MetCleaning v1.1.0

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

2016-11-25

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, *MetClean* and *MetStat* in *MetCleaning* utilizing demo data.

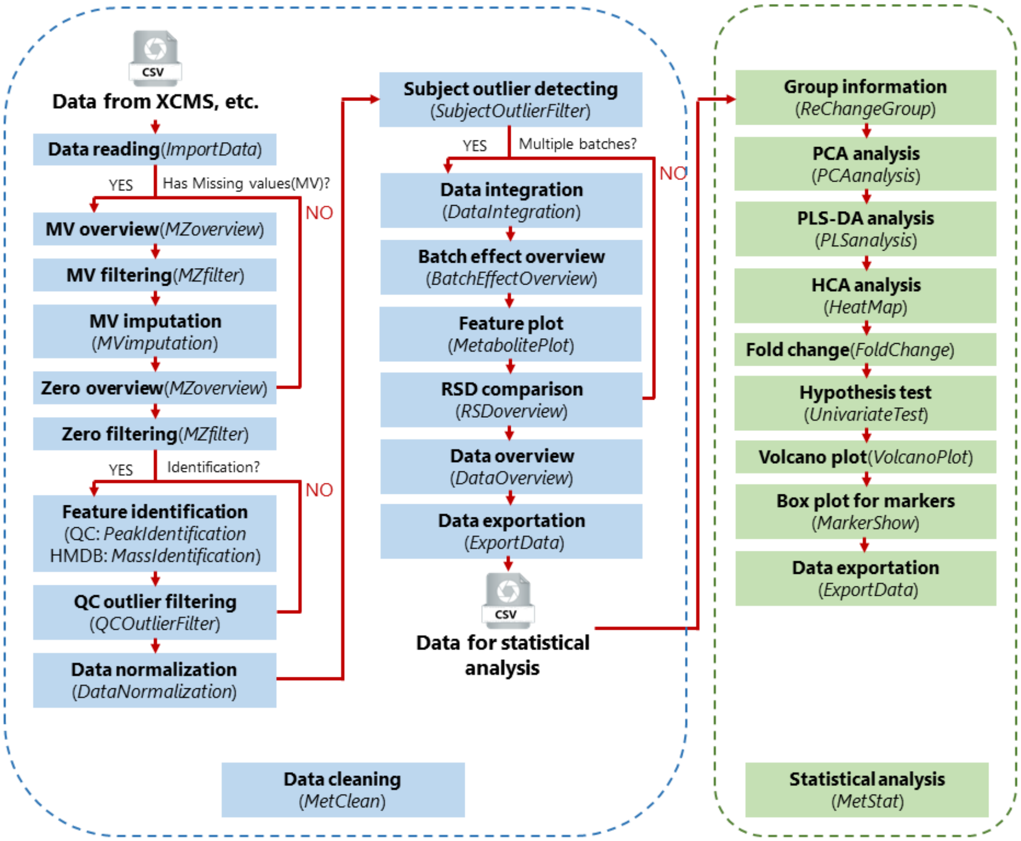

CSV
Data from XCMS, etc.
Data reading(ImportData)
YES
Has Missing values(MV)?
NO
MV overview(MZoverview)
MV filtering(MZfilter)
MV imputation (MVimputation)
Zero overview(MZoverview)
Zero filtering(MZfilter)
YES
Identification?
NO
Feature identification (QC: PeakIdentification HMDB: MassIdentification)
QC outlier filtering (QCOutlierFilter)
Data normalization (DataNormalization)
Subject outlier detecting (SubjectOutlierFilter)
YES
Multiple batches?
NO
Data integration (DataIntegration)
Batch effect overview (BatchEffectOverview)
Feature plot (MetabolitePlot)
RSD comparison (RSDoverview)
Data overview (DataOverview)
Data exportation (ExportData)
CSV
Data for statistical analysis
Data cleaning (MetClean)
Group information (ReChangeGroup)
PCA analysis (PCAanalysis)
PLS-DA analysis (PLSanalysis)
HCA analysis (HeatMap)
Fold change(FoldChange)
Hypothesis test (UnivariateTest)
Volcano plot(VolcanoPlot)
Box plot for markers (MarkerShow)
Data exportation (ExportData)
Statistical analysis (MetStat)

Figure1 Workflow of MetCleaning

安装及查看帮助文档

MetCleaning 存储在 github 上 (link). 可以通过 github 进行安装.

code 1: 安装 *MetCleaning*

```
##pcaMethods,impute 和 pathifier 需要通过 bioconductor 安装.
##pcaMethods
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")
```

数据清洗

在 *MetCleaning* 中数据清洗被整合为一个函数 *MetClean*. 以包自带的数据作为例子进行练习. 拷贝下面的代码并粘贴在 R 操作台中.

