



Version 1.2 (Mar, 2019)

User Guide for SMART software

Yu-Jen Liang, Yu-Ting Lin, Chia-Wei Chen, Chien-Wei Lin, Kun-Mao Chao, Wen-Harn Pan and Hsin-Chou Yang*

*Correspondence: Hsin-Chou Yang (hsinchou@stat.sinica.edu.tw), Institute of Statistical Science, Academia Sinica. No 128, Academia Road, Section 2, Nankang, Taipei 115, Taiwan.

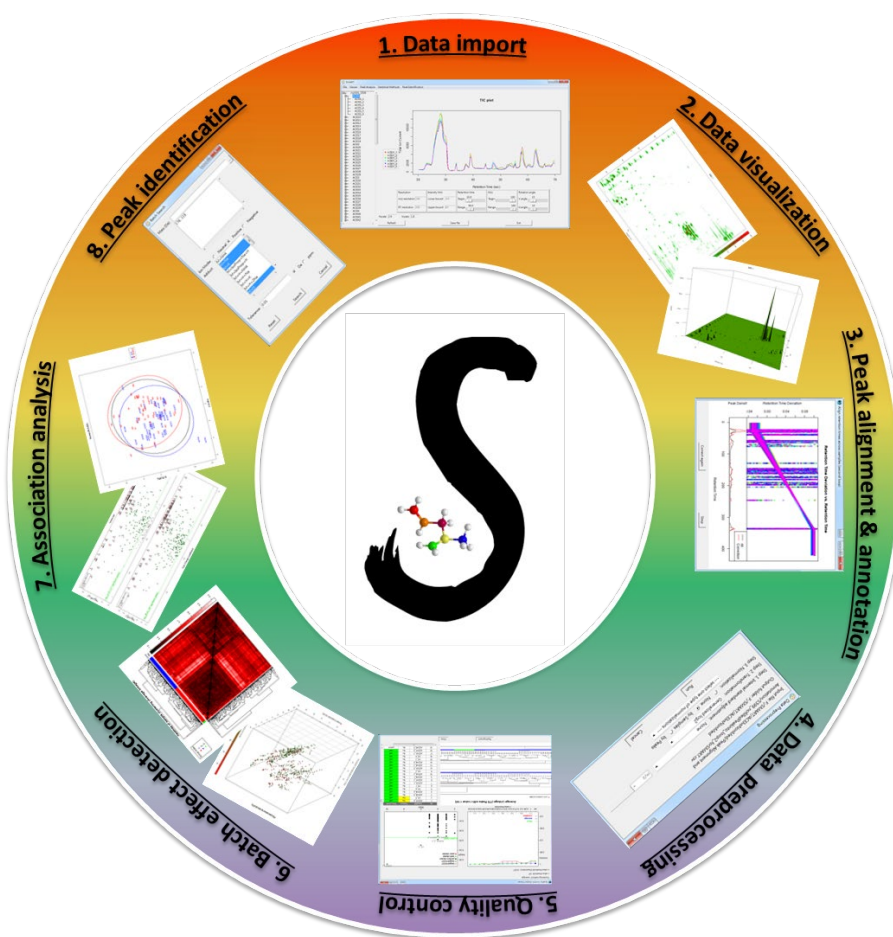


Table of Contents

1. SMART License	3
2. Overview	3
3. Software Download and Installation	4
3.1 SMART	4
3.2 R program.....	4
3.3 ActiveTcl.....	4
3.4 ProteoWizard.....	5
3.5 R packages.....	5
4. SMART Initialization	5
5. SMART Interfaces, Tools, and Operating Procedures	7
5.1 Data import	8
5.2 Data visualization	10
5.3 Peak alignment and annotation.....	15
5.4 Data preprocessing	18
5.5 Quality control	21
5.6 Re-alignment and annotation	24
5.7 Batch effect detection	25
5.8 Association analysis.....	29
5.9 Peak identification	36
6. Example.....	37
7. Application Programming Interface.....	38
8. References.....	40

1. SMART License

All copyright are reserved by authors of **SMART**. **SMART** are released under GPL_v2 license. We welcome any noncommercial use of **SMART** for your own research. Commercial use of **SMART** should be directed to hsinchou@stat.sinica.edu.tw. For free software **SMART**, we assume no warranty and no responsibility for the results of analyses. If publications are based on the results from the use of **SMART**, please cite the following reference:

Yu-Jen Liang, Yu-Ting Lin, Chia-Wei Chen, Chien-Wei Lin, Kun-Mao Chao, Wen-Harn Pan and Hsin-Chou Yang (2016). SMART: Statistical Metabolomics Analysis – An R Tool. *Analytical Chemistry*, 88 (12), pp 6334–6341.

2. Overview

SMART written in R and R GUI has been developed as user-friendly software for integrated analysis of metabolomics data. **SMART** streamlines the complete analysis flow from initial data preprocessing to downstream association analysis, consisting of analyzing different data file formats (e.g., .raw, .d, and mzXML), visually representing various types of data features (e.g., total ion chromatogram (TIC) and mass spectra), implementing peak alignment, conducting quality control for samples and peaks, exploring batch effects (e.g., known experimental conditions, unknown latent groups (LGs), or hidden substructures), performing association analysis, and accomplish peak identification (**Figure 2.1**).

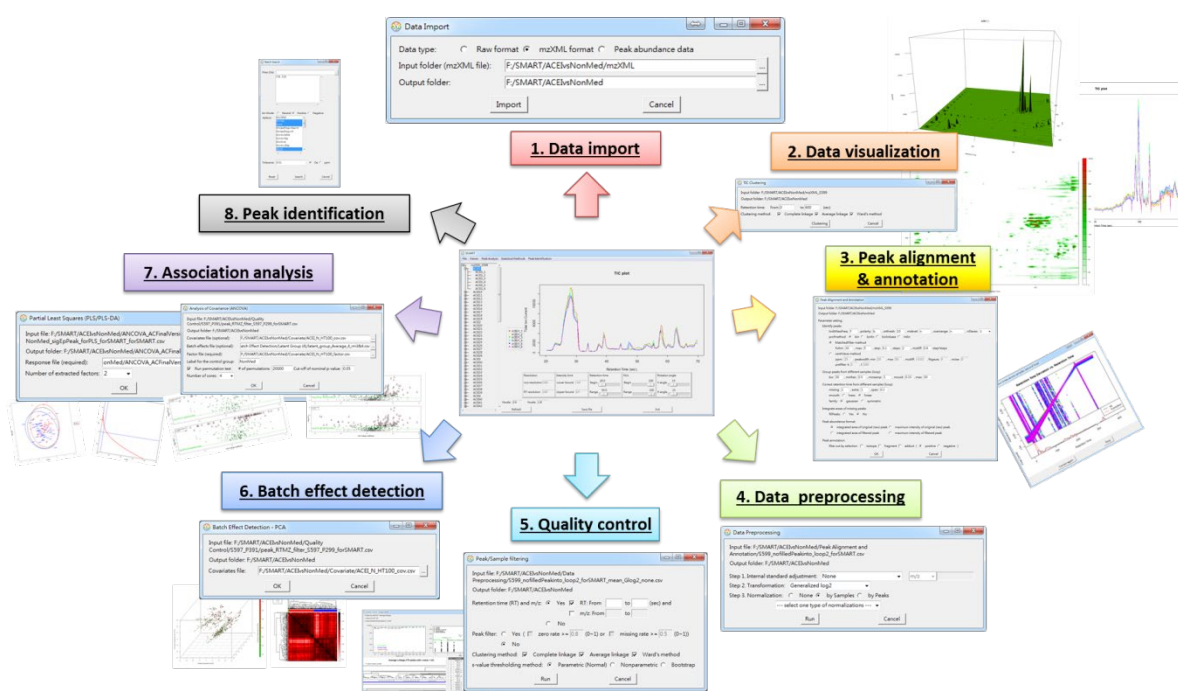


Figure 2.1. Overview of **SMART**.

3. Software Download and Installation

Execution of **SMART** requires installation of **SMART** program, R program, ActiveTcl, ProteoWizard, and some R packages.

3.1 SMART

We provide **SMART** programs for 32-bit and 64-bit Windows operating systems (Windows 7, Windows 8, and Windows 10) and Mac operating systems (OS X 10.6 or later). The programs, user guide, and examples can be downloaded from the **SMART** website at: <http://www.stat.sinica.edu.tw/hsinchou/metabolomics/SMART.htm>.

3.2 R program

Users can download R program from the website of “The R Project for Statistical Computing” at <http://www.r-project.org/>. Users click “CRAN” (Comprehensive R Archive Network) in the left of the page and then select a suitable mirror site to download R. Select a platform (Windows or Mac OS X) for R execution in your end. For Windows system users, click the hyperlink “base” and select “Download R 3.5.1 for Windows”. Then execute the file “R-3.5.1-win.exe” to install R to “C:\Program Files\R\R-3.5.1”. After finishing the installation of R, doubly click the icon “R i386 3.5.1” or “R x64 3.5.1” to initialize R in a 32-bit or 64-bit system, respectively. A window “RGui” with a sub-window “R Console” jumps up await for the subsequent analysis action. Users are suggested to update packages in R. They can select “Packages” in the tool bar, click “Uppdate packages” and then select a suitable mirror site to update packages. A window “CRAN mirror” jumps up and the icon “OK” is clicked to update packages. For Mac system users, click the hyperlink “R-3.3.0.pkg” to download R-3.3.0.pkg for Mac version R and click the hyperlink “XQuartz” to download XQuartz.dmg for Mac version X. Execute XQuartz.dmg to obtain XQuartz.pkg. Execute R-3.3.0.pkg and XQuartz.pkg to initialize Mac version R and X, respectively.

3.3 ActiveTcl

Program ActiveTcl Community Edition can be downloaded from the website of ActiveState. For 64-bit Windows system users, please download ActiveTcl from the following hyperlink <https://www.activestate.com/products/activetcl/downloads/thank-you/?dl=https://downloads.activestate.com/ActiveTcl/releases/8.6.8.0/ActiveTcl-8.6.8.0-MSWin32-x64.exe>. For Mac OS, please download ActiveTcl from the following hyperlink

http://www.activestate.com/activetcl/downloads/thank-you?dl=http://downloads.activestate.com/ActiveTcl/releases/8.5.18.0/ActiveTcl8.5.18.0.298892-macosx10.5-i386-x86_64-threaded.dmg.

3.4 ProteoWizard

ProteoWizard can be downloaded from the website at <http://proteowizard.sourceforge.net/downloads.shtml>. Please select “Windows installer (able to covert vendor files)”, choose a license agreement, and click “I agree to the licensing terms” to download and then install ProteoWizard. **Note that ProteoWizard does not provide a version for Mac operating systems.**

3.5 R packages

The analyses provided by **SMART** require several additional R packages, consisting of boot, car, fields, graphics, gtools, limma, mzR, png, tcltk2, tkrplot, scatterplot3d, snow, xcms, CAMERA, and ropls. **SMART** checks whether these R packages have been installed in R. If NOT, the packages will be automatically downloaded if users are using a latest version of R, e.g., R-3.5.1, in an environment with internet access. **Note that program R-3.5.1 or a version of R program newer than version R-3.5.0 is recommended for running SMART.**

4. SMART Initialization

For Windows system users, once all of the software and packages have been installed, **SMART** can be initialized by doubly clicking the executable file of **SMART**. When **SMART** is initialized first time, users will be asked to provide the paths for where R program, ActiveTcl, and ProteoWizard were installed (see **Figure 4.1**). **SMART** will save the path setting for next use. **Please wait for a moment when using SMART first time because it may take time to install the required R packages. The details about package installation can be found in the log file.** Once all packages are ready, the interface of **SMART** will jump up (see **Figure 4.2**).

```

  SMART v2
  *****
  S-----S
  | Welcome to SMART software | v2.0 | 22/Aug/2018 |
  |-----|
  | (C) 2018, GNU General Public License, v1.2 |
  |-----|
  | For documentation, citation & example instructions: |
  | http://www.stat.sinica.edu.tw/hsinchou/metabolomics/SMART.htm |
  S-----S
  *****

  *** Please provide the path of required software .....

  Please enter the path of your R software: <e.g., C:\Program Files\R>

```

Figure 4.1. Path setting of **SMART** in Window systems.

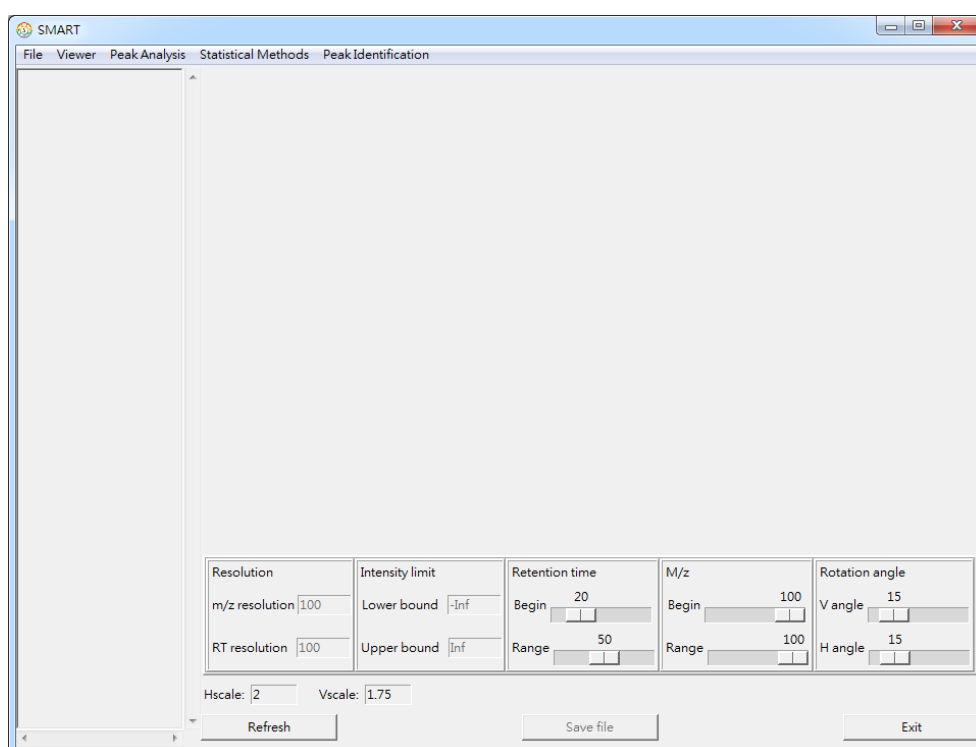


Figure 4.2. Initial interface of **SMART** in Window systems.

For Mac users, please execute **SMART** by the following four steps: **(under construction)**

Step 1: Save three R programs, “SMART_V1.2_Mac.r”, “SMART_V1.2_Mac_Gui.r” and “SMART_V1.2_Mac_Sub.r”, in the same folder (e.g., ~/Documents/SMART).

Step 2: Initialize R as mentioned in Section 3.2.

