Exploratory Analysis of 16S Bacterial Metagenomic Data

Przemysław Kiljan

Winter semester 2019/20

Abstract

Report focuses on data obtained from exploratory analysis of 16S sequencing data gathered under Earth Microbiome Project (EMP). Samples were collected from locations all over the world and by different research groups under various research papers. Data is contained in form of Operational Taxonomy Unit Table (OTU Table), which contains abundance profiles of chosen sites. Exploratory analysis was focused on better understating of aforementioned data and drawing conclusions on its heterogeneous nature.

Introduction

Considering prices of sequencing techniques that are getting lower each year it is understandable that amount of that kind of data will be increasing accordingly. Current biomedical trends points to increasing interest of people that are seemingly unrelated to nature sciences in microbial aspects of their life. Whetеher someone is interested in microorganisms outside or inside their bodies, proper data has to be gathered in order to assess one's health condition or suppoosed disease development. Sequencing techniques have been steadily improving over last few decades (Kchouk, Gibrat, and Elloumi 2017), which allows scientists to produce large datasets.

It is worth noting that despite growing interest in health benefits that sequencing can provide us, many researchers are actively looking for other ways to utilize newly acquired methods. High-throughput techniques allows us to not only get to know and understand human body better, but also the enivornment around us. And given the opportunity to actually look into our every day surroundings, scienitst were able to determine what is roughly microbial footprint that we leave wherever we go (Afshinnekoo et al. 2015).

That being said it is important to acknowledge how much effort goes into proper sample collection and overall data preparation. That is why one of the main goals in order to improve future research is to define standardized sample gathering techniques (Mason et al. 2017). Same is being done with benchmarking bioinformatics tools used for data preparation and later its statistical analysis (McIntyre et al. 2017),(Mangul et al. 2018). Importance of those kind of review papers unfolds once someone performs meta-analysis of given results and by

using standarized methods is able to compare those evenly.
Same standarization of sample gathering was performed for the need of Earth Microbiome Project, idea of which was born during meeting of "leading microbial ecology, computation, bioinformatics and statistics researchers" in Snowbird, Utah (USA) in 2010 (Gilbert et al. 2010). On the EMP website it is listed that over 500 investigators were involved in this project so far, and more than 200 000 samples were gathered.
Data obtained from those samples is sequencing data that has to properly processed. First tool used for general preprocessing of data from EMP is QIIME but it's worth noting that there are many alternative programs used for aforementioned process e.g. Mothur or PICRUSt. Output from this software may be in two forms, depending on what kind of information is needed to obtain i.e. microbiome abundance or microbiome function. Analysis performed in this project were tables containing information about species abundances in samples gathered from soil. Samples were selected to serve as a supplementary data in one of the challenges presented at this year's Conference on Critical Assesment of Massive Data Analysis (CAMDA). Metagenomic Geolocation Challenge mainly focuses on developing tools for efficient metagenomic profiling of different places in the world. Main data for this challenge comes from MetaSUB which is a project focusing on gathering samples from urban biome, and then sequencing them with Whole Genome Sequencing technique (WGS). EMP subset consists of 3043 soil samples collected by various research groups under EMP, and it is mainly to provide research groups engaged in this challenge with some additional context for possible geospatial identification of metagenomic data.
Default abundance output file format in QIIME is *.biom. It is used due to its high compression rate when it comes to tables that may contain high percentage of empty spaces, in form of zeroes. Abundance tables tend to be sparse, usually with over 90% of data to be zeroes. The abundance table contains information about quantitative bacterial content in given samples. Specific taxonomic ranks and their members are identified by finding referenced sequences inside metagenomic data. Those groups are called Operational Taxonomic Unit (OTU) and abundance tables are usually called OTU-tables since row names consist of specific OTU IDs, which then later can be identified by referencing entries of database used in preprocessing.
Considering size and nature of this data one could expect that proper analysis might be extremly difficult. Fortunately analysis tools are being developed on par with new techniques. Statistical analysis of metagenomic data can be performed in many different ways by various packages available for R.

