LiraFlame_4cer_preprocessing

Asger_Wretlind

24/2/2022

```
#import raw ms peaks from MZmine step 12 export
data_raw_peaks <-
    vroom::vroom(here::here("data-raw/0083_LiraFlame_non-annotated_export_01042022.csv"))
```

```
## New names:
## Rows: 7179 Columns: 257
## -- Column specification
## -------------------------------------------------------- Delimiter: ";" dbl
## (255): row ID, row m/z, row retention time, LiraFlame_12_P1-B3_Sol_3.mzM... lgl
## (2): row identity, ...257
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...257`
```

```
# #View data to check if it is correctly loaded
# View(data_raw_peaks)
```

```
#Select peaks based on specific IDs
#input peak ID based on viewed chromatograms
ID_list <- c("Target_Cer17_H_ISTD" = 1258,
                 "Target_Cer17_H2O_ISTD" = 1255,
                 "Target_Cer16_H" = 9757,
                 "Target_Cer16_H2O" = 3856,
                 "Target_Cer18_H" = 5688,
                 "Target_Cer18_H2O" = 4173,
                 "Target_Cer20_H" = 435,
                 "Target_Cer20_H2O" = 365,
                 #"Target_Cer22_H" = 287,
                 "Target_Cer22_H2O" = 284,
                 "Target_Cer24:0_H" = 2696,
                 "Target_Cer24:0_H2O" = 2700,
                 "Target_Cer24:1_H" = 2889,
                 "Target_Cer24:1_H2O" = 2915)

ID_list <- ID_list[order(ID_list)]

data <- data_raw_peaks %>%
    filter(data_raw_peaks$`row ID` %in% ID_list) %>%
    arrange(`row ID`) %>%
    mutate(`row identity` = names(ID_list)) %>%
    arrange(`row identity`)
```

```r
rm(ID_list, data_raw_peaks)
```

```r
#NOTE This is project specific cleaning

#Remove weird artifact at the final data column
data <- data[,-length(data)]

#Keep only H2o adducts, remove RT, mz, row ID a and pivot table
data <- data %>%
    filter(grepl("H2O", `row identity`)) %>%
    mutate(ID = gsub("_H2O", "", `row identity`)) %>%
    select(-c("row m/z", "row retention time", "row identity", "row ID")) %>%
    pivot_longer(cols = -ID, names_to = "Steno ID") %>%
    pivot_wider(names_from = ID, values_from = value)


#Separate Steno ID into a new columns
data <- data %>%
    mutate(Remade = grepl("remade", `Steno ID`)) %>%
    separate(`Steno ID`,
        c(  "Project",
            "Run_nr",
            "Plate_nr",