Step 3: Drag and drop “SMART_V1.2_Mac.r” to the window “R Console” to initialize **SMART**. Or you can type the command, `source(“~/Documents/SMART/SMART_V1.2_Mac.r”)`, in the window “R Console”.

Step 4: Type the full path name (e.g., ~/Documents/SMART) and press the Enter key, and then the interface of **SMART** will jump up (see **Figure 4.3**).

There are five main functions in **SMART**: (1) File, (2) Viewer, (3) Peak Analysis, (4) Statistical Methods, and (5) Peak Identification. Descriptions about these functions will be introduced in Section 5.

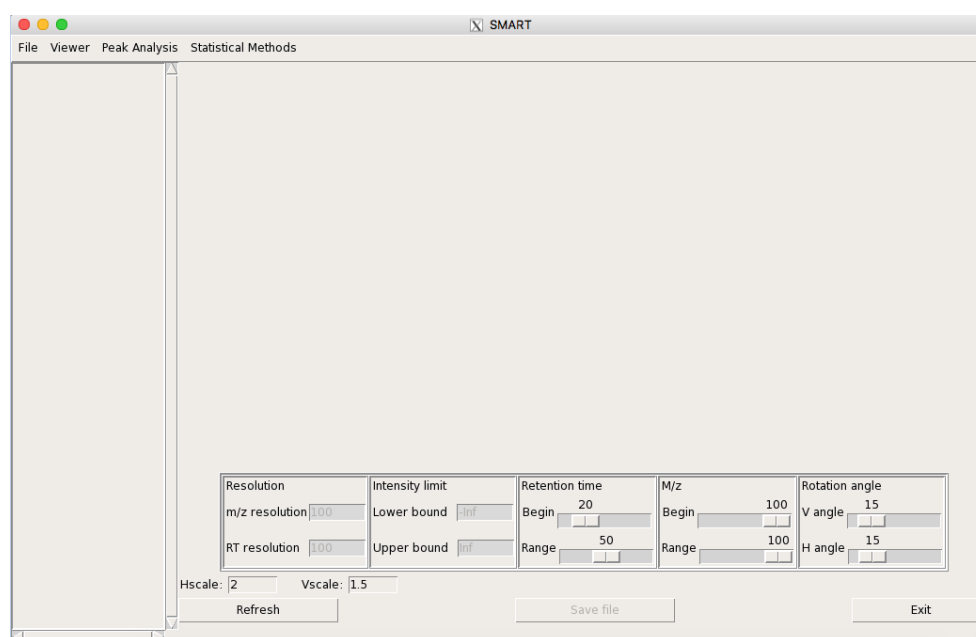


Figure 4.3. Initial interface of **SMART** in Mac systems.

5. SMART Interfaces, Tools, and Operating Procedures

The analysis procedures of **SMART** in Windows and Mac systems are almost the same. In this and next sections, we illustrate the operating environments and procedures in Windows systems.

5.1 Data import

SMART supports multiple input file formats (see **Figure 5.1.1**). The first format is the raw spectrum data format, such as .d files from Agilent Technologies (Santa Clara, CA, USA) and .raw files from Waters (Milford, MA, USA). The second format is the mzXML file format, which is an Extensible Markup Language (XML) file format for MS data. The third format is the comma-delimited peak abundance data of mass spectra. The peak abundance data file contains the mass-to-charge ratio (m/z), chromatographic retention time (RT, in seconds or minutes), and peak abundance or metabolite concentration of all replicate samples of a study subject.

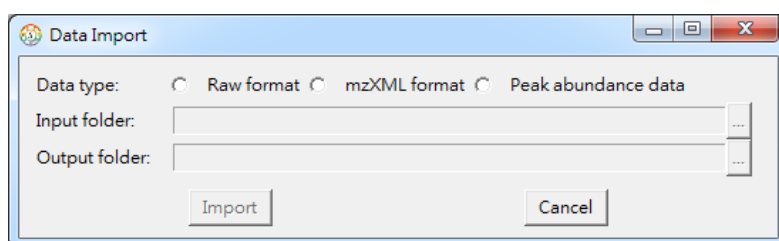


Figure 5.1.1. Interface of data import.

- raw spectrum data file:
Users follow the procedure to import data: [Click “File”](#) → [Click “Data Import”](#) → [Choose “Raw format”](#) → [Specify the I/O path](#). Note that only raw spectrum data files can be saved in the specified input folder. Note that this function is only available for Windows systems.
- mzXML data file:
Users follow the procedure to import data: [Click “File”](#) → [Click “Data Import”](#) → [Choose “mzXML format”](#) → [Specify the I/O path](#). Note that only mzXML files can be saved in the specified input folder.
- comma-delimited peak abundance data file:
Users follow the procedure to import data: [Click “File”](#) → [Click “Data Import”](#) → [Choose “Peak abundance data”](#) → [Specify the I/O path](#). An example of a comma-delimited peak abundance data file is provided (see **Table 5.1.1**). The first three columns are peak index, m/z , RT, and followed by peak abundance data of all replicate samples of all subjects. Each row indicates a peak/metabolite. The subject nomenclature consists of the subject name followed by the replicate sample name(s). For example, ACEI1_1 and ACEI1_2 represent the first and second replicate samples of subject ACEI1.

Table 5.1.1. The comma-delimited peak abundance data file.

Peak_ Index	mz	Ret_ time.sec	ACEI1_1	ACEI1_2	ACEI2_1	ACEI2_2	N1_1	N1_2	N2_1	N2_2
1	22.98998	27.7059	7.0783	6.6935	6.5680	6.4190	7.0669	6.6445	7.0774	6.5764
2	55.93659	336.5655	6.9994	7.9379	NA	8.4114	8.3721	8.3248	8.0697	8.3277
3	56.96345	335.7621	6.3181	6.0349	6.8107	6.7456	6.4947	6.3940	6.2704	6.0403
4	57.93642	335.8248	6.2205	6.1592	6.8686	6.5870	6.5836	6.4859	6.4517	6.4039
5	60.07343	29.2071	6.2056	NA	6.6890	6.5365	6.2876	6.1709	6.1499	6.2357
6	68.98333	24.0914	6.4834	6.5592	7.4892	7.4233	6.4490	6.4999	6.4278	6.4466
7	69.07036	53.4821	NA	NA	7.4339	6.9873	5.6892	5.2228	5.4749	5.2144
8	70.06573	32.3258	NA	NA	8.3018	8.0467	6.0851	5.7031	NA	NA
9	72.08130	64.1186	NA	NA	7.8959	7.6318	7.1014	7.1676	NA	NA
10	72.08131	37.8573	7.7268	7.3779	10.2431	10.0360	8.3697	8.2444	8.2197	7.9276
11	74.93966	335.9213	6.7730	7.0727	7.6475	7.3816	7.1578	7.3676	7.3392	7.2886
12	77.03906	58.1514	6.7476	6.6204	9.4728	9.5908	6.8912	6.9771	6.8814	6.7502
13	79.05363	58.1659	6.0404	5.4665	8.3506	8.3917	5.9861	6.1343	5.7149	5.9090
14	82.01453	24.0855	5.9334	6.0697	7.2531	7.0337	5.8216	6.0698	5.9923	5.5052
15	82.01411	335.6980	NA	NA	NA	10.1490	NA	7.3956	NA	NA
16	84.04471	51.8347	6.5344	6.5024	7.6371	7.5025	7.3956	7.1509	6.2767	NA
17	84.04468	30.1110	7.4784	7.5663	10.1490	10.1662	8.7341	8.6003	7.9926	7.7599
18	84.08100	25.1120	NA	NA	9.2761	9.1194	7.0255	6.7111	5.2853	NA
19	84.96001	24.1023	9.7988	9.6337	10.8931	10.7249	9.6305	9.5511	9.5060	9.3717
20	84.95969	335.6133	10.0044	10.1116	10.8662	10.4914	10.4841	10.3043	10.4317	10.1453

After raw spectrum data or mzXML data import, a sample tree diagram will be shown in the left-hand panel of the interface of **SMART** (see **Figure 5.1.2**).

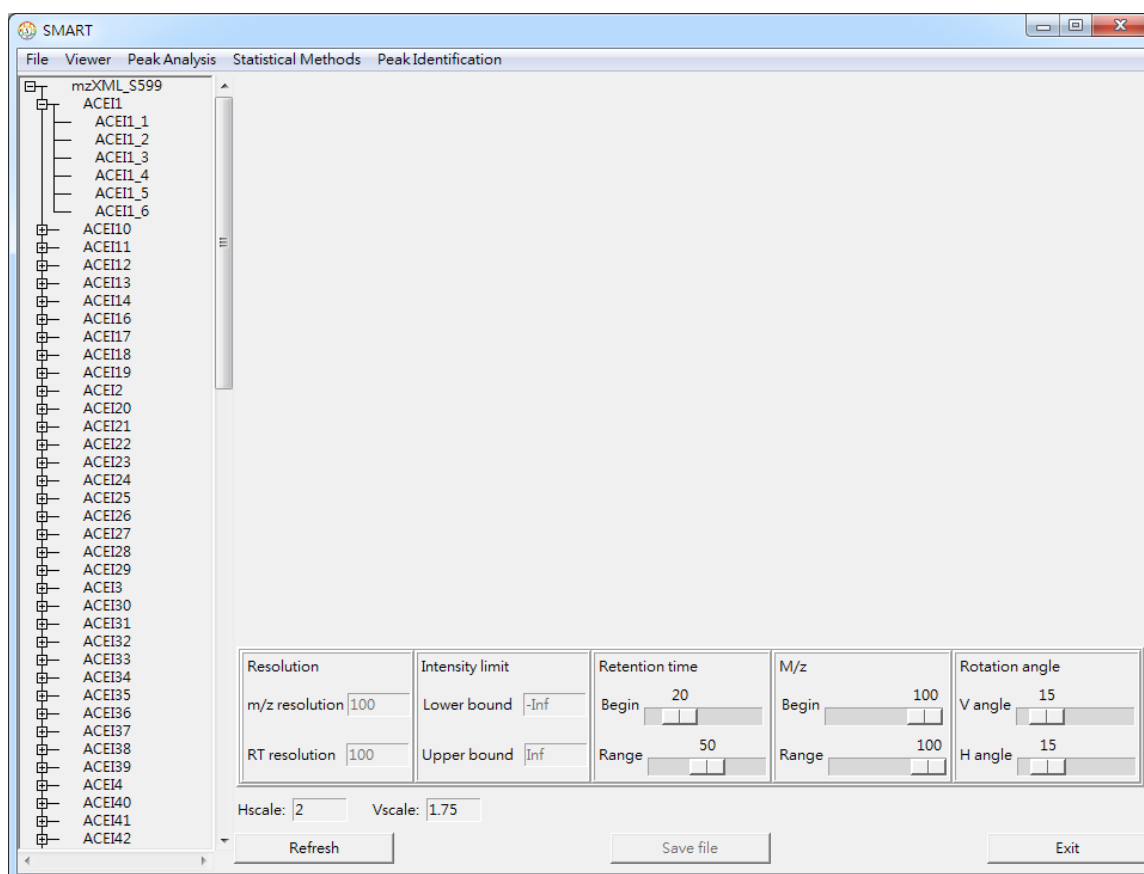


Figure 5.1.2. Interface after data import.

5.2 Data visualization

SMART supports visualization of mzXML files. First, users follow the procedure to visually represent two-dimensional (2D) spectrum data in any user-specified m/z and RT region for the replicate sample(s) of interest (See **Figure 5.2.1**): *Select one specific sample in sample tree diagram in the left-hand panel → Click “Viewer” → Choose “2D Plot”*.

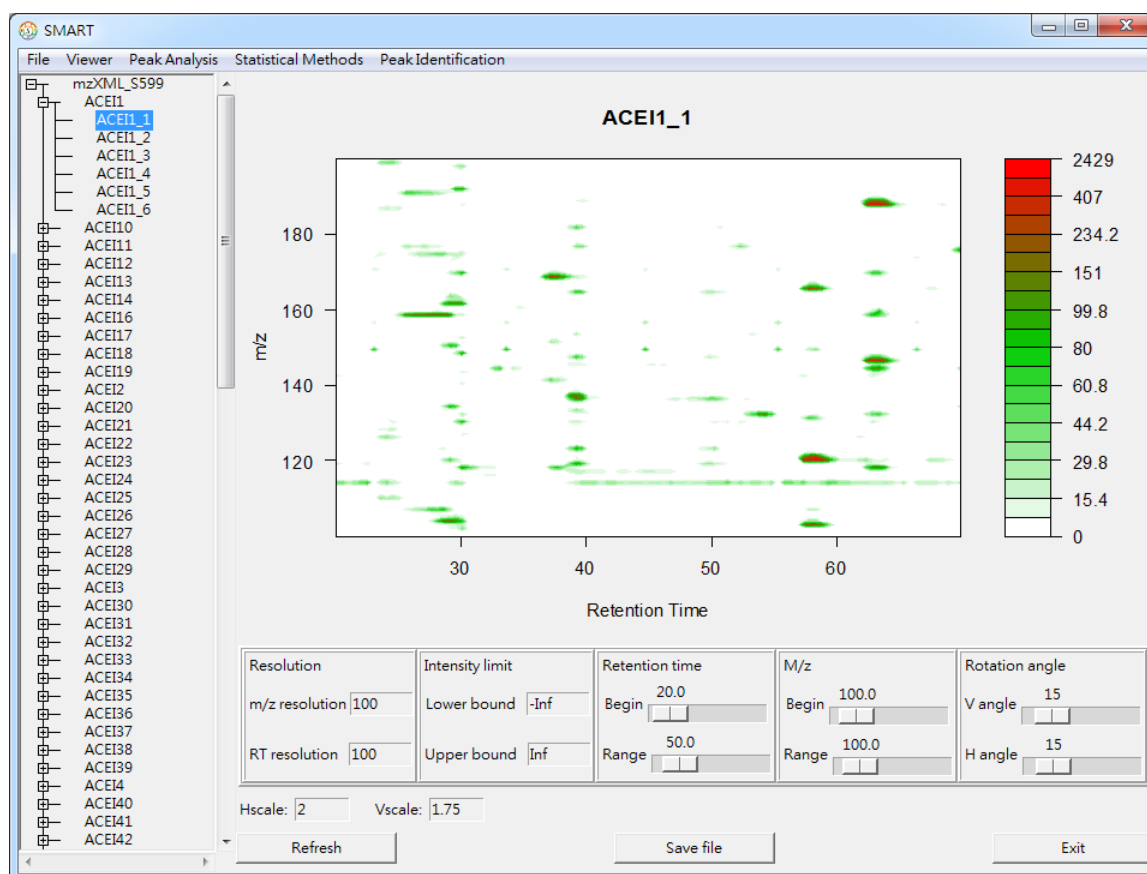


Figure 5.2.1. 2D spectrum plot.

