MetCleaning v0.99.3

Xiaotao Shen(shenxt@sioc.ac.cn) and Zhengjiang Zhu

2017-03-21

Introduction

***MetCleaning* provides a comprehensive pipeline for data cleaning and statistical analysis of large-scale mass spectrometry (MS) based-metabolomics data. It includes missing value (MV) filtering and imputation, zero value filtering, detection of sample outliers, data normalization, data integration, data quality assessment, and common statistical analysis such as univariate and multivariate statistical analysis. This document describes the step-by-step processing metabolomics data using *MetCleaning*.**

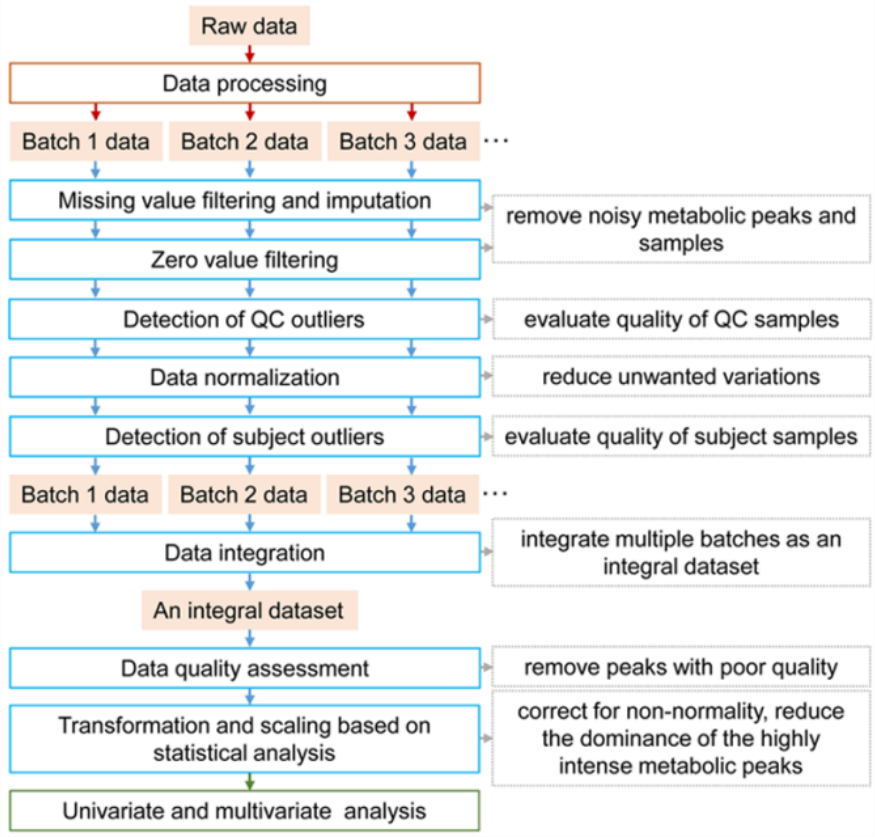

Figure1 Workflow of instruction

Figure1. The detailed data cleaning pipeline for large-scale mass spectrometry-based untargeted metabolomics using the R package MetCleaning.

Installation and help

MetCleaning is published in github. So you can install it via to github.

code 1: Installation of *MetCleaning*

```
##pcaMethodsand impute should be installed form bioconductor
##pcaMethos
source("http://bioconductor.org/biocLite.R")
    biocLite("pcaMethods")
##impute
source("http://bioconductor.org/biocLite.R")
    biocLite("impute")
 if(!require(devtools)) {
  install.packages("devtools")
 }
 library(devtools)
 install_github("jaspershen/MetCleaning")
 library(MetCleaning)
 help(package = "MetCleaning")
```

Demo data

Demo data in *MetCleaning* is from a study to discover metabolite biomarkers for screening of esophagus cancer (EC). The participants were screened using endoscope and iodine staining for EC (golden standard for diagnosis of EC). The participants were divided into two classes according to their reaction to iodine staining: screening positive and screening negative.

In *MetCleaning* package, we selected a two-batch dataset as an example. The dataset contains 1401 metabolic peaks and 606 samples (291 subject samples and 34 QC samples). See the detailed information in Table 1. In *MetCleaning*, metabolomics data is named as "data.csv" and sample information is named as "sample.information.csv".

Table1. The basic information of demo data in MetCleaning.

