Version 2.0 (June, 2025)

User Guide for SMART 2.0

Yu-Jen Liang, Chih-Ting Yang, Chia-Wei Chen, Yin-Chun Lin, Shu-Yao Lin, Yi Sheng Wang, 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.

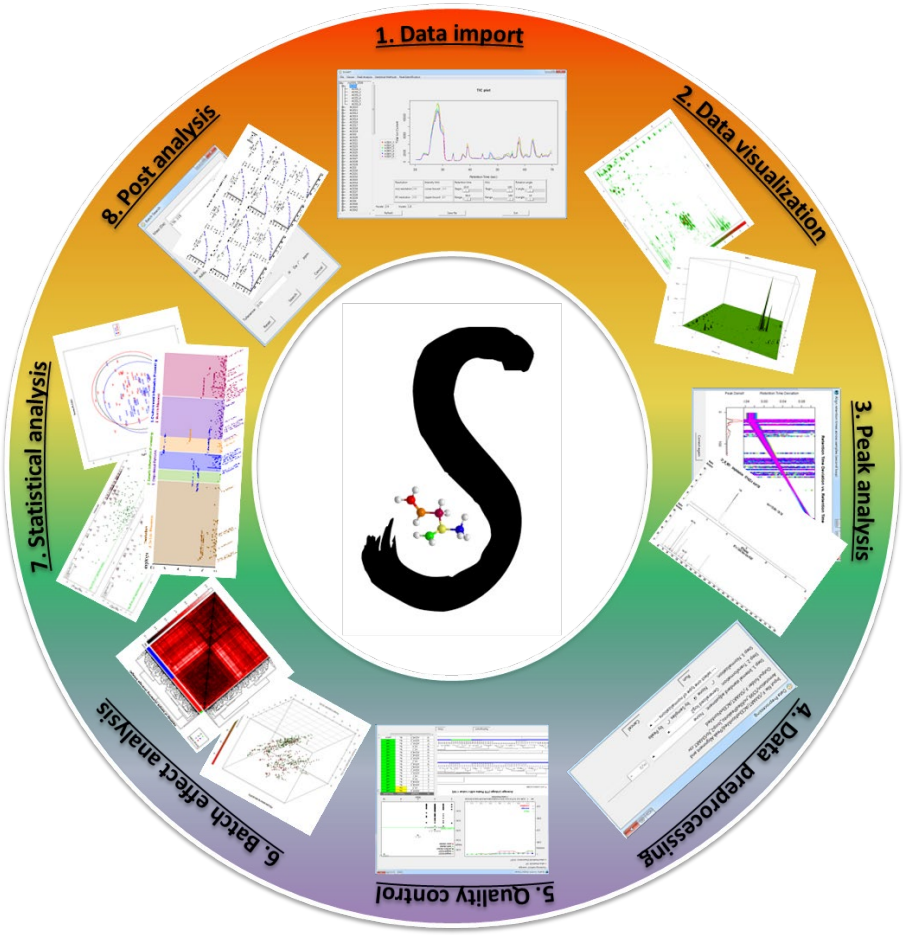


Table of Contents

1. SMART 2.0: New features and enhancements

- **Expanded Analysis Modules:** **SMART 2.0** extends the analysis modules, including peak analysis for targeted data, advanced data preprocessing normalization methods such as Pareto scaling (PS) and rank-based inverse normal transformation (INT), quality control signal-to-noise ratio (S/N) calculation, partial least squares or partial least squares discriminant analysis (PLS/PLS-DA), and integrative omics pathway analysis (IOPA) for statistical analysis, as well as peak identification and concentration calibration in post-analysis. The modules highlighted in yellow indicate the updated parts in **Figure 1.1**.
- **Enhanced Analytical Capabilities:** **SMART 2.0** introduces more analytical methods, allowing users to explore metabolomics data more comprehensively and providing deeper statistical insights.
- **Simplified Workflow:** **SMART 2.0** also improves the user interface to better guide users through the analytical workflow, making it simpler and more intuitive.
- **Document Updates:** The documentation for **SMART 2.0** has been updated to reflect changes in new features, methods, and interfaces, and it provides new user guides. Newly updated features will be highlighted in orange.

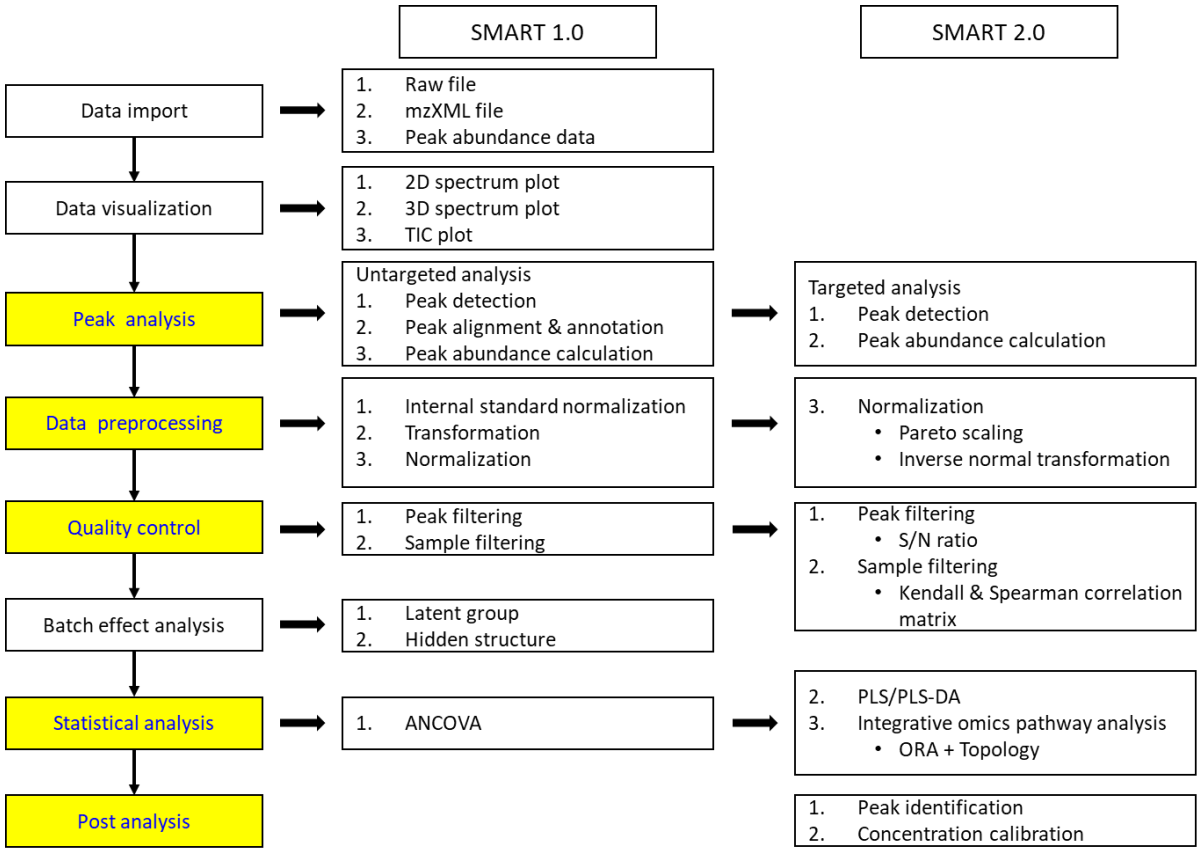


Figure 1.1. Overview of **SMART 1.0** and **2.0**.

2. SMART License

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.

Yu-Jen Liang, Chih-Ting Yang, Chia-Wei Chen, Yin-Chun Lin, Shu-Yao Lin, Wen-Harn Pan, Yi Sheng Wang and Hsin-Chou Yang (2024). SMART 2.0: Statistical Metabolomics Analysis – An R Tool 2.0. (Submitted)

3. 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 analysis for both untargeted data and target data, conducting quality control for samples and peaks, exploring batch effects (e.g., known experimental conditions, unknown latent groups (LGs), or hidden substructures), performing statistical analysis (ANCOVA, PLS/PLS-DA) and integrative omics pathway analysis (IOPA), and accomplish post analysis including peak identification and concentration calibration (**Figure 3.1**).

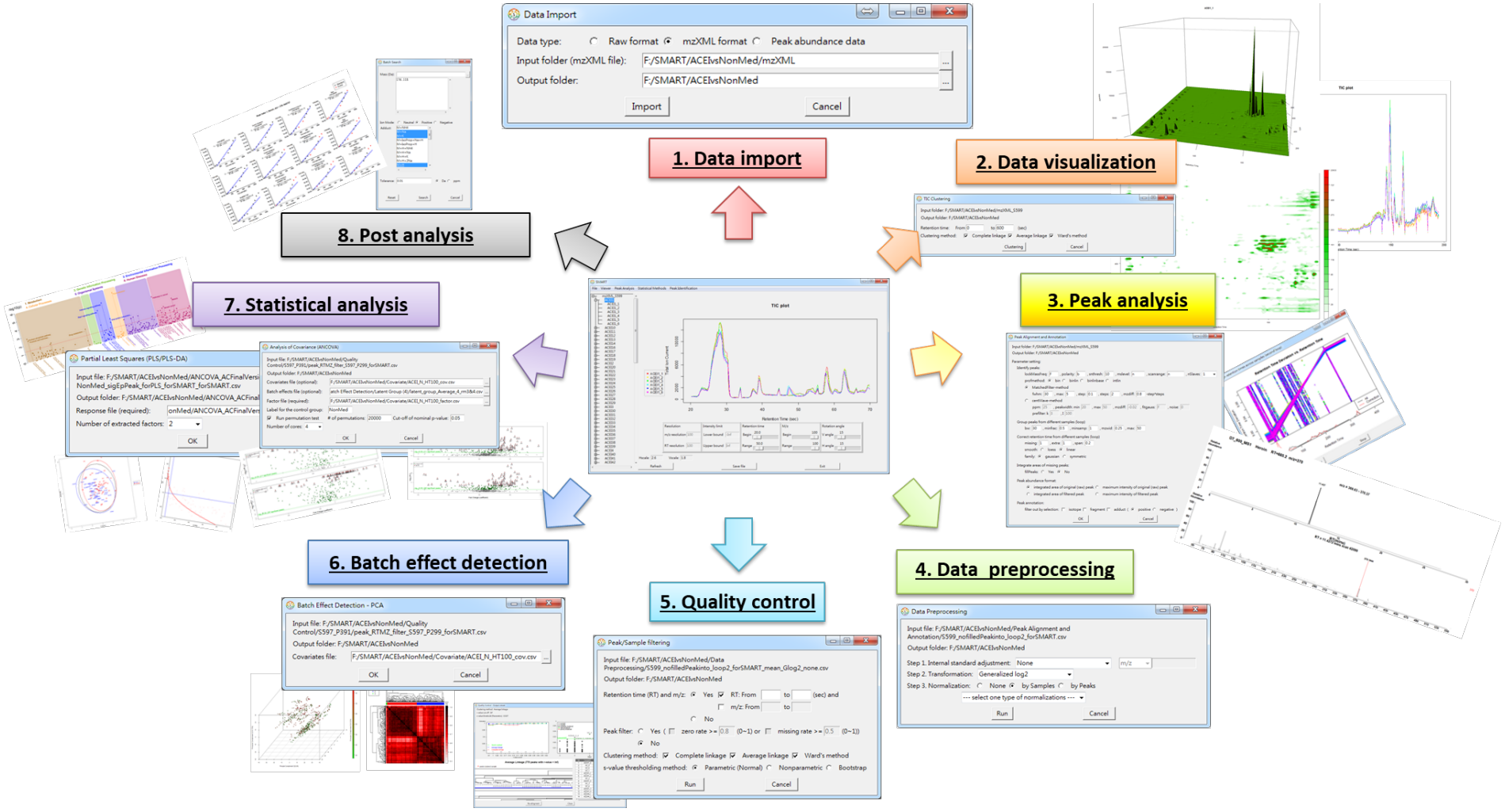


Figure 3.1. Overview of **SMART**.

4. Software Download and Installation

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

4.1 SMART

- **SMART 1.0~1.2:**

We provide the executive 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).

- **SMART 2.0:**

We provide **SMART 2.0** R source codes. It can perform on Windows operating systems (Windows 10 or later) and Mac operating systems (based on R language).

DOWNLOAD:

The programs, user guide, and examples can be downloaded from the SMART website at http://www.stat.sinica.edu.tw/hsinchou/metabolomics/SMART.htm or from GitHub at https://github.com/YuJenL/SMART.

4.2 R program

- **SMART 1.0~1.2:**

Users can download the R program (R3.3.0) from the official R Project website: http://www.r-project.org/. On the homepage, click “CRAN” (Comprehensive R Archive Network) and select a suitable mirror site to download R.

For Windows Users:

1. Select “Windows” as the platform.
2. Click on “base” and choose “Download R 3.3.0 for Windows”.
3. Run the downloaded file “R-3.3.0-win.exe” to install R in “C:\Program Files\R\R-3.3.0”.
4. After installation, double-click the “R i386 3.3.0” icon for 32-bit or “R x64 3.3.0” icon for 64-bit systems to start R. The “RGui” window with “R Console” will open for analysis.

To update packages, go to the “Packages” menu in the toolbar, select “Update packages,” choose a mirror site, and click “OK” in the “CRAN mirror” window.

For Mac Users:

1. Click on “R-3.3.0.pkg” to download R for Mac.
2. Download “XQuartz.dmg” for XQuartz, then execute “XQuartz.dmg” to get “XQuartz.pkg”.
3. Run “R-3.3.0.pkg” to install R and “XQuartz.pkg” to initialize XQuartz.

- **SMART 2.0:**

SMART 2.0 development began with **R version 4.3**, so please use this version or later. You can install R or RStudio to run the software. All other details remain the same as mentioned above.