            "Plate_pos",
            "Sample_ID",
            "Time_point"))
```

```
## Warning: Expected 6 pieces. Additional pieces discarded in 252 rows [1, 2, 3, 4, 5, 6,
## 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].
```

```r
#Repair inconsistent ID
data[data[,"Sample_ID"] == 14,"Sample_ID"] <- "014"
data[data[,"Sample_ID"] == 51,"Sample_ID"] <- "051"
data[data[,"Sample_ID"] == 84,"Sample_ID"] <- "084"

#Substitute remade samples
data <- data %>%
    mutate(IDxTime = paste0(Sample_ID, "_", Time_point)) %>%
    slice(-which(IDxTime %in% IDxTime[Remade] & !Remade))
```

```r
#Run_nr_plot function
Run_nr_plot <- function(Data, Target, Filter_out = FALSE) {
    Target_pos <- which(colnames(Data) %in% Target)
    Target <- sym(Target)

    tmp_bounds <- data.frame("Mean" = NA, "SD" = NA, "lower" = NA, "upper" = NA)

    Data %>%
        filter(!grepl(Filter_out, `Sample_ID`)) %>%
        pull(Target) %>%
        mean() -> tmp_bounds$Mean
```

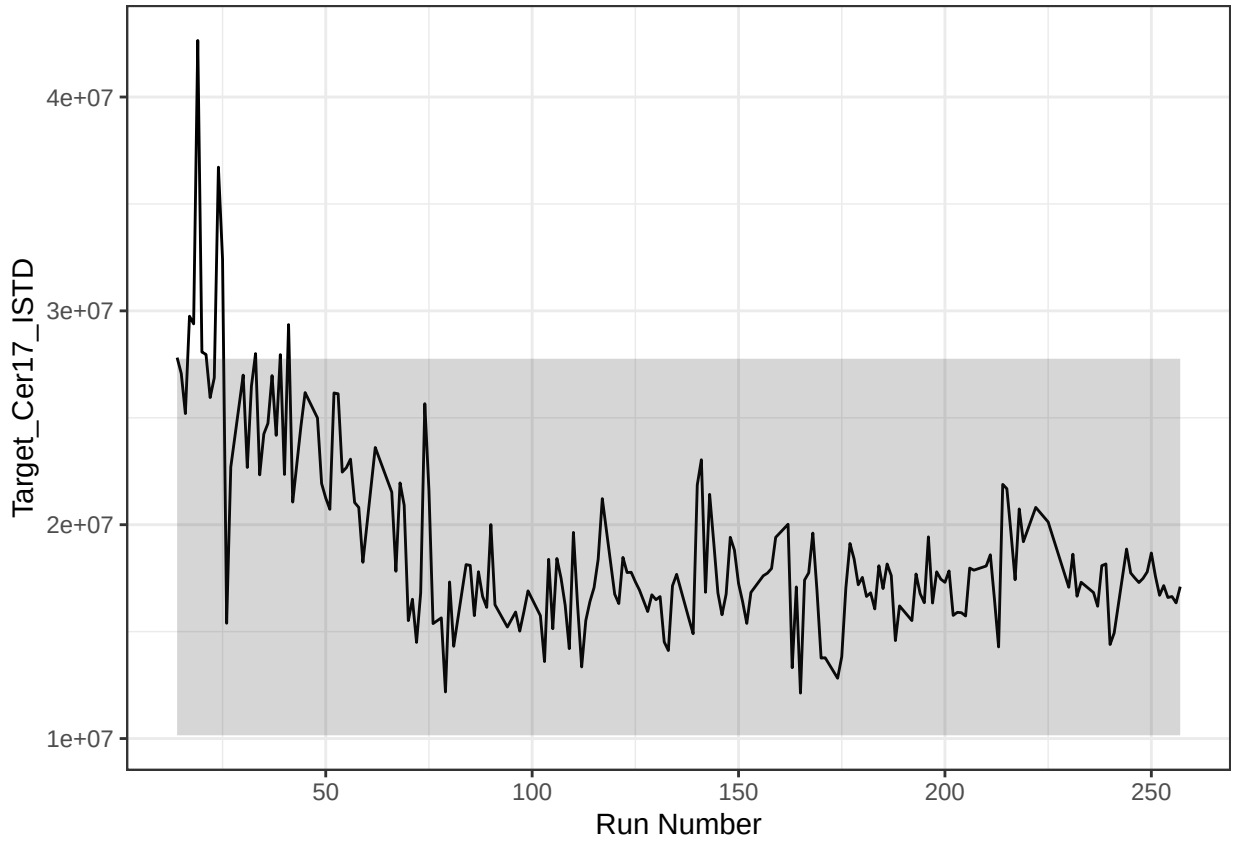
```
    Data %>%
        filter(!grepl(Filter_out, `Sample_ID`)) %>%
        pull(Target) %>%
        sd() -> tmp_bounds$SD

    tmp_bounds$lower <- tmp_bounds$Mean - 2*tmp_bounds$SD
    tmp_bounds$upper <- tmp_bounds$Mean + 2*tmp_bounds$SD

    p <- Data %>%
        filter(!grepl(Filter_out, `Sample_ID`)) %>%
        ggplot(aes(x = as.numeric(Run_nr), y = !!Target)) +
        geom_line() +
        geom_ribbon(aes(ymin = tmp_bounds$lower, ymax = tmp_bounds$upper),
                    alpha = 0.2) +
        xlab(label = "Run Number")+
        theme_bw()

    return(p)
}

#ISTD Run nr plot
Run_nr_plot(Data = data, Target = "Target_Cer17_ISTD", Filter_out = "^[A-Za-z]+")
```




```
# data %>%
#     filter(Sample_ID == "PO") %>%
```

```
#     ggplot(aes(x = as.numeric(Run_nr), y = Target_Cer17_ISTD)) +
#                geom_line()
#
# #cer 16 Run nr plot, w/o QC samples
# Run_nr_plot(Data = data_new_new, Target = "Target_Cer16", Filter_out = "^[A-Za-z]+")
#
# data_new_new %>%
#     filter(Sample_ID == "PO") %>%
#     summarise("Cer16_median" = median(Target_Cer16),
#               "Cer17_median" = median(Target_Cer17_ISTD))
#               #"Cer16_norm" = median(Target_Cer16/Target_Cer17_ISTD))
#
```

```
##NOTE Normalizing to ISTD should no be carried out if using proxy conc.

