1. First look at the metabolite data

Anu Joshi

May 16, 2019

The data in third sheet (VolNormImp) has been:
* Volume normalized: in order to account for samples that had different volumes while performing LC-MS.
* Imputed: missing data has been replaced with the smallest value.

```
metabolite = read_excel("data/MSSM-01-18ML+ CLIENT DATA TABLE 9.25.18.XLSX",
                        sheet = "VolNormImpData",
                        col_names = FALSE)
```

The table has multiple identifiers associated with each metabolites. Here we extract all the identifiers for later use, as all of them are not required at the same time.

```
metabolite_IDs = metabolite[10:1317, 1:13] %>%
  set_colnames(metabolite[9, 1:13])

head(metabolite_IDs)
```

```
## # A tibble: 6 x 13
##   `PATHWAY SORTOR~ BIOCHEMICAL `SUPER PATHWAY` `SUB PATHWAY` `COMP ID`
##   <chr>            <chr>       <chr>           <chr>         <chr>
## 1 1441             (12 or 13)~ Lipid           Fatty Acid, ~ 38293
## 2 1444             (14 or 15)~ Lipid           Fatty Acid, ~ 38768
## 3 1448             (16 or 17)~ Lipid           Fatty Acid, ~ 38296
## 4 4600             (2,4 or 2,~ Xenobiotics     Food Compone~ 62533
## 5 1668             (R)-3-hydr~ Lipid           Fatty Acid M~ 43264
## 6 1669             (S)-3-hydr~ Lipid           Fatty Acid M~ 52984
## # ... with 8 more variables: PLATFORM <chr>, `CHEMICAL ID` <chr>,
## #   RI <chr>, MASS <chr>, PUBCHEM <chr>, CAS <chr>, KEGG <chr>, `Group
## #   HMDB` <chr>
```

```
# Save the values in a separate table for use later
# write.csv(metabolite_IDs, "data/1_metabolite-details.csv", row.names = FALSE)
```

Of all the metabolite identifiers, the "COMP ID" has the most complete data. It is an arbitary value assigned to the metabolites for identification. But since it is the most complete data, we are going to use it to extract our metabolite information.

```
# transpose the information because this is going to be used as column names
compID = metabolite[c(9:1317), 5] %>% t()
```

We are using the Assay ID as participant identifier, this will be later matched with studyID from the covariate data to merge and create a complete file.
In genetic datasets the participants IDs are in rows and the metabolites are in columns. So, we tranpose our dataset to follow that convention. This is the general rule when the number of variables are greater than the number of participants (n > p).

```
metabolite_compID = metabolite[c(3, 10:1317), 14:513] %>%
    t() %>%
    as_tibble() %>%
    set_colnames(compID)
colnames(metabolite_compID)[1] = "Assay"
```

```r
# save the data in a csv file
write.csv(metabolite_compID, "data/1_PRISM-metabolites.csv", row.names = FALSE)
```