Variable	Screen negative	Screen positive	QC number	Total
Sample number	141	141	34	291
Batch 1	141	116	34	291

Data cleaning

Data cleaning is integrated as a function named as *MetCleaning*. We use the demo data as the example. Copy the code below and paste in you R console.

code 2: Demo data of *MetCleaning*

```
library(MetCleaning)
##demo data
data(data, package = "MetCleaning")
data(sample.information, package = "MetCleaning")
##demo work directory
dir.create("Demo for MetCleaning")
setwd("Demo for MetCleaning")
##write files
write.csv(data, "data.csv", row.names = FALSE)
write.csv(sample.information , "sample.information.csv", row.names = FA
LSE)
```

The demo data have been added in your work directory and organized as Figure 2 shows. It contains two files, "data.csv" and "sample.information.csv".

1. "data.csv" is the raw metabolomics dataset. Rows are metabolic peaks, and columns are metabolic peak abundance of samples and information of metabolic peaks. The information of metabolic peaks must contain "name" (peak name), "mz" (mass to change ratio) and "rt" (retention time). Other information of metabolic peaks is optional, for example "isotopes" and "adducts". The name of sample can contain ".", but cannot contain "-" and space. And the start of sample name cannot be number. For example, "A210.a" and "A210a" are valid, but "210a" or "210-a" are invalid.
2. "sample.information.csv" is sample information for metabolomics dataset. Column 1 is "sample.name" which is the name of subject and QC samples. Please confirm that the sample names in "sample.information.csv" and "data.csv" are completely same. Column 2 is "injection.order" which is the injection order of QC and subject samples. Column 3 is "class", which is used to distinguish "QC" and "Subject" samples. Column 4 is "batch" to provide acquisition batch information for samples. Column 5 is "group", which is used to label the group of subject sample, for example, "control" and "case". The "group" of QC samples is labeled as "QC".

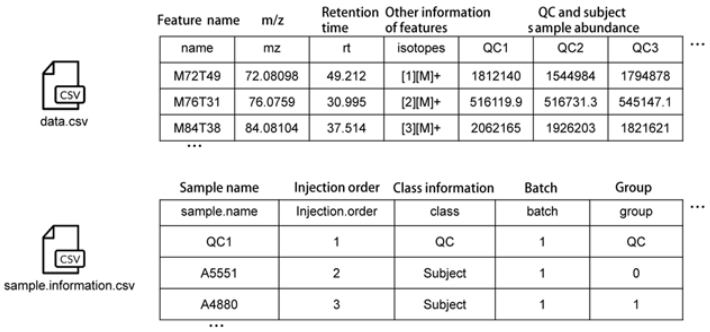

Figure2 Data organization of MetCleaning

Figure2. Data organization and data format of MetCleaning.

Then you can run *MetCleaning* function to do data cleaning of data. All the arguments of *MetCleaning* can be found in MetCleaning. You can use help(package = "MetCleaning") to see the help page of *MetCleaning*.

code 3: Running of *MetCleaning*

```
##demo data
library(MetCleaning)
MetCleaning(polarity = "positive")
```

Running results of *MetCleaning*

1. Missing or zero values filtering. In the missing or zero value filtering step, if there are samples which beyond the threshold you set, you should decide to remove them or not. We recommend removing all of them as Figure 3 shows.

```
Missing values filter...
No QC should be removed.
X257 A5546 X231 sholud be removed!!!
Subject shoulde be removed are:
65 160 177
Which subject you want to remove(please type the index of subject sample,65,160,177
        and separate them using comma,
        if you don't want to remove any subject, please type n):
```

Figure3 Missing or zero value filtering

Figure3. Missing or zero value filtering.

2. Detection of sample outliers. In the detection of QC or subject sample outlier step (based on PCA), if there are samples which beyond the threshold you set, you should decide to remove them or not. We don't recommend to remove them as Figure 4 shows, because they should be considered combined other information.