4.3 ActiveTcl

ActiveTcl is a commercial distribution of Tcl (Tool Command Language) by ActiveState, used for rapid prototyping, embedded systems, test automation, GUI programming, and more. Program ActiveTcl can be downloaded from the website of ActiveState (https://www.activestate.com/).

4.4 ProteoWizard

ProteoWizard is an open-source software suite for processing mass spectrometry data. Its key feature, msconvert, allows users to convert data between different formats, facilitating streamlined data handling and analysis in proteomics and related fields. ProteoWizard can be downloaded from the website at https://proteowizard.sourceforge.io/download.html. 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.

4.5 R packages The analysis provided by **SMART** requires several additional R packages. **SMART** will automatically check for and download any missing packages if they are not already installed. For **SMART** versions 1.0 to 1.2, R versions must be between R-3.1.0 and R-3.3.0, along with their corresponding package versions. For **SMART 2.0**, R version 4.2.1 or later is required, with the appropriate package versions. Initial startup of SMART may take longer as it installs these packages, so please be patient.

5. SMART Initialization

- **SMART 1.0~1.2:**

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 5.1**). **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.

Figure 5.1. Path setting of **SMART 1.2** in Window systems.

For Mac users, please execute **SMART** by the following four steps (take SMART 1.0 as an example):

Step 1: Save three R programs, "SMART_V1.0_Mac.r", "SMART_V1.0_Mac_Gui.r" and "SMART_V1.0_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_V1.0_Mac.r" to the window "R Console" to initialize **SMART**. Or you can type the command, **source('~/Documents/SMART/SMART_V1.0_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 5.2**).

There are four main functions **in SMART 1.0**: (1) File, (2) Viewer, (3) Peak Analysis, and (4) Statistical Methods.

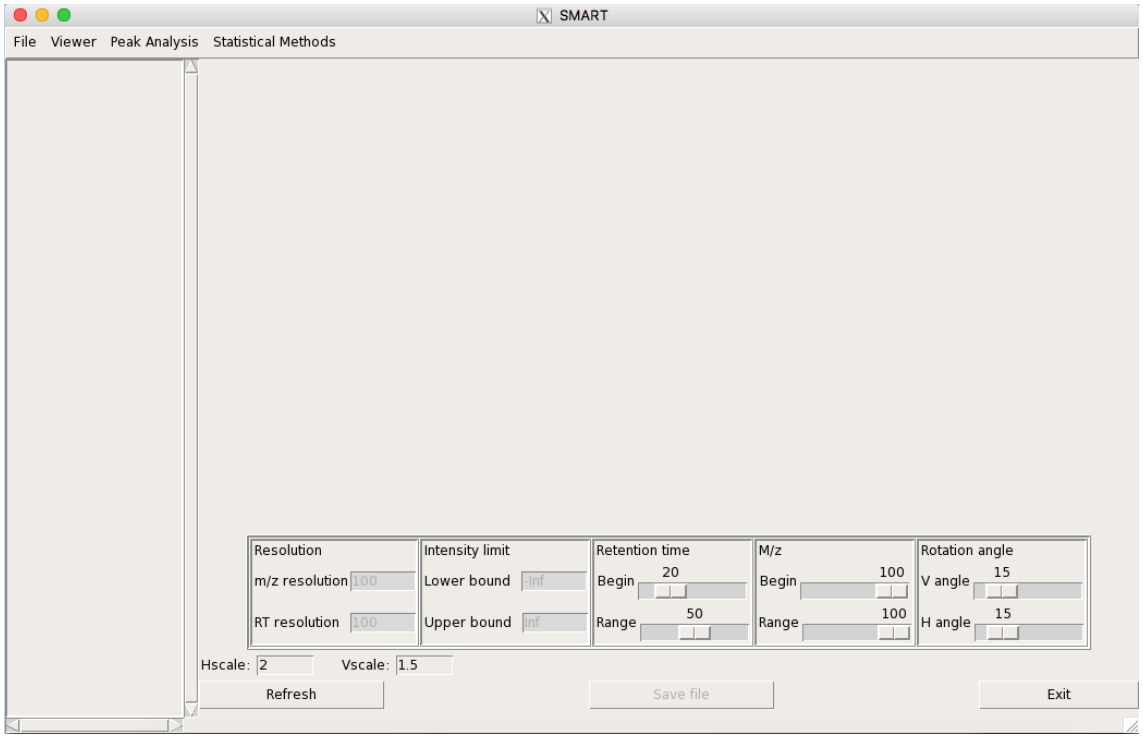


Figure 5.2. Initial interface of **SMART** in Mac systems

- **SMART 2.0:**

SMART 2.0 includes 8 main programs, an initialization program (SMART_main.R), 4 Rdata files, and an icon, as shown in **Figure 5.3**. Please place these files in the same directory, such as "G:/SMART2/Code/SMART2.0". All the programs are included in the compressed file SMART2.0.zip.

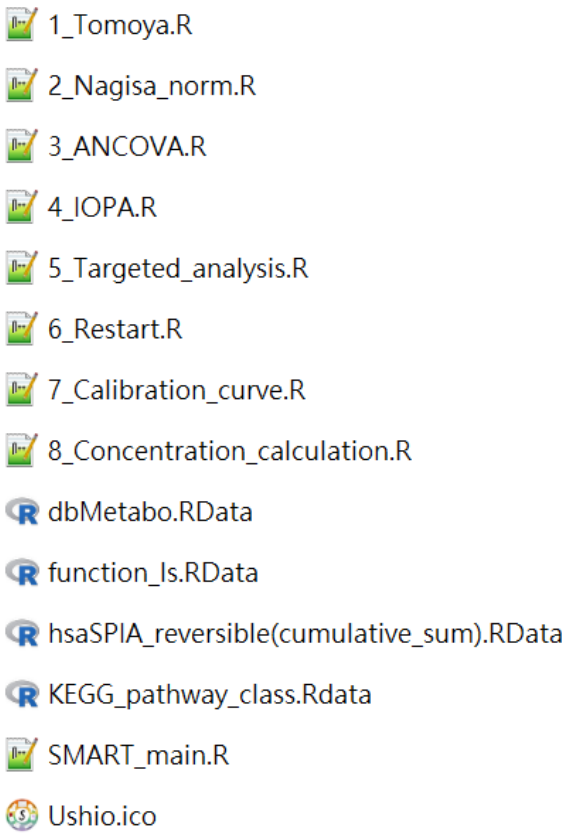


Figure 5.3. **SMART 2.0**'s core programs and components.

Steps:

1. Open R and Set Working Directory:

   Launch R and set the working directory within the file "SMART_main.R" using the setwd() command, e.g., setwd("G:/SMART2/Code/SMART2.0").

2. Run Initialization Program:

   Next, execute "SMART_main.R" in R to initialize SMART 2.0.

Once all packages are ready, the interface of **SMART 2.0** will jump up (see **Figure 5.4**). There are five main functions **in SMART 2.0**: (1) File, (2) Viewer, (3) Peak Analysis, (4) Statistical Analysis, and (5) Post Analysis. Descriptions about these functions will be introduced in next Section.

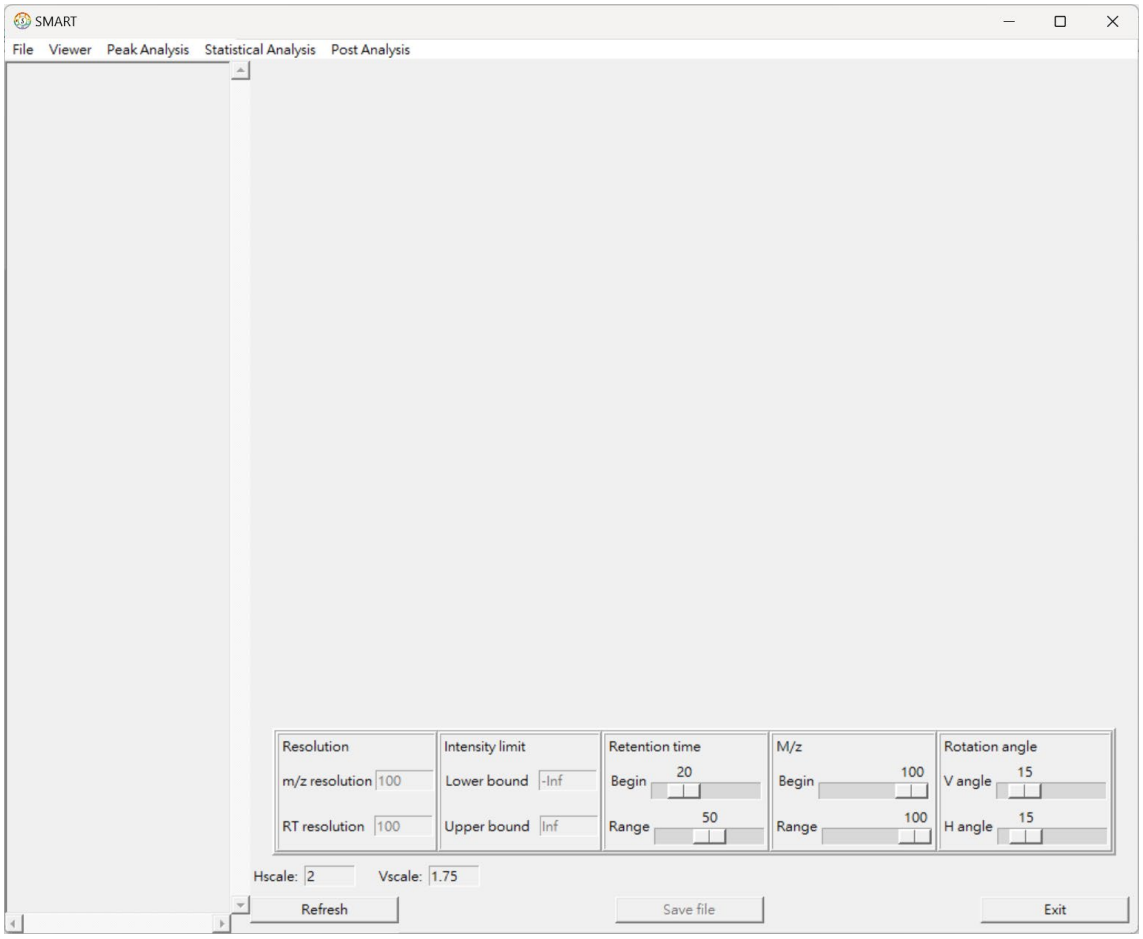
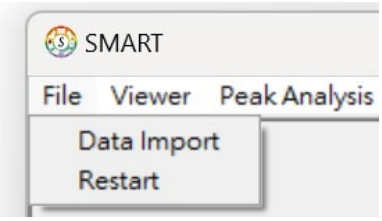


Figure 5.4. Initial interface of **SMART 2.0**.

Please be patient when using **SMART 2.0** for the first time, as the initial setup may take a few minutes. This is because SMART needs to install the required R packages to function correctly. The installation process may vary in duration depending on your system configuration and internet speed.

If a problem occurs during analysis, you can directly re-enable SMART: *Click "File"* ➔ *Click "Restart"*.



6. 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.

6.1 Data import

SMART supports multiple input file formats (see **Figure 6.1.1**). 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.

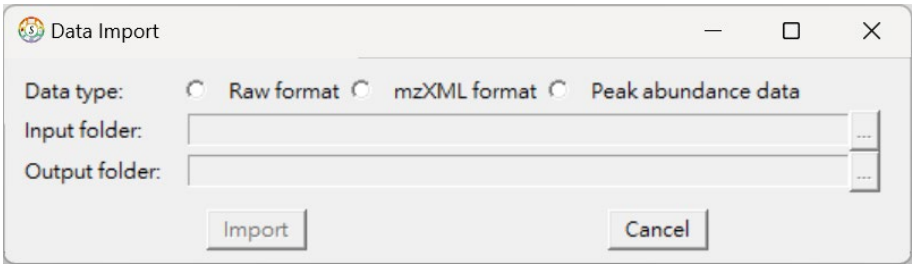


Figure 6.1.1. Interface of data import.

- 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.

  SMART uses ProteoWizard's *msconvert.exe* to convert raw data. The installation path for ProteoWizard must be specified, such as *C:/Users/yujen/AppData/Local/Apps*, which is the default path for Windows 11, where '*yujen*' is the username (see **Figure 6.1.2**).

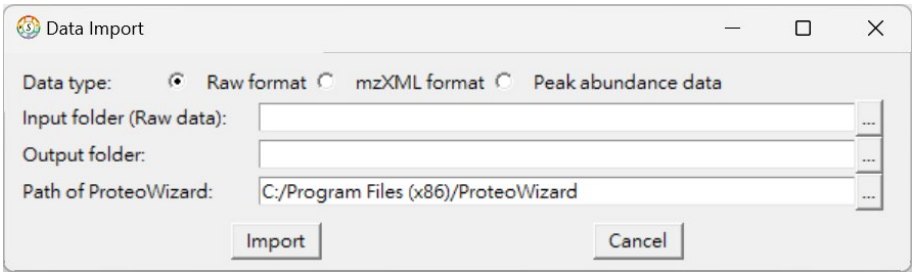


Figure 6.1.2. Interface of raw spectrum data import.

- 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 6.1.1**). 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 6.1.1. 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

Click "Continue" to proceed with data field settings (see **Figure 6.1.3**) and input the corresponding column information from the data file.

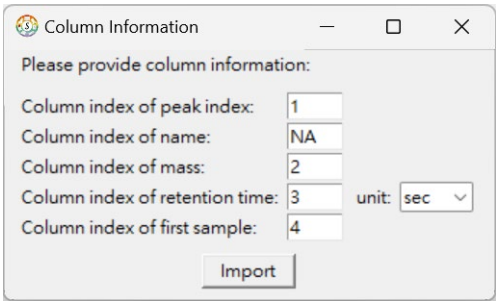


Figure 6.1.3. Interface of peak abundance data import (Column Information).

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 6.1.4**).

