libra

The toolbox for analysing LC-MS based metabolomics data (cleaning, calibration)

User manual
version 0.1.1

20, Apr, 2020

Installation

- devtools::install_github(“tuhulab/libra”)

Workflow

Overview

- libra workflow is streamlined by following the philosophy of tidyverse design. The cleaning and calibration process can be realised with one line of code:
- data_table %>% rm_feature.blank() %>% rm_feature.bird() %>% imputeNA() %>% calibrateBatch.inter.rlm()
  - Use calibrateBatch.intra.rlm() instead for intra-batch calibration

Workflow

Before start

- Data transformation
  - The simplicity of the workflow sacrifices the flexibility of the input data structure.
  - The data from up-stream analysis needs to be transformed to the specific format. But libra is highly compatible with XCMS/CAMERA, therefore it is very easy to transform data generated by XCMS/CAMERA.

Workflow

Data transformation

- Two tables are needed for libra: data_table, sample_table.
  - data_table stores feature (ion) information and intensity for each sample.
    - Each row stores information for one feature; while each column stores information for one sample.
    - The mandatory columns are mz, rt, Xi (X1, X2, X3, …, Xn)
      - libra assumes the sequence of Xi is the sequence of injection.
      - Minute (minute) is the preferred unit for rt.

Workflow

Data transformation

```
> data_table
# A tibble: 650 x 62
      mz    rt pcgroup adduct     X1      X2     X3     X4      X5     X6     X7     X8     X9     X10    X11     X12     X13     X14     X15     X16     X17     X18     X19     X20
   <dbl> <dbl> <chr>   <chr>   <dbl>   <dbl>  <dbl>  <dbl>   <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
 1  69.0  0.4   25      "[M+H… NA      NA      NA      217. 1.53e+2  170.   295.  1.54e2 1.67e2 3.53e2  150.  3.03e+2 2.66e2  163.    192.    152.    156.  1.05e2  119.  8.85e1
 2  76.1  0.51 256      ""      NA      NA      NA      321. 4.06e+2  114.   332.  2.89e2 1.48e2 2.97e2  152.  5.71e+2 1.79e2  306.     91.9    50.9   255.  6.76e1  571.  1.11e2
 3  77.0  3.06 3        ""      NA     188.    110.    1197. 4.67e+2  498.  1447.  4.86e2 6.76e2 1.67e3  659.  1.57e+3 1.06e3  909.    472.    375.    890.  2.47e2 1124.  3.35e2
 4  81.0  1.72 250      ""      NA      NA      NA       109. 5.03e-1  344.    11.2 4.98e0 1.31e0 2.01e0   11.8 7.86e-1 9.19e0   21.6   208.    345.     63.9 1.40e2   33.1 1.83e0
 5  81.1  3.46 68       "[M+H…  0.507   NA      NA      187. 1.93e+1   72.2  104.  1.95e1 4.06e1 8.45e2   26.8 1.77e+2 1.78e2   18.4    15.7    11.6   100.  6.01e0   29.5 7.39e1
 6  81.5  0.41 365      ""      0.540   NA       2.16   317. 1.53e+2  314.   515.  2.82e2 9.23e1 5.57e1  201.  1.56e+2 3.03e2  284.    142.    144.    186.  4.24e1  339.  4.03e0
 7  82.0  0.4   25      "[M+H…  2.30    0.511   36.9   2828. 1.97e+3 2127.  3745.  2.35e3 2.40e3 7.41e3 2084.  3.94e+3 3.37e3 2219.   2370.   1988.   2627.  1.40e3 2394.  1.29e3
 8  83.0  0.41 25       ""      NA      0.512    1.02   478. 3.25e+2  351.   635.  3.85e2 4.06e2 1.23e3  368.  6.74e+2 6.06e2  370.    422.    348.    462.  2.40e2  394.  2.21e2
 9  83.0  2.6  153      "[M+H… 10.8     NA      NA      226. 5.41e+1   65.2  367.  7.71e1 1.80e2 3.74e2  212.  3.60e+2 4.65e2  314.    128.     93.6   183.  1.67e1  348.  1.26e1
10  83.1  3.49 63       ""      NA      NA      NA      397. 4.44e+1  176.   189.  1.01e2 6.49e1 2.40e3   72.3 2.58e+2 4.38e2   26.4    36.5    28.0   312.  5.33e0   45.2 1.36e2
# … with 640 more rows, and 38 more variables: X21 <dbl>, X22 <dbl>, X23 <dbl>, X24 <dbl>, X25 <dbl>, X26 <dbl>, X27 <dbl>, X28 <dbl>, X29 <dbl>, X30 <dbl>, X31 <dbl>,
#   X32 <dbl>, X33 <dbl>, X34 <dbl>, X35 <dbl>, X36 <dbl>, X37 <dbl>, X38 <dbl>, X39 <dbl>, X40 <dbl>, X41 <dbl>, X42 <dbl>, X43 <dbl>, X44 <dbl>, X45 <dbl>, X46 <dbl>,
#   X47 <dbl>, X48 <dbl>, X49 <dbl>, X50 <dbl>, X51 <dbl>, X52 <dbl>, X53 <dbl>, X54 <dbl>, X55 <dbl>, X56 <dbl>, X57 <dbl>, X58 <dbl>
```

pcgroup and adduct are not mandatory columns.