Second, users follow the procedure to visually represent three-dimensional (3D) spectrum data in any user-specified m/z and RT region for the replicate sample(s) of interest (See **Figure 5.2.2**): *Select one specific sample in sample tree diagram in the left-hand panel → Click “Viewer” → Choose “3D Plot”.*

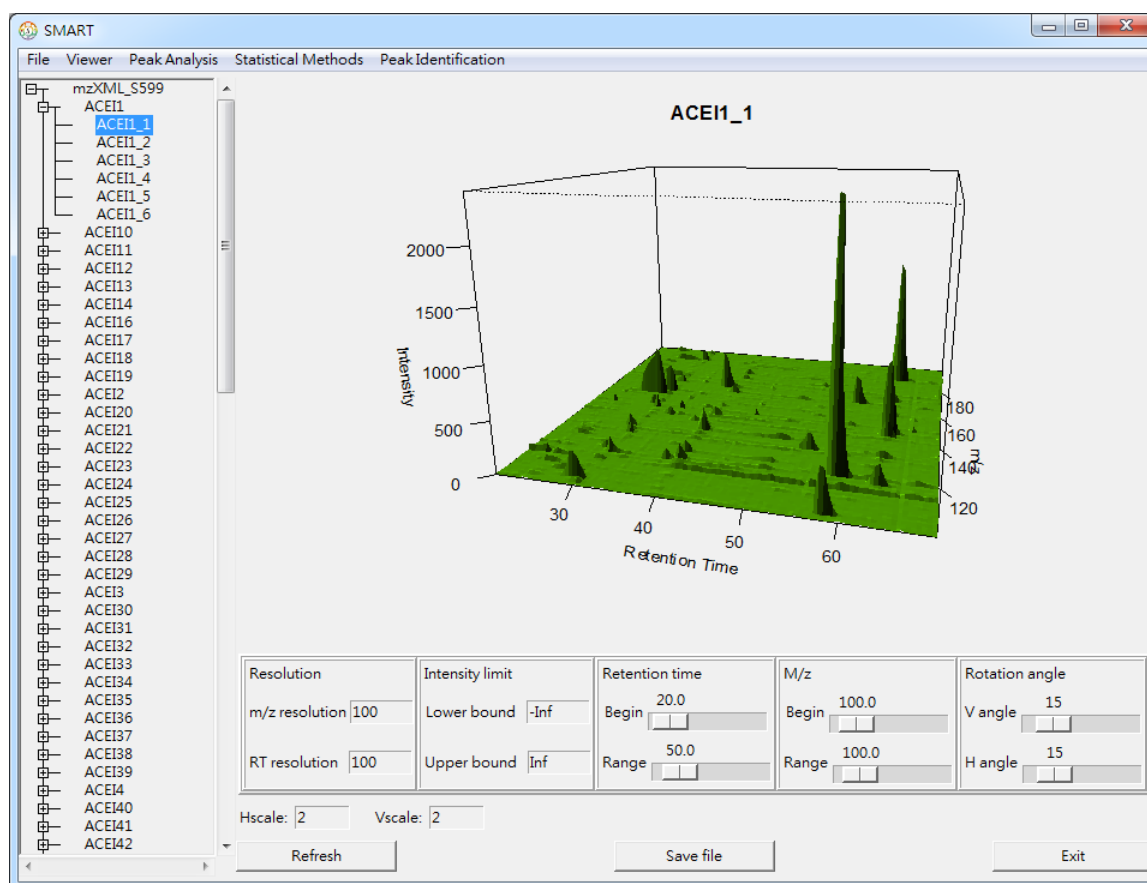


Figure 5.2.2. 3D spectrum plot.

Third, users follow the procedure to visually represent TIC in any user-specified m/z and RT region for the replicate sample(s) of interest (See **Figure 5.2.3**): *Select one specific sample → Click “Viewer” → Choose “TIC plot”*. Note that users can doubly click any point on the TIC curve (i.e., at a fixed RT scan) to look into a spectrum plot (See **Figure 5.2.4**). Under a fixed RT, the spectrum plot displays peak intensities for peaks across all m/z in the data.

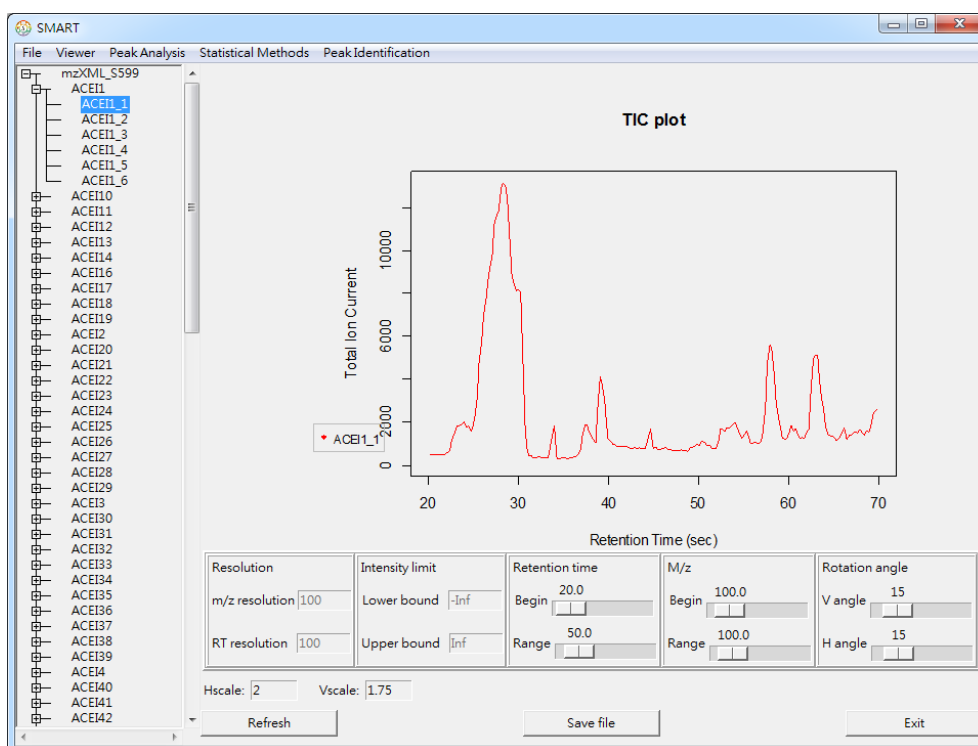


Figure 5.2.3. TIC plot.

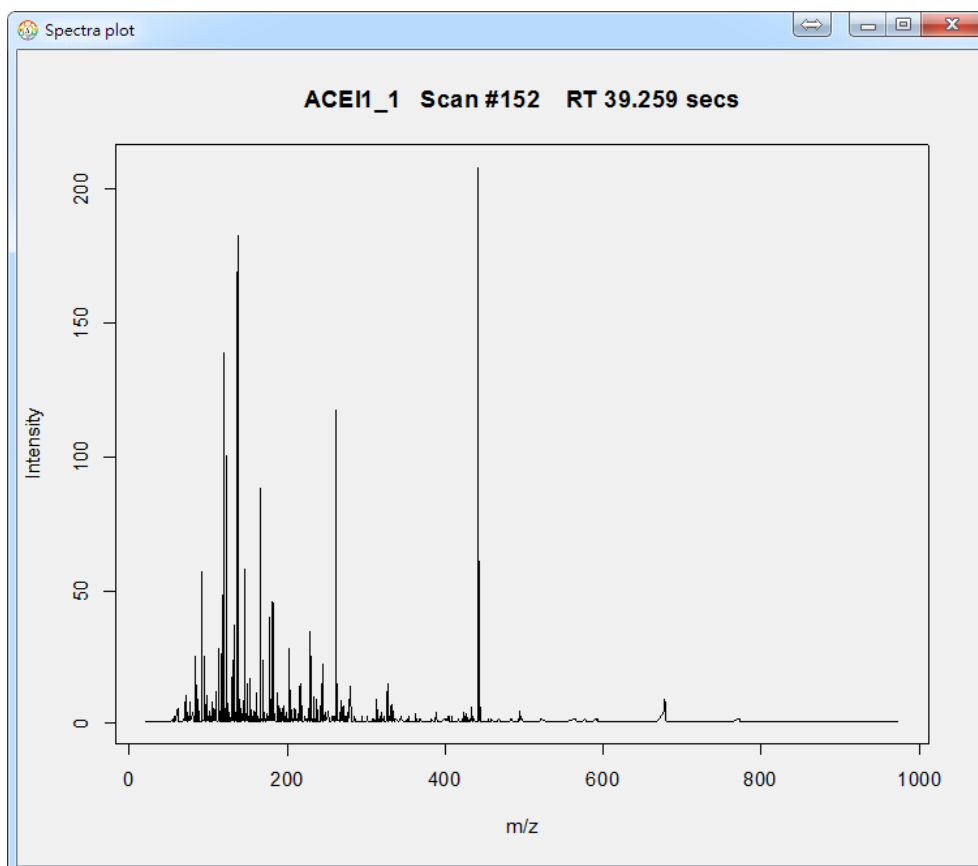


Figure 5.2.4. Spectrum plot at a fixed RT scan.

Fourth, users follow the procedure to visually represent TIC for multiple replicate samples and subjects (See **Figure 5.2.5**): *Select multiple replicate samples and subjects* → *Click “Viewer”* → *Choose “TIC plot”*.

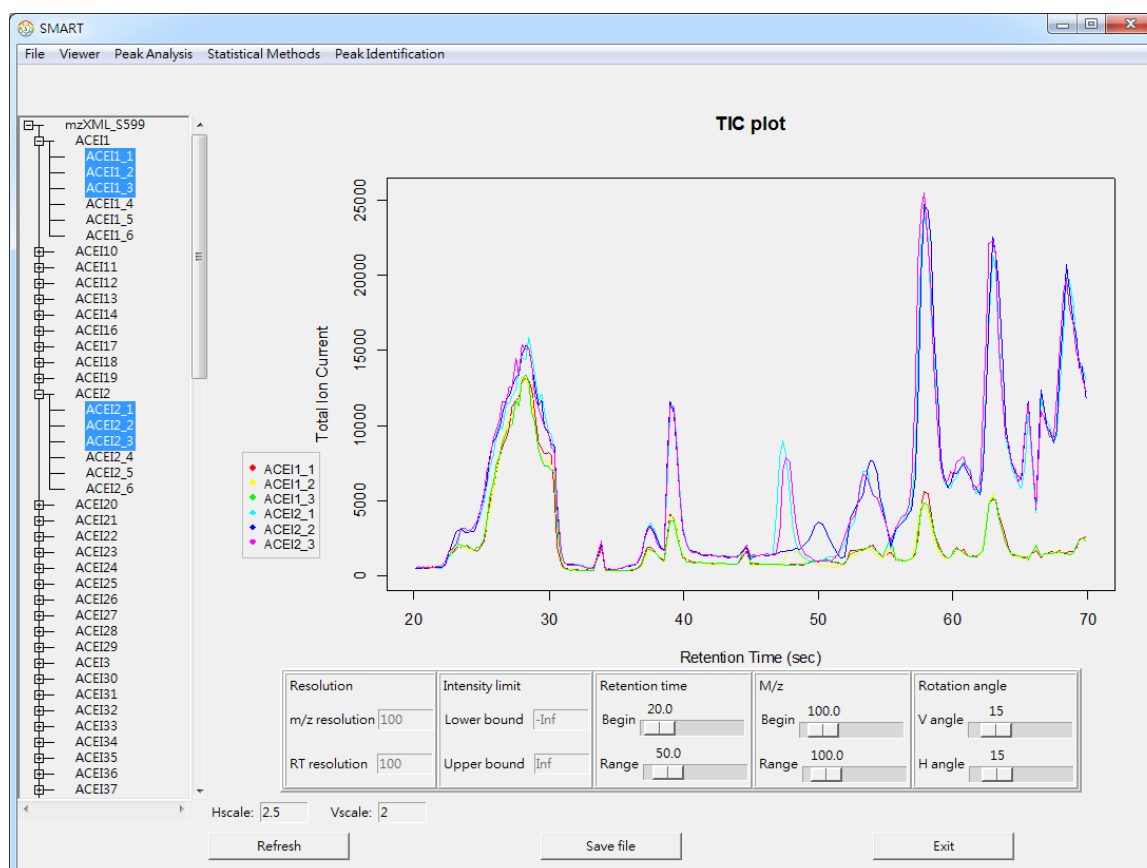


Figure 5.2.5. Cross-replicate-sample and cross-subject TIC overlay plots.

SMART provides five figure display options, consisting of (1) “Resolution” – figure resolution for densities of m/z and RT, (2) “Intensity limit” – the lower and upper bounds of peak intensity, (3) “Retention time” – the beginning RT and the range of RT, (4) “M/z” – the beginning m/z and the range of m/z, and (5) “Rotation angle” – rotation angles for a 3D spectrum plot. Users can modify the parameters as their need. In addition, users can reset the height and width of the **SMART** interface by filling in numbers and clicking “Refresh”. Users can save figures in the **SMART** interface by clicking “Save file” and escape the **SMART** environment any time by clicking “Exit”.

Finally, users follow the procedure to visually represent a TIC cluster diagram of all replicate samples of all subjects (See **Figure 5.2.6**): *Click “Viewer”* → *Choose “TIC*

Clustering” → Fill in retention time range of interest → Choose clustering method (Complete linkage, Average linkage, or Ward’s method) (See **Figure 5.2.7**).

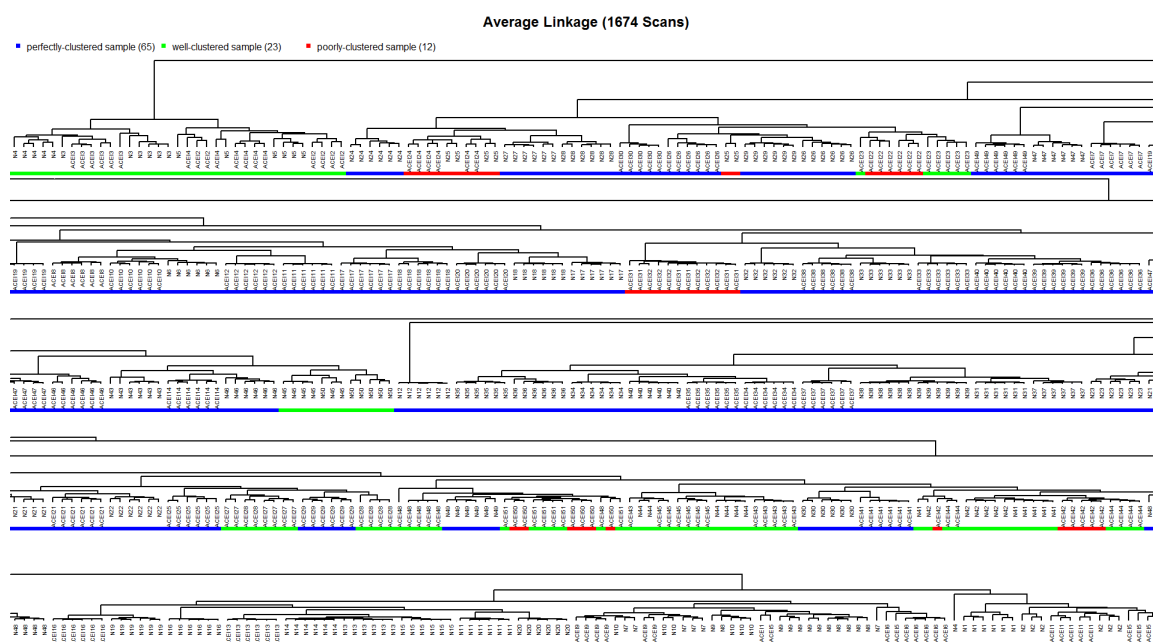


Figure 5.2.6. TIC cluster tree diagram.