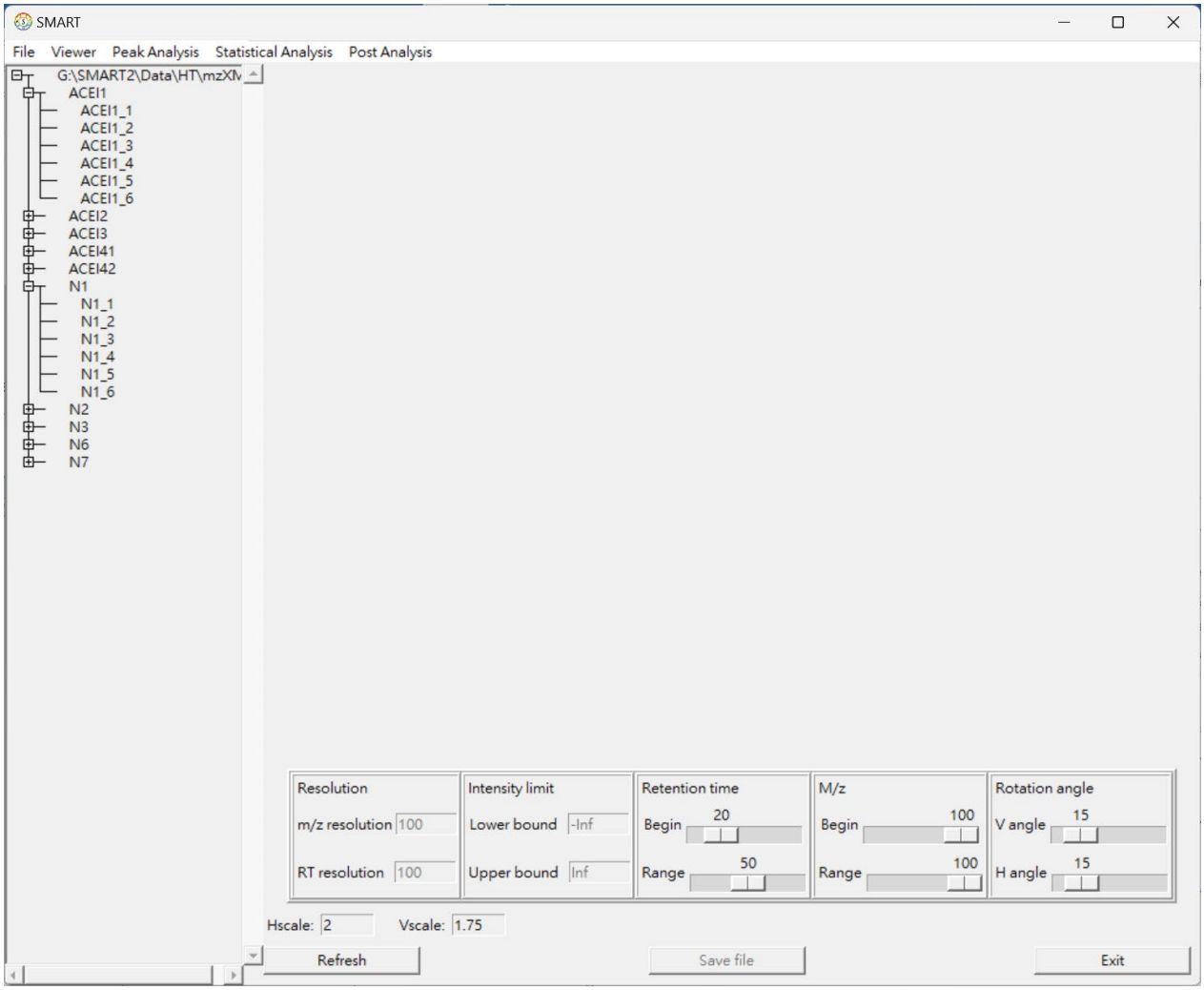


Figure 6.1.4. The window interface displayed after data import.

6.2 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 6.2.1**): *Select one specific sample in sample tree diagram in the left-hand panel ➔ Click "Viewer" ➔ Choose "2D Plot"*.

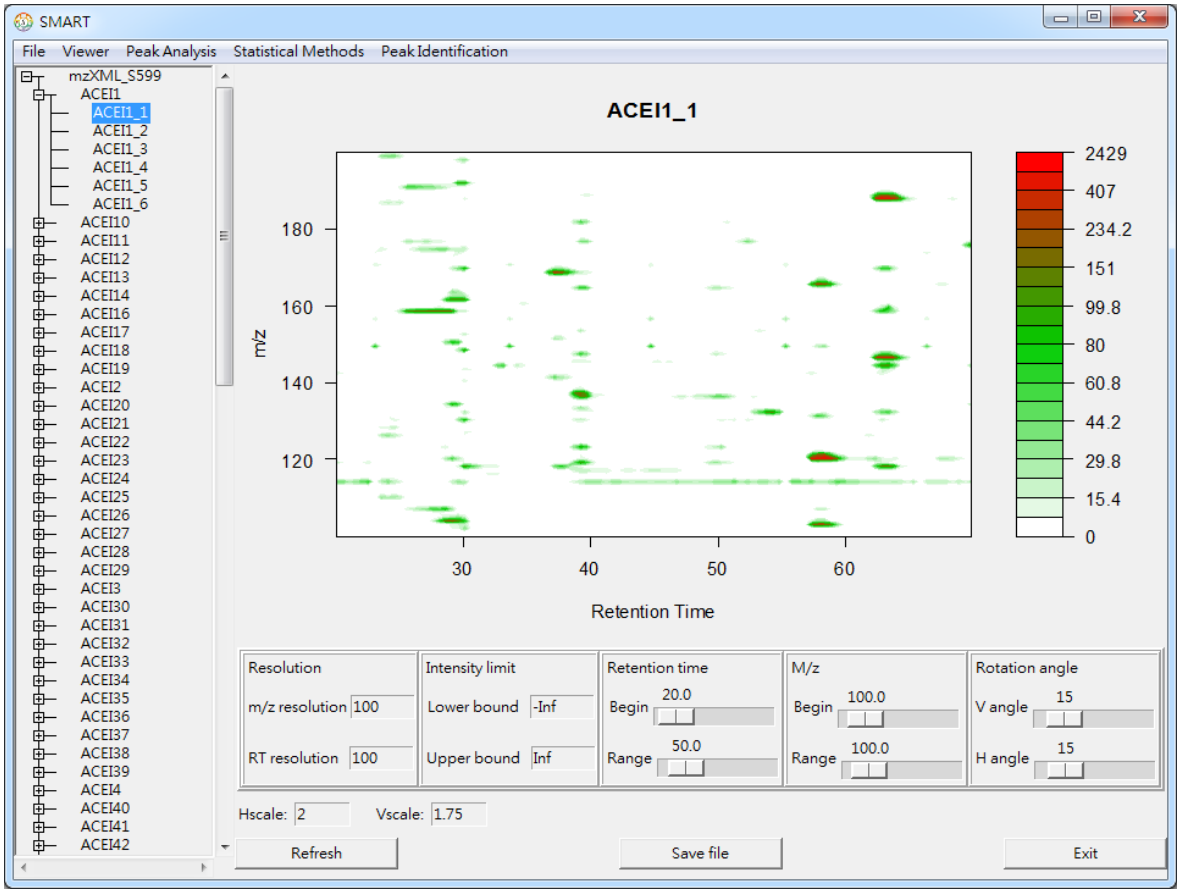


Figure 6.2.1. 2D spectrum plot.

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 6.2.2**): *Select one specific sample in sample tree diagram in the left-hand panel ➔ Click "Viewer" ➔ Choose "3D Plot"*.

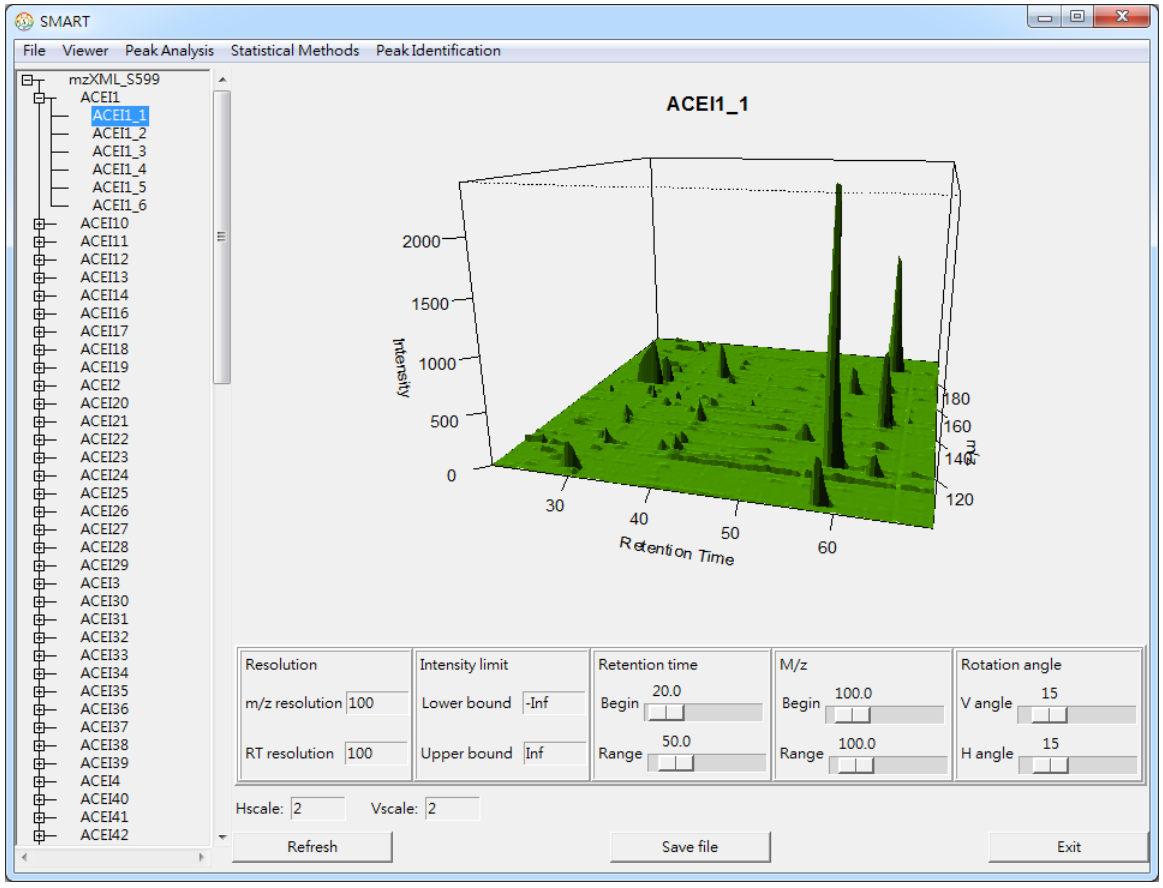


Figure 6.2.2. 3D spectrum plot.

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 6.2.3**): *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 6.2.4**). Under a fixed RT, the spectrum plot displays peak intensities for peaks across all m/z in the data.

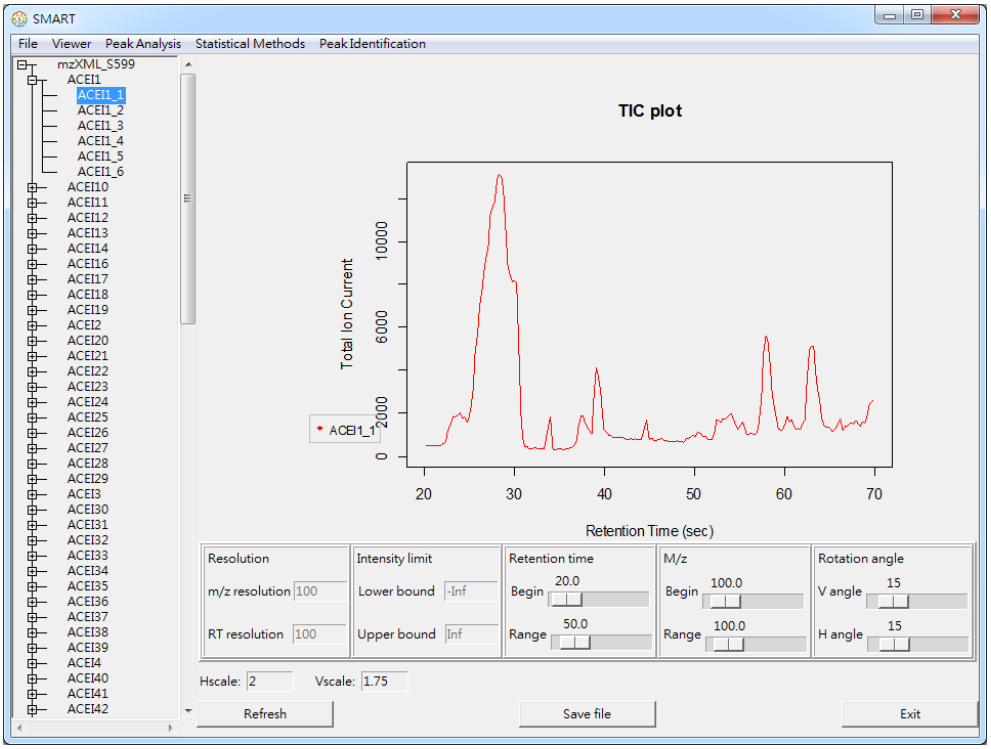


Figure 6.2.3. TIC plot.