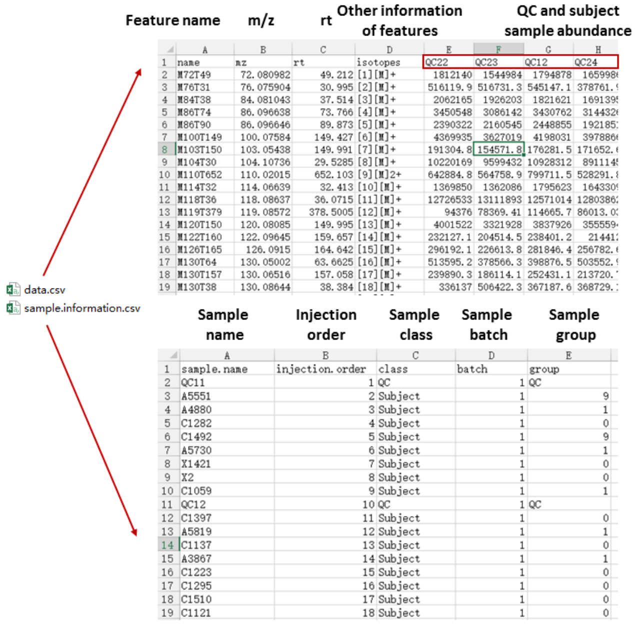
code 2: *MetClean* 示例数据

```
##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)
```

如图一所示, 示例数据这时候就已经放在了你的工作路径中. 包括两个文件, "data.csv"和"sample.information.csv". **如果需要做鉴定, 则需要另外新建立一个文件夹, 命名为"peak identification", 在里面放入二级鉴定的数据, 且文件命名必须含有"ms2"字样.**

1. "data.csv" 是要处理的数据. 行是 feature, 列是样品名以及 feature 的信息. Feature 的信息必须包括"name" (feature 名字), "mz" (质核比) and "rt" (保留时间). 其他的信息是可选的, 比如"isotopes"和"adducts". 样品名可以含有"."或者"_", 但是不能含有"-"或者空格. 而且名字的开头不能是数字, 必须是字母. 比如, "A210.a"和"A210a"有效, 而 "210a" or "210-a"无效.

2. "sample.information.csv" 是要处理的数据的样品信息. 第一列是"sample.name", 是 subject 和 QC 样品的名字. 请一定要确认"sample.information.csv"和"data.csv"中的样品名字完全一致. 第二列是"injection.order", 是 QC 和 subject samples 的进样顺序. 第三列是"class", 是用来表明样品是属于"QC"还是"Subject", **请一定注意大小写.** 第四列是"batch", 用来表明样品的批次信息. 第五列是"group", 用来表明 subject 样品的分组. 例如, "control" and "case". QC 样品的"group"信息标为"QC".