#Relative Standard deviation (RSD) of Pooled samples before normalization
data %>%
    filter(Sample_ID == "PO") %>%
    summarise(across(starts_with("Target_"), ~ sd(.)/mean(.)*100, .names = "RSD_{.col}"))
```

```
## # A tibble: 1 x 7
##   RSD_Target_Cer16 RSD_Target_Cer17_ISTD RSD_T~1 RSD_T~2 RSD_T~3 RSD_T~4 RSD_T~5
##              <dbl>                 <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1             16.9                  28.2    35.6    60.8    8.92    78.0    67.9
## # ... with abbreviated variable names 1: RSD_Target_Cer18, 2: RSD_Target_Cer20,
## #   3: RSD_Target_Cer22, 4: `RSD_Target_Cer24:0`, 5: `RSD_Target_Cer24:1`
```

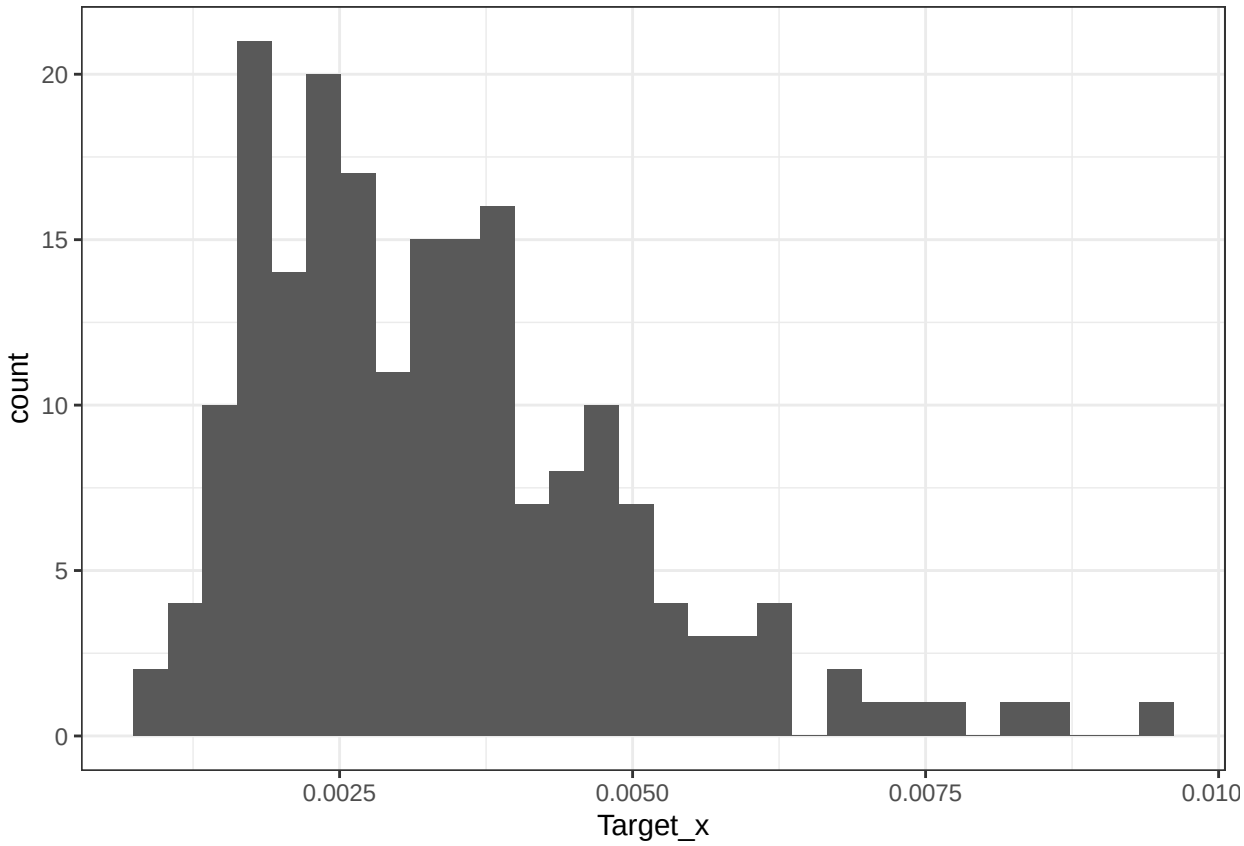
```
#Normalize to ISTD
data <- data %>%
    mutate(across(starts_with("Target") & !contains("ISTD"),
                  ~ ./Target_Cer17_ISTD))

#Relative Standard deviation (RSD) of Pooled samples after normalization
data %>%
    filter(Sample_ID == "PO") %>%
    summarise(across(starts_with("Target_"), ~ sd(.)/mean(.)*100, .names = "RSD_{.col}"))
```

```
## # A tibble: 1 x 7
##   RSD_Target_Cer16 RSD_Target_Cer17_ISTD RSD_T~1 RSD_T~2 RSD_T~3 RSD_T~4 RSD_T~5
##              <dbl>                 <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1             15.0                  28.2    16.8    41.0    25.8    48.5    39.5
## # ... with abbreviated variable names 1: RSD_Target_Cer18, 2: RSD_Target_Cer20,
## #   3: RSD_Target_Cer22, 4: `RSD_Target_Cer24:0`, 5: `RSD_Target_Cer24:1`
```