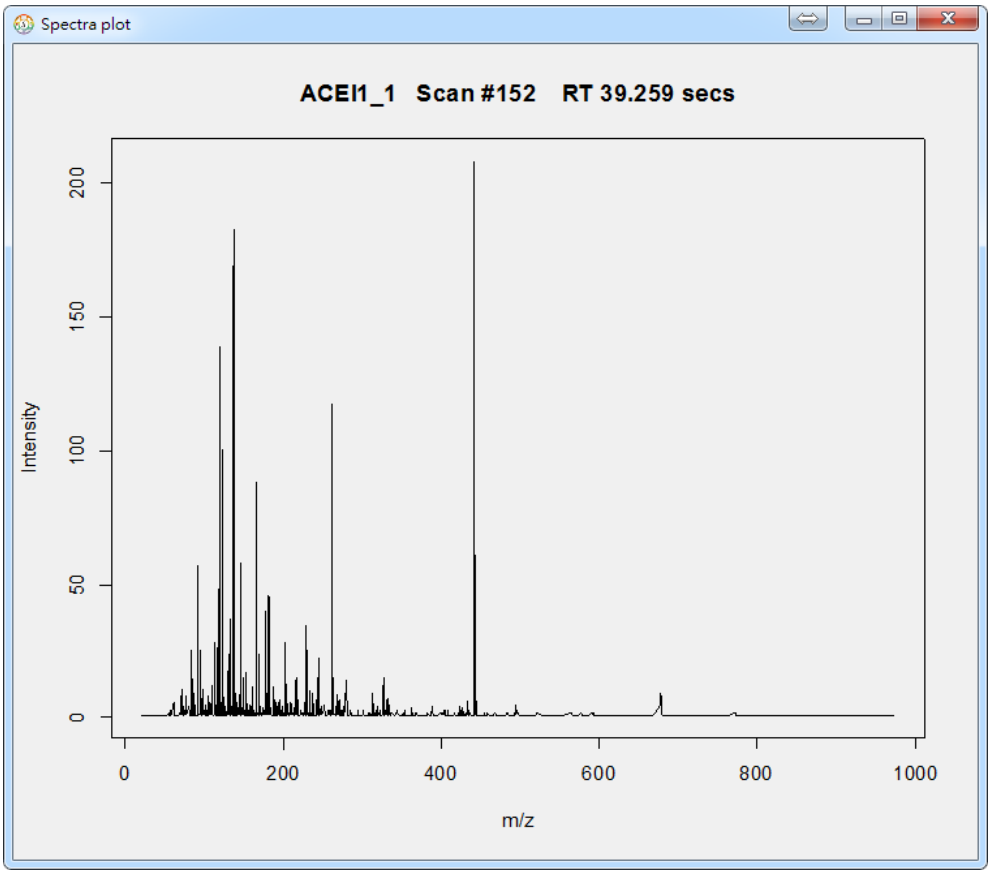


Figure 6.2.4. Spectrum plot at a fixed RT scan.

Fourth, users follow the procedure to visually represent TIC for multiple replicate samples and subjects (See **Figure 6.2.5**): *Select multiple replicate samples and subjects ➔ Click "Viewer" ➔ Choose "TIC plot".*

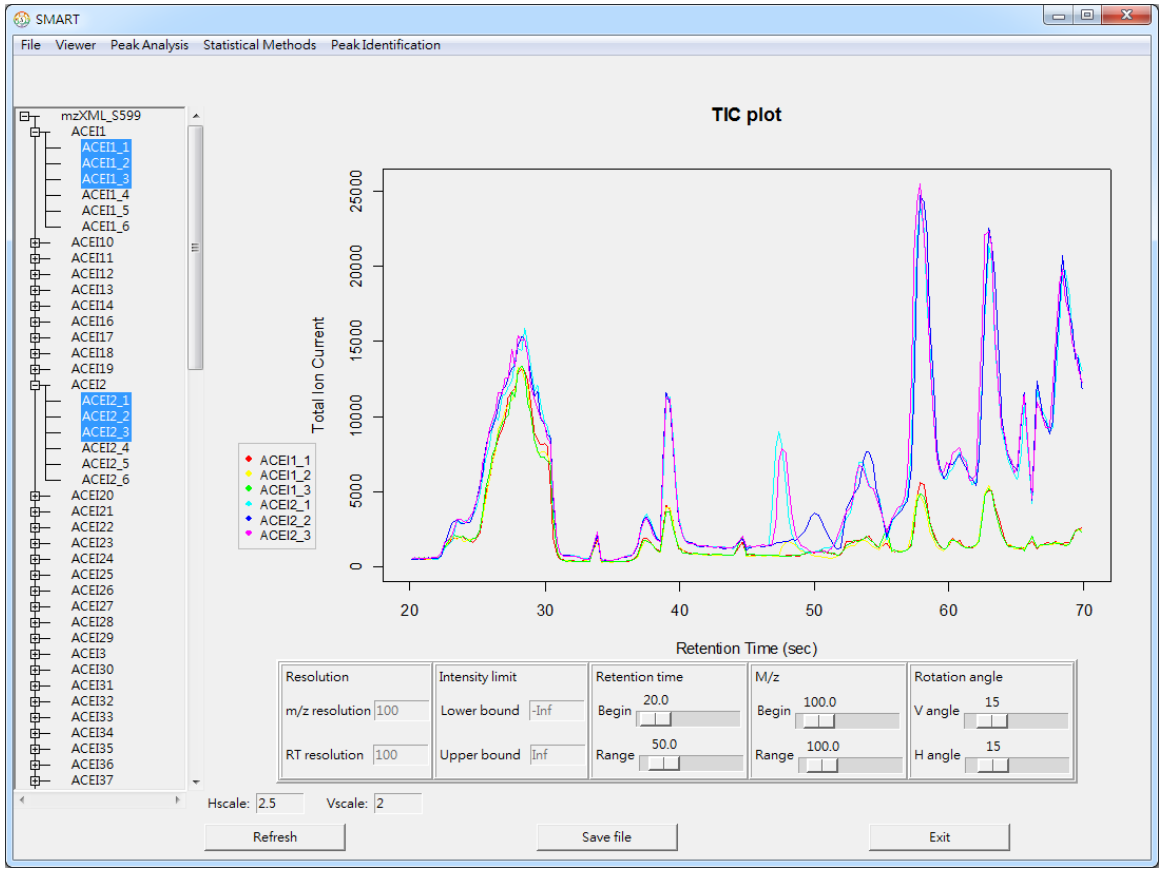


Figure 6.2.5. Cross-replicate-sample and cross-subject TIC overlay plots.

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 6.2.6**): *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 6.2.7**).

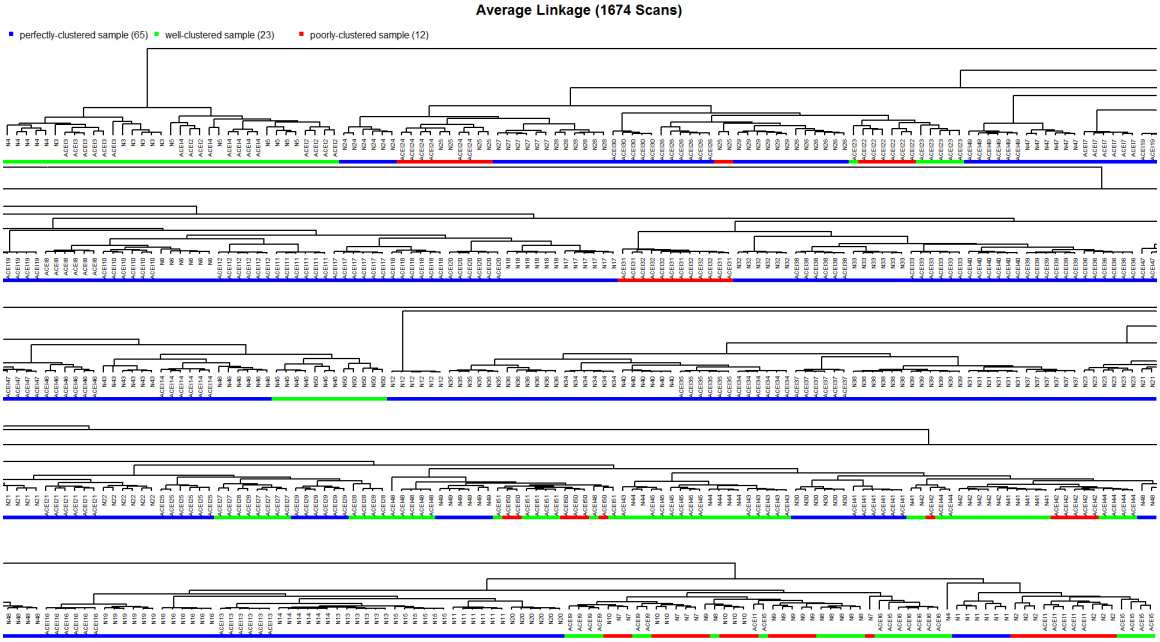


Figure 6.2.6. TIC cluster tree diagram.

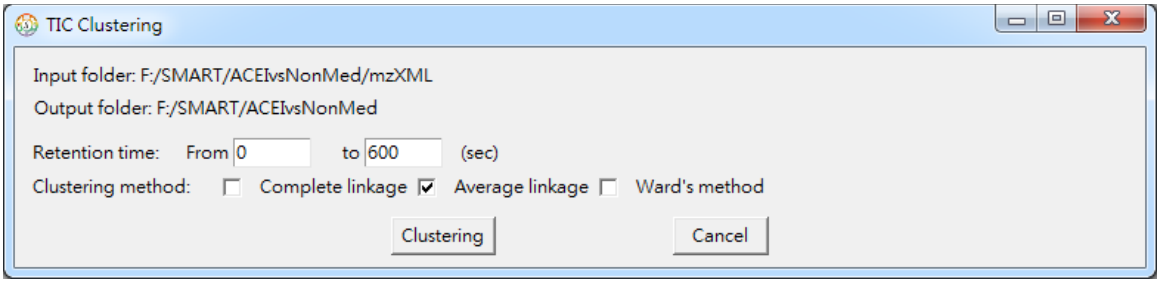


Figure 6.2.7. Interface of TIC clustering.

6.3 Peak analysis

6.3.1 Untargeted peak analysis

SMART carries out peak alignment (i.e., peak detection and RT alignment) by incorporating the matched filtration and centWave in XCMS [1] (see **Figure 6.3.1**) 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 Detection and Abundance Calculation" ➔ Choose "Untargeted Analysis"*. Then a window about a citation of XCMS and CAMERA will pop up.

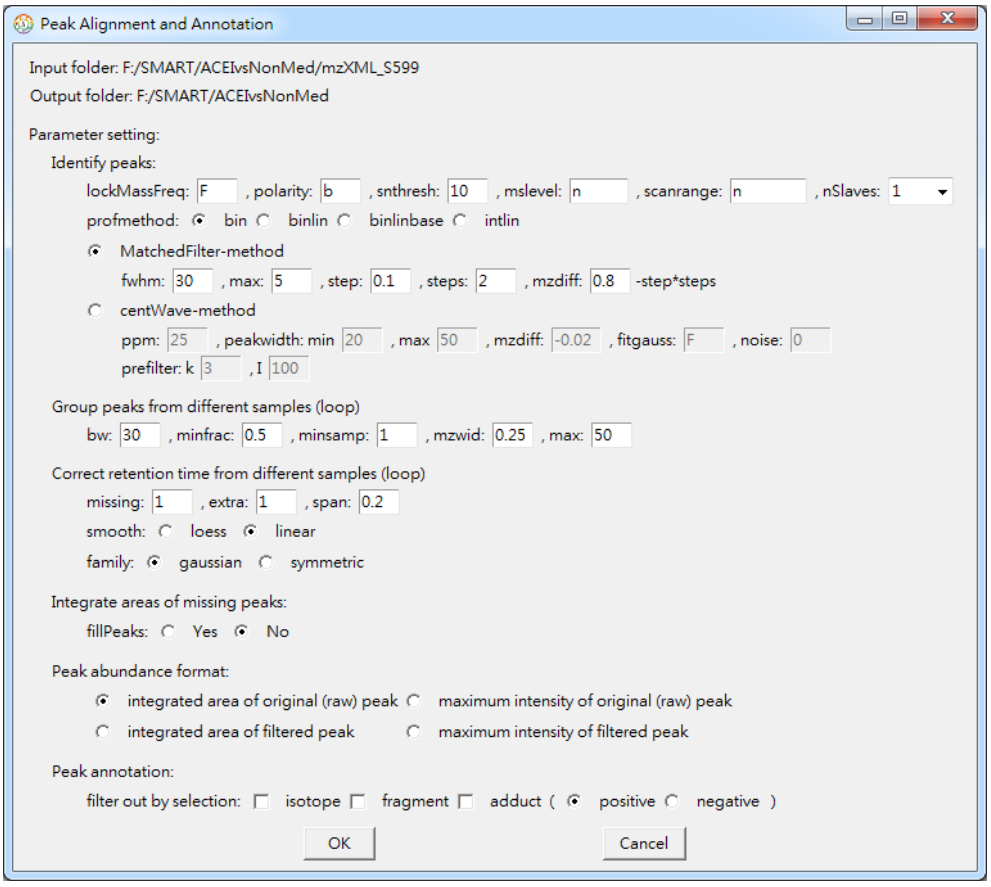


Figure 6.3.1. Interface and default settings of peak alignment and annotation.

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 6.3.1**). In addition, **SMART** exports the peak annotation, including the peak index, m/z, RT (seconds), isotope, adduct, and peak group (pcgroup) (see **Table 6.3.2** as an example).

Table 6.3.1. 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 6.3.2. 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-H20]+ 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 6.3.2** 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 6.3.2** 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 6.5**) because those peaks provide useful information for correlation estimation and quality control.

6.3.2 Targeted peak analysis

SMART supports targeted peak analysis, allowing for peak detection and abundance calculation for known compounds. Users should prepare mzXML files, which can be either MS1 or MS1+MS2, and organize them into separate folders. Additionally, unknown data can be stored in subfolders (see **Figure 6.3.2**). It is essential to provide the accurate m/z information of the known targeted compounds and store them as a CSV file (see **Tables 6.3.3** and **Table 6.3.4**).