	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]+	1812140	1544984	1794878	165998
3	M76T31	76.075904	30.995	[2][M]+	516119.9	516731.3	545147.1	378761.
4	M84T38	84.081043	37.514	[3][M]+	2062165	1926203	1821621	169139
5	M86T74	86.096638	73.766	[4][M]+	3450548	3086142	3430762	314432
6	M86T90	86.096646	89.873	[5][M]+	2390322	2160545	2448855	192185
7	M100T149	100.07584	149.427	[6][M]+	4369935	3627019	4198031	397886
8	M103T150	103.05438	149.991	[7][M]+	191304.8	154571.8	176281.5	171652.
9	M104T30	104.10736	29.5285	[8][M]+	10220169	9599432	10928312	891114
10	M110T652	110.02015	652.103	[9][M]2+	642884.8	564758.9	799711.5	528291.
11	M114T32	114.06639	32.413	[10][M]+	1369850	1362086	1795623	164330
12	M118T36	118.08637	36.0715	[11][M]+	12726533	13111893	12571014	1280386
13	M119T379	119.08572	378.5005	[12][M]+	94376	78369.41	114665.7	86013.0
14	M120T150	120.08085	149.995	[13][M]+	4001522	3321928	3837926	355559
15	M122T160	122.09645	159.657	[14][M]+	232127.1	204514.5	238401.2	21441
16	M126T165	126.0915	164.642	[15][M]+	296192.1	226613.8	281846.4	256782.
17	M130T64	130.05002	63.6625	[16][M]+	513595.2	378566.3	398876.5	503552.
18	M130T157	130.06516	157.058	[17][M]+	239890.3	186114.1	252431.1	213720.
19	M130T38	130.08644	38.384	[18][M]+	336137	506422.3	367187.6	368729.

	A	B	C	D	E
1	sample.name	injection.order	class	batch	group
2	QC11	1	QC	1	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	1	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 organization of MetCleaning

然后就可以使用 *MetClean* 函数来进行数据的清洗.

code 3: Running of *MetClean*

```
MetClean(data = "sz2016005 urine pos ms1.csv",
         sample.information = "sample information.csv",
         polarity = "positive",
         obs.zero.cutoff = 0.5,
         var.zero.cutoff = 0.5,
         method = "svr",
         threads = 2,
         hmdb.matching = FALSE,
         mass.tolerance = 30,
         mz.tolerance = 30,
         rt.tolerance = 180,
         met.plot = TRUE)
```

参数说明

(1) data: 要处理的数据的名字, 例如"data.csv".

(2) sample.information: 样品信息的文件名, 比如"sample.information.csv".

(3) polarity: 样品的极性, "positive", "negative". 如果是 MRM 数据, 可以设置为"none".

(4) obs.zero.cutoff: 零值筛选的样品阈值, 如果样品中的零值比例超过 obs.zero.cutoff 值, 那么这个样品就会被舍去, 默认为 0.5, 是指一个样品中零值比例超过 50%, 则该样品会被除去, 如果设置为 100%, 是指一个样品中的所有峰都是 0, 才会被舍去.

(5) var.zero.cutoff: 和 obs.zero.cutoff 相同, 是指对峰进行零值筛选.

(6) method: 数据标准化方法, 默认为 "svr".

(7) threads: 数据标准化时的多线程个数, 默认为 2, 服务器上可以适当改大一些.

(8) hmdb.matching: 是否使用精确分子质量在 HMDB 数据库中进行匹配.

(9) mass.tolerance: HMDB 数据库精确分子质量匹配的 tolerance.

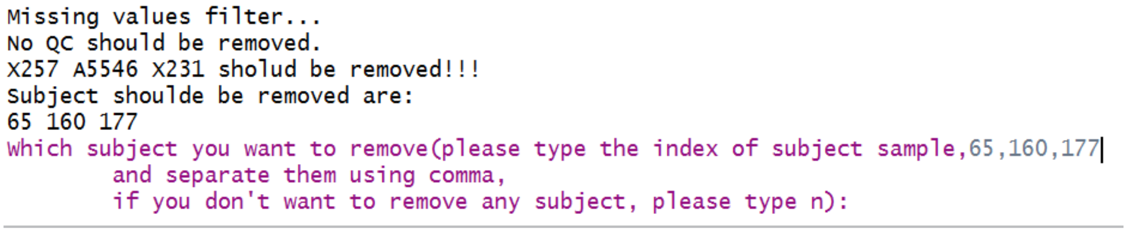
(10)mz.tolerance:一级二级数据匹配时 mz 的 tolerance.

(11)rt.tolerance:一级二级数据匹配时 rt 的 tolerance.

(12)mt.plot: 是否将所有 feature 的 plot 画出来.

