

# MetCleaning v0.99.0

Xiaotao Shen([shenxt@sioc.ac.cn](mailto:shenxt@sioc.ac.cn)) and Zhengjiang Zhu

2017-01-03

## Introduction

*MetCleaning* provides an integrated and automatic pipeline for data cleaning and statistical analysis of large scale mass spectrometry (MS) based-metabolomic data. It includes missing value (MV) filtering and imputation, zero value filtering, data normalization, data integration, data quality assessment, univariate statistical analysis, multivariate statistical analysis such as PCA and PLS-DA, potential marker selection and show. This document describes how to use the integrated functions, *MetCleaning* and *MetStat* in *MetCleaning* utilizing demo data.

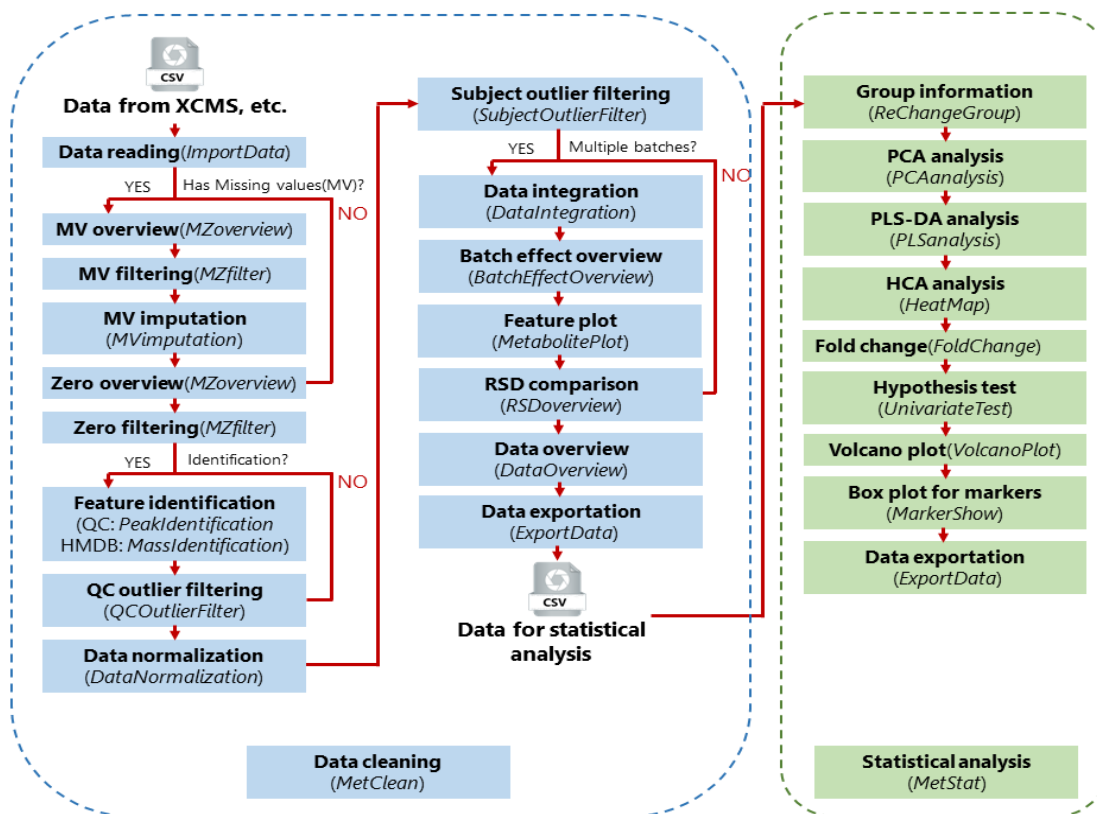


Figure1 Workflow of MetCleaning

## Installation and help

---

*MetCleaning* is published in github ([link](#)). So you can install it via to github.