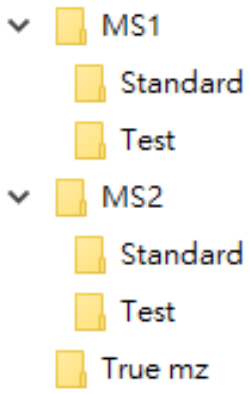


Figure 6.3.2. Directory setting for targeted peak data.

Table 6.3.3. True m/z information of MS1 data.

Name	mz	Ret_Time.sec	Tol_time.sec
Heroin	370	685.2	5
Morphine	286	147	5
Cocaine	304	702	5
Thebaine	312	678.6	5
delta-9 THC	315	1513.8	5
Amphetamine	136	422.4	20
MA	150	495.6	10
MDMA	194	539.4	5
Love Drug	180	506.4	10
Ketamine	238	600	5
FM2	314	1044	5
Nimetazepam	296	1041	5

Here, *Name* refers to the known compound name, *mz* is the parent m/z of the compound, *Ret_Time.sec* is the LC or GC retention time of the compound, and *Tol_time.sec* is the allowed time tolerance for peak detection, in seconds.

Table 6.3.4. True m/z information of MS2 data.

Name	Parent_ion	Daughter_ion	Tol_time.sec
Heroin	370	370;328;268; 211;193;58	10
Morphine	286	286; 268; 229; 201	10
Cocaine	304	304; 182; 150; 105; 82	10
Thebaine	312	312; 281; 266; 251; 221; 58	10
delta-9 THC	315	315; 259; 193; 135; 123; 107; 93	10
Amphetamine	136	136; 119; 91	20
MA	150	150; 119; 91	10
MDMA	194	194; 163; 135; 133; 105	10
Love Drug	180	180; 163; 135; 133; 105	10
Ketamine	238	238; 220; 179; 163; 152; 125	10
FM2	314	314; 200; 286; 268	15
Nimetazepam	296	296; 268; 250; 222; 193; 165	10

Here, *Name* refers to the known compound name, *Parent_ion* is the parent m/z of the compound, *Daughter_ion* is the fragment m/z for MS2 experiments, and *Tol_time.sec* is the allowed time tolerance for peak detection, in seconds.

To perform target peak analysis, users must first import mzXML data of MS1 (see **Section 6.1**). Next, users should follow the outlined procedure to conduct the target peak analysis: *Click "Peak Analysis" ➔ Choose "Peak Detection and Abundance Calculation" ➔ Choose "Targeted Analysis"*. The first step is to set the parameters. For this, please refer to the instructions provided in the *XCMS* package documentation. The interface is shown in **Figure 6.3.3**. (As an example, consider a concentration of 800 ppb.)

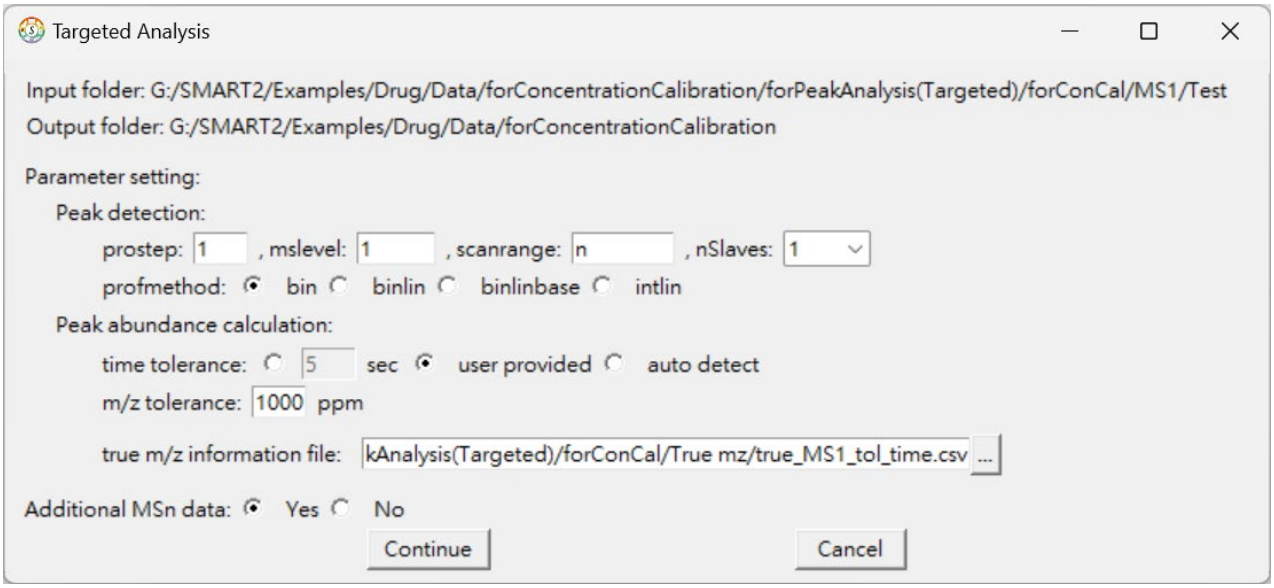


Figure 6.3.3. Interface of targeted peak analysis for MS1.

For peak detection, it is essential to set the mslevel parameter. Set mslevel=1 for MS1 data and mslevel=2 for MS2 data.

When calculating peak abundance, specify the retention time (RT) range (time tolerance) and m/z range (m/z tolerance) for accurate abundance calculations. It is crucial to provide a true m/z information file, which should contain the MS1 spectra of known compounds. This file should be in CSV format (see **Table 6.3.3**).

If MS2 data is available, *click "Continue"* to proceed with the MS2 parameter settings (see **Figure 6.3.4**). Set mslevel=2 and specify the m/z and RT tolerances. Lastly, provide the mass spectrometry information for the known compounds in MS2 format (see **Table 6.3.4**).

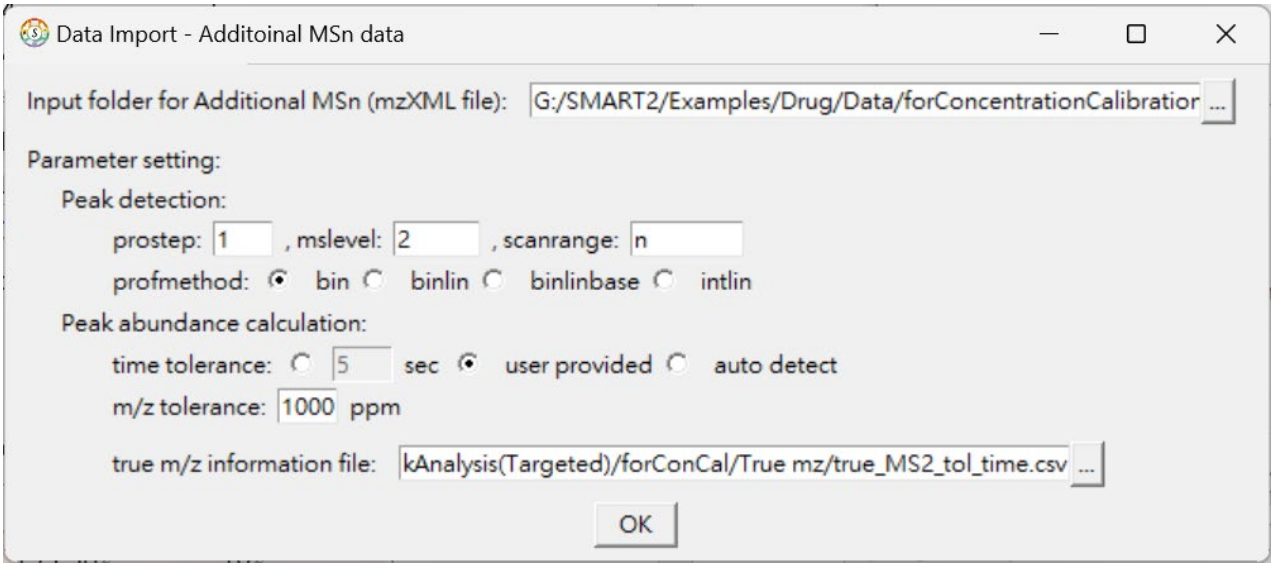


Figure 6.3.4. Interface of targeted peak analysis for MS2.

After the analysis is completed, data for all known compounds in each sample will be generated and presented in a peak abundance data (see **Table 6.1.1**). For both MS1 and MS2, the output will include m/z, retention time (RT), abundance, signal-to-noise (S/N) ratio, and other relevant information, along with all MS1 and MS2 spectra (see **Figure 6.3.5**, **Figures 6.3.6** and **Figure 6.3.7**). Additionally, the relative abundance of the parent ion corresponding to MS2 will be provided.

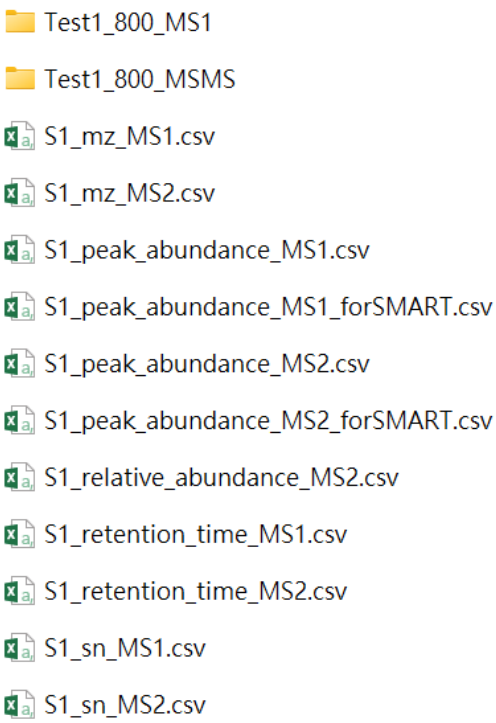


Figure 6.3.5. Outputs of targeted peak analysis.

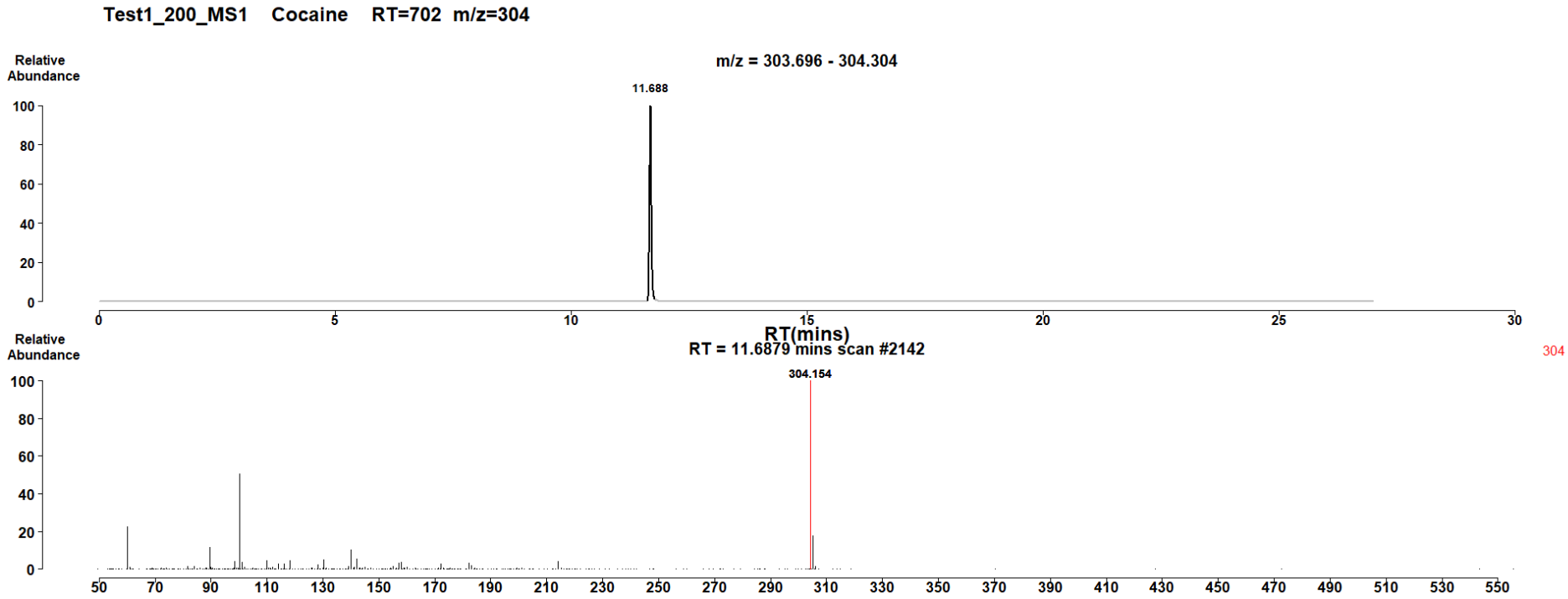


Figure 6.3.6. Cocaine's MS1 spectrum includes the total ion chromatogram (TIC) and the m/z distribution at the optimal retention time (RT).