Running results of *MetClean*

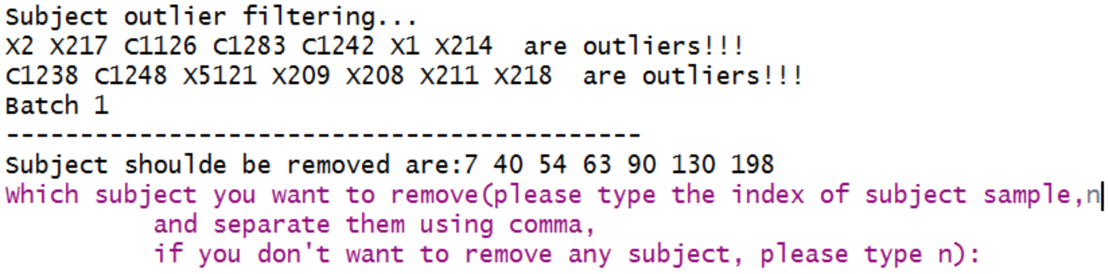
1. Missing or zero values filtering. In the missing or zero value filtering step, if there are samples which beyond the threshold you set, you should decide to filter them or not. We recommend to remove all of them as Figure 3 shows.



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

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 shoulde 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 MetClean 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.

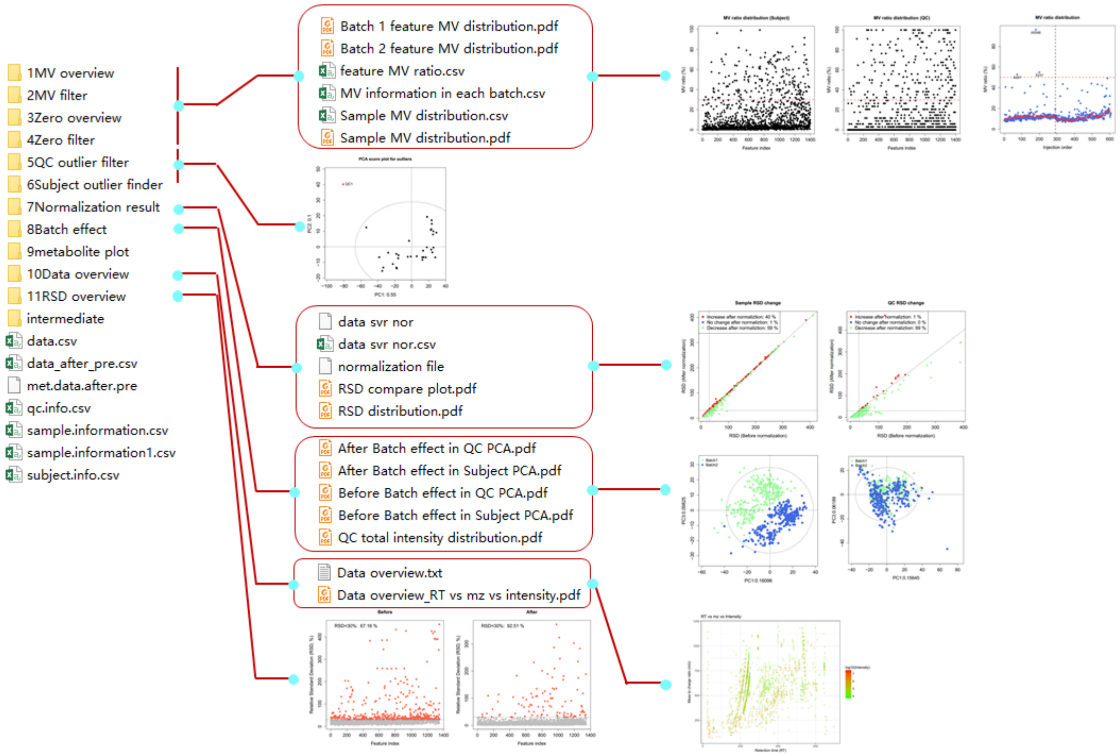


Figure5 Output files of MetClean

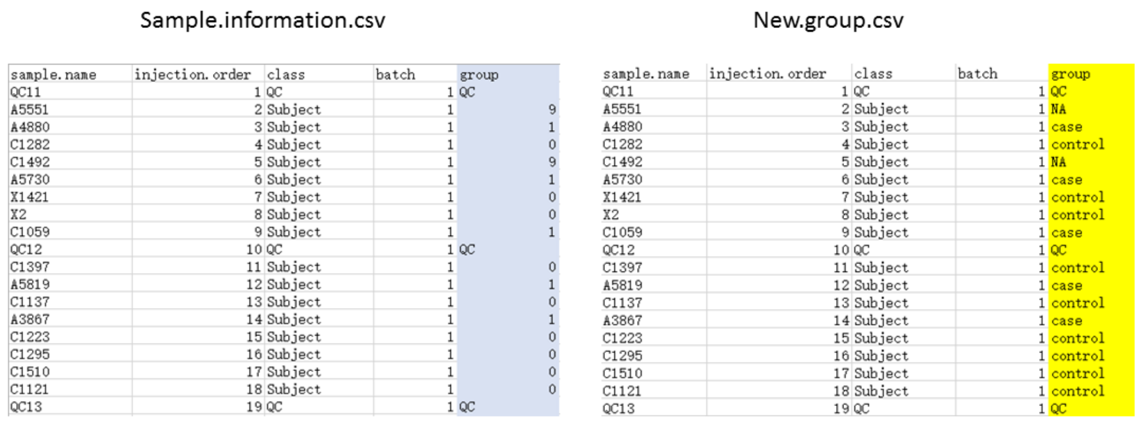