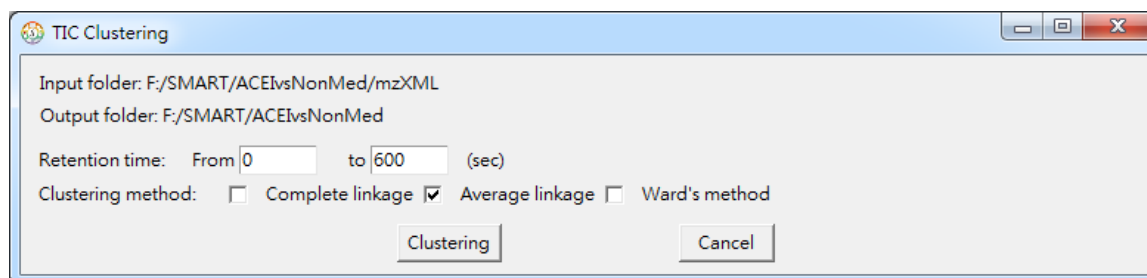


Figure 5.2.7. Interface of TIC clustering.

5.3 Peak alignment and annotation

SMART carries out peak alignment (i.e., peak detection and RT alignment) by incorporating the matched filtration and centWave in XCMS [1] (see **Figure 5.3.1**) and provides peak annotation by incorporating package CAMERA [2]. After importing raw spectrum or mzXML data file(s), users follow the procedure to perform peak alignment: Click “Peak Analysis” → Choose “Peak Alignment and Annotation”. Then a window about a citation of XCMS and CRAMEA will pop up.

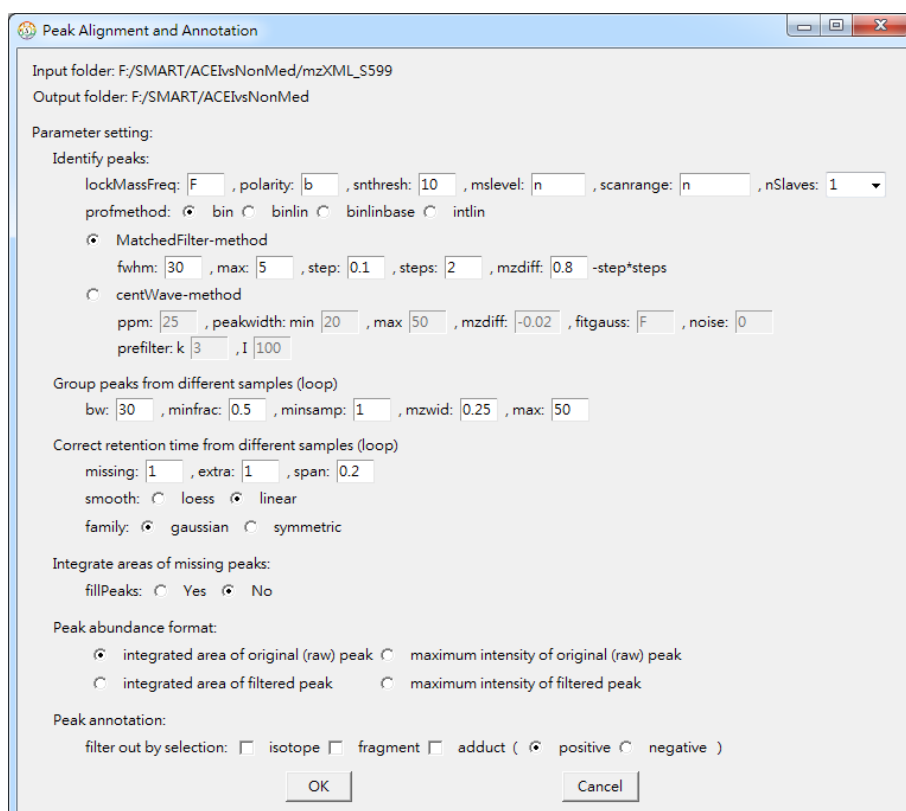


Figure 5.3.1. Interface and default settings of peak alignment and annotation.

After setting parameters and running peak alignment and annotation, **SMART** exports the comma-delimited peak abundance data, including the peak index, m/z, RT (seconds), and peak abundance of every replicate sample (see **Table 5.3.1**). In addition, **SMART** exports the peak annotation, including the peak index, m/z, RT (seconds), isotope, adduct, and peak group (pcgroup) (see **Table 5.3.2** as an example).

Table 5.3.1. Data after peak alignment

Peak _Index	mz	Ret_ time.sec	ACEI1_1	ACEI1_2	ACEI2_1	ACEI2_2	N1_1	N1_2	N2_1	N2_2
1	22.98998	27.7059	7.0783	6.6935	6.5680	6.4190	7.0669	6.6445	7.0774	6.5764
2	55.93659	336.5655	6.9994	7.9379	NA	8.4114	8.3721	8.3248	8.0697	8.3277
3	56.96345	335.7621	6.3181	6.0349	6.8107	6.7456	6.4947	6.3940	6.2704	6.0403
4	57.93642	335.8248	6.2205	6.1592	6.8686	6.5870	6.5836	6.4859	6.4517	6.4039
5	60.07343	29.2071	6.2056	NA	6.6890	6.5365	6.2876	6.1709	6.1499	6.2357
6	68.98333	24.0914	6.4834	6.5592	7.4892	7.4233	6.4490	6.4999	6.4278	6.4466
7	69.07036	53.4821	NA	NA	7.4339	6.9873	5.6892	5.2228	5.4749	5.2144
8	70.06573	32.3258	NA	NA	8.3018	8.0467	6.0851	5.7031	NA	NA
9	72.08130	64.1186	NA	NA	7.8959	7.6318	7.1014	7.1676	NA	NA
10	72.08131	37.8573	7.7268	7.3779	10.2431	10.0360	8.3697	8.2444	8.2197	7.9276
11	74.93966	335.9213	6.7730	7.0727	7.6475	7.3816	7.1578	7.3676	7.3392	7.2886
12	77.03906	58.1514	6.7476	6.6204	9.4728	9.5908	6.8912	6.9771	6.8814	6.7502
13	79.05363	58.1659	6.0404	5.4665	8.3506	8.3917	5.9861	6.1343	5.7149	5.9090
14	82.01453	24.0855	5.9334	6.0697	7.2531	7.0337	5.8216	6.0698	5.9923	5.5052
15	82.01411	335.6980	NA	NA	NA	10.1490	NA	7.3956	NA	NA
16	84.04471	51.8347	6.5344	6.5024	7.6371	7.5025	7.3956	7.1509	6.2767	NA
17	84.04468	30.1110	7.4784	7.5663	10.1490	10.1662	8.7341	8.6003	7.9926	7.7599
18	84.08100	25.1120	NA	NA	9.2761	9.1194	7.0255	6.7111	5.2853	NA
19	84.96001	24.1023	9.7988	9.6337	10.8931	10.7249	9.6305	9.5511	9.5060	9.3717
20	84.95969	335.6133	10.0044	10.1116	10.8662	10.4914	10.4841	10.3043	10.4317	10.1453

Table 5.3.2. Peak annotation

Peak_Index	mz	Ret_Time.sec	isotope	adduct	pcgroup
50	104.10134	149.0014			1
316	267.64494	149.0323			1
390	312.02142	149.0211			1
402	314.03463	149.0197			1
412	316.03100	149.0111			1
668	478.34350	149.0298	[67][M]+	[M+H-H2O]+ 495.35	1
670	479.34696	149.0197	[67][M+1]+		1
699	496.35508	149.0359	[75][M]+	[M+H]+ 495.35	1
700	497.35911	149.0335	[75][M+1]+		1
703	499.36087	149.0323			1
731	518.34134	149.0651	[82][M]+	[M+Na]+ 495.35	1
734	519.34629	149.0789	[82][M+1]+		1
786	552.28134	149.0298			1
992	771.99371	148.9999			1
993	772.99560	148.9977			1

If users choose to filter out redundant isotopic peaks, unwanted adducts, and/or daughter ion fragments, **SMART** will export the files of the peak abundance data and annotation after the peak filter. We use the peaks in **Table 5.3.2** as examples. First, isotope information is contained in the fourth column “isotope”. Peaks 668 ([67][M]+) and 670 ([67][M+1]+) are isotopes, peaks 699 ([75][M]+) and 700 ([75][M+1]+) are isotopes, and peaks 731 ([82][M]+) and 734 ([82][M+1]+) are isotopes. If users choose to filter out only redundant isotopic peaks, peaks 670, 700, and 734 will be removed (i.e., peak with [M]+ will be remained). Second, adduct information is contained in the fifth column “adduct”. If users choose to filter out those adducts which are NOT M+H, then peaks 668 and 731 will be removed. Finally, fragment information is contained in the final column “pcgroup”. Parent and daughter ion fragments will be grouped together and assigned the same value in pcgroup by CAMERA [2]. All peaks in **Table 5.3.2** are in the same group (pcgroup = 1). If users choose to filter out fragments, then, except for the parent peak 699 which has the largest average abundance 25481.9445, adduct M+H and not an isotopic peak, other daughter peaks will be removed. [Note that we do not suggest users to filter out the isotopes, adducts or fragments before the sample quality control procedure \(see Section 5.5\) because those peaks provide useful information for correlation estimation and quality control.](#)

5.4 Data preprocessing

SMART provides three procedures for data transformation and normalization (see **Figure 5.4.1**). Users follow the procedure to perform data preprocessing: [Click “Peak Analysis” → “Data Preprocessing”](#). The first procedure is abundance adjustment by using an internal standard. Users can choose “None”, “Scaling to the mean of internal standard”, or “Scaling to the median of internal standard”. Users must provide the m/z or RT of the internal standard in users’ peak abundance data if an international standard adjustment is requested (see **Figure 5.4.2**).

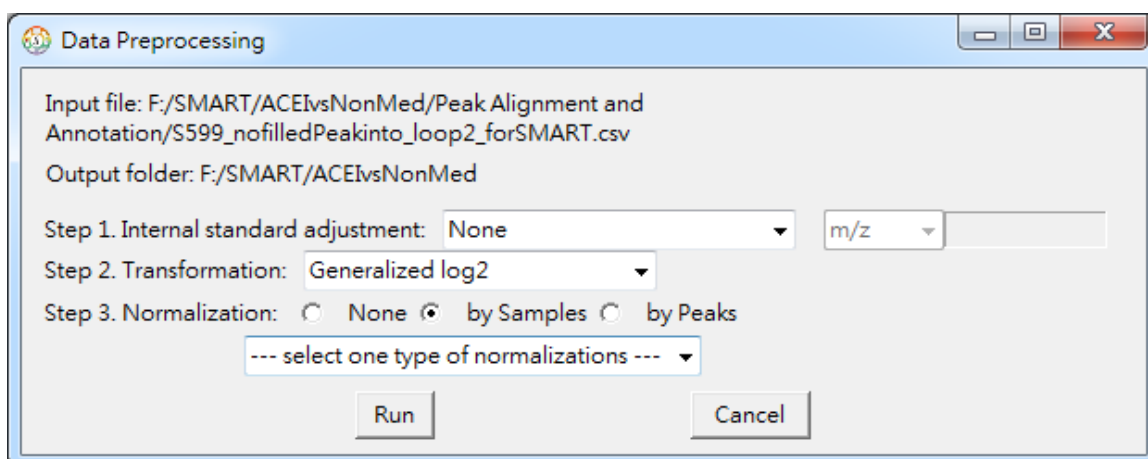


Figure 5.4.1. Interface of data preprocessing.

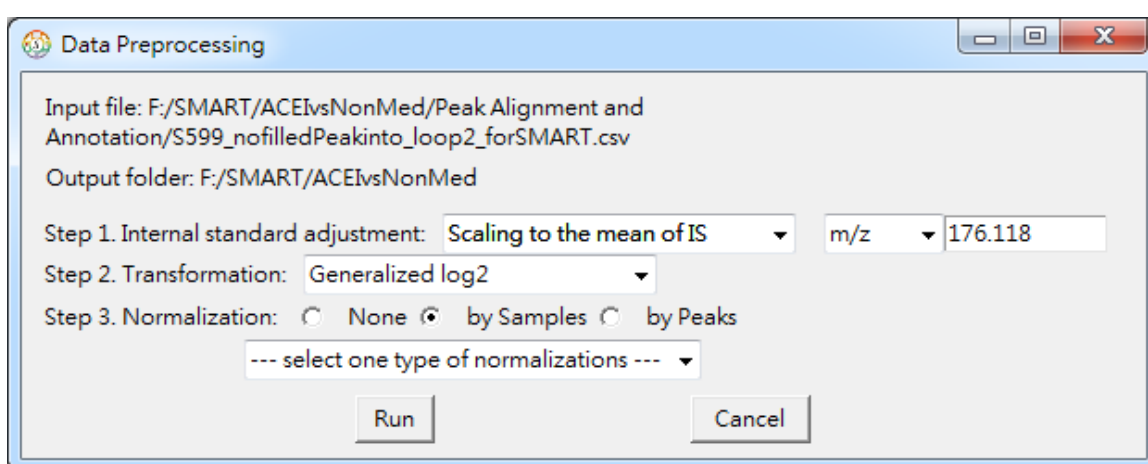


Figure 5.4.2. Internal standard adjustment.

The second procedure is to consider variable transformation. Users can choose “None” or “Generalized log2” [3]. The latter transformation is a variance-stabilization procedure and can avoid the problem of log transformation of a zero value (see **Figure 5.4.3**). Note that if generalized log2 was performed then a zero value in the original data will be transformed as 2.

