Metadata curation for LassalleF_2017

Marisa Isabell Metzger

Oktober 2018

This markdown creates the curated and standardised metadata table for the curatedMetagenomicData package of the study LassalleF_2017, which provides metagenomic data for 24 oral microbiomes from hunter-gatherers and traiditional farmers living on the Philippines.

load required packages

```
library(readxl)
library(tibble)
library(dplyr)
library(tidyr)
library(readr)
```

import Data

the "study" table comes directly from the LassalleF_2017 study supplementary information. It contains mostly information about the sex, age, lifestyle information of the study participants. The "mapping" and the "metadata" tables were created by the python script "ncbi_downloader_dev.py". This script downloads the raw data from NCBI and maps the sampleID with the NCBI accession code ("mapping"), in addition it creates a table with the init metadata information from the study with information like isolation source, geographic location, size,...

There is no unqiue approach to generate the metadata for a study. The distribution of the metadata in those three tables described above is not the same for each study. Some information about metadata can also be found in the text of the paper. Therefore, the curator for cmd has to manually curate the metadata and standardise and has to start the approach of curation for each study all over again.

In general, the tables have one row for each sample in the study and the information about the samples are arranged in the columns

```
study <- read_excel("C:/Users/Marisa/Documents/Biologie_Studium/Master/Segata_Lab_Rotation/Literature/La

mapping <- read_table2("C:/Users/Marisa/Documents/Biologie_Studium/Master/Segata_Lab_Rotation/Literature

metadata <- read_delim("C:/Users/Marisa/Documents/Biologie_Studium/Master/Segata_Lab_Rotation/Literature
head(study)
```

```
## # A tibble: 6 x 13
##   Sample Population Sex   Age   Run   `Number reads` `Number reads, ~
##   <chr>  <chr>      <chr> <chr> <chr>          <dbl>            <dbl>
## 1 Ae10   Aeta       M     23    1          454127918         30923276
## 2 Ae12   Aeta       F     30    1          328005076         55733807
## 3 Ae61   Aeta       F     26    1          434480393         40939249
## 4 Ae08   Aeta       M     24    1          365011123         26075608
## 5 Cae01  Zambal     F     na    1          469673252        101962267
## 6 CAe42  Zambal     M     na    1          296596746         32889119
## # ... with 6 more variables: `Microbiome read fraction` <dbl>,
## #   lifestyle <chr>, Year <dbl>, Island <chr>, `Average Village/Camp GPS
## #   Coordinates` <chr>, `EBI Metagenomics Run_ID` <chr>
```

```
head(mapping)
```

```
## # A tibble: 6 x 2
##   X3           X5
##   <chr>        <chr>
## 1 SID700171428 ERR1474612
## 2 SID700161820 ERR1474611
## 3 SID700037591 ERR1474610
## 4 SID700023710 ERR1474609
## 5 SID700023346 ERR1474608
## 6 SID700023122 ERR1474607
```

```
head(metadata)
```

```
## # A tibble: 6 x 31
##   `ENA-FIRST-PUBL~ `ENA-LAST-UPDAT~ `Sample Name` Studysampleid
##   <date>           <date>           <chr>         <chr>
## 1 NA               NA               <NA>          SID700171428
## 2 NA               NA               <NA>          SID700023710
## 3 NA               NA               <NA>          SID700023346
## 4 NA               NA               <NA>          SID700023122
## 5 NA               NA               <NA>          SID700021297
## 6 2017-11-15       2017-11-09       ERS1202885    SAMEA4031775
## # ... with 27 more variables: `analyte type` <chr>, `biospecimen
## #   repository` <chr>, `biospecimen repository sample id` <int>,
## #   `colection date` <int>, `environment (biome)` <chr>, `environment
## #   (feature)` <chr>, `environment (material)` <chr>, gap_accession <chr>,
## #   gap_consent_code <int>, gap_consent_short_name <chr>,
## #   gap_sample_id <int>, gap_subject_id <int>, `geographic location
## #   (country and/or sea)` <chr>, host_sex <chr>, `human oral environmental
## #   package` <chr>, `investigation type` <chr>, isolation_source <chr>,
## #   latitude <dbl>, longitude <dbl>, `project name` <chr>, `sequencing
## #   method` <chr>, size <dbl>, `study design` <chr>, `study name` <chr>,
## #   `submitted sample id` <int>, `submitted subject id` <int>, `submitter
## #   handle` <chr>
```