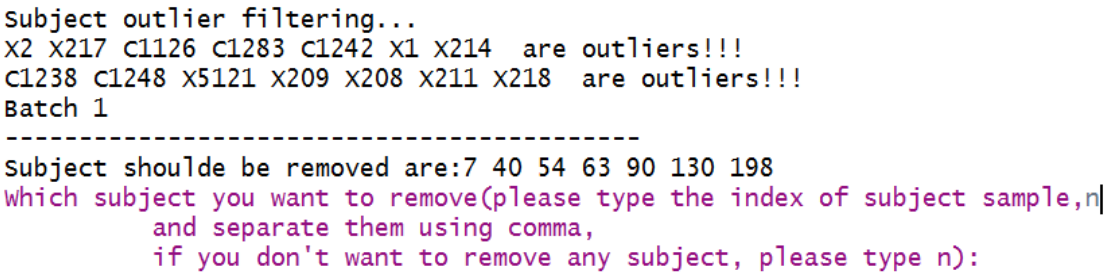

Figure4 Sample filtering

Figure4. The detection of sample outliers step in MetCleaning.

3.Output files. Output files of *MetCleaning* are listed as Figure 5 shows.

(1) "1MV overview", "2MV filter", "3Zero overview" and "4Zero filter" are missing and zero values filtering information.

(2) "5QC outlier filter" and "6Subject outlier filter" are sample filtering based on PCA information.

(3) "7Normalization result" is the data normalization information for each batch.

(4) "8Batch effect" is the batch effect both in before and after data cleaning.

(5) "9metabolite plot" is the scatter plot for each feature.

(6) "10Data overview" is the overview of data.

(7) "11RSD overview" is the RSD distribution for each batch both before and after data cleaning.

(8) **"data_after_pre.csv", "qc.info.csv" and "subject.info"** are the data and sample information after data cleaning.

(9) "intermediate" is the intermediate data during processing.

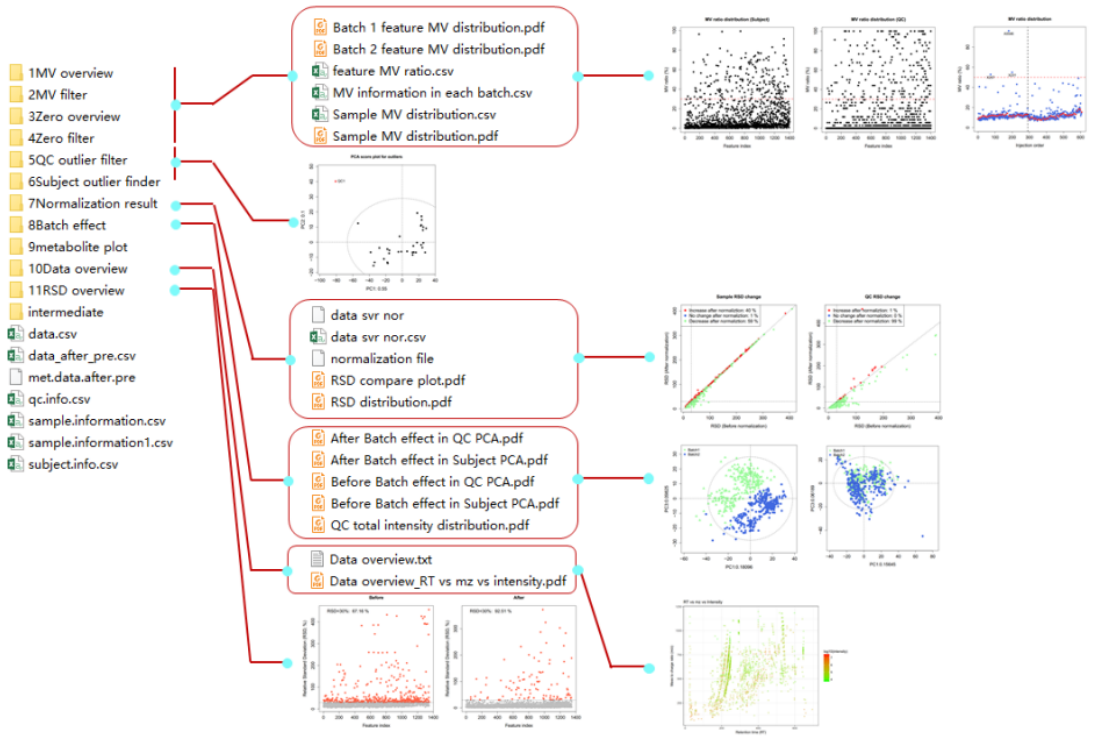

Figure5 Output files of MetCleaning

Figure5. Output files of *MetCleaning*

Statistical analysis

Data statistical analysis is integrated as a function named as MetStat in *MetCleaning*. We use the demo data as the example. Please note that now *MetStat* can only process two class data. Copy the code below and paste in you R console.

code 4: Demo data of *MetStat*

```
data(new.group, package = "MetCleaning")
##create a folder for MetStat demo
dir.create("Demo for MetStat")
setwd("Demo for MetStat")
## export the demo data as csv
write.csv(new.group, "new.group.csv", row.names = FALSE)
```

The demo data have been added in your work directory. "new.group.csv" is a sample.information which has been changed the group information you want to use for statistical analysis. For the sample which you don't want to use them for statistical analysis, you can set they group information as NA like Figure 6 shows.

