Multivariate statistical analysis of metagenomic data

Boris Shilov

Introduction

The aim of this analysis is to explore the metagenomic dataset assembled by Forslund et al. (n.d.).

Data format

```r
library(tidyverse)
data = read_tsv("data.r")
head(data)
```

```
## # A tibble: 6 x 6
##   Sample Dataset Status  FeatureType Feature Abundance
##   <chr>  <chr>   <chr>   <chr>       <chr>       <dbl>
## 1 MH0443 MHD     ND CTRL GMM         MF0119        902
## 2 MH0443 MHD     ND CTRL GMM         MF0090        302
## 3 MH0443 MHD     ND CTRL GMM         MF0097          0
## 4 MH0443 MHD     ND CTRL GMM         MF0092          0
## 5 MH0443 MHD     ND CTRL GMM         MF0011       1076
## 6 MH0443 MHD     ND CTRL GMM         MF0083          3
```

We can see that the dataset consists of six columns. The first is sample ID, which are not a unique key since every sample may have multiple features. The second is the dataset of origin, which is either Danish (MHD), Swedish (SWE) or Han Chinese (CHN). The third column is treatment status of a particular patient, where they are either healthy (ND CTRL), are having type 2 diabetes treated with metformin included (T2D metformin+), are having type 2 diabetes treated without metformin (T2D metformin-) or are having type 1 diabetes treated (T1D). The fourth column specifies the type of feature in the fifth column. This is either gut microbial/metabolic module (GMM), SEED database annotation, bacterial family, bacterial genus or metagenomic operational taxonomic units (Motu). The fifth column contains the feature names which are coded differently depending on the feature type. Finally, the sixth column contains the abundance of a given feature.

Bibliography

Forslund, Kristoffer, Falk Hildebrand, Trine Nielsen, Gwen Falony, Le ChatelierEmmanuelle, Shinichi Sunagawa, Edi Prifti, et al. n.d. “Disentangling Type 2 Diabetes and Metformin Treatment Signatures in the Human Gut Microbiota.” *Nature* 528: 262 EP. https://doi.org/10.1038/nature15766.