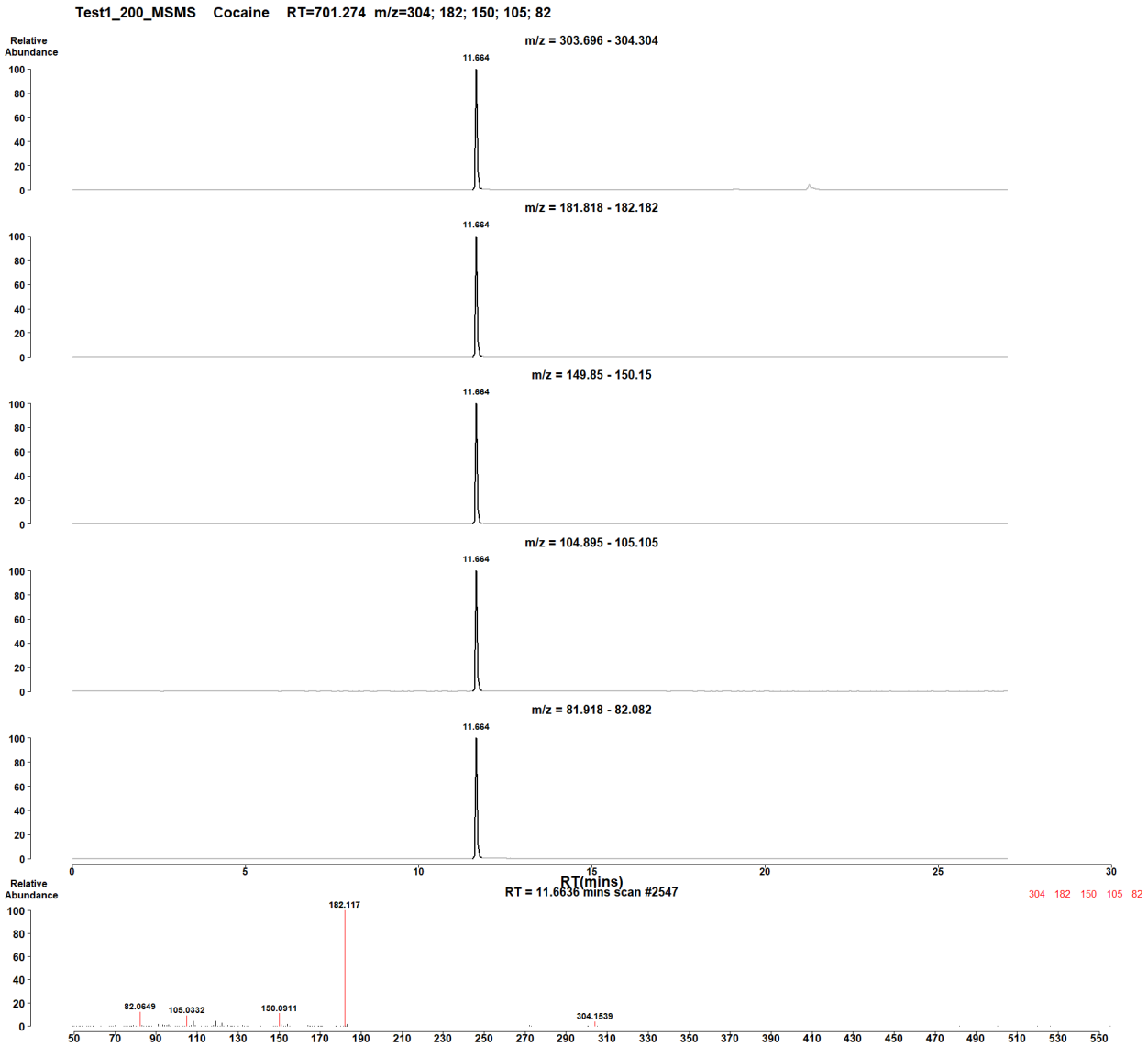


Figure 6.3.7. Cocaine's MS2 spectrum includes the quantity of each MS2 fragment at the optimal retention time (RT) and the relative abundance of all fragments at this RT, normalized to the fragment with the highest abundance.

Tips in the instructions

Targeted peak analysis uses MS1 and MS2 data from the same sample group, such as Standard or Test (see **Figure 6.3.2**). Both MS1 and MS2 data must be from the specific group being analyzed. Note that SMART only requires the sample file directory information.

6.4 Data preprocessing

SMART provides three procedures for data transformation and normalization (see **Figure 6.4.1**). 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 6.4.2**).

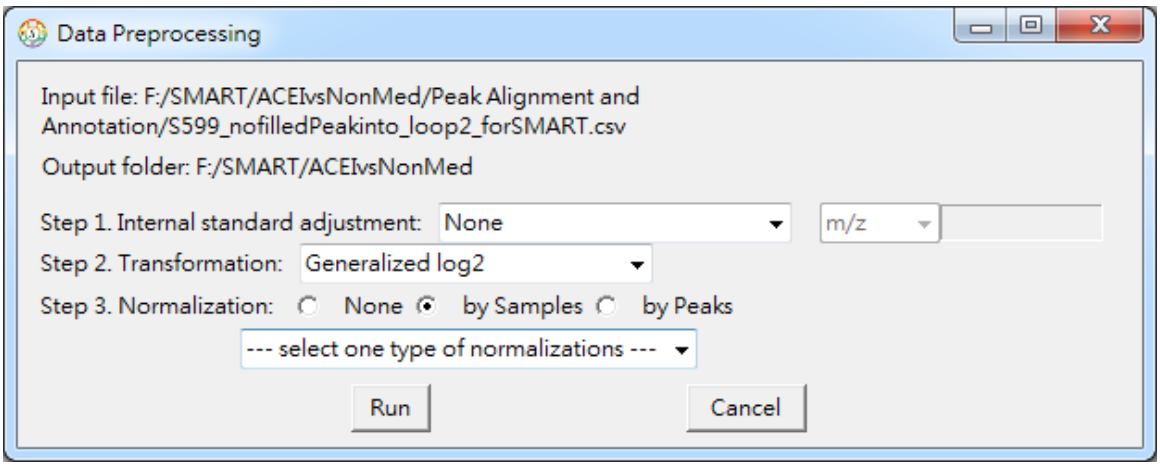


Figure 6.4.1. Interface of data preprocessing.

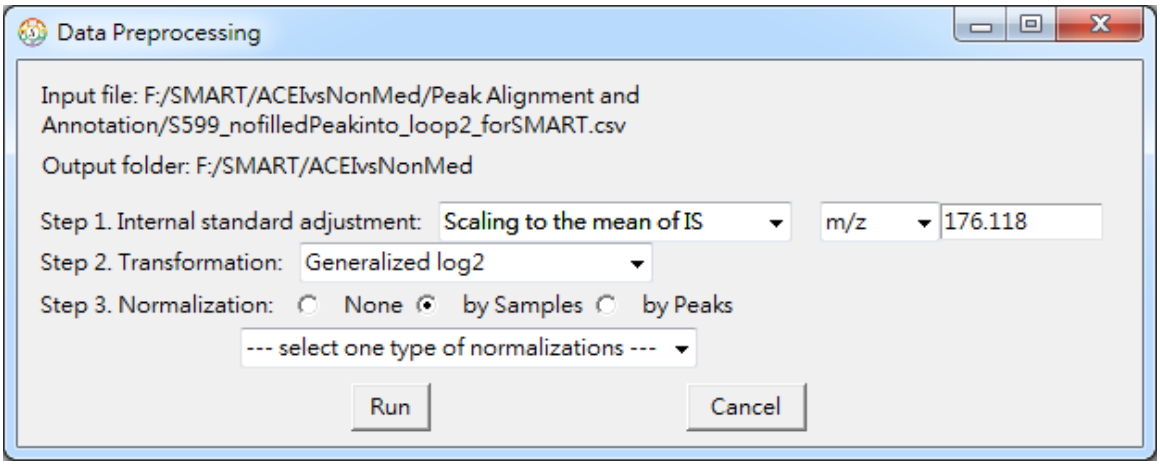


Figure 6.4.2. Internal standard adjustment.

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 6.4.3**). Note that if generalized log2 was performed then a zero value in the original data will be transformed as 2.

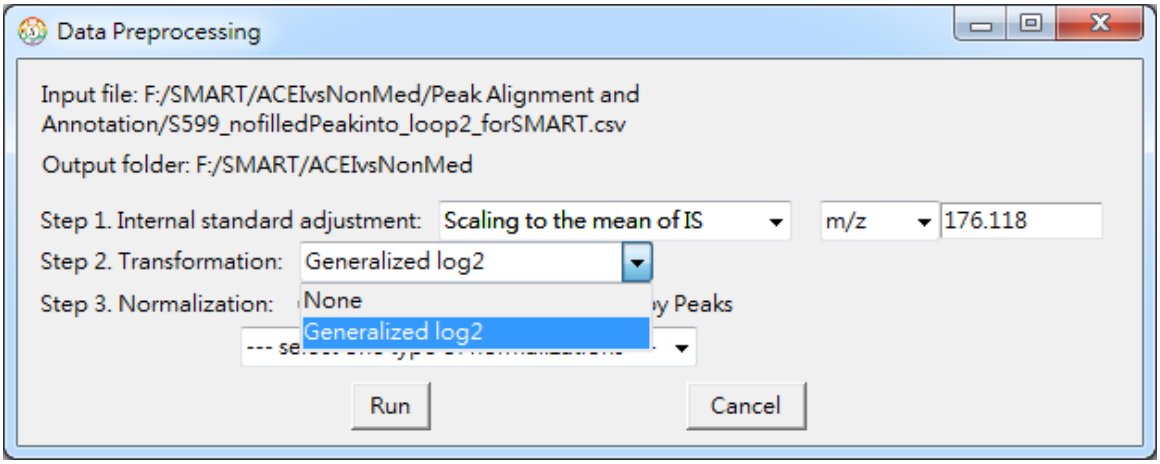


Figure 6.4.3. Transformation.

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”, “Pareto scaling”, and “Inverse normal transformation” (see **Figure 6.4.4**). The peak-based data normalization includes “None”, “Scale normalization – Mean”, “Scale normalization – Median”, “Standardization”, “Pareto scaling”, and “Inverse normal transformation” (see **Figure 6.4.5**). 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**.

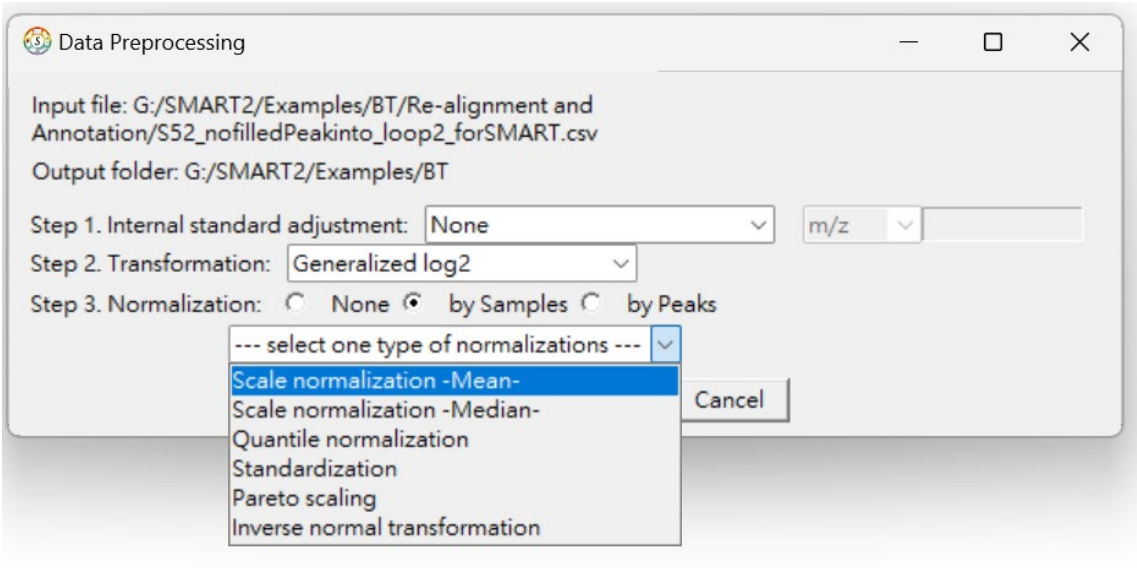


Figure 6.4.4. Sample-based data normalization.

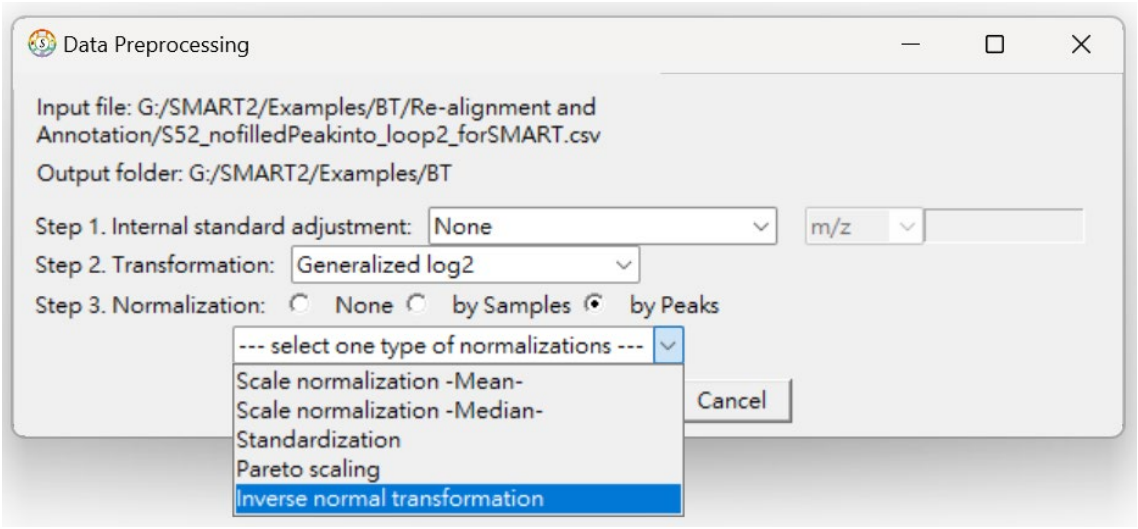


Figure 6.4.5. Peak-based data normalization.

6.5 Quality control

SMART provides quality control for peak filtering and sample filtering (see **Figure 6.5.1**). 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.