Workflow

Data transformation

- The mandatory columns of sample_table are
  - code: X1, X2, … Xn (corrsponding to the column name in data_table)
  - sample_name
    - Regular expression (pattern) of sample_name is used to
      - Distinguish blank, pool for cleaning features
      - Remove samples
  - batch
    - Use 1, 2, 3, … n to represent the batch

Workflow

Data transformation

```
> sample_table
# A tibble: 58 x 8
   sample_type  subject  time mode  drink  code  batch sample_name
   <chr>          <int> <int> <chr> <fct>  <chr> <dbl> <chr>
 1 blank             NA    NA pos   NA     X1        1 Blank
 2 blank+intstd      NA    NA pos   NA     X2        1 Blank+intstd
 3 metstd            NA    NA pos   NA     X3        1 Metstd
 4 ps                NA    NA pos   NA     X4        1 Urinepool
 5 sample          5761     5 pos   Water  X5        1 5761-2-5
 6 sample          5757     2 pos   Coffee X6        1 5757-1-2
 7 sample          5764     2 pos   Water  X7        1 5764-2-2
 8 sample          5761     2 pos   Water  X8        1 5761-2-2
 9 sample          5755     0 pos   Coffee X9        1 5755-1-0
10 sample          5765     0 pos   Water  X10       1 5765-2-0
# … with 48 more rows
```

sample_type, subject, time, mode, drink are not mandatory columns

Workflow

Remove features

- Remove noise features
  - The main purpose is to increase the statistical power (discriminating ability of the statistical model).
  - Based on blank samples and early-/late- eluting compounds (birds)

Workflow

Remove features

- rm_feature.blank()
  - Assumption: If a feature has **a high prevalence in blank samples** and **relative higher intensity in blank samples than in pooled samples**, this feature can be considered as a noise feature.
  - The prevalence is defined by **threshold.prevalence.** The default value is 0.6 - if the feature is presented in 60% of blank samples, the feature is marked as a potential noise feature.
  - The intensity is defined by **threshold.intensity.fraction.** The default value is 0.67 - if the average intensity of the feature in blank samples is higher than **2/3** of the mean intensity in pooled samples, the feature is marked as a potential noise feature.
  - If **both** credentials are fulfilled, the features will be removed.

Workflow

Remove features

- rm_feature.blank()
  - How does libra know which samples are blank/pool? libra learns from the sample_name column of sample_table
  - pattern.blank = “Blank$ | Assay blank"
  - pattern.pool = "Global pool IS$ | Urinepool | urinepool | pool”

Workflow

Remove features

- rm_feature.bird()
  - bird.litmit.low defines the lower-limit of the noise feature.
  - bird.limit.high defines the upper-limit of the noise feature.
  - col.rt: the column name of data_table to store retention time.
    - The default value is rt. If another variable name is used, such as “retention_time” or “RT”, libra will use the specific col.rt
    - libra keeps improving its flexibility, e.g. supporting custom defined variable names :-)

Workflow

Removing samples

- The main purpose is to improve the calibration effects.
- rm_sample()
  - sample.df
    - The name of sample_table. The default value is sample_table. If the alternate name is used, e.g. sample_table_pos, libra will search code (X1, X2, …, Xn) from sample_table_pos
  - sample_rm: these samples will be removed prior to batch effect calibration.
    - sample_rm <- c("Blank", "Blank+intstd", "Blank_IS", "Assay_blank", "MetStd")

Workflow

Impute missing values

- imputateNA()
  - XCMS will return a missing value (NA) if it failed to detect a peak. imputateNA can introduce a randomised intensity (between zero to 2/3 of the lowest detected intensity).
  - None variable needs to be specified for this function. This function supports a non-standard variable name, intensity, which can be specified by the user.
  - Imputate is a miss spelled word and will be corrected in later version.

Workflow

Batch effect calibration - intra or inter

- calibrateBatch.inter.rlm()
  - This function calibrates inter-batch effect. The following non-standard variables are accepted: feature, batch, intensity.
- calibrateBatch.intra.rlm()
  - This function calibrates intra-batch effect. The following non-standard variable is accepted: intensity.

Summary

- libra provides a streamlined workflow for cleaning and calibrating the batch effect of metabolomics data. The advantages of libra include:
  - open-sourced
  - The compatibility with XCMS/CAMERA
  - Every parameter is tuneable.
- The limitations of the current version libra include:
  - The limited number of functionality
  - The documentation has a large space to improve.