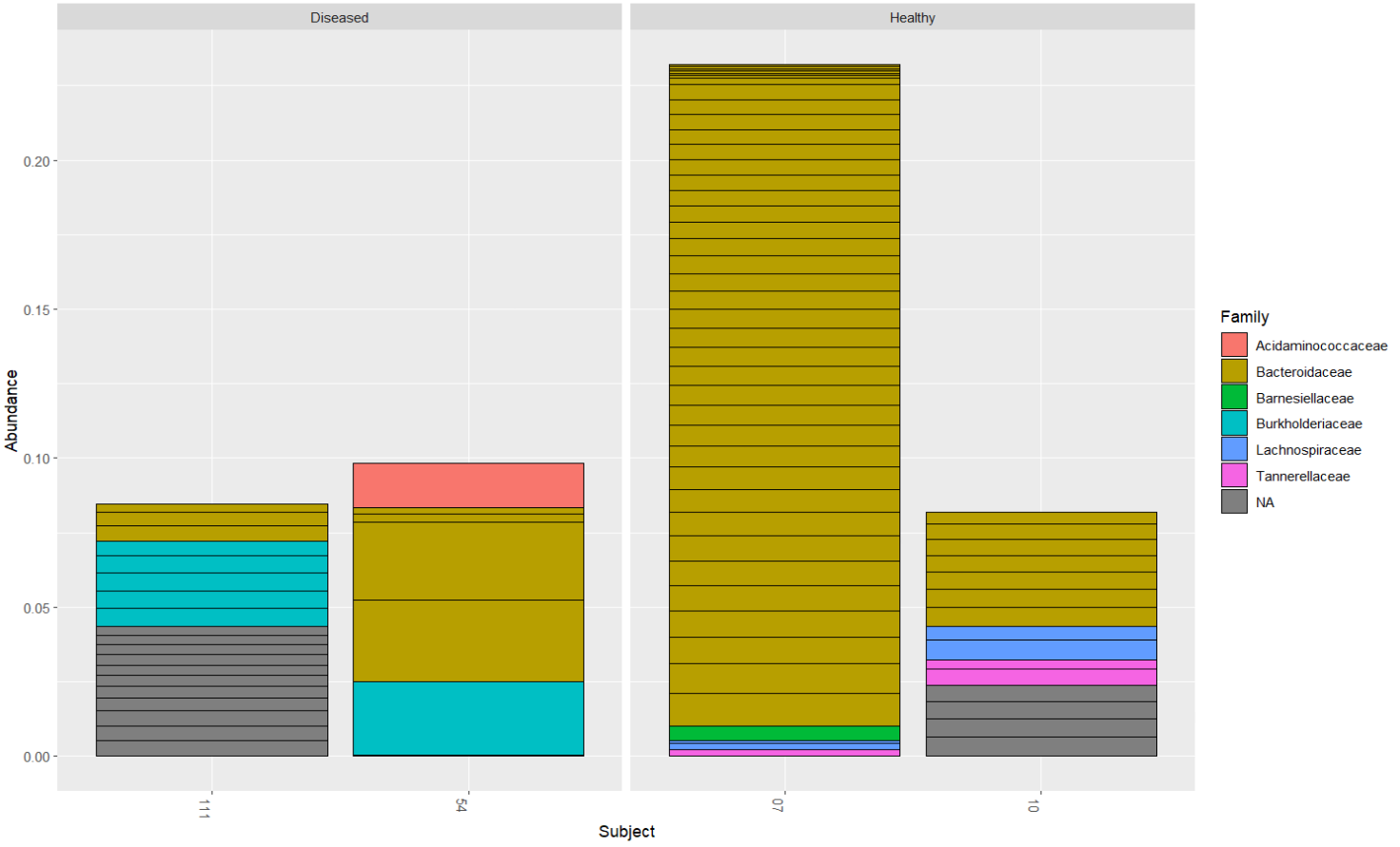
Top 70 family was taken from taxonomy and plot a bar plot .

CODE:



**Result:**

Abundance plot shows bar plot based on family



It shows that the “***Bacteriodaceae***” family was observed more in Healthy patients as compare to Dieased patient and also “***Burkholderiaceae***” family found in more frequently in Healthy as compared to disease patient.