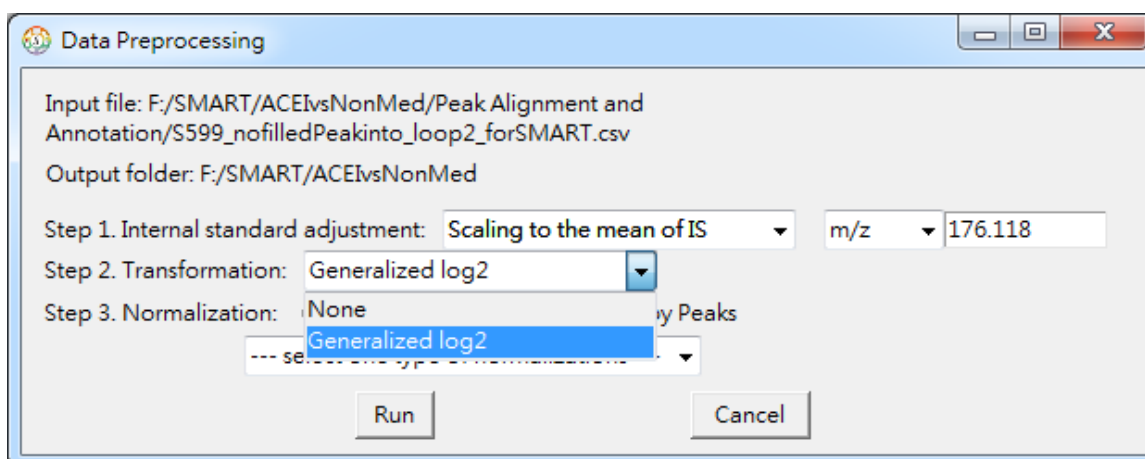


Figure 5.4.3. Transformation.

The third procedure is sample-based or peak-based data normalization. The sample-based normalization includes “None”, “Scale normalization – Mean”, “Scale normalization – Median”, “Quantile normalization”, “Standardization”, and “Pareto scaling” (see **Figure 5.4.4**). The peak-based data normalization includes “None”, “Scale normalization – Mean”, “Scale normalization – Median”, “Standardization”, and “Pareto scaling” (see **Figure 5.4.5**). Users choose one(s) of the aforementioned three preprocessing procedures according to their data properties. If data preprocessing has been done before, users can skip the transformation or normalization step in **SMART**.

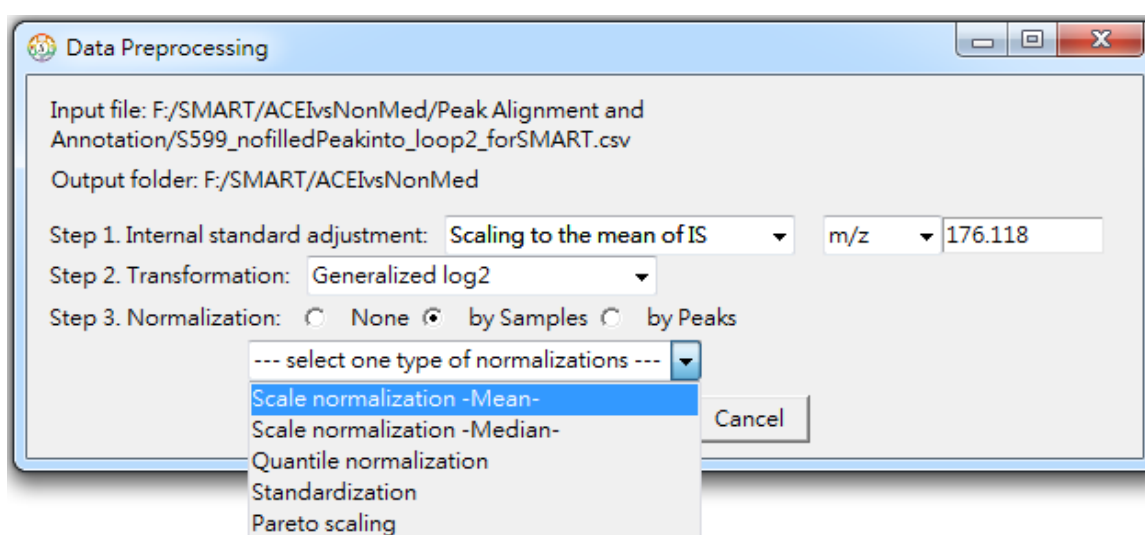


Figure 5.4.4. Sample-based data normalization.

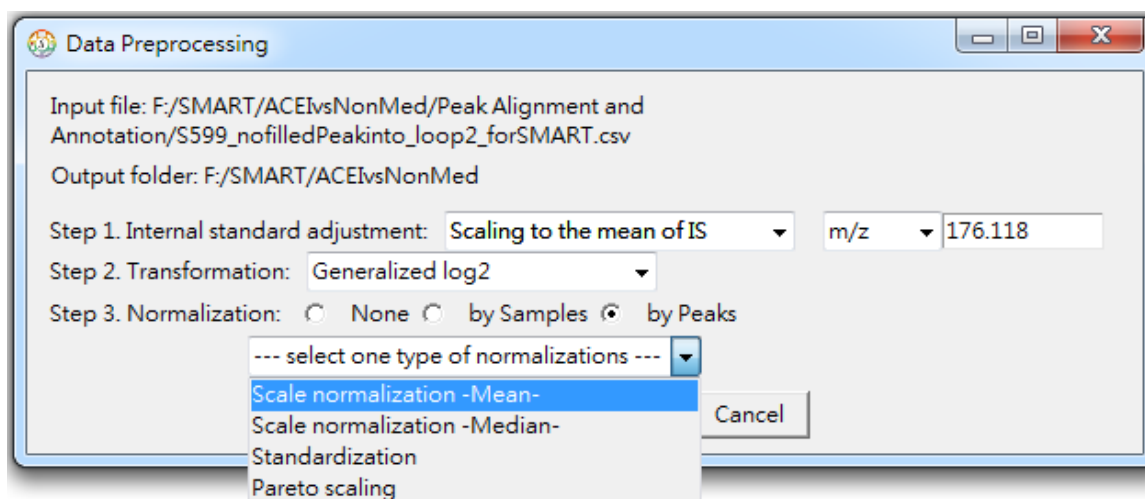


Figure 5.4.5. Peak-based data normalization.

5.5 Quality control

SMART provides quality control for peak filtering and sample filtering (see **Figure 5.5.1**). Users follow the procedure to perform quality control: [Click “Peak Analysis”](#) → [Choose “Quality Control”](#) → [Choose “Peak/Sample Filtering”](#). The first step is to specify the ranges of RT and m/z for quality control.

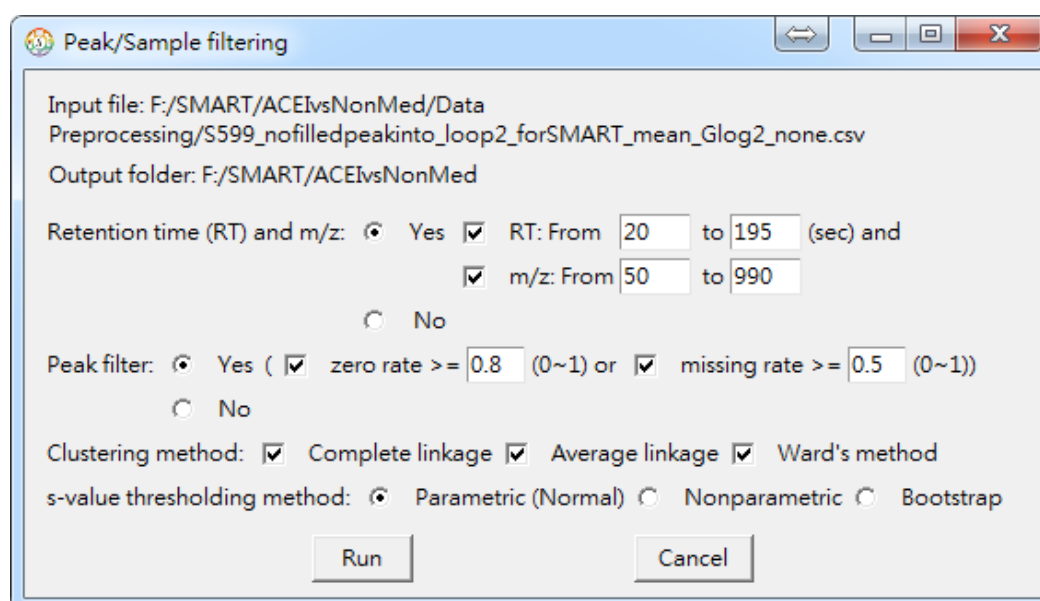


Figure 5.5.1. Peak and sample filtering.

The second step is to specify cutoffs for the zero rate (0 ~ 1) and missing rate (0 ~ 1). If users check the option to filter out peaks with a high zero rate, **SMART** requires users to input what value corresponds to “zero” in the data (see **Figure 5.5.2**). In other words, if users did not perform generalized log2 transformation, please input 0; if users performed

generalized log2 transformation in data preprocessing, please input 2. For peak filtering, if the missing or zero rate of a peak is higher than the cutoff, the peak is removed.

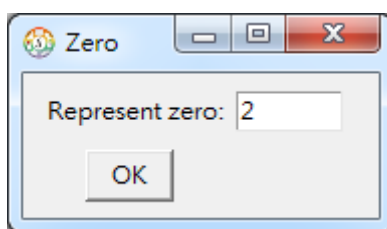


Figure 5.5.2. Interface of zero-value indication.

SMART calculates an *r-value* as the sum of squares within a subject (SSW) divided by the sum of squares between subjects (SSB) based on the aligned peak abundance data. The *r-value* is a quality index for peaks; the higher the *r-value* is, the poorer the peak quality. **SMART** calculates quantiles of the *r-value* (see **Figure 5.5.3**). Sample accuracy at the quantiles will be calculated and shown later. Users can also specify preferred (nonnegative) values as cutoffs of the *r-value*.

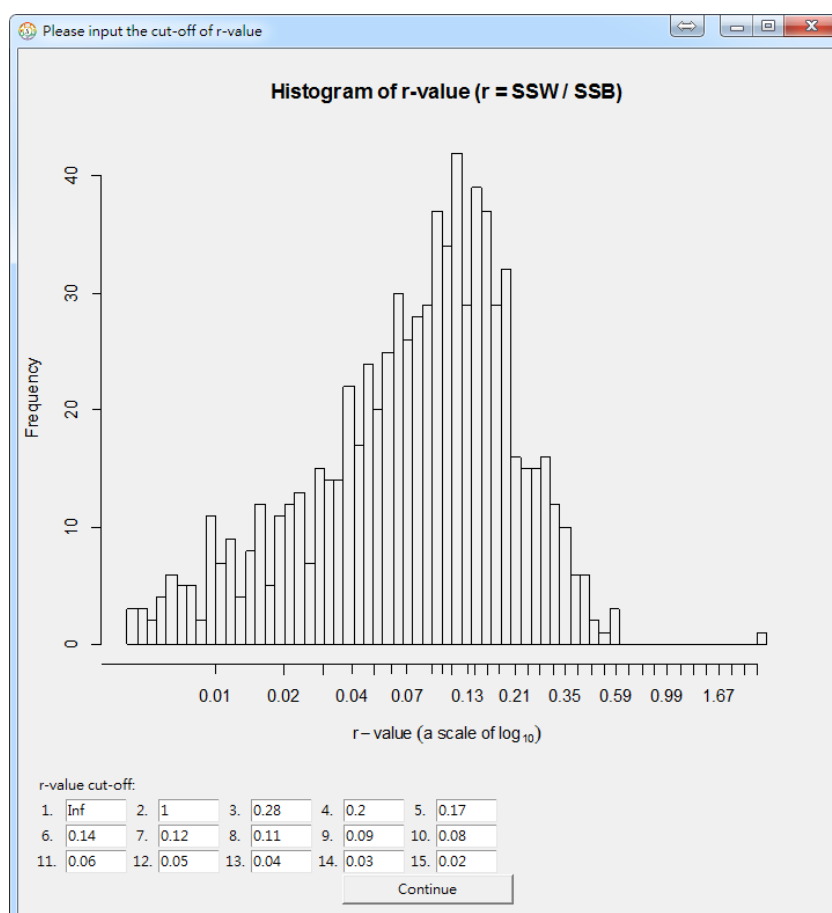


Figure 5.5.3. Distribution and quantiles of the *r-value*.

For sample filtering, quality of the replicate samples and subjects is evaluated only relying on the remaining high-quality peaks that have an *r-value* lower than a specified cutoff. Interactive window of the sample filtering comprises four plots (**Figure 5.5.4**). The upper-left plot is a sample accuracy plot. Note that users can click the legend of a clustering method (green: Ward's method; blue: Average-linkage method; red: Complete-linkage method) to hide the results of the cluster method and doubly click the legend to show the results again. Users follow the procedures to perform quality control: Users follow the procedure to start quality control: *Move the cursor to a point on a sample accuracy curve* → *Click to specify the cluster method and the cutoff of *r-value**. For example, users may choose the average linkage method and *r-value* cutoff = Inf because the setting attains the highest sample accuracy (**Figure 5.5.4**).

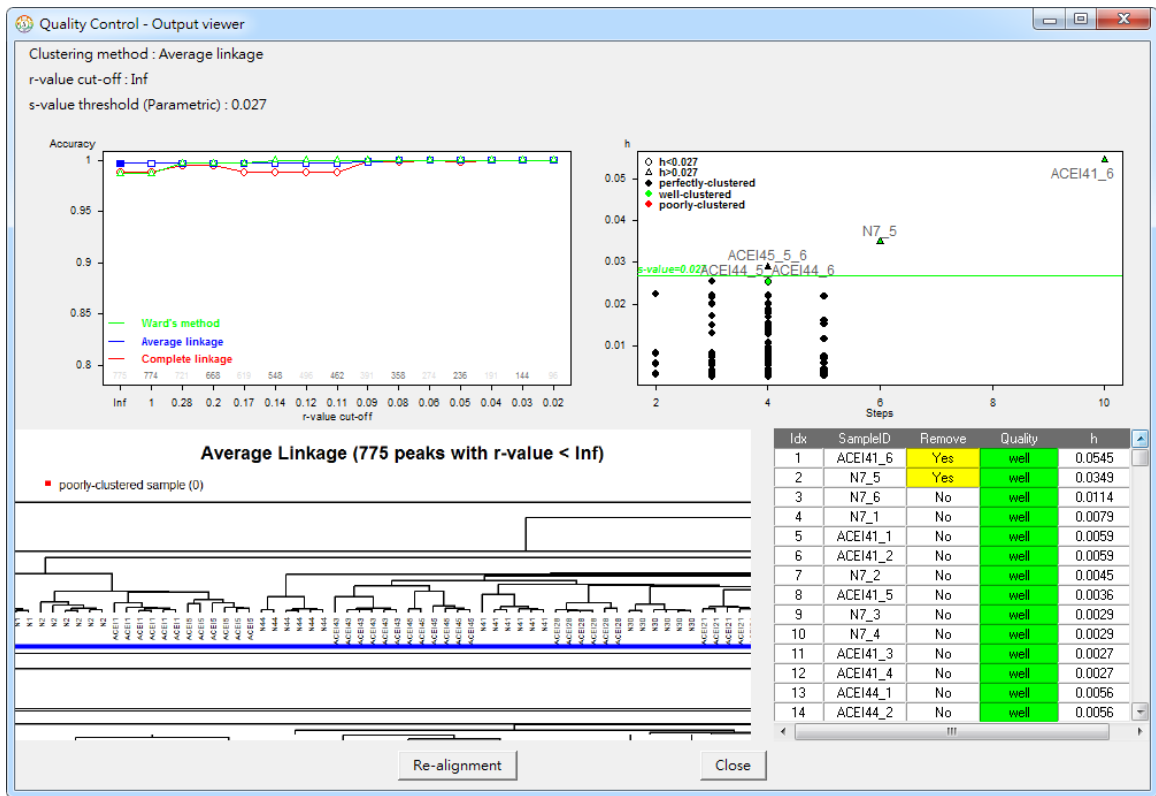


Figure 5.5.4 Output of quality control.