Analysis

Importing data

First and foremost required libraries have to be imported:

```
library(phyloseq)
library(ggplot2)
library(vegan)
```

```
## Warning: package 'vegan' was built under R version 3.6.2
```

`phyloseq` package is main package that will be used, since it was specifically made for analysis of microbiome data. It also allows for proper integration of all essential data. `vegan` is a library that consists of many tools for stastical analysis of community ecology data. Originally data provided for this project sized over 35 Gb but integrating it with taxonomy table resulted in vast reduction of its volume. Since its in *.tsv format an apropriate way to import it would be to read it as a data table.

```
pre_otu<-read.table("mini_emp.tsv",sep="\t",header=T,row.names=1)
```

Since sample names starts with number, importing it into R renamed column (sample) names to start with letter 'X'. Command below ensures that names are the same as originally.

```
colnames(pre_otu) <- gsub("X","",colnames(pre_otu))
```

`phyloseq` requires otu data in form of matrix.

```
otu <- as.matrix(pre_otu)
```

Once data is in matrix it is possible to create first `phyloseq`-specific object.

```
abundance = otu_table(otu,taxa_are_rows = TRUE)
```

Next, taxonomy data has to be imported. Originally file that contained taxonomy information was very poorly formated. Short script in Python together with small manual adjustment, were able to reformat it into proper *.csv file.

```
import re
fn = '97_otu_taxonomy.txt'
pre = open(fn, 'r').readlines()
sep = [i.strip('\n').replace(';','').replace('\t',' ').\
      split(' ', 7) for i in pre]
presto = [[re.sub(re.compile('.__'),'',j) if len(j) !=3\
      else re.sub(re.compile('.__'),'-',j) for j in i] for i in sep]

print ('OTU_ID,Kingdom,Phylum,Class,Order,Family,Genus,Species')

for i in presto:
        print (','.join(i))
```

Import taxonomy table and create `phyloseq` object.

```
taxonomy <- as.matrix(read.table("Tax.csv",
                                 sep=",",
                                 header = T,
                                 row.names = 1))
tax_final = tax_table(taxonomy)
```

Last object to create is object containing samples metadata that are also stored in .tsv file.

```
meta_data <- read.table('CAMDA_2019_EMP_metainformation.tsv',
                        sep='\t',
                        header=T,row.names=1)
row.names(meta_data) = tolower(row.names(meta_data))
meta_final <- sample_data(meta_data)
```

Once all three objects are ready it is finally possible to integrate them into one, complex object. Calling newly composed object by its name displays its general characteristics. From which one can deduct that limiting factor for number of OTUs in otu_table is possibly number of rows in taxonomy table.

```
ds <- merge_phyloseq(abundance,tax_final,meta_final)

ds
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 16325 taxa and 3043 samples ]
## sample_data() Sample Data:       [ 3043 samples by 75 sample variables ]
## tax_table()   Taxonomy Table:    [ 16325 taxa by 7 taxonomic ranks ]
```

Rarefication step can be ommited, especially due to potential significance of low-count taxa.

```
ds.rarefied = rarefy_even_depth(ds,
                                rngseed=2137,
                                sample.size=0.1*min(sample_sums(ds)),
                                replace=F)
```

Plotting abundance

Due to multitude of samples from different studies it's important to select specfific sample feature for further analysis. In this case environmental biome 2 (envo_biome_2) was chosen from samples metadata. That is one of few features that are shared between all samples. Next step is merging samples of the same envo_biome_2 within range of each study. First new phyloseq object has to be initialized, only then it is possible to add up merged abundances from other studies.

```
subset_1= subset_samples(ds,
                         sample_data(ds)$"study_id"==
                           sort(unique(sample_data(ds)$study_id))[1])
merged_1= merge_samples(subset_1, "envo_biome_2")
```