### code 1: Installation of *MetCleaning*

```
##pcaMethods and 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")
```

## Data cleaning

---

Data cleaning is integrated as a function named as *MetCleaning* in *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*

```
##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 = FALSE)
```

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

1. "data.csv" is the metabolomic dataset you want to process. Rows are features and columns are feature abundance of samples and information of features. The information of features must contain "name" (feature name), "mz" (mass to change ratio) and "rt" (retention time). Other information of features are 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, and "210a" or "210-a" are invalid.

2. "sample.information.csv" is sample information for metabolomic dataset. Column 1 is "sample.name" which is the names 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".

	Feature name		m/z	rt	Other information of features		QC and subject sample abundance			
	A	B			C	D	E	F	G	H
1	name	mz	rt		isotopes	QC22	QC23	QC12	QC24	
2	M72T49	72.080982	49.212	[1]	[M] <sup>+</sup>	1812140	1544984	1794878	165998	
3	M76T31	76.075904	30.995	[2]	[M] <sup>+</sup>	516119.9	516731.3	545147.1	378761.1	
4	M84T38	84.081043	37.514	[3]	[M] <sup>+</sup>	2062165	1926203	1821621	169139	
5	M86T74	86.096638	73.766	[4]	[M] <sup>+</sup>	3450548	3086142	3430762	314432	
6	M90T90	86.096646	89.873	[5]	[M] <sup>+</sup>	2390322	2160545	2448955	192185	
7	M100T149	100.07584	149.427	[6]	[M] <sup>+</sup>	4369935	3627019	4198031	397886	
8	M103T150	103.05438	149.991	[7]	[M] <sup>+</sup>	191304.8	154571.8	176281.5	171652.4	
9	M104T30	104.10736	29.5285	[8]	[M] <sup>+</sup>	10220169	9599432	10928312	891114	
10	M110T652	110.02015	652.103	[9]	[M] <sup>2+</sup>	642884.8	564758.9	799711.5	528291.1	
11	M114T32	114.06639	32.413	[10]	[M] <sup>+</sup>	1369850	1362086	1795623	164330	
12	M118T36	118.08637	36.0715	[11]	[M] <sup>+</sup>	12726533	13111893	12571014	1280386	
13	M119T379	119.08572	378.5005	[12]	[M] <sup>+</sup>	94376	78369.41	114665.7	86013.0	
14	M120T150	120.08085	149.995	[13]	[M] <sup>+</sup>	4001522	3321928	3837926	355559	
15	M122T160	122.09645	159.657	[14]	[M] <sup>+</sup>	232127.1	204514.5	238401.2	21441	
16	M126T165	126.0915	164.642	[15]	[M] <sup>+</sup>	296192.1	226613.8	281846.4	256782.4	
17	M130T64	130.05002	63.6625	[16]	[M] <sup>+</sup>	513595.2	378566.3	398876.5	503552.4	
18	M130T157	130.06516	157.058	[17]	[M] <sup>+</sup>	239890.3	186114.1	252431.1	213726.1	
19	M130T38	130.08644	38.384	[18]	[M] <sup>+</sup>	336137	506422.3	367187.9	368729.1	

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

Figure2 Data organisation of MetCleaning

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

#### code 3: Running of *MetCleaning*

```
##demo data
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 filter them or not. We recommend to remove all of them as Figure 3 shows.

```
Missing values filter...
No QC should be removed.
X257 A5546 X231 should be removed!!!
Subject should 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*

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

```
Subject outlier filtering...
X2 X217 C1126 C1283 C1242 X1 X214 are outliers!!!
C1238 C1248 X5121 X209 X208 X211 X218 are outliers!!!
Batch 1
-----
Subject should be removed are:7 40 54 63 90 130 198
which subject you want to remove(please type the index of subject sample,n|
and separate them using comma,
if you don't want to remove any subject, please type n):
```