For each sample, the statistics were calculated with the python script fna_len.py (bitbucket repositry / PyPhlAn /Source) The "statistic" table contains the jointed information about the number of bases, number of reads and the minimum, mean, median and maximum read length.

```
# the working directory were set for this junk to import the statistic information. By default the work
setwd("C:/Users/Marisa/Documents/Biologie_Studium/Master/Segata_Lab_Rotation/Literature/LassalleF_2017/
temp <- list.files(pattern="*.stats*")

statistic <- NULL
for (i in 1:length(temp)){
  stats_tmp <- read_delim(temp[i], "\t",trim_ws=TRUE)
  stats_tmp <- dplyr::bind_cols(ID = rep(temp[i], nrow(stats_tmp)),
                             stats_tmp)
  statistic <- dplyr::bind_rows(statistic, stats_tmp)
}
head(statistic)
```

```
## # A tibble: 6 x 8
##   ID    `#samplename` n_of_bases n_of_reads min_read_len median_read_len
##   <chr> <chr>              <dbl>      <int>        <int>           <dbl>
## 1 SAME~ stdin_fastq   1354644126   14236398           43              97
## 2 SAME~ stdin_fastq   2202468578   24523600           40              94
## 3 SAME~ stdin_fastq   2428708630   25284648           43              98
## 4 SAME~ stdin_fastq   1288478907   15465158           42              91
## 5 SAME~ stdin_fastq   8014713960   83782478           44              98
## 6 SAME~ stdin_fastq   1549241066   16758352           43              97
## # ... with 2 more variables: mean_read_len <dbl>, max_read_len <int>
```

the three tables metadata, mapping and study were merged to one table by common columns to ensure the correct matching of cohesive rows (samples). The column “X3” in the table mapping is equal to the “”studysampleid" column in the metadata table. And the column “X5” from the mapping table corresponds to the “EBI Metagenomics Run_ID” column.
mapping (X3) = metadata (studysampleid)
mapping (X5) = study(EBI Metagenomics Run_ID)

```
metadata <- left_join(metadata, mapping, by = c("Studysampleid" = "X3"))
metadata <- left_join(metadata, study, by = c("X5" = "EBI Metagenomics Run_ID"))
```

Curate the Metadata

In addition to the 24 samples from humans living on the Philippines, the study uses control samples from healthy subjects in the HMP. This samples can be removed. In addition, we removed columns which gave us no information about the Philippine samples.

```
metadata <- subset(metadata, metadata$`environment (biome)`== "human")
metadata <- metadata[,colSums(is.na(metadata)) !=nrow(metadata)]
```

some columns has to be renamed to fit into the cmd standardisation.

```
metadata <- plyr::rename(metadata, replace = c(
                  "X5" = "NCBI_accession",
                   "Sex" = "gender",
                   "environment (feature)" = "body_site",
                   "geographic location (country and/or sea)" = "country",
                   "sequencing method" = "sequencing_platform",
                   "Sample" = "sampleID",
                   "Island" = "location",
                   "Age" = "age",
                   "Population" = "population"))
```

some columns have information, which is not necessary for the cmd. This columns are removed

```
delete <- c("ENA-FIRST-PUBLIC",
            "ENA-LAST-UPDATE",
            "colection date",
            "environment (biome)",
            "environment (material)",
            "human oral environmental package",
            "investigation type",
            "latitude",
            "longitude",
            "project name",
            "Year",
            "Average Village/Camp GPS Coordinates",
            "Number reads",
            "Number reads, human screened out",
            "Microbiome read fraction",
            "Run",
            "size",
            "Sample Name")

metadata <- metadata[,!(colnames(metadata) %in% delete), drop = FALSE]
```