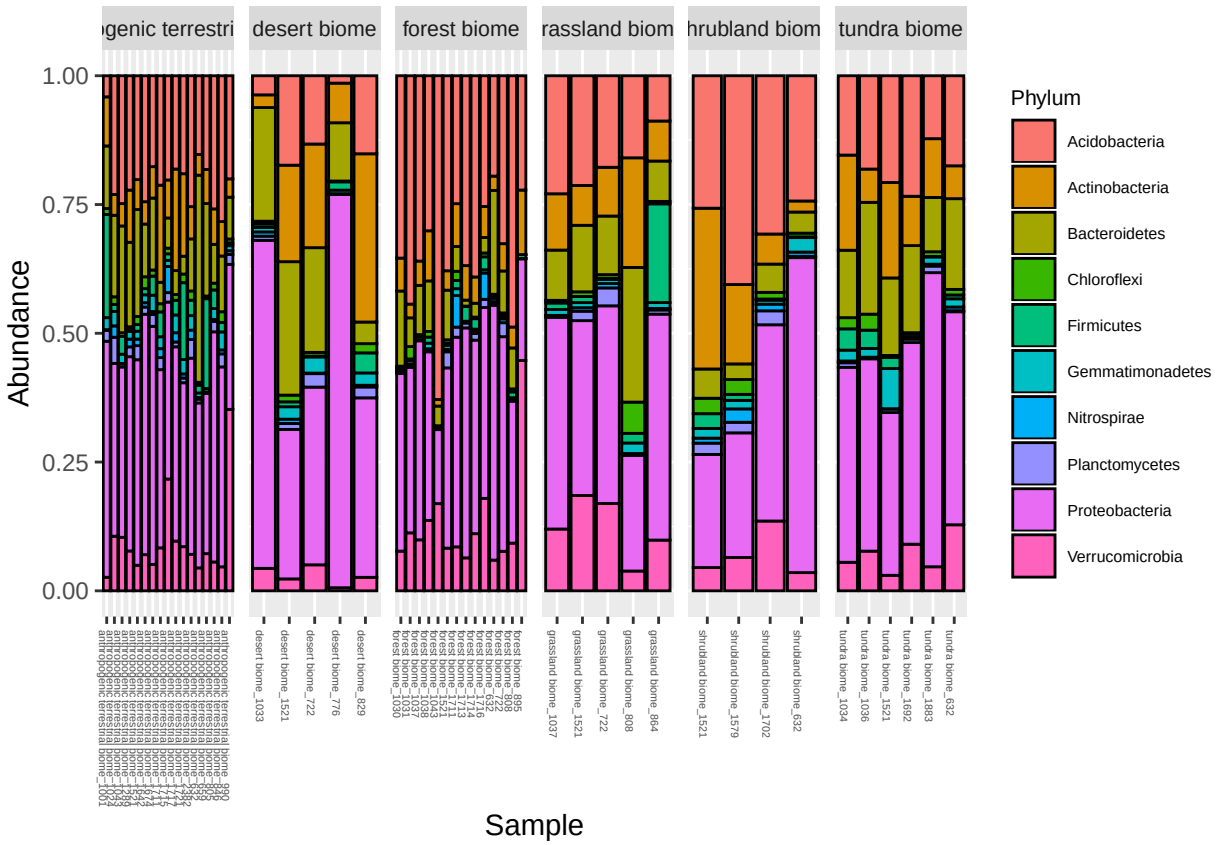
```
merged_1@sam_data$"envo_biome_2" = sample_names(merged_1)
sample_names(merged_1) = paste(sample_names(merged_1),
                               sample_data(merged_1)$"study_id"[1],
                               sep = "_")
phylo_fill = merged_1

for(i in sort(unique(sample_data(ds)$study_id))
    [2:length(sort(unique(sample_data(ds)$study_id)))]) {
  subset_i= subset_samples(ds, sample_data(ds)$"study_id"== i)
  merged_i= merge_samples(subset_i, "envo_biome_2")
  merged_i@sam_data$"envo_biome_2" = sample_names(merged_i)
  sample_names(merged_i) = paste(sample_names(merged_i),
                                 sample_data(merged_i)$
                                   "study_id"[1], sep = "_")
  phylo_fill = merge_phyloseq(phylo_fill, merged_i)

}
```