Sample.information.csv

sample.name	injection.order	class	batch	group
QC11	1	QC	1	QC
A5551	2	Subject	1	9
A4880	3	Subject	1	1
C1282	4	Subject	1	0
C1492	5	Subject	1	9
A5730	6	Subject	1	1
X1421	7	Subject	1	0
X2	8	Subject	1	0
C1059	9	Subject	1	1
QC12	10	QC	1	QC
C1397	11	Subject	1	0
A5819	12	Subject	1	1
C1137	13	Subject	1	0
A3867	14	Subject	1	1
C1223	15	Subject	1	0
C1295	16	Subject	1	0
C1510	17	Subject	1	0
C1121	18	Subject	1	0
QC13	19	QC	1	QC

New.group.csv

sample.name	injection.order	class	batch	group
QC11	1	QC	1	QC
A5551	2	Subject	1	NA
A4880	3	Subject	1	case
C1282	4	Subject	1	control
C1492	5	Subject	1	NA
A5730	6	Subject	1	case
X1421	7	Subject	1	control
X2	8	Subject	1	control
C1059	9	Subject	1	case
QC12	10	QC	1	QC
C1397	11	Subject	1	control
A5819	12	Subject	1	case
C1137	13	Subject	1	control
A3867	14	Subject	1	case
C1223	15	Subject	1	control
C1295	16	Subject	1	control
C1510	17	Subject	1	control
C1121	18	Subject	1	control
QC13	19	QC	1	QC

Figure6 new group information

Figure6. Group information for statistical analysis.

code 5: Running of *MetStat*

```
MetStat(MetFlowData = met.data.after.pre, new.group = TRUE)
```

Running results of *MetStat*

1.Sample removing. Firstly, you need to confirm the samples which you want to remove form dataset as Figure 7 shows.

```
Change group information
The samples you want to remove from dataset are:
A5551 C1492 FA13 A5134 A3820 M135 M134 M133 C1262 C1442 X1 A5520 C1458
7 C1485 M139 A5636 C1430 M132 A3657 M140 A4994 C1371
Right(y) or wrong(n)?y
```

Figure7 sample removing confirmation

Figure7. The confirmation of the samples you want to remove.

2.The selection of best number of component in PLS-DA analysis. In PLS-DA analysis, you should manually select the best choice of the number of component. When the console show "How many comps do you want to see?", you can type 10 and hit "Enter" key. Then a MSE plot is showing, and the best number of component is the one has the smallest CV values. So type the number (in this example is 4) and hit "Enter" key.

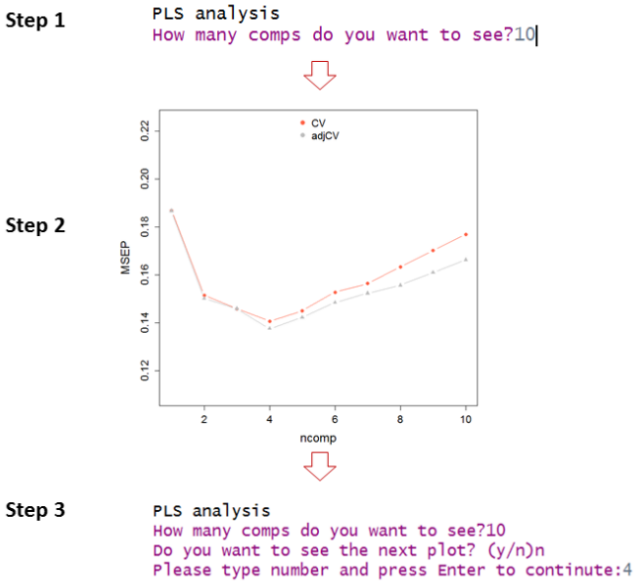

Figure8 Number of component selection in PLS-DA analysis

Figure8. The selection of best number of component in PLS-DA analysis.

3.Output files. Output files of *MetStat* are listed as Figure 9 shows.