Statistical analysis

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 their 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

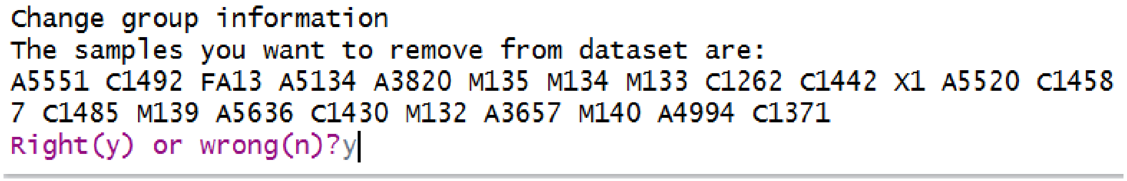
Figure6 new group information

code 5: Running of *MetStat*

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

Running results of *MetStat*

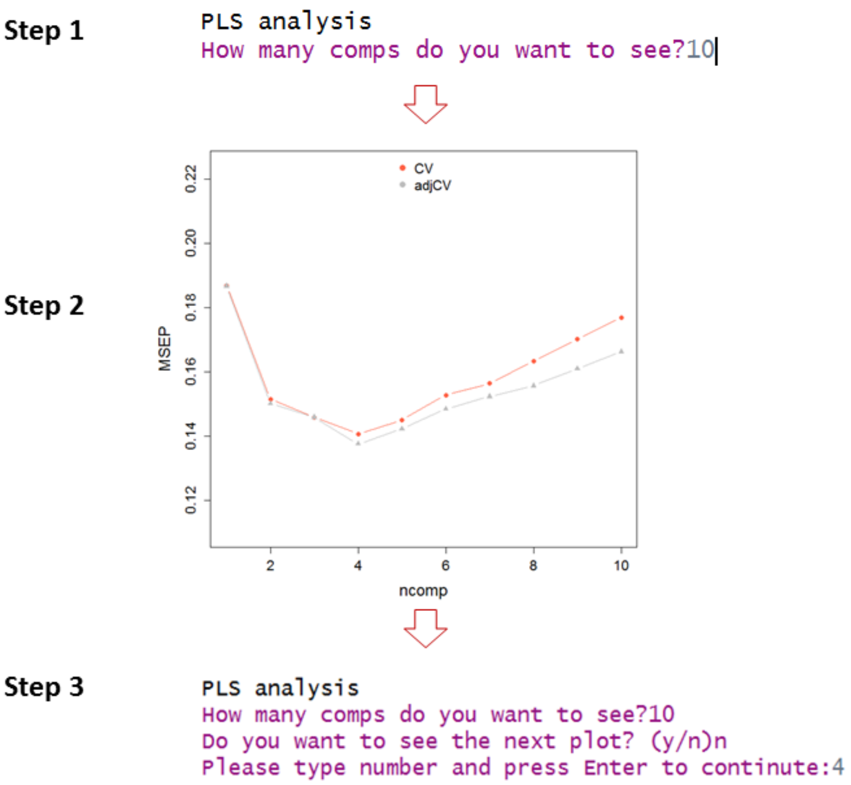
1. Sample removing. Firstly, you need to confirm the samples which you want to remove form dataset as Figure 7 shows.



```
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

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 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.



Step 1