After merger it is important to normalize data to avoid overrepresentation of examples created from multiple samples.

```
phylo_fill_prop = transform_sample_counts(phylo_fill,
                                          function(x) x/sum(x))
```

Once table of proportions is obtained it is possible to plot all abundances. Merging OTUs at higher taxonomic rank and displaying top 10 most abundant phylum allows for more clear representation.

```
phylo_fill_prop = transform_sample_counts(phylo_fill,
                                          function(x) x/sum(x))
phylo_fill_prop.phylum = tax_glom(phylo_fill_prop,
                                  taxrank = "Phylum", NArm = FALSE)
top_10 <- names(sort(taxa_sums(phylo_fill_prop.phylum), TRUE)[1:10])
phylo_top_10 <- prune_taxa(top_10, phylo_fill_prop.phylum)
plot_bar(transform_sample_counts(phylo_top_10,
                                 function(x) x/sum(x)), fill="Phylum") +
  facet_wrap(~envo_biome_2,
             scales= "free_x", nrow=1) +
  theme(strip.text.x = element_text(size = 8),
        axis.text.x=element_text(size=rel(0.5)),
        legend.title = element_text(size = 8),
        legend.text = element_text(size = 6))
```



Measuring diversity

- There are two measures of diversity of metagenomic data:
  - Alpha divesity - which describes microbial richness within sample
  - Beta diversity - which represent differences in bacterial composition between samples

Alpha diversity

There are two main indicators of alpha diversity: richness of the sample and its evenness. First one can be measured using Chao1 index which takes into account that species that are more rare in given environment can be omitted when sampled. Chao1 measure is defined as

$$R_{Chao1} = R_{obs} + \frac{f_1(f_1 - 1)}{2(f_2 + 1)}$$

where f_1 is the number of observed sepcies that were spotted only once and f_2 signifies number of species that were observed twice.
Evenness can be measured using Shannon index which can be represented as

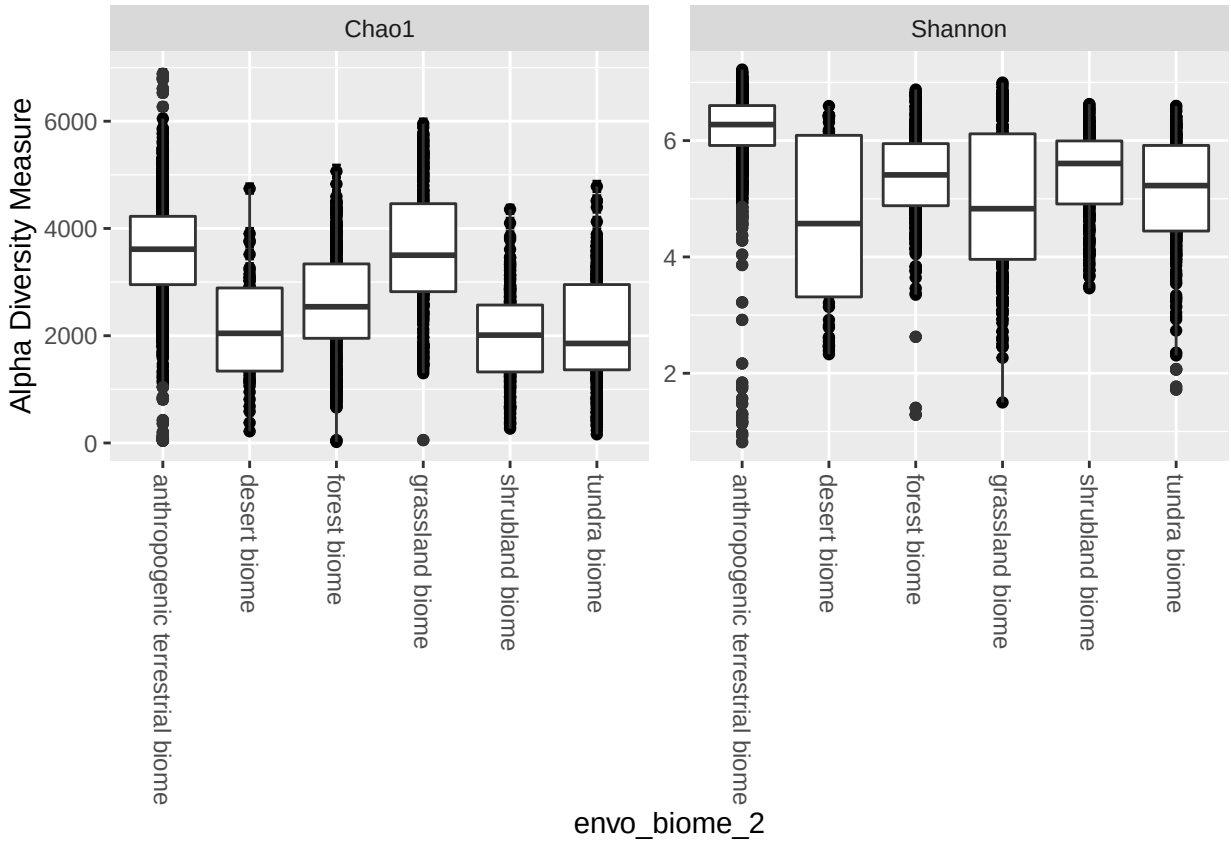
$$R_{Shannon} = -\sum_{i=1}^{k} p_i \log(p_i)$$