(1) "12PCA analysis" is the PCA score plot.
(2) "13PLS analysis" contains the PLS-DA results.
(3) "14heatmap" is the heatmap.
(4) "15marker selection" contains the information of markers, volcano plot and boxplots of markers.
(5) **"data_after_stat.csv", "qc.info.csv" and "subject.info"** are the data and sample information after statistical analysis.
(6) "intermediate" is the intermediate data during processing.

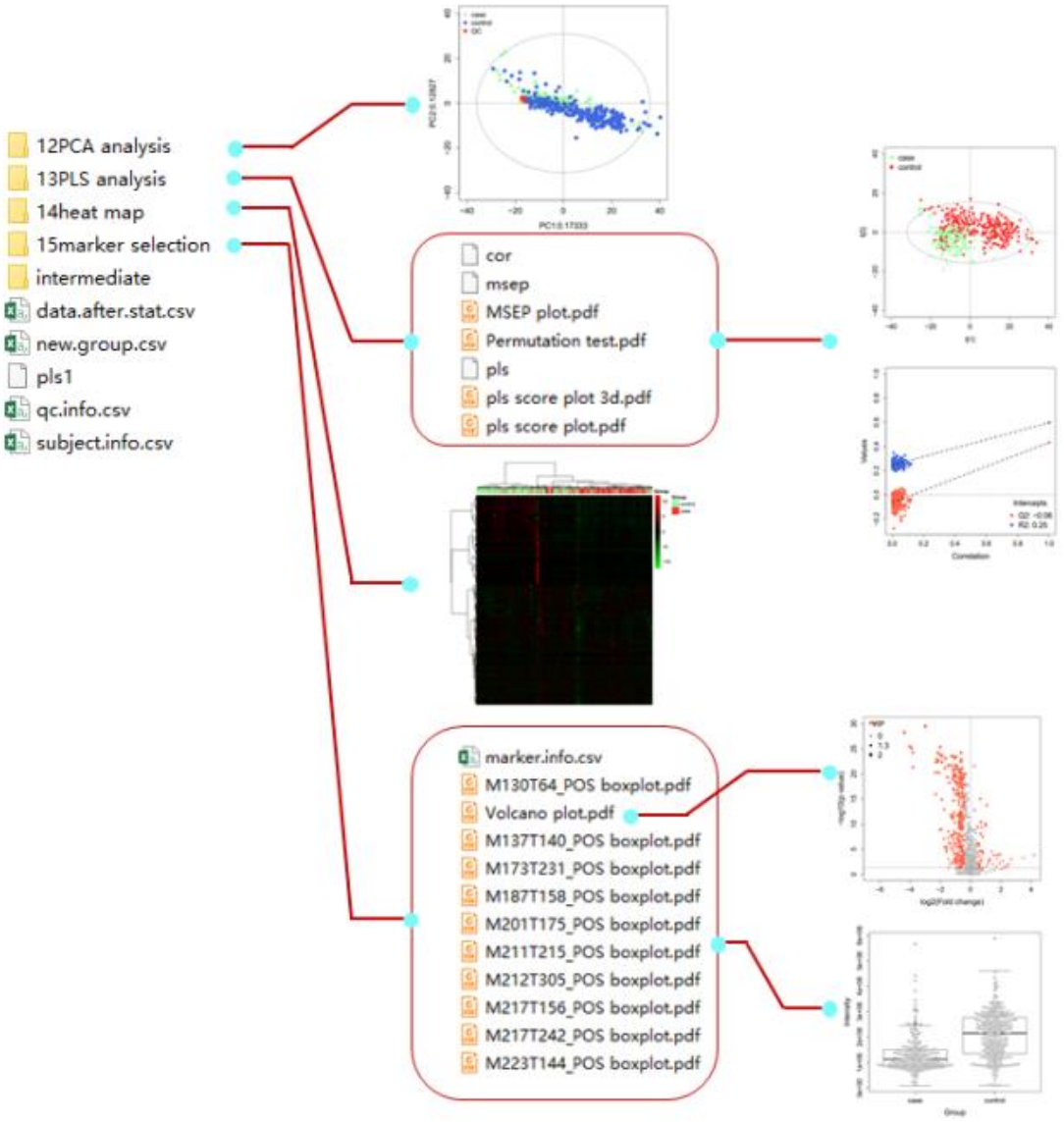

Figure9 Output files of MetStat

Figure9. Output files of *MetStat*.