```
PLS analysis
How many comps do you want to see?10
```

Step 2

Step 3

```
PLS analysis
How many comps do you want to see?10
Do you want to see the next plot? (y/n)n
Please type number and press Enter to contiinute:4
```

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.

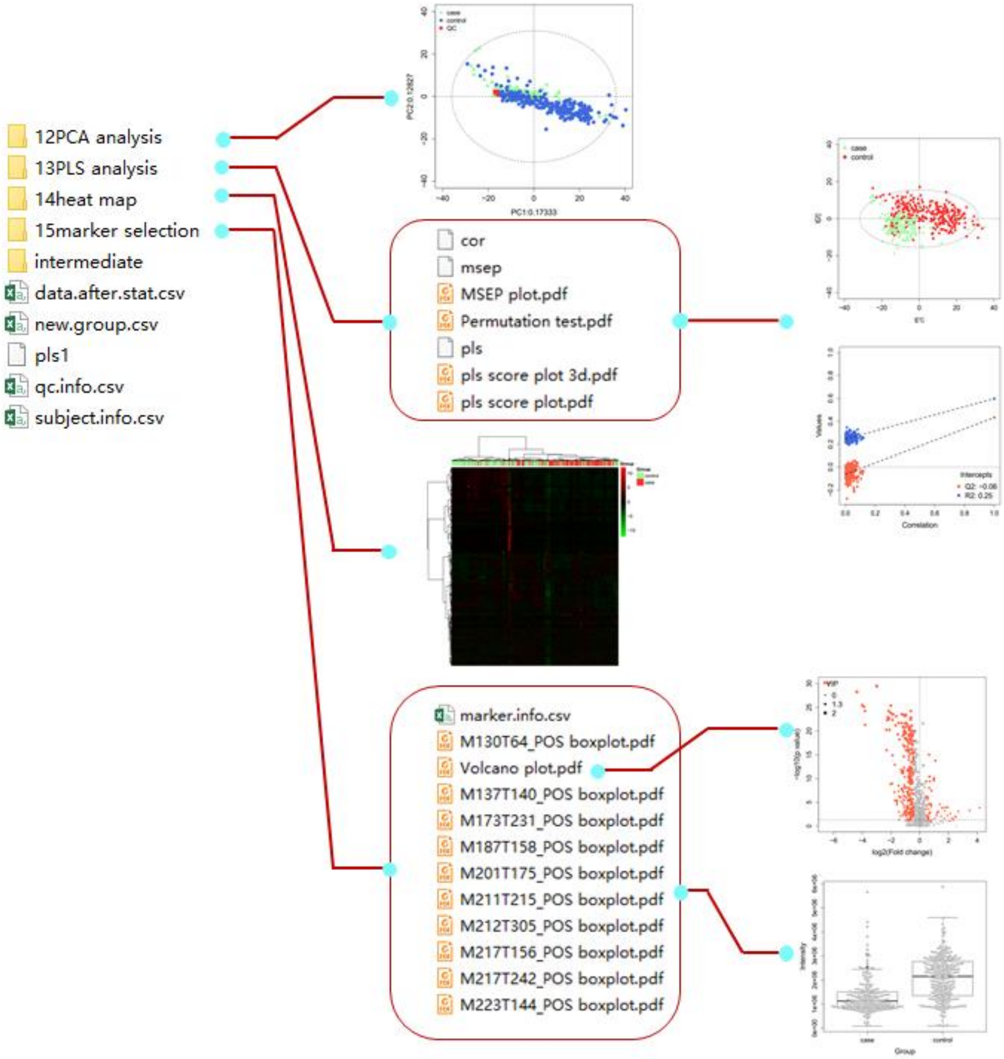


Figure9 Output files of MetStat