Metadata curation

Ece Kartal

5.10.2022

Metadata is in several excel sheet, we first combine them to work with it.

```
# Set parameters
PARAM <- list()
PARAM$folder.R <- paste0(getwd(), "/")
PARAM$folder <- gsub("src/", "", PARAM$folder.R)
PARAM$folder.input <- paste0(PARAM$folder, "input/")
PARAM$folder.Rdata <- paste0(PARAM$folder, "Rdata/")

library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.0     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(readxl)
```

```
# read meta files
Metadata_PROMISE <- read_excel(paste0(PARAM$folder.input,"/metadata/clean/Metadata_PROMISE.xlsx"))
meta1 <- read_excel(paste0(PARAM$folder.input, "metadata/Summary_vital_status_PROMISE_20231001.xlsx"), s
meta2 <-  read_excel(paste0(PARAM$folder.input, "metadata/Summary_vital_status_PROMISE_20231001.xlsx"),
meta3 <-  read_excel(paste0(PARAM$folder.input, "metadata/Summary_vital_status_PROMISE_20231001.xlsx"),

## TODO ##
meta4 <-  read_excel(paste0(PARAM$folder.input, "metadata/Summary_vital_status_PROMISE_20231001.xlsx"),
meta5 <-  read_excel(paste0(PARAM$folder.input, "metadata/Summary_vital_status_PROMISE_20231001.xlsx"),
meta6 <-  read_excel(paste0(PARAM$folder.input, "metadata/Summary_vital_status_PROMISE_20231001.xlsx"),
meta7 <-  read_excel(paste0(PARAM$folder.input, "metadata/Summary_vital_status_PROMISE_20231001.xlsx"),


# load read counts
mqc_kraken <- read_delim(paste0(PARAM$folder, "data/metag/multiqc_metag/mqc_kraken-top-n-plot_Species.ti
                         delim = "\t", escape_double = FALSE, trim_ws = TRUE)
```

```
## Rows: 335 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (1): Sample
## dbl (2): Homo sapiens, Unclassified
##
## 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.
```

```
colnames(mqc_kraken) <- c("Filename", "ReadCount (Homo)", "ReadCount (Microbiome)")
# Replace "-kraken2-report" with an empty string in the specified column
mqc_kraken$Filename <- gsub("-kraken2-report", "", mqc_kraken$Filename)
# Replace "-" with "." in the "Filename" column
mqc_kraken$Filename <- gsub("-", ".", mqc_kraken$Filename)
```

```
# there are a lot of duplicates, rmeove duplicate rows
meta2=distinct(meta2)
meta3=distinct(meta3)
# Combine the two dataframes by StudentID
# Pivot the long table into a wider format
# wide_data <- meta2 %>%
#   pivot_wider(names_from = Timepoint, values_from = Note)
df1 <- Metadata_PROMISE %>%
  left_join(meta2, by = c("StudienID", "Timepoint"))
 # left_join(meta3, by = c("StudienID", "Timepoint"))

df2  <- df1 %>%
  left_join(meta1, by = c("StudienID"))
df3 <- df2 %>%
  full_join(mqc_kraken, by = c("Filename"))
```

Response variable can not be added because so many patients have double coded respoonse that does not make sense. For e.g. SD (Stable disease) and PD (Progressive disease) for the same patient and same timepoint

```
meta=df3
meta$`ReadCount (Microbiome)` = as.integer(meta$`ReadCount (Microbiome)`)
meta$ID <- paste0(meta$SampleName, "_", meta$Timepoint, "_", meta$Sequencing_batch, "_",meta$Resubmissi

# recode all the T2 extended as T3 after taslking with Inka
meta <- meta %>%
  # mutate(Timepoint = ifelse(Timepoint == "T2 extended", "T3", Timepoint)) %>%
  mutate(Timepoint = ifelse(Timepoint == "0", "Control", Timepoint)) %>%
  filter(Sequencing_batch %in% c(0,1,2,3))
```