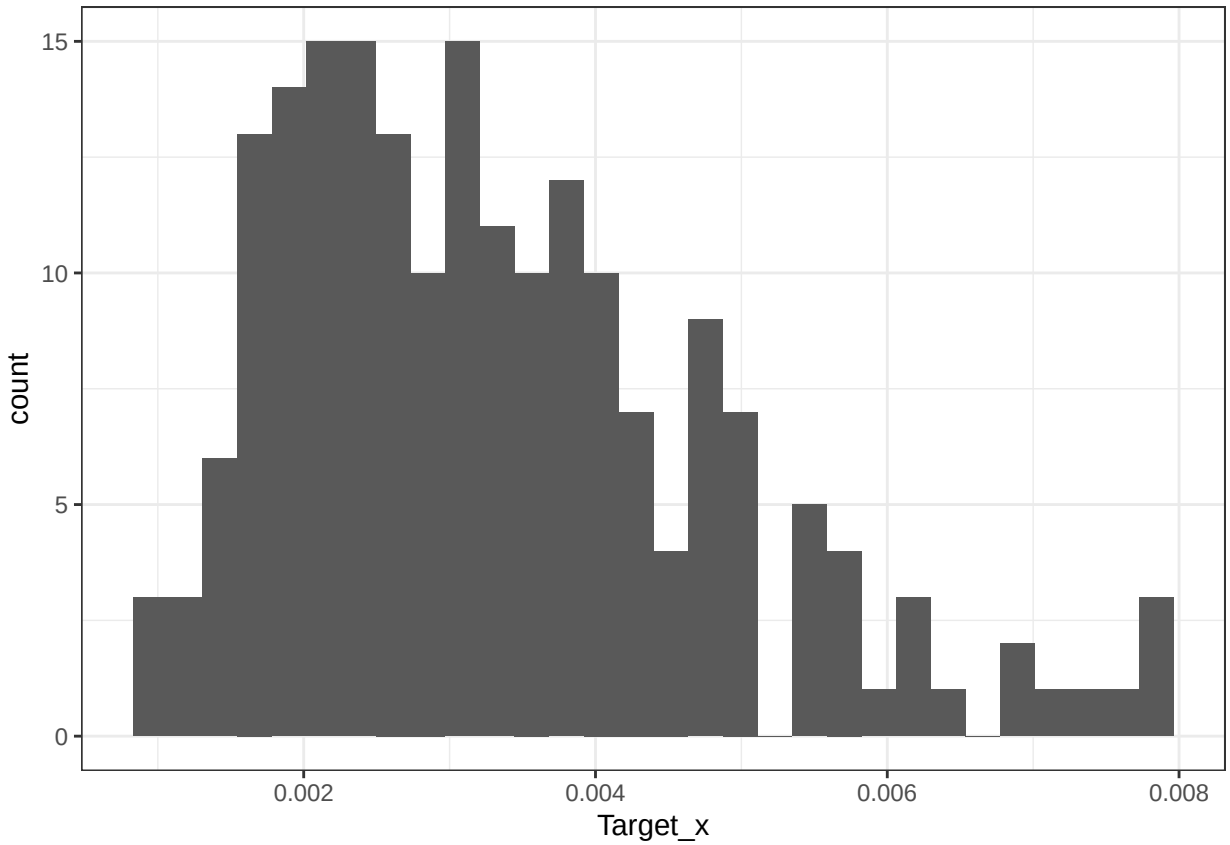
```
#Remove QC samples and unnecessary features
data <- data %>%
    filter(!grepl("^[A-Za-z]+", `Sample_ID`)) %>%
    select(-c(Project, Run_nr, Plate_nr, Plate_pos, Remade, IDxTime)) %>%
    select(-Target_Cer17_ISTD)
```

```r
#plot distribution
data %>%
    mutate(Target_x = Target_Cer18 ) %>%
    #mutate(Target_x = log10(Target_x)) %>%
    #filter(!Target_x > median(Target_x) + 3 * sd(Target_x)) %>%
    #filter(!Target_x < median(Target_x) - 3 * sd(Target_x)) %>%
    ggplot(aes(x = Target_x)) +
    geom_histogram(bins = 30) +
    theme_bw()
```




```r
#Truncate outliers (outside median +- 3*sd) to median +- 3*sd
data <- data %>%
  mutate(across(starts_with("Target"),
  ~ ifelse(. > median(.)+3*sd(.), median(.)+3*sd(.), .))) %>%
  mutate(across(starts_with("Target"),
  ~ ifelse(. < median(.)-3*sd(.), median(.)-3*sd(.), .)))

#plot distribution
data %>%
    mutate(Target_x = Target_Cer18 ) %>%
    #mutate(Target_x = log10(Target_x)) %>%
    ggplot(aes(x = Target_x)) +
    geom_histogram(bins = 30) +
    theme_bw()
```




```
#import clinical measurements
data_raw_clinical <-
    readxl::read_xlsx(path = here::here("data-raw/Liraflame_clinical_data.xlsx"))

#Remove 38 from the patient ID to be able to match Steno ID
data_clinical <- data_raw_clinical %>%
    mutate(PTID = str_sub(PTID, 3, 5))


#Import fixed clinical measurements for Chol, LDL, HDL and Trig