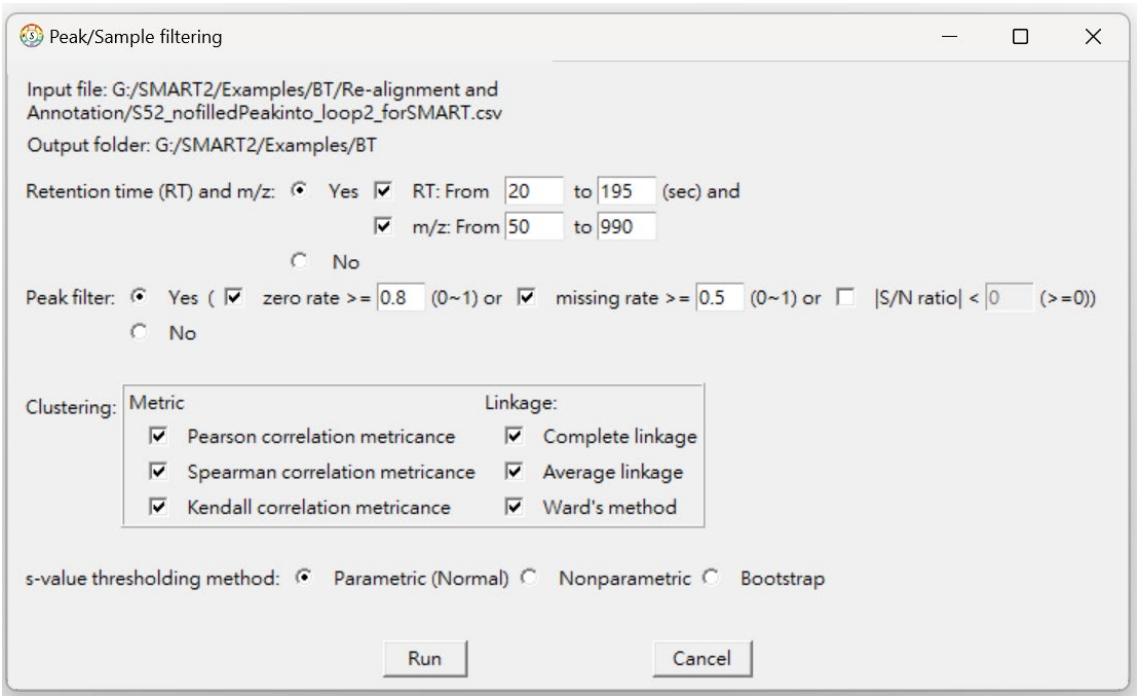


Figure 6.5.1. Peak and sample filtering.

The second step is to define the cutoffs for the zero rate (ranging from 0 to 1), the missing rate (ranging from 0 to 1), and the signal-to-noise (S/N) ratio for each peak. 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 6.5.2**). 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.

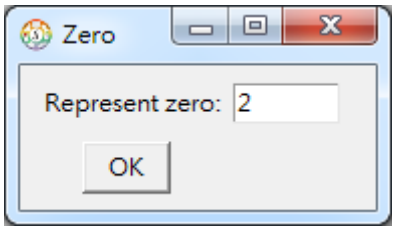


Figure 6.5.2. Interface of zero-value indication.

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 6.5.3**). Sample accuracy at the quantiles will be calculated and shown later. Users can also specify preferred (nonnegative) values as cutoffs of the *r-value*.

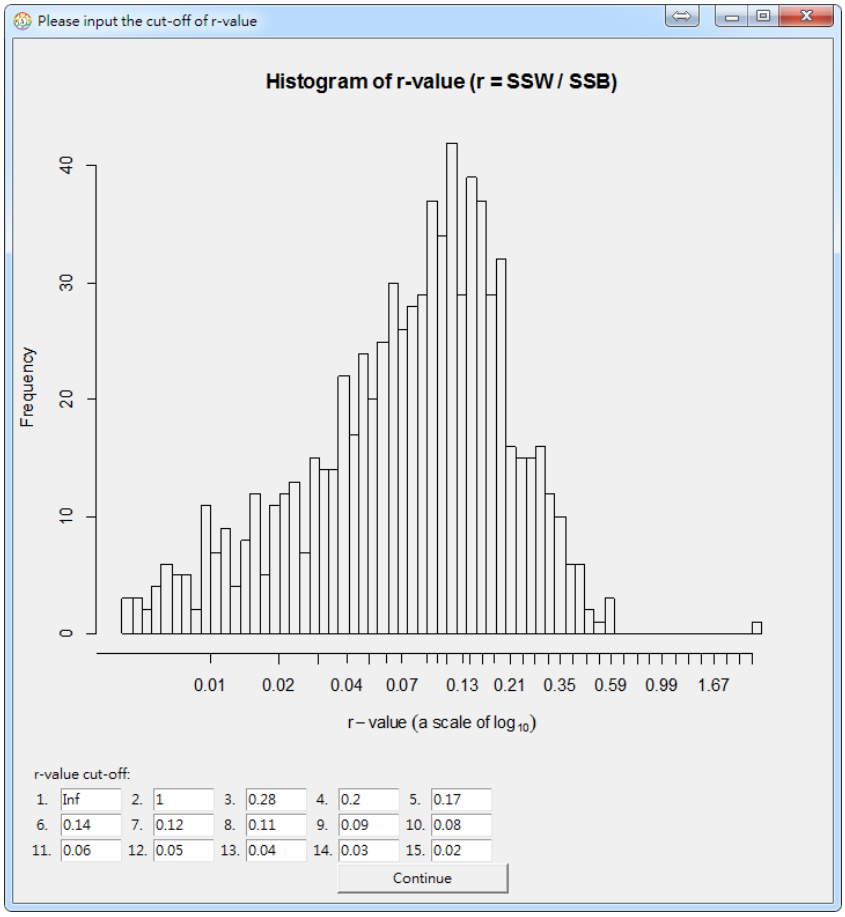


Figure 6.5.3. Distribution and quantiles of the *r-value*.

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 (see **Figure 6.5.4**). 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 (see **Figure 6.5.4**).

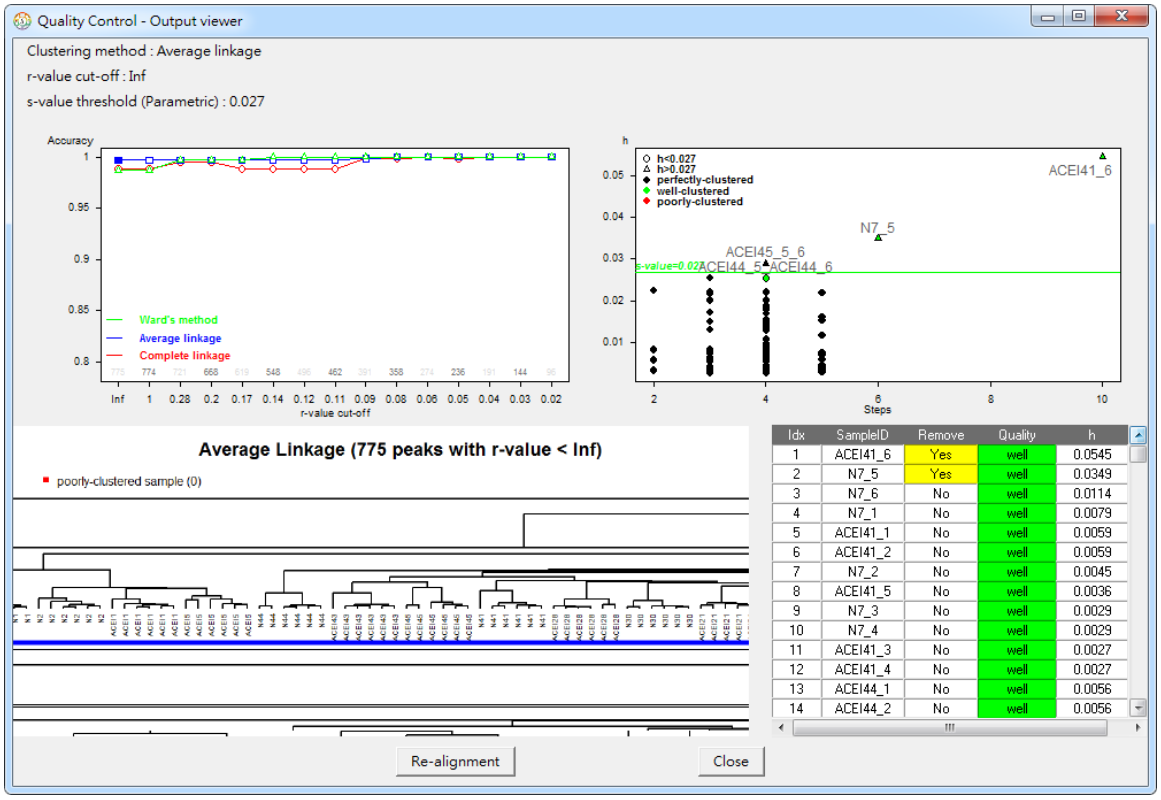


Figure 6.5.4 Output of quality control.

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*.

6.6 Re-alignment and annotation (only for untargeted peak analysis)

If users decide to do realignment (see **Figure 6.6.1**) 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 6.3**. 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).

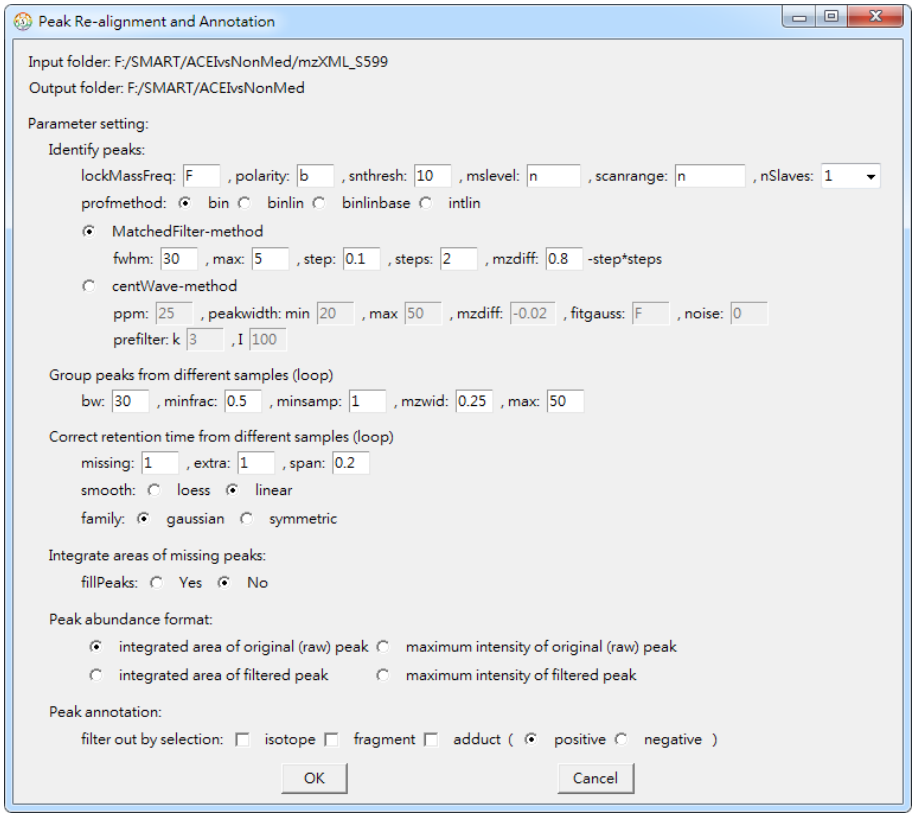


Figure 6.6.1. Interface of re-alignment and annotation

6.7 Batch effect analysis

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 Analysis” ➔ 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 6.7.1**) should be provided if some covariate(s) will be adjusted for (see **Figure 6.7.1** for a PCA and **Figure 6.7.2** 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 6.7.1**).

Table 6.7.1. 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

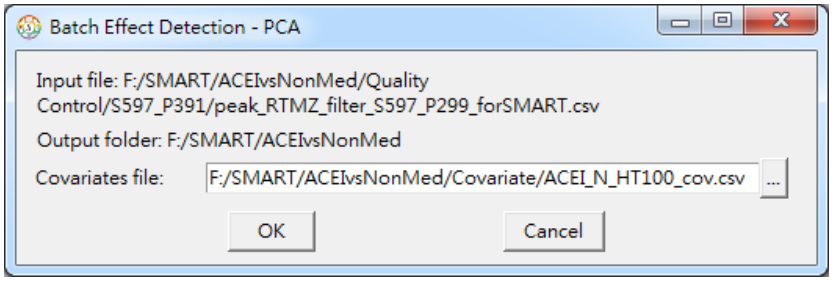


Figure 6.7.1. Interface of principal component analysis.

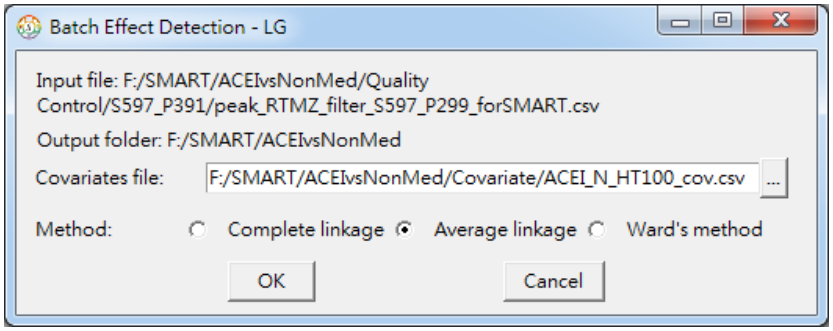


Figure 6.7.2. Interface of latent group analysis.

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) (see **Figure 6.7.3**). Finally, users can drag and drop the button "Number of PCs" in the bottom of **Figure 6.7.3** 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 (see **Figure 6.7.4**). 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.