The lower-left plot is a cluster tree diagram of all replicate samples. Users can click the right button in mouse to zoom out and doubly click the left button to zoom in the tree diagram. The upper-right plot displays the measures of the replicate sample quality and *s-value*. The lower-right table is a summary result table for sample quality control. Users can

drag and drop the label of sample to avoid labels overlapped. The fourth column lists the categories of replicate samples: perfectly clustered (red), well clustered (green), and poorly clustered (black). The third column lists the replicate samples that must be removed. The results will change according to the parameters setting in the sample accuracy plot. After quality control, users can click “Re-alignment” to realign the data or click “Close” to finish the quality control. The results after quality control will be exported as an R data file. After the output window is closed, if users would like to see the output again, users can follow the procedure to view the outputs again: [Click “Peak Analysis”](#) → [Choose “Quality Control”](#) → [Choose “Output Viewer”](#) → [Specify the output R data file.](#)

5.6 Re-alignment and annotation

If users decide to do realignment (**Figure 5.6.1**) after removing poor-quality peaks and samples, users just specify the folder where they saved their mzXML files, and other procedures are the same as introduced in Section 5.3. Users can optionally choose to filter out the redundant isotopic peaks, unwanted adducts, and daughter ion fragments from the subsequent analysis. This can be done by checking “isotope”, “fragment”, and “adduct” and specifying “positive” and/or “negative” ion mode(s).

Peak Re-alignment and Annotation

Input folder: F:/SMART/ACEIvsNonMed/mzXML_S599
Output folder: F:/SMART/ACEIvsNonMed

Parameter setting:

Identify peaks:

lockMassFreq: F , polarity: b , snthresh: 10 , mslevel: n , scanrange: n , nSlaves: 1

profmethod: ☒ bin ☐ binlin ☐ binlinbase ☐ intlin

☒ MatchedFilter-method

fwHM: 30 , max: 5 , step: 0.1 , steps: 2 , mzdiff: 0.8 -step*steps

☐ centWave-method

ppm: 25 , peakwidth: min 20 , max 50 , mzdiff: -0.02 , fitgauss: F , noise: 0

prefilter: k 3 , I 100

Group peaks from different samples (loop)

bw: 30 , minfrac: 0.5 , minsamp: 1 , mzwid: 0.25 , max: 50

Correct retention time from different samples (loop)

missing: 1 , extra: 1 , span: 0.2

smooth: ☐ loess ☒ linear

family: ☒ gaussian ☐ symmetric

Integrate areas of missing peaks:

fillPeaks: ☐ Yes ☒ No

Peak abundance format:

☒ integrated area of original (raw) peak ☐ maximum intensity of original (raw) peak

☐ integrated area of filtered peak ☐ maximum intensity of filtered peak

Peak annotation:

filter out by selection: ☐ isotope ☐ fragment ☐ adduct (☒ positive ☐ negative)

OK Cancel

Figure 5.6.1. Interface of re-alignment and annotation

5.7 Batch effect detection

SMART can evaluate batch effects caused by known experimental conditions, unknown latent groups (LGs), or hidden substructures. Users follow the procedure to perform batch effect detection: [Click “Statistical Methods”](#) → [Choose “Batch Effect Detection”](#) → [Choose “Principal Component Analysis \(PCA\)” or “Latent Group \(LG\)”](#). Because **SMART** will check collinearity of batch effect variables and covariates in the subsequent analysis of covariance (ANCOVA), a covariate file (see **Table 5.7.1**) should be provided if some covariate(s) will be adjusted for (see **Figure 5.7.1** for a PCA and **Figure 5.7.2** for an LG analysis). In a covariate file, the first column is sample id and followed by covariate data with comma delimited. Name of a discrete covariate should be prefixed with “D” and a continuous covariate with “C”. For example, there are four discrete covariates (Dataset, Gender, Date, and Month) and two continuous covariates (Age and BMI) in this example (see **Table 5.7.1**).

Table 5.7.1. Covariate table.

SampleID	DDataset	DGender	CAGE	CBMI	DDate	DMonth
ACEI1	1	1	46	26.08	D0914	M09
ACEI2	1	2	32		D0914	M09
ACEI3	2	2	46	28.36	D1005	M10
ACEI4	2	1	32	23.79	D1005	M10
ACEI5	2	1	39	23.66	D1005	M10
N1	1	2	37	25.23	D0914	M09
N2	1	1	38	25.35	D0914	M09
N3	2	2	39	24.20	D1005	M10
N4	2	1	32	25.60	D1005	M10
N5	2	1	23	22.69	D1005	M10

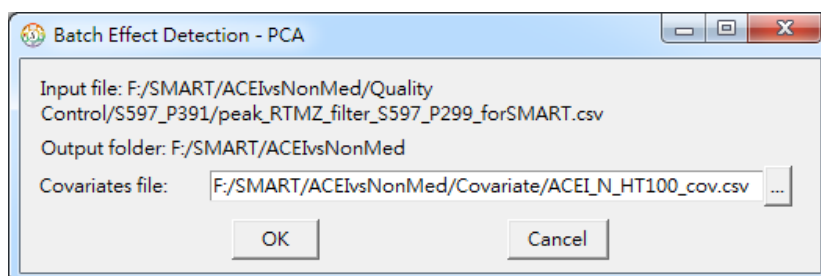


Figure 5.7.1. Interface of principal component analysis.

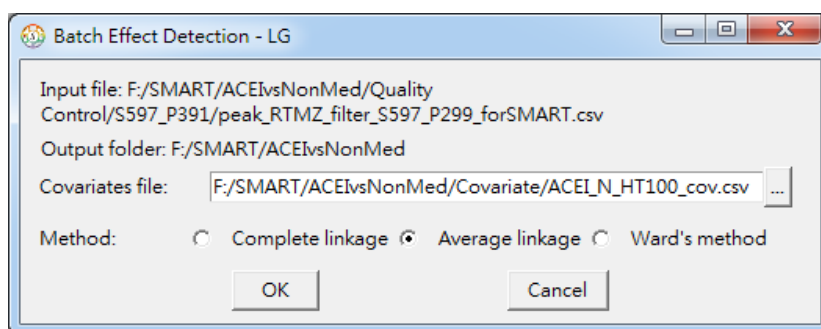


Figure 5.7.2. Interface of latent group analysis.

For a PCA, users can determine the number of PCs by a scree plot (left) or according to the number of PCs with an eigenvalue of ≥ 1 or with a proportion of variance explained of $>1\%$ in a variation-explained plot (right) (**Figure 5.7.3**). Finally, users can drag and drop the button "Number of PCs" in the bottom of **Figure 5.7.3** and click the button OK to assign the number of PCs. **SMART** provides PCA plots, where the PCs are colored according to known covariates and used to evaluate the relationship between the known covariates and the PCs (**Figure 5.7.4**). For example, if the pattern of the experiment date is matching to that of the pattern of LGs, then it reveals that the experiment date is one of the major batch effects. After a PCA, **SMART** exports PC scores for all replicate samples of all subjects for adjusting for the batch effects in the subsequent association analysis.

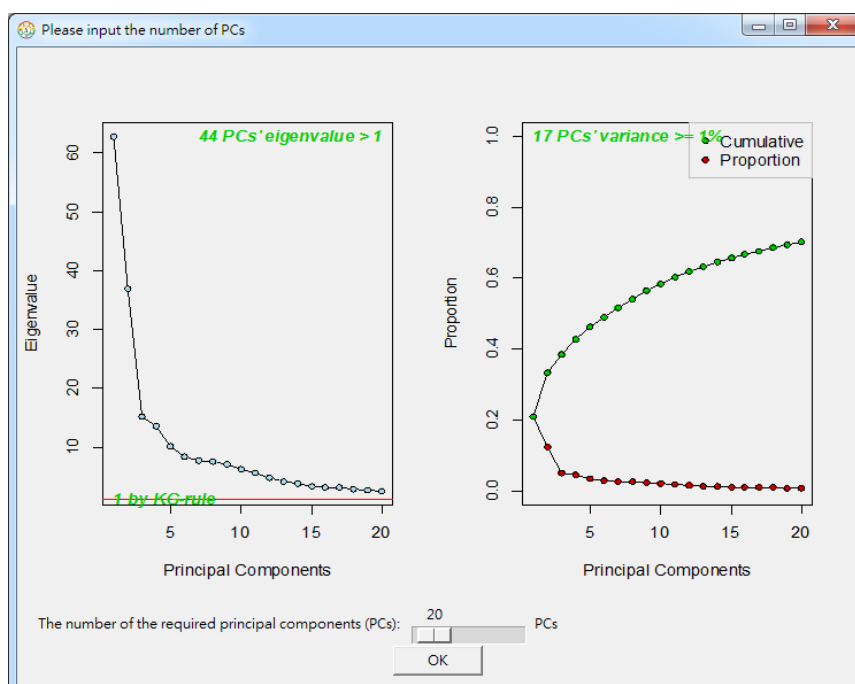


Figure 5.7.3. Interface of scree plot (left) and a variation-explained plot (right).

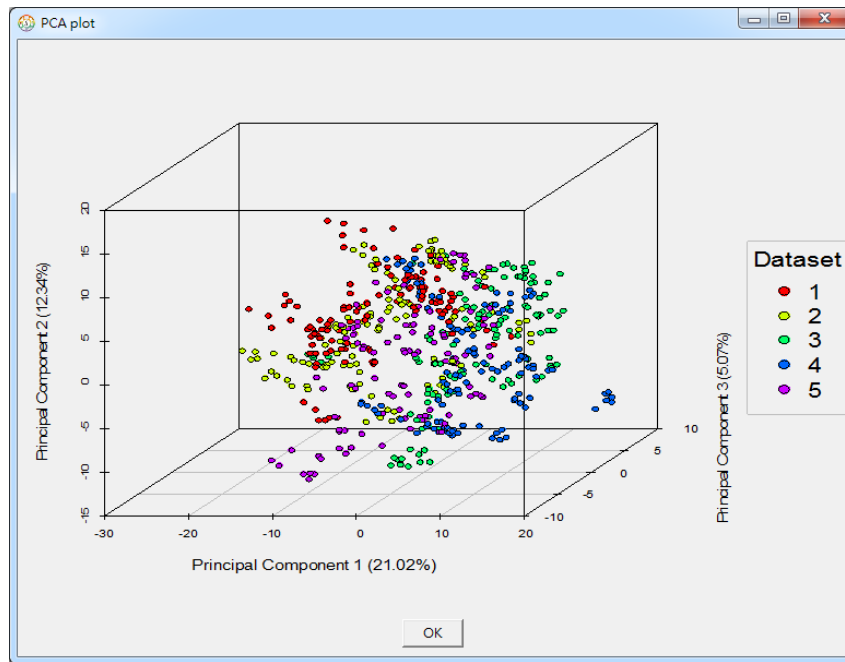


Figure 5.7.4. PCA plot.

For an LG analysis, users can determine the number of LGs and evaluate the patterns between known covariates and LGs by a heat map and cluster tree diagram (**Figure 5.7.5**). For example, a heat map and cluster tree diagram showed a substructure of three LGs. Then users key in the number of LGs in the bottom of the interface of heat map and cluster tree diagram. Because the third LG contains only two patients, **SMART** can remove the third LG by specifying the index of LG in the bottom of the interface. After an LG analysis, **SMART** exports index of LG for all replicate samples of all subjects for adjusting for batch effects in the subsequent association analysis.

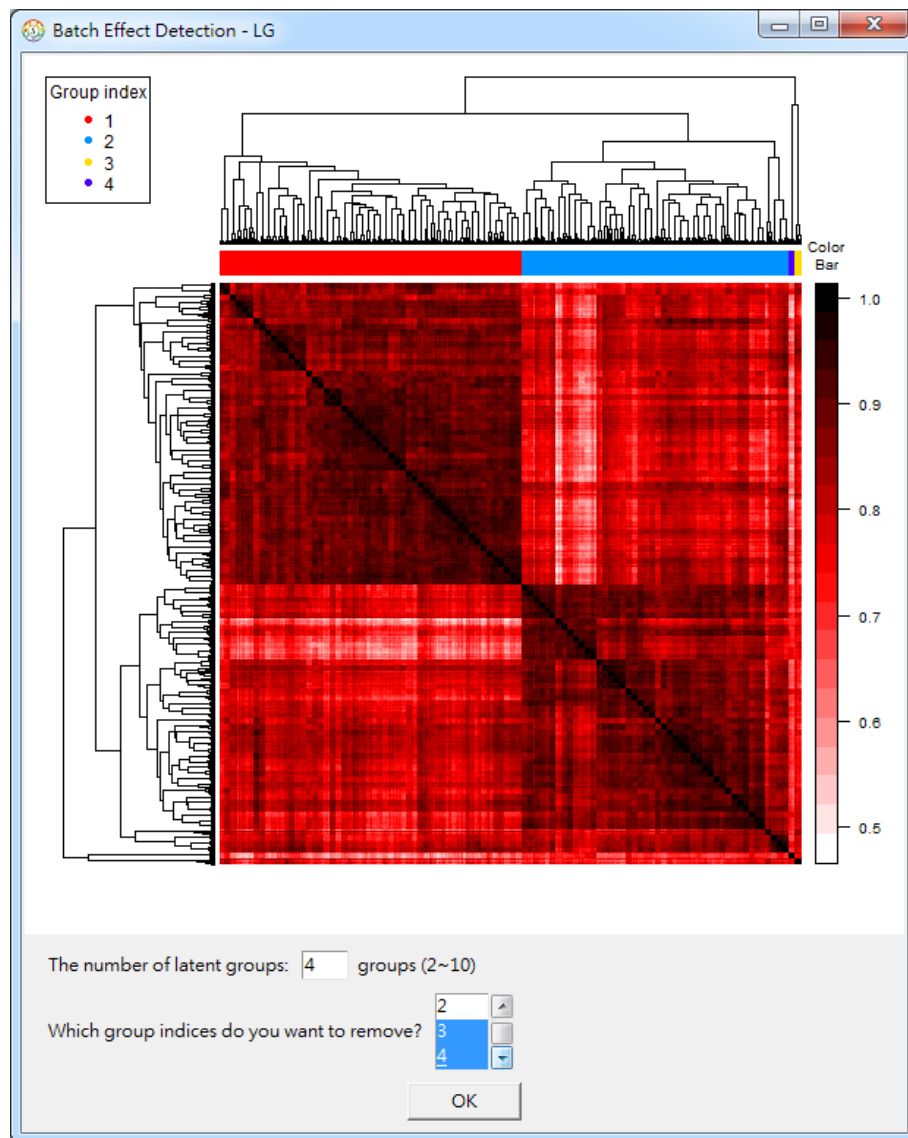


Figure 5.7.5. Heat map and cluster tree diagram.

