Report

Background:

Solve the problems R program was used. In R, majorly “dplyr”, “tidyr”, and “reshape2” packages were used for data wrangling. And “ggplot2” and “ggbiplot” packages were used for graph visualization. For natural abundance correction package “accucor” was used for isoptop abundance correction. The usage details are given in the following sections.

library(dplyr)

library(ggplot2)

library(ggbiplot)

library(tidyr)

library(reshape2)

**Steps taken to manipulate data:**

The first step was to find the PCA for all given samples. PCA would broadly tell us if there is a difference between the samples, if so, then which samples would exhibit these differences.

Since PCA is a global approach, therefore it was wise (also suggested) to look at the changes in the metabolite levels at a gross level, that is at the pool total levels.

In order to find the pool totals,

* The files were read with the following names:
  + metadata - sample\_metadata.csv
  + maven\_data - Maven\_processed.csv
* the unnecessary columns were dropped, and only columns with sample names and compound names were retained.
* The rows were grouped according to the compound name, and rows of each of the column according to the group was added together to get the total and data frame called pool\_total was made.
* To calculate fractional enrichment columns, compound, note, and samples, were selected.
* This wide format data was converted into long format data, using “melt” function.
* The samples were given their Phenotype and Time by joining this dataframe with metadata dataframe. ---1
* Also, pool\_total dataframe was converted from wide format to long format
* The metadata was assigned to pool\_total converted dataframe ---2
* Dataframes 1 and 2 were then joined to have all values and metdata. Meaning, all information including measurements, phenotype, time and totals were now present in one dataframe—fraction\_enrichment
* Next to calculate fraction enrichment—column with measurements (value) was divided by Total column and multiplied by 100 and put in a new column called fraction\_enrich.

Two dataframes – pool\_total and fraction\_enrichment, were used for all analyses.

To plot the PCA pool\_total dataframe was used. In order to perform PCA the data has to normally distributed and more importantly, the variation amongst all the variables (samples in our case) need to same. If the variation is not taken care of a variable with huge values will lead to huge variations, where as a variable with low values will cause no variation. These variations can skew the Principal components being calculated. Thus we need to check if the data is normalized or not.

I check the data it is normal and in the same scale for all columns in pool\_total (graphs not saved but code present in the file). None of the data were normal or to scale. I then tried normalizing the data from 0-1. This normalized the variation, however, the scales for each variable were still different. Moreover, there was a clear skew of data towards the extremes, indicating a non-normal distribution. Thus I went for log transformation of the data. Many of the variables now showed normal distribution (visually, not checked with statistics). I then calculated the PCA based on the log transformed data.