*Figure4 Sample filtering*

### 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.

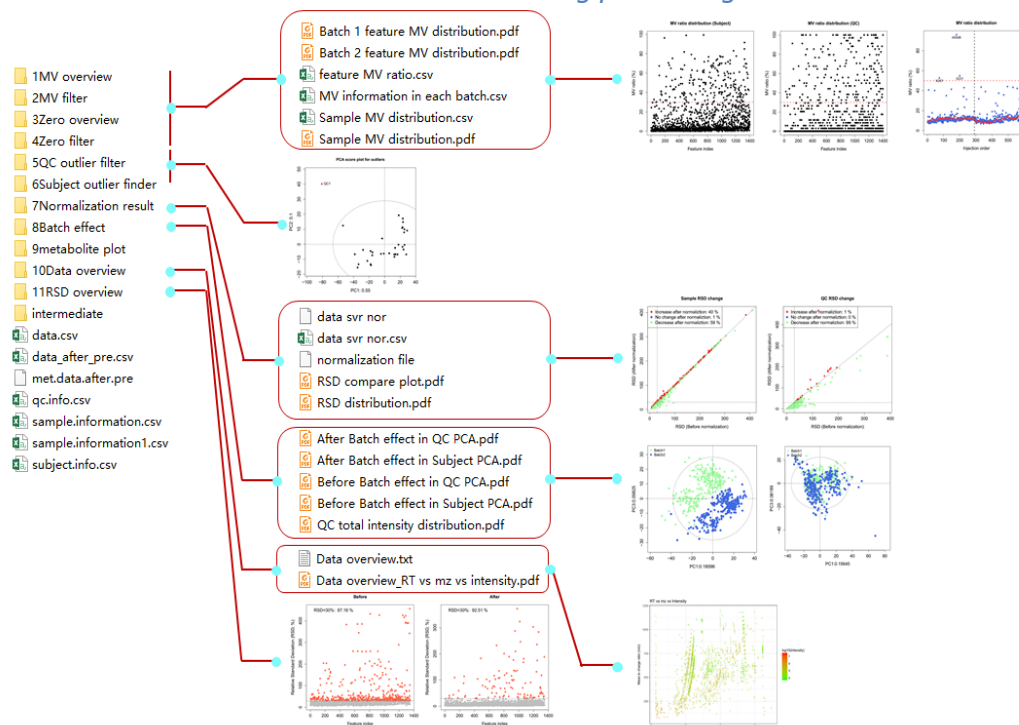


Figure5 Output files of *MetCleaning*

## 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("met.data.after.pre", package = "MetCleaning")
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					New.group.csv				
sample.name	injection.order	class	batch	group	sample.name	injection.order	class	batch	group
QC11	1	QC	1	QC	QC11	1	QC	1	QC
A5551	2	Subject	1	9	A5551	2	Subject	1	NA
A4880	3	Subject	1	1	A4880	3	Subject	1	case
C1282	4	Subject	1	0	C1282	4	Subject	1	control
C1492	5	Subject	1	9	C1492	5	Subject	1	NA
A5730	6	Subject	1	1	A5730	6	Subject	1	case
X1421	7	Subject	1	0	X1421	7	Subject	1	control
X2	8	Subject	1	0	X2	8	Subject	1	control
C1059	9	Subject	1	1	C1059	9	Subject	1	case
QC12	10	QC	1	QC	QC12	10	QC	1	QC
C1397	11	Subject	1	0	C1397	11	Subject	1	control
A5819	12	Subject	1	1	A5819	12	Subject	1	case
C1137	13	Subject	1	0	C1137	13	Subject	1	control
A3867	14	Subject	1	1	A3867	14	Subject	1	case
C1223	15	Subject	1	0	C1223	15	Subject	1	control
C1295	16	Subject	1	0	C1295	16	Subject	1	control
C1510	17	Subject	1	0	C1510	17	Subject	1	control
C1121	18	Subject	1	0	C1121	18	Subject	1	control
QC13	19	QC	1	QC	QC13	19	QC	1	QC

Figure6 new group information

### 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 from dataset as Figure 7 shows.

Change group information

The samples you want to remove from dataset are:

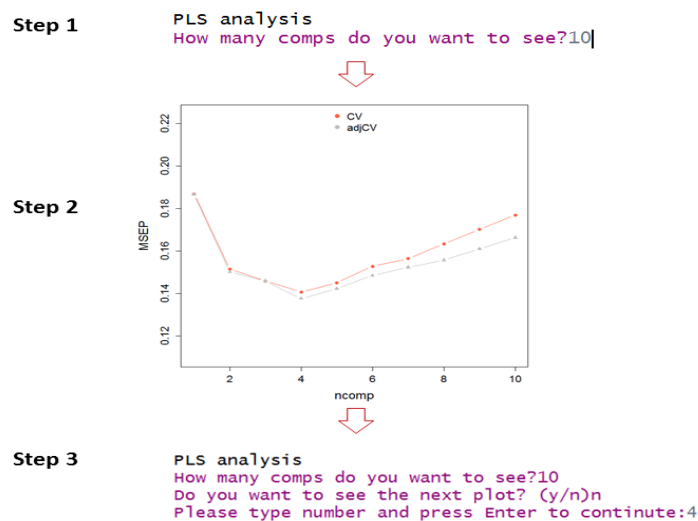
A5551 C1492 FA13 A5134 A3820 M135 M134 M133 C1262 C1442 X1 A5520 C1458  
7 C1485 M139 A5636 C1430 M132 A3657 M140 A4994 C1371

Right(y) or wrong(n)?y|

---

*Figure7 sample removing confirmation*

2. Number of component selection 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 enter "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 enter "Enter" key.



*Figure8 Number of component selection 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.

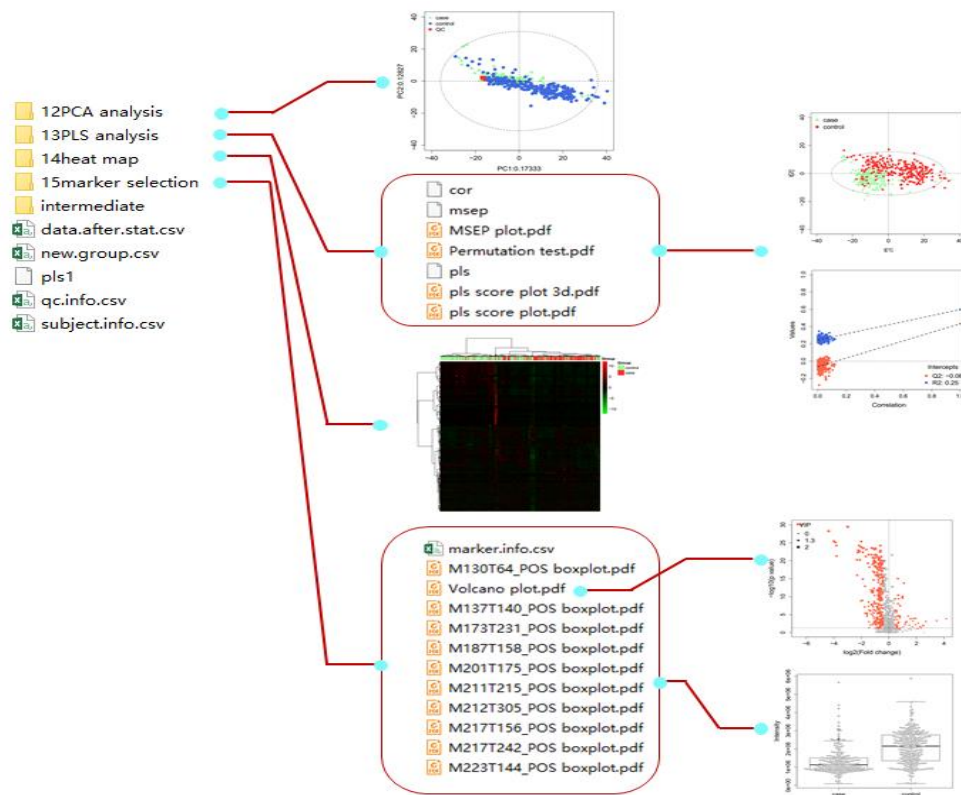


Figure9 Output files of *MetStat*