where p_i represents the relative abundances of the i-th taxon (Calle 2019).

Both can be represented in form of boxplot, using in-built `phyloseq` function.

```
plot_richness(ds,
              x="envo_biome_2",
              measures=c("Shannon", "Chao1")) + geom_boxplot()
```

```
## Warning: Removed 3043 rows containing missing values (geom_errorbar).
```



Further it can be tested whether both of those richness indices significantly differ between environmental biomes they were sampled in. For this purpose non-paramteric , the Wilcoxon rank-sum test can be performed (aka Mann-Whitney U test) for both indexes.

```
alph_rich = estimate_richness(ds, measures = c("Chao1","Shannon"))
pairwise.wilcox.test(alph_rich$Shannon,
                     sample_data(ds)$envo_biome_2,
                     p.adjust.method = 'hochberg')
```

```
##
##  Pairwise comparisons using Wilcoxon rank sum test
##
## data:  alph_rich$Shannon and sample_data(ds)$envo_biome_2
```

```
##
##                 anthropogenic terrestrial biome desert biome forest biome
## desert biome    8.7e-15                         -            -
## forest biome    < 2e-16                         0.01170      -
## grassland biome < 2e-16                         0.16899      1.3e-05
## shrubland biome < 2e-16                         0.01127      0.19301
## tundra biome    < 2e-16                         0.18759      0.00150
##                 grassland biome shrubland biome
## desert biome    -               -
## forest biome    -               -
## grassland biome -               -
## shrubland biome 3.6e-05         -
## tundra biome    0.19301         0.00074
##
## P value adjustment method: hochberg
```

```
pairwise.wilcox.test(alph_rich$Chao1,
                     sample_data(ds)$envo_biome_2,
                     p.adjust.method = 'hochberg')
```

```
##
##  Pairwise comparisons using Wilcoxon rank sum test
##
## data:  alph_rich$Chao1 and sample_data(ds)$envo_biome_2
##
##                 anthropogenic terrestrial biome desert biome forest biome
## desert biome    < 2e-16                         -            -
## forest biome    < 2e-16                         0.012        -
## grassland biome 0.782                           6.7e-14      < 2e-16
## shrubland biome < 2e-16                         0.740        6.7e-14
## tundra biome    < 2e-16                         0.782        1.6e-09
##                 grassland biome shrubland biome
## desert biome    -               -
## forest biome    -               -
## grassland biome -               -
## shrubland biome < 2e-16         -
## tundra biome    < 2e-16         0.519
##
## P value adjustment method: hochberg
```

Beta diversity

Beta diversity of two or more samples can be performed by tools provided in `vegan` package for R. It is measured by so called distance which indices differences in microbiome composition in given sample. There are few available distnace measurement techniques

e.g.weighted or unwieghted UniFrac, Aitchison and Bray-Curtis. To measure Beta diversity in this project was used last of those aforementioned - Bray-Curtis distance, which can be defined as:

$$d_{BC}(p_i, p_2) = \frac{\sum_{i=1}^{k} |p_{1i} - p_{2i}|}{\sum_{i=1}^{k} (p_{1i} + p_{2i})}$$

where $p_1 = (p_{11}, ..., p_{1k})$ and $p_2 = (p_{21}, ..., p_{2k})$ denote the microbiome relative abundance for two different samples (Calle 2019).