5.8 Association analysis

To discover the association between metabolites and variables of interest, **SMART** provides a general Analysis of Covariance (ANCOVA) model for MWASs. Users follow the procedure to perform association analysis: [Click “Statistical Methods”](#) → [Choose “Analysis of Covariance \(ANCOVA\)”](#). In this model, the dependent variable is the peak abundance and the independent variables include the factor groups (e.g., case vs. control) or quantitative traits of interest (e.g., blood pressure), covariates, and batch effects. Users should specify their covariate file (optional), batch effect file (optional), and factor file (required) (see **Figure 5.8.1**). Covariate data format has been introduced in **Table 5.7.1**. In a batch effect file, the first column is sample id and followed by the data of batch effect variables (e.g., index of LG or PC scores). In a factor file, the first column is sample id and followed by a multinomial variable (i.e., factor group) or a continuous variable (i.e., quantitative trait(s)).

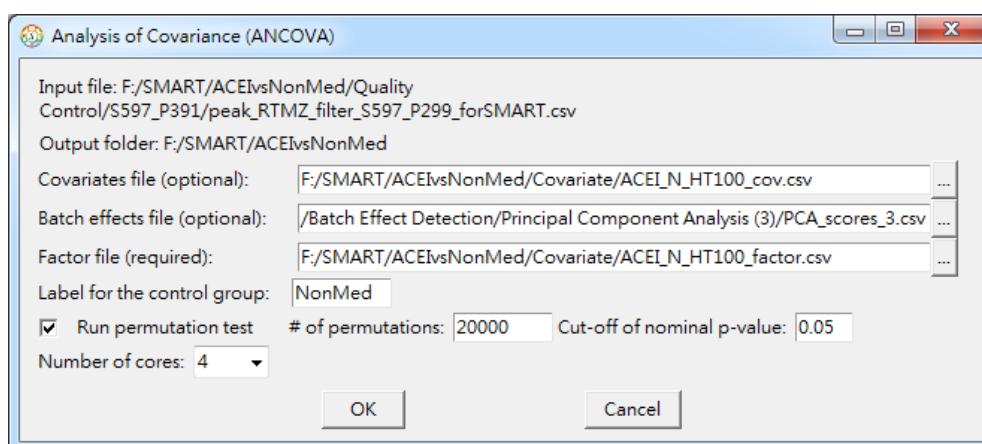


Figure 5.8.1. Interface of analysis of covariance (ANCOVA).

Next, “Label for the control group:” is used to specify a control group. If it is specified, **SMART** performs an ANCOVA analysis that the dependent variable is peak abundance and the main independent variable is a factor group variable. Here we use an antihypertensive pharmacometabolomics study as an example to demonstrate the operation. We are interested in association between peak abundance and a multinomial factor group ($X = 1, 2, 3, 4$, and 5 indicates “no medication”, “ACEi medication”, “ARB medication”, “CCB medication”, and “Diuretics” respectively). There are six covariates: Dataset, Gender, Age, BMI, Date, and Month. The control group is treated as a reference and other groups will be compared with the control group individually. For example, if

“NonMed” is specified in “Label for the control group:”, then **SMART** performs four ANCOVA analyses to compare means of peak abundance for ACEi vs. NonMed, ARB vs. NonMed, CCB vs. NonMed, and Diuretics vs. NonMed, separately. If “Label for the control group:” is not specified, then **SMART** performs a regression analysis, where the dependent variable is peak abundance and the main independent variable is a quantitative trait of interest (e.g., blood pressure).

Each of the ANCOVA analyses in the aforementioned example can consider different covariates. **SMART** automatically generates six covariates for each factor group. For example, in the ACEi group, the six covariates are named as “Dataset (ACEi vs. NonMed)”, “Gender (ACEi vs. NonMed)”, “Age (ACEi vs. NonMed)”, “BMI (ACEi vs. NonMed)”, “Date (ACEi vs. NonMed)”, and “Month (ACEi vs. NonMed)” (see **Figure 5.8.2**). Users can click some or all of the six covariates in the left-hand side window and press the >> button to include the covariates into the ANCOVA model in a comparison of ACEi vs. NonMed (see **Figure 5.8.2**). Similar procedures are applied to comparisons between NonMed and other medications.

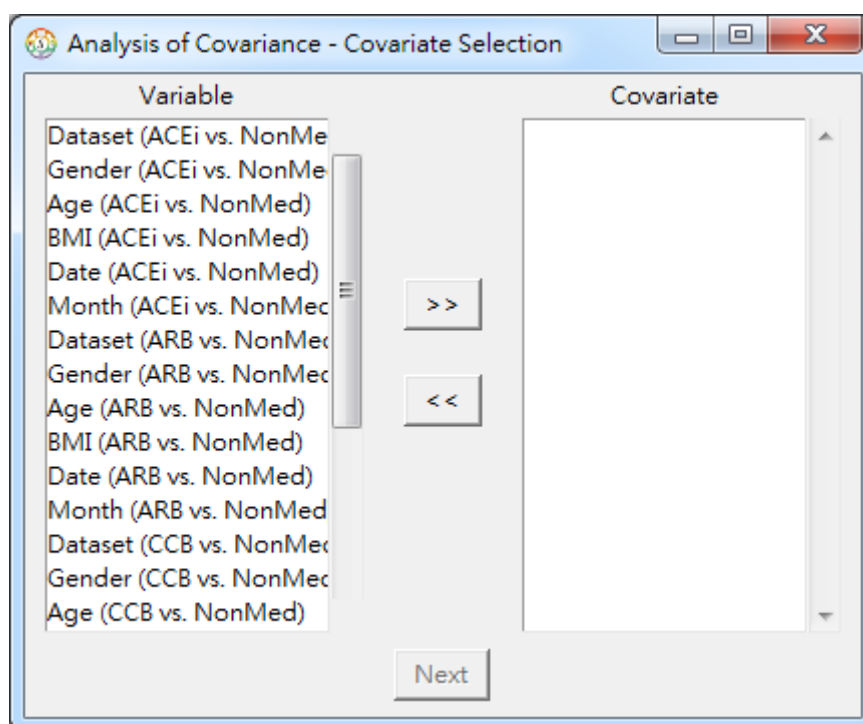


Figure 5.8.2. Interface of covariate selection in ANCOVA.

To avoid multicollinearity among the independent variables in ANCOVA, **SMART** calculates the variance inflation factor (VIF) for evaluating multicollinearity between batch

effects and factor groups and between batch effects and covariates. Users can input a VIF upper bound (see **Figure 5.8.3**). The default cutoff of VIF is 10. An LG or PC is removed from the ANCOVA model if its VIF is greater than the cutoff.

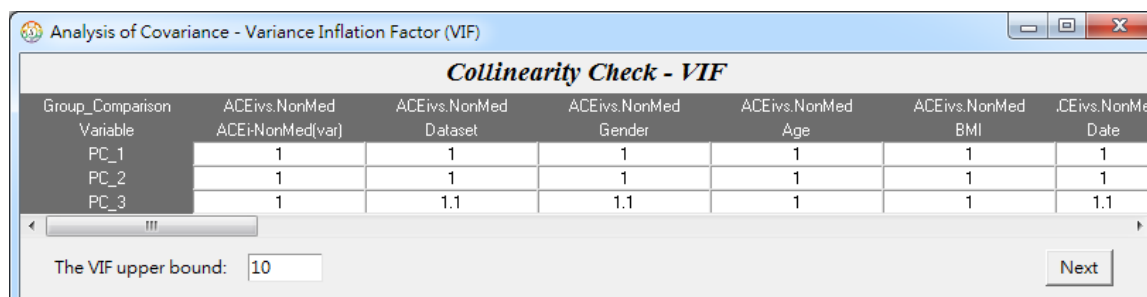


Figure 5.8.3. Interface of collinearity check.

Finally, users should specify the fitted ANCOVA models (see **Figure 5.8.4**). For example, “ACEi vs. NonMed: Group + Batch Effect + Group:Rep + Gender + Age” is the first ANCOVA model. This model is fitted to examine the difference of peak abundance in ACEi medication group and NonMed group. Group variable and batch effect variable are forced to include. Term “Group:Rep” considers replicate samples nested in the factor group under a nested study design. Term “Gender + Age” means that covariates gender and age are adjusted for.

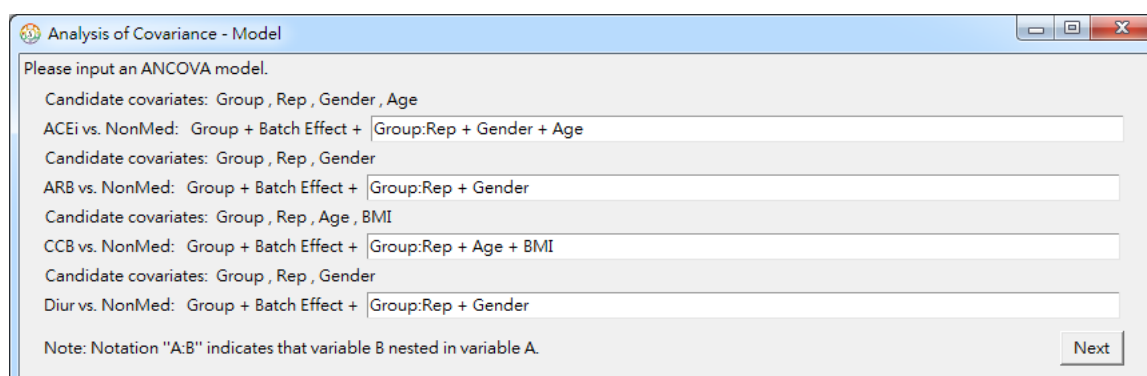


Figure 5.8.4. Interface of ANCOVA model construction.

SMART provides two methods for evaluating the statistical significance of the association between the metabolites and the factor groups. When the peak abundance follows a normal distribution, the F test is used; otherwise, a permutation test, which randomly shuffles the values of the factor group(s) or quantitative trait, can be used (see **Figure 5.8.1**). Nominal p -values (pv), adjusted p -values after false discovery rate (FDR)

adjustment (pFDR), empirical p -values (epv), and empirical p -values after FDR adjustment (epFDR) are exported. A volcano plot is generated to present the results (see **Figure 5.8.5**).

Figure 5.8.5. Volcano plot of nominal p -values.

To understand a quantitative relationship between multivariate responses and high-dimensional explanatory variables, SMART provides PLS regression [45, 76]. PLS regression can be combined with discriminant analysis (DA) (i.e., PLS-DA) for sample classification and biomarker selection [61]. SMART provides PLS and PLS-DA by applying the `ropls` package [64] (see **Figure 5.8.6**).

The second part is the response variable file for PLS or PLS-DA analysis. If the response variable file has only one continuous response variable whose name should have a prefix “C” (e.g., Cbp), **SMART** will perform PLS analysis. If the response variable file has only one discrete response variable whose name should have a prefix “D” (e.g., DGroup), **SMART** will perform PLS-DA analysis. If the response variable file contains multiple response variables, **SMART** will treat all as the continuous response variables and only perform PLS (see **Table 5.8.1**, **Table 5.8.2**, and **Table 5.8.3**).

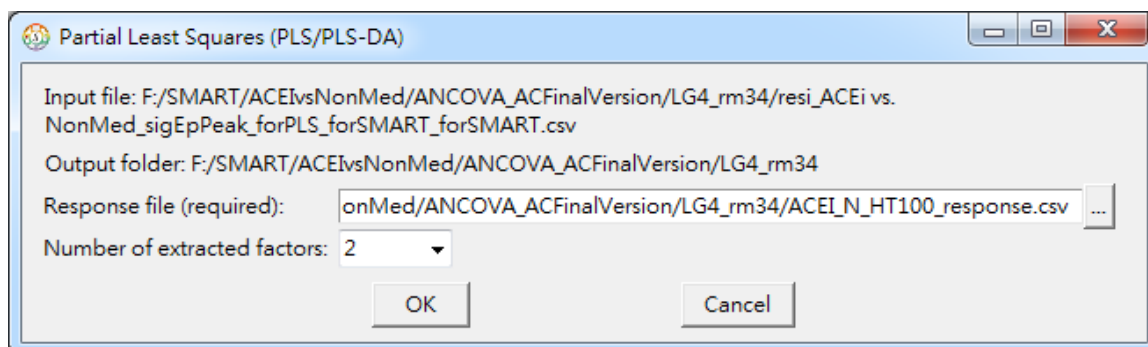


Figure 5.8.6. Interface of PLS/PLS-DA.

Table 5.8.1. Response variable file with a discrete response variable.

SampleID	DGroup
ACEI1	ACEi
ACEI2	ACEi
ACEI3	ACEi
ACEI4	ACEi
ACEI5	ACEi
N1	NonMed
N2	NonMed
N3	NonMed
N4	NonMed
N5	NonMed

Table 5.8.2. Response variable file with a continuous response variable.

SampleID	Cbp
ACEI1	126
ACEI2	110
ACEI3	112
ACEI4	121
ACEI5	125
N1	135
N2	140

N3	139
N4	142
N5	151

Table 5.8.3. Response variable file with more than one response variables.

SampleID	Cbp	CBMI
ACEI1	126	26.08
ACEI2	110	
ACEI3	112	28.36
ACEI4	121	23.79
ACEI5	125	23.66
N1	135	25.23
N2	140	25.35
N3	139	24.20
N4	142	25.60
N5	151	22.69

The third part is the number of PLS factors. Users have to decide how many PLS factors do you want to calculate. The maximum number is 15. For each peak, **SMART** will calculate the variable importance for projection (VIP) value and the p-value from a correlation test. The correlation test is to test the association between peak abundance and a response variable. For a continuous response variable, **SMART** will measure the correlation by calculating Pearson's correlation coefficient. For a discrete response variable, **SMART** will calculate Kendall's tau. **SMART** exports the plots of p-value vs. VIP (see **Figure 5.8.7**). In addition, a score plot of subjects of the first PLS factor vs. the second PLS factor colored by the response variable will be provided (see **Figure 5.8.8**). In **Figure 5.8.8**, there are 100 subjects, and 50 of them were treated with ACEi and the remainders were non-treated. The discrete response variable represents the ACEi group and nonmedication group. Black ellipse is the 95% confidence band of 100 subjects, blue ellipse is the 95% confidence band of 50 ACEi treated subjects, and blue ellipse is the 95% confidence band of 50 nonmedication subjects. The results can be used to judge if the identified peaks are informative to subject classification or subgrouping. The summary results, regression coefficients, x loadings, x scores, and x weights are exported.