data_raw_clinLip <-
    readxl::read_xlsx(path = here::here("data-raw/Liraflame_clinical_lipid_fix.xlsx"))

#Prepare data_raw_clinLip for merging with data_clinical
data_raw_clinLip <- data_raw_clinLip %>%
    select(PTID, HDL_v2, HDL_v5, LDL_v2, LDL_v5, CHOL_v2, CHOL_v5, TRIG_v2, TRIG_v5) %>%
    mutate(PTID = str_sub(PTID, 3, 5))

#Substitute fixed lipid measures in data_clinical
data_clinical <- data_clinical %>%
    select(-c(starts_with("V1"), starts_with("V3"), starts_with("V4"),
           contains("CHOL"), contains("LDL"), contains("HDL"), contains("TRIG"))) %>%
    left_join(., data_raw_clinLip)
```

```
## Joining with `by = join_by(PTID)`
```

```r
#Create Log10MeanUAER
data_clinical <- data_clinical %>%
    mutate(V2_U_LogMeanUAER = log10((V2_U1_UALB+V2_U2_UALB)/2)) %>%
    mutate(V5_U_LogMeanUAER = log10((V5_U1_UALB+V5_U2_UALB)/2))

rm(data_raw_clinical, data_raw_clinLip)
```

```r
#Merge clinical data and peak data
data <- data %>%
    left_join(x = .,
              y = data_clinical,
              by = c("Sample_ID" = "PTID"))

#Format column type and order
data <- data %>%
    rename(Treatment = kategori) %>%
    relocate(, c(Time_point, Treatment), .after = 1)

rm(data_clinical)
```

```r
#Remove "Target_" from metabolite names
data <- data %>%
    rename_with(~gsub("Target_", "", .))
```

```r
#Export data
# vroom::vroom_write(data, here::here("data/0083_liraflame_4cer_data_preprocessed.csv"))
```