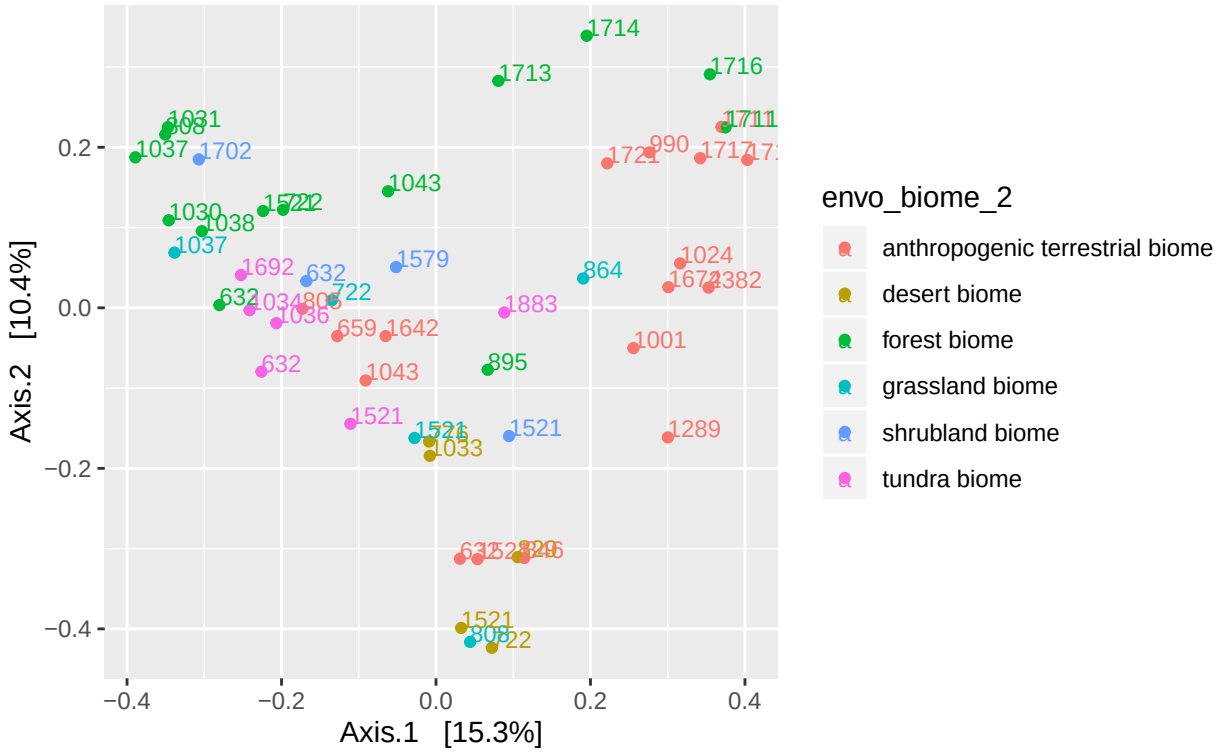
```
bray_dist = phyloseq::distance(phylo_fill_prop, method="bray")
```

Furthermore, to properly visualize beta diversity, ordination plot has to be generated. Most commonly used method of ordination is principal coordinates analysis (PCoA) which is an extension of widely known Principal Component Analysis (PCA). With distance measure, (D),PCoA performs eigenvalue decomposisiton of D_c‘D_c where D_c is the centered distance matrix. If D is the Euclidean distance, PCoA returns same results as those produced by PCA technique.

```
ordination = ordinate(phylo_fill_prop, method="PCoA", distance=bray_dist)
```

Generating ordination plot with both features of interest denoted i.e. envo_biome_2 as colors and study_id in form of labels.

```
plot_ordination(phylo_fill_prop,
                ordination, color="envo_biome_2") +
                theme(aspect.ratio=1) + geom_point(size=1) +
                geom_text(aes(label=study_id),
                          size=3,
                          hjust=0,
                          vjust=0)
```



From plot above it is unclear whether diversity in samples derive from different environmental biomes where the samples were taken or it depends on the study under which samples were collected.