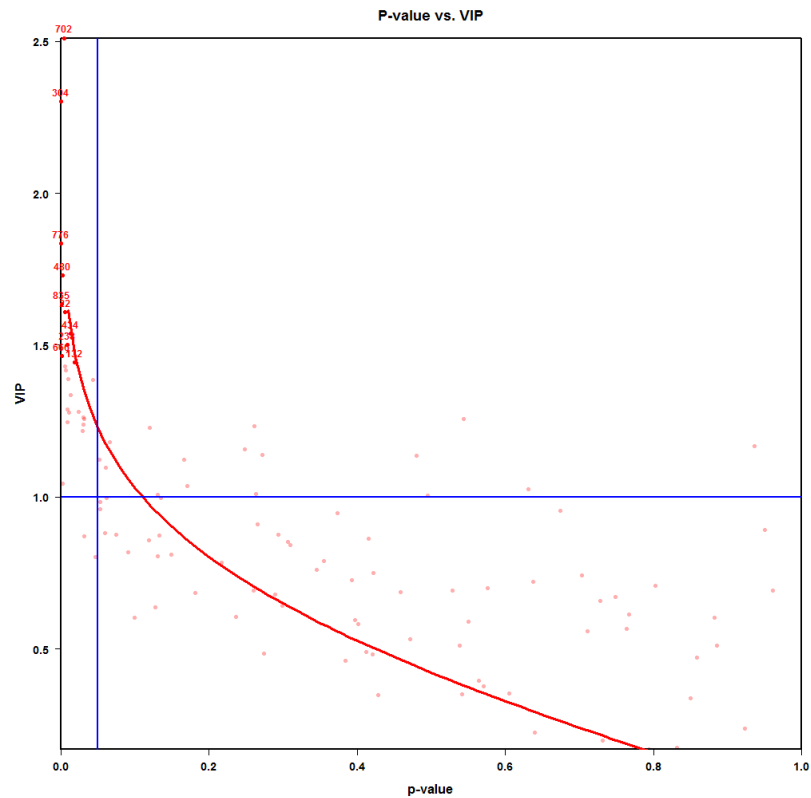


Figure 5.8.7. P-value vs. VIP plot.

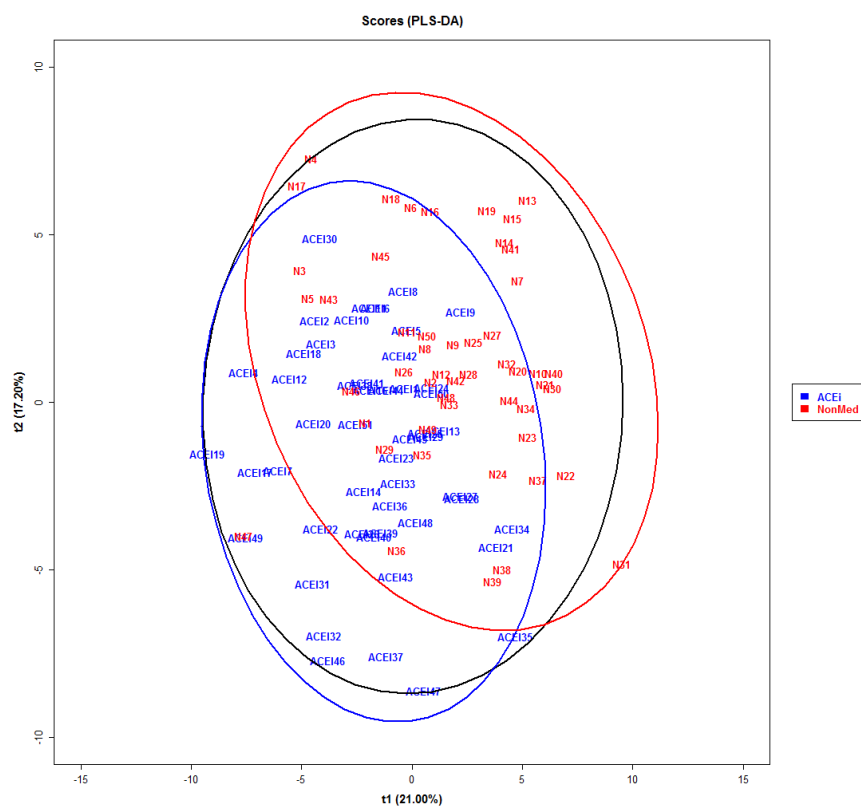


Figure 5.8.8. Score plot of the first two PLSs.

5.9 Peak identification

For peak identification, **SMART** provides a batch search interface to query the m/z from HMDB and MassBank (see **Figure 5.9.1**). The metabolite data of HMDB in XML format released on 2016-06-26 were downloaded from <http://www.hmdb.ca/downloads> and the metabolite data of MassBank in txt format released on 2013-02-01 were downloaded from <http://www.massbank.jp/en/download.html>.

Users follow the procedure to perform peak identification: Click “Peak Identification” → Choose “Batch Search”. In the interface of peak identification (see **Figure 5.9.1**), users can load a csv file that the first three columns are peak index, m/z, and RT (e.g., peak abundance data). Or users provide m/z to perform peak identification. Three ion modes, “Neutral”, “Positive”, and “Negative”. can be chosen. The candidates of possible adducts should be specified. Finally, users set the tolerance for the acceptable m/z difference.

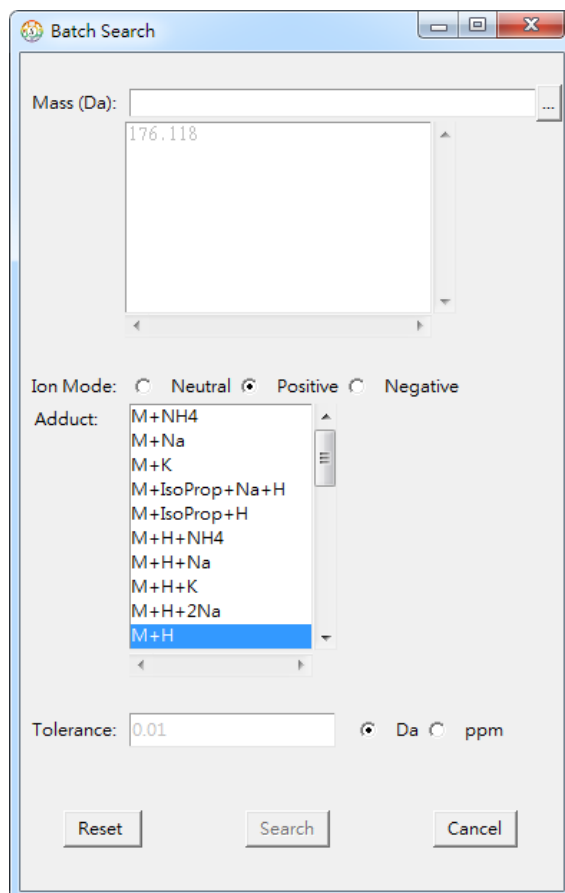
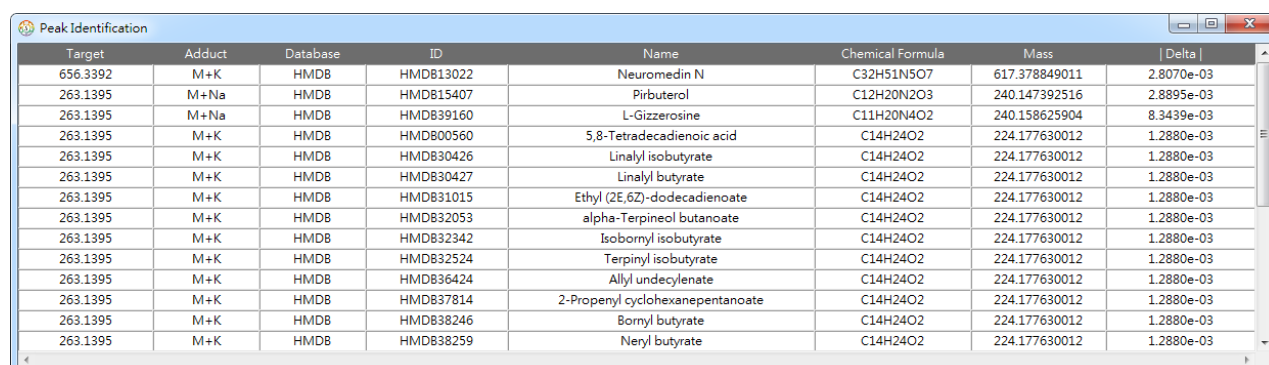


Figure 5.9.1. Interface of Volcano plot of nominal *p*-values.

After clicking Search, **SMART** shows a table including “Target” (i.e., the m/z we queried), “Adduct”, “Database”, “ID”, “Name”, Chemical Formula”, “Mass” (i.e., M: the

monoisotopic molecular weight), and “Delta” (see **Figure 5.9.2**). Also, the peak identification table is exported as a csv file.



Target	Adduct	Database	ID	Name	Chemical Formula	Mass	Delta
656.3392	M+K	HMDB	HMDB13022	Neuromedin N	C32H51N5O7	617.378849011	2.8070e-03
263.1395	M+Na	HMDB	HMDB15407	Pirbuterol	C12H20N2O3	240.147392516	2.8895e-03
263.1395	M+Na	HMDB	HMDB39160	L-Gizzerosine	C11H20N4O2	240.158625904	8.3439e-03
263.1395	M+K	HMDB	HMDB00560	5,8-Tetradecadienoic acid	C14H24O2	224.177630012	1.2880e-03
263.1395	M+K	HMDB	HMDB30426	Linalyl isobutyrate	C14H24O2	224.177630012	1.2880e-03
263.1395	M+K	HMDB	HMDB30427	Linalyl butyrate	C14H24O2	224.177630012	1.2880e-03
263.1395	M+K	HMDB	HMDB31015	Ethyl (2E,6Z)-dodecadienoate	C14H24O2	224.177630012	1.2880e-03
263.1395	M+K	HMDB	HMDB32053	alpha-Terpineol butanoate	C14H24O2	224.177630012	1.2880e-03
263.1395	M+K	HMDB	HMDB32342	Isobornyl isobutyrate	C14H24O2	224.177630012	1.2880e-03
263.1395	M+K	HMDB	HMDB32524	Terpinyl isobutyrate	C14H24O2	224.177630012	1.2880e-03
263.1395	M+K	HMDB	HMDB36424	Allyl undecylenate	C14H24O2	224.177630012	1.2880e-03
263.1395	M+K	HMDB	HMDB37814	2-Propenyl cyclohexanepentanoate	C14H24O2	224.177630012	1.2880e-03
263.1395	M+K	HMDB	HMDB38246	Bornyl butyrate	C14H24O2	224.177630012	1.2880e-03
263.1395	M+K	HMDB	HMDB38259	Neryl butyrate	C14H24O2	224.177630012	1.2880e-03

Figure 5.9.2. Peak identification table.

6. Example

SMART provides a real example from our antihypertensive pharmacometabolomics study. The data can be downloaded from the **SMART** website. In this example, there are 10 young-onset hypertensive patients from the Academia Sinica Multi-Center Young-Onset Hypertension Study [4]. The patients are divided into two groups: the medication group comprises 5 patients who were treated with ACEis and the nonmedication group (NonMed) comprises 5 patients who were not treated with antihypertensive medicine. Six technical replicate samples were used for each of the 10 patients.

7. Application Programming Interface

This section provides brief description about the application programming interface (API) that users can use them to develop their own packages easily (see **Table 7.1**).

Table 7.1. Main APIs in **SMART**.

Function (no arguments and no return value)	Input		Output	
	Variable	Description	Variable	Description
Quality control QualityControl()	QCtable	data.frame: peak abundance data	QCtable_temp	data.frame: post-QC peak abundance data
	method	numerical vector: 1 = complete linkage 2 = average linkage 3 = ward method	r	numerical vector: peak's r-value
	height_method	a numerical value: 1 = parametric method 2 = nonparametric method 3 = bootstrap	acc	list: \$height = heights for different linkage methods \$height_thres = s-value for different linkage methods
Batch effect detection (PCA) pcaStat()	pcatable	data.frame: peak abundance data	pcainfo	list: \$scores = PC scores \$scree = eigenvalues \$varprop = proportion of explained variation \$varpropcum = cumulative proportion of explained variation
	Covariate	data.frame: covariate data		
	BEtable	data.frame:	hr	cluster object:

Batch effect detection (LG) LatentGroup()		peak abundance data		hierarchical clustering result
	Covariate	data.frame: covariate data		
ANCOVA ANCOVA()	peakabun	data.frame: peak abundance data	pvalue	matrix: p-values of factor, covariates, and batch effects (per row a peak) matrix: FDR-adjusted p-values of factor, covariates, and batch effects (per row a peak) matrix: estimates of beta of factor, covariates, and batch effects (per row a peak)
	Factor	data.frame: factor information	pfdr	
	ctrl_label	a character value: control label	beta_temp	
	ncpu	a numeric value: number of threads		
	Covariate	data.frame: covariate data		
	batch_effect	data.frame: batch effect data		

8. References

- [1] Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal Chem*. 2006; 78(3):779-87.
- [2] Kuhl C, Tautenhahn R, Böttcher C, Larson TR, Neumann S. CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal Chem*. 2012; 84(1):283-9.
- [3] Durbin BP, Hardin JS, Hawkins DM, Rocke DM. A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*. 2002; 18 Suppl 1:S105-10.
- [4] Yang HC, Liang YJ, Wu YL, Chung CM, Chiang KM, Ho HY, Ting CT, Lin TH, Sheu SH, Tsai WC, Chen JH, Leu HB, Yin WH, Chiu TY, Chen CI, Fann CS, Wu JY, Lin TN, Lin SJ, Chen YT, Chen JW, Pan WH. Genome-wide association study of young-onset hypertension in the Han Chinese population of Taiwan. *PLoS One*. 2009; 4(5):e5459.
- [5] Nicholson JK, Lindon JC. Systems biology: Metabonomics. *Nature*. 2008; 455(7216):1054-6.
- [6] Szymańska E, Saccenti E, Smilde AK, Westerhuis JA. Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies. *Metabolomics*. 2012; 8(Suppl 1):3-16.
- [7] Thévenot EA, Roux A, Xu Y, Ezan E, Junot C. Analysis of the Human Adult Urinary Metabolome Variations with Age, Body Mass Index, and Gender by Implementing a Comprehensive Workflow for Univariate and OPLS Statistical Analyses. *J Proteome Res*. 2015; 14(8):3322-35.

9. News & updates

- 27 March 2019
 - ✓ **SMART** version 1.2 released.
 - Add “Pareto scaling” method to sample-based normalization and add “Scale normalization – Mean”, “Scale normalization – Median”, “Standardization”, and “Pareto scaling” to peak-based normalization.
 - Output the residual data from ANCOVA that factor variable is not included and batch effects and covariates are included; the file can be used as an input file for PLS or PLS-DA analysis.
 - Add PLS and PLS-DA to the module of association analysis.
 - Add a function for peak identification.