Adding columns with general information for the samples found in the text of the paper (non-westernized, study_condition, disease) or the information became necessary by the curation (curator, PMID)

```
metadata <- mutate(metadata,curator="Marisa_Metzger",
                   PMID="29165844",
                   non_westernized = "yes",
                   study_condition = "control",
                   disease = "healthy"
                   )

metadata$age_category <- ifelse(metadata$age >= 19 , "adult", NA )
```

the content of some columns in the metadata are necessary and important, but the description does not fit to our cmd standards. Here, we change the content of those columns.

```
metadata <- within (metadata, body_site[body_site == "oral"] <- "oralcavity")
metadata <- within (metadata, country[country == "Philippines"] <- "PHL")
metadata <- within (metadata, sequencing_platform[sequencing_platform == "Illumina HiSeq"] <- "Illumina
metadata <- within (metadata, gender[gender == "F"] <- "female")
metadata <- within (metadata, gender[gender == "M"] <- "male")
metadata <- within (metadata, age[age == "na"] <- NA)
```

Adding statistical information to metadata

the statistic table gets modified to suit our standars and to merge it afterwars with the metadata table.

```
statistic$`#samplename` <- NULL
statistic$mean_read_len <- NULL
statistic$max_read_len <- NULL
statistic <- separate(statistic, ID, into= c("Studysampleid", "Stat"), sep=".sta")
```

```
statistic$Stat <- NULL
statistic <- plyr::rename(statistic, replace = c(
   "n_of_bases" = "number_bases",
   "n_of_reads" = "number_reads",
   "min_read_len" = "minimum_read_length",
   "median_read_len" = "median_read_length"))
```

merge the statistic table to the metadata table by the column “Studysampleid”

```
metadata <- dplyr::left_join(metadata, statistic, by= "Studysampleid")
```

checking for unique sampleID

the sampleID has to be unique, the following command will check if this is the case.

```
metadata <- plyr::rename(metadata, replace = c(
  "sampleID" = "subjectID",
  "Studysampleid" = "sampleID"))
```

put the sampleID column on the first place

```
col_idx <- grep("sampleID", names(metadata))
col_idx2 <- grep("subjectID", names(metadata))
metadata <- metadata[,c(col_idx, col_idx2,(1:ncol(metadata))[-c(col_idx, col_idx2)])]
```

```
ns <- unique(metadata$sampleID)

if(length(ns) == nrow(metadata)) {
  print("sampleID is unique")
} else {
  print("WARNING: sampleID is not unique")
}
```

```
## [1] "sampleID is unique"
```

```
print(metadata)
```

```
## # A tibble: 24 x 21
##    sampleID subjectID body_site country sequencing_plat~ NCBI_accession
##    <chr>    <chr>     <chr>     <chr>   <chr>            <chr>
##  1 SAMEA40~ CAg34     oralcavi~ PHL     IlluminaHiSeq    ERR1474587
##  2 SAMEA40~ CAg32     oralcavi~ PHL     IlluminaHiSeq    ERR1474586
##  3 SAMEA40~ CAg09     oralcavi~ PHL     IlluminaHiSeq    ERR1474585
##  4 SAMEA40~ Cag05     oralcavi~ PHL     IlluminaHiSeq    ERR1474584
##  5 SAMEA40~ Ag57      oralcavi~ PHL     IlluminaHiSeq    ERR1474583
##  6 SAMEA40~ Ag44      oralcavi~ PHL     IlluminaHiSeq    ERR1474582
##  7 SAMEA40~ Ag27      oralcavi~ PHL     IlluminaHiSeq    ERR1474581
##  8 SAMEA40~ Ag09      oralcavi~ PHL     IlluminaHiSeq    ERR1474580
##  9 SAMEA40~ B99       oralcavi~ PHL     IlluminaHiSeq    ERR1474579
## 10 SAMEA40~ B73       oralcavi~ PHL     IlluminaHiSeq    ERR1474578
## # ... with 14 more rows, and 15 more variables: population <chr>,
## #   gender <chr>, age <chr>, lifestyle <chr>, location <chr>,
## #   curator <chr>, PMID <chr>, non_westernized <chr>,
## #   study_condition <chr>, disease <chr>, age_category <chr>,
## #   number_bases <dbl>, number_reads <int>, minimum_read_length <int>,
## #   median_read_length <dbl>
```

save the table

Finally, the metadata table can be saved and included to the cmd.

```
write.table(metadata, file="C:/Users/Marisa/Documents/Biologie_Studium/Master/Segata_Lab_Rotation/Litera
sep = "\t",
quote = FALSE,
col.names = TRUE,
row.names = FALSE
)
```

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.