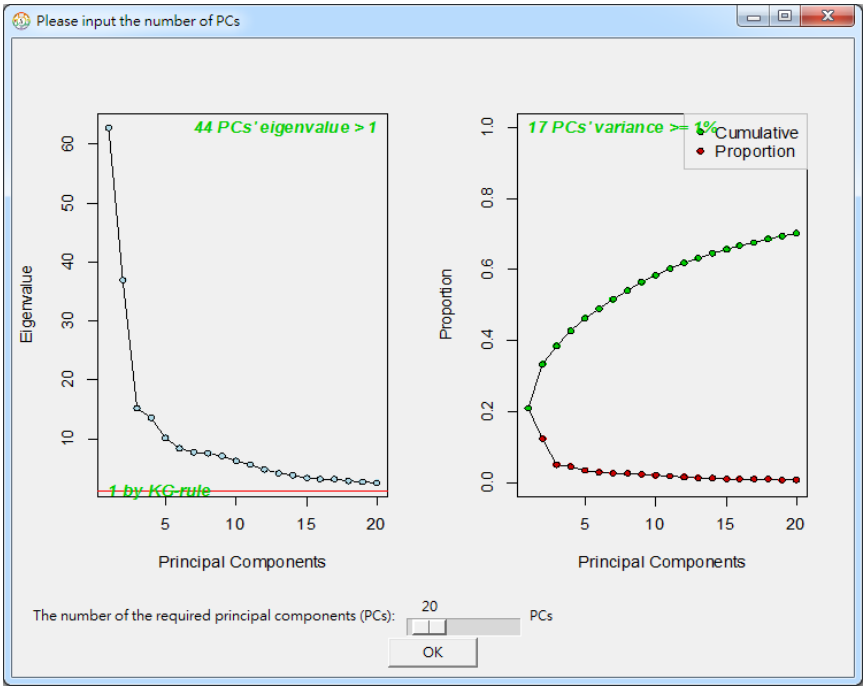


Figure 6.7.3. Interface of scree plot (left) and a variation-explained plot (right).

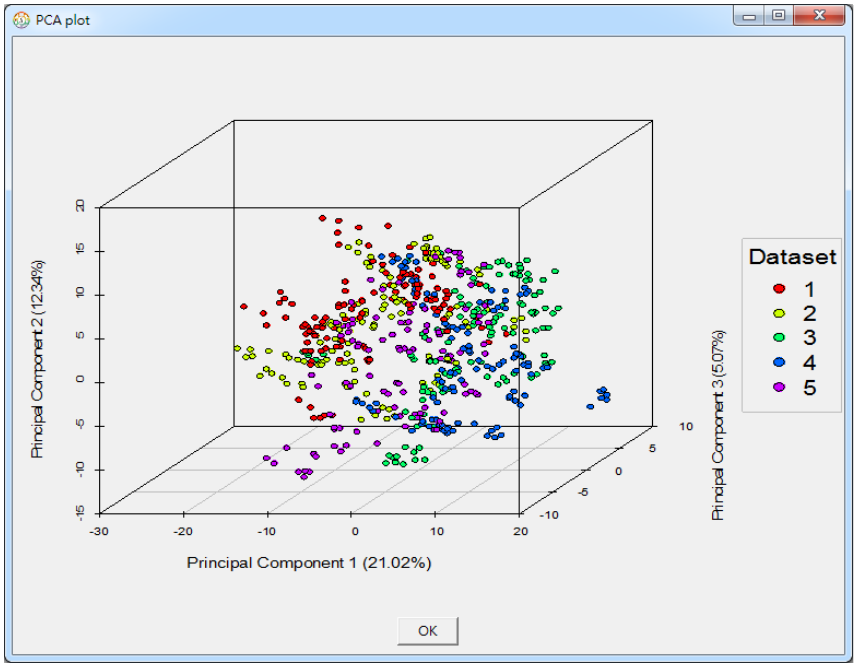


Figure 6.7.4. PCA plot.

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 (see **Figure 6.7.5**). 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.

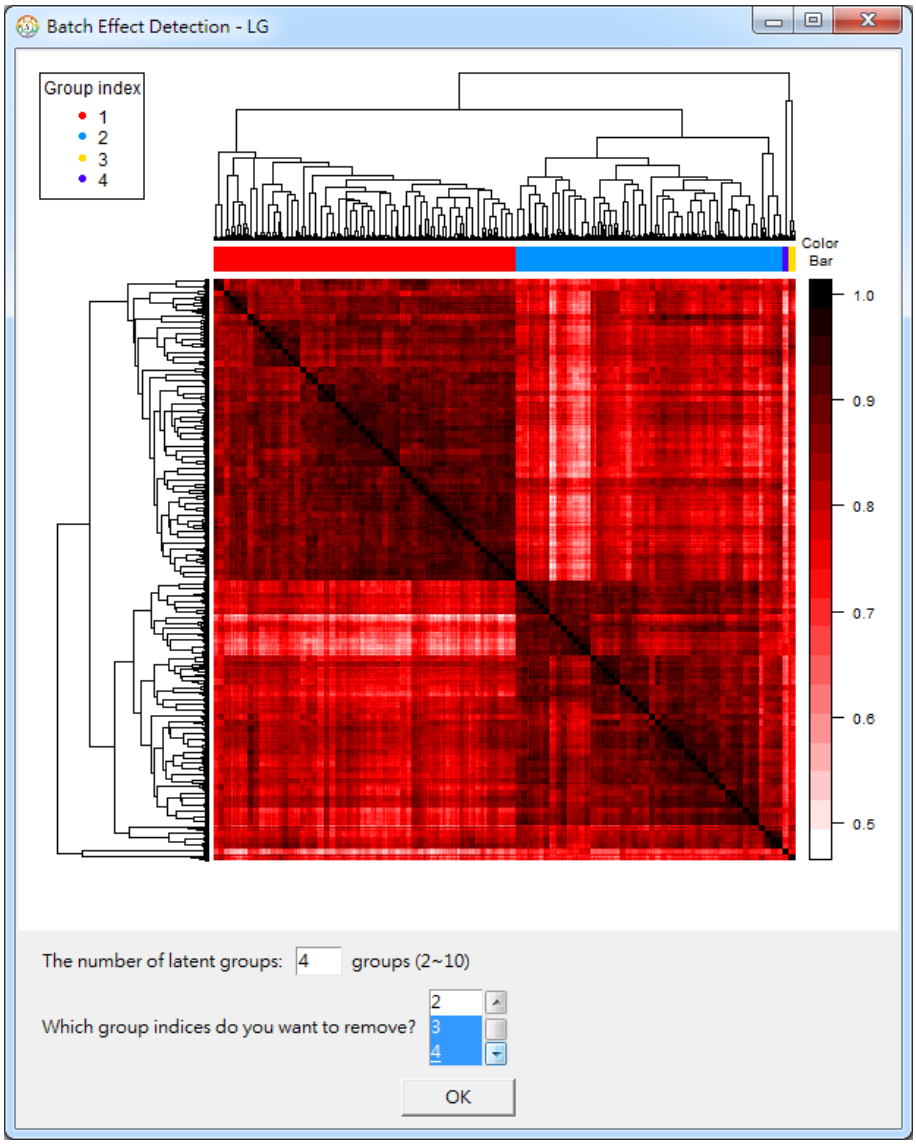


Figure 6.7.5. Heat map and cluster tree diagram.

6.8 Statistical analysis

6.8.1 ANCOVA

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 Analysis" ➔ 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 6.8.1**). Covariate data format has been introduced in **Table 6.7.1**. 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)).

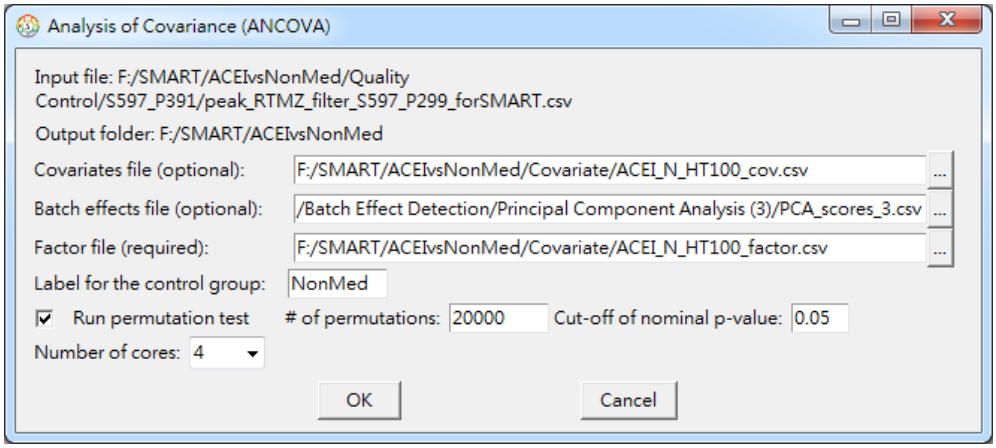


Figure 6.8.1. Interface of analysis of covariance (ANCOVA).

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 6.8.2**). 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 6.8.2**). Similar procedures are applied to comparisons between NonMed and other medications.

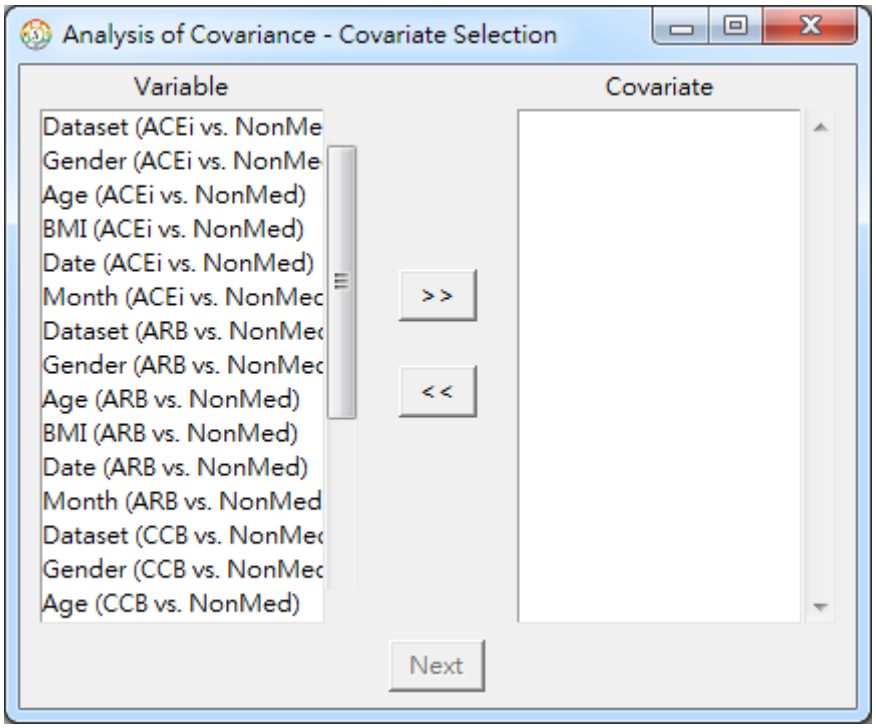


Figure 6.8.2. Interface of covariate selection in ANCOVA.

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 6.8.3**). 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.

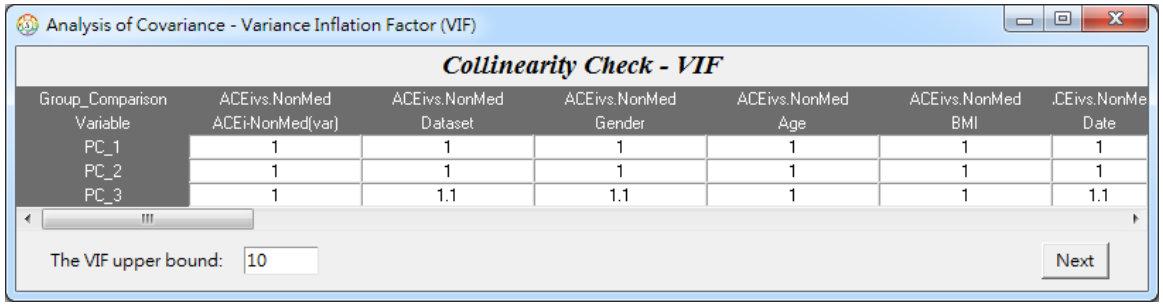


Figure 6.8.3. Interface of collinearity check.

Finally, users should specify the fitted ANCOVA models (see **Figure 6.8.4**). 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.

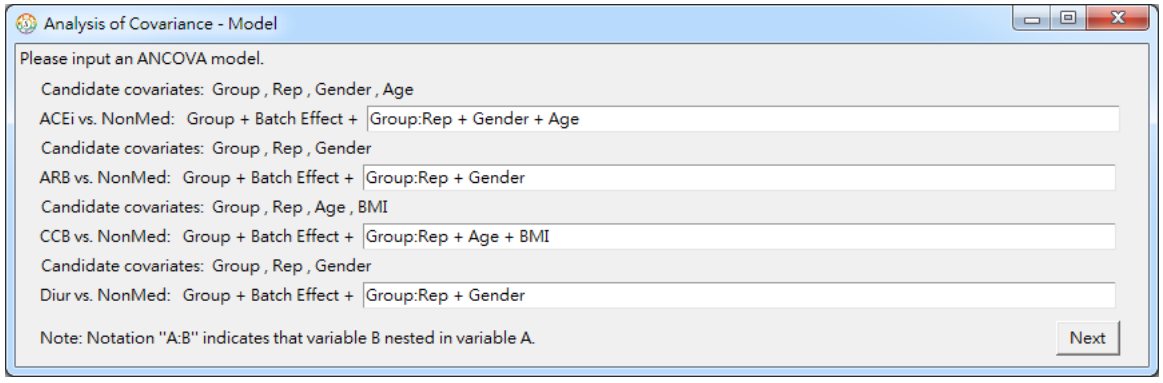


Figure 6.8.4. Interface of ANCOVA model construction.

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 6.8.1**). 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 6.8.5**).