Multivariate differential abundance testing

To recieve more specific statistical data there has to be multivariate differential abundance testing performed. Package `vegan` provides useful tool in this manner, called `anosim`. It allows for distance-based approach analysis of similarities between sample groups (Calle 2019). Once again, two features of interest were compared.

```
anosim(otu_table(phylo_fill_prop),
       sample_data(phylo_fill_prop)$envo_biome_2)
```

```
##
## Call:
## anosim(x = otu_table(phylo_fill_prop), grouping = sample_data(phylo_fill_prop)$envo_b
## Dissimilarity: bray
##
## ANOSIM statistic R: 0.3534
##       Significance: 0.001
##
```

```
## Permutation: free
## Number of permutations: 999
```

```
anosim(otu_table(phylo_fill_prop),
       sample_data(phylo_fill_prop)$study_id)
```

```
##
## Call:
## anosim(x = otu_table(phylo_fill_prop), grouping = sample_data(phylo_fill_prop)$study_
## Dissimilarity: bray
##
## ANOSIM statistic R: 0.5297
##       Significance: 0.002
##
## Permutation: free
## Number of permutations: 999
```

Basing on R value it is clear that samples abundances are more likely to be related to their scientific source, rather than their enivornmental features.

Discussion

While above analysis were not as in-depth as other, performed in frequently cited papers related to study of microbiome dependencies. Nevertheless, working on this data was as insightful as informative for someone that considers using this data as point of reference in future studies. It is quite possible that machine learning approach to whole dataset could perhaps bring more informative results.

Bibliography

Main article that was used in process of making this project was one by Mrs Calle, published in 2019.

Afshinnekoo, Ebrahim, Cem Meydan, Shanin Chowdhury, Dyala Jaroudi, Collin Boyer, Nick Bernstein, Julia M Maritz, et al. 2015. "Geospatial Resolution of Human and Bacterial Diversity with City-Scale Metagenomics." *Cell Systems* 1 (1): 72–87.

Calle, M Luz. 2019. "Statistical Analysis of Metagenomics Data." *Genomics & Informatics* 17 (1).

Gilbert, J. A., F. Meyer, D. Antonopoulos, P. Balaji, C. T. Brown, C. T. Brown, N. Desai, et al. 2010. "Meeting report: the terabase metagenomics workshop and the vision of an Earth microbiome project." *Stand Genomic Sci* 3 (3): 243–48.

Kchouk, Mehdi, Jean-François Gibrat, and Mourad Elloumi. 2017. "Generations of Sequencing Technologies: From First to Next Generation." *Biology and Medicine* 9 (3).

Mangul, Serghei, Lana S Martin, Brian Hill, Angela Ka-Mei Lam, Margaret Distler, Alex Zelikovsky, Eleazar Eskin, and Jonathan Flint. 2018. "Towards Reproducible, Transparent, and Systematic Benchmarking of Omics Computational Tools."

Mason, Christopher E, Ebrahim Afshinnekoo, Scott Tighe, Shixiu Wu, and Shawn Levy. 2017. "International Standards for Genomes, Transcriptomes, and Metagenomes." *Journal of Biomolecular Techniques: JBT* 28 (1): 8.

McIntyre, Alexa BR, Rachid Ounit, Ebrahim Afshinnekoo, Robert J Prill, Elizabeth Hénaff, Noah Alexander, Samuel S Minot, et al. 2017. "Comprehensive Benchmarking and Ensemble Approaches for Metagenomic Classifiers." *Genome Biology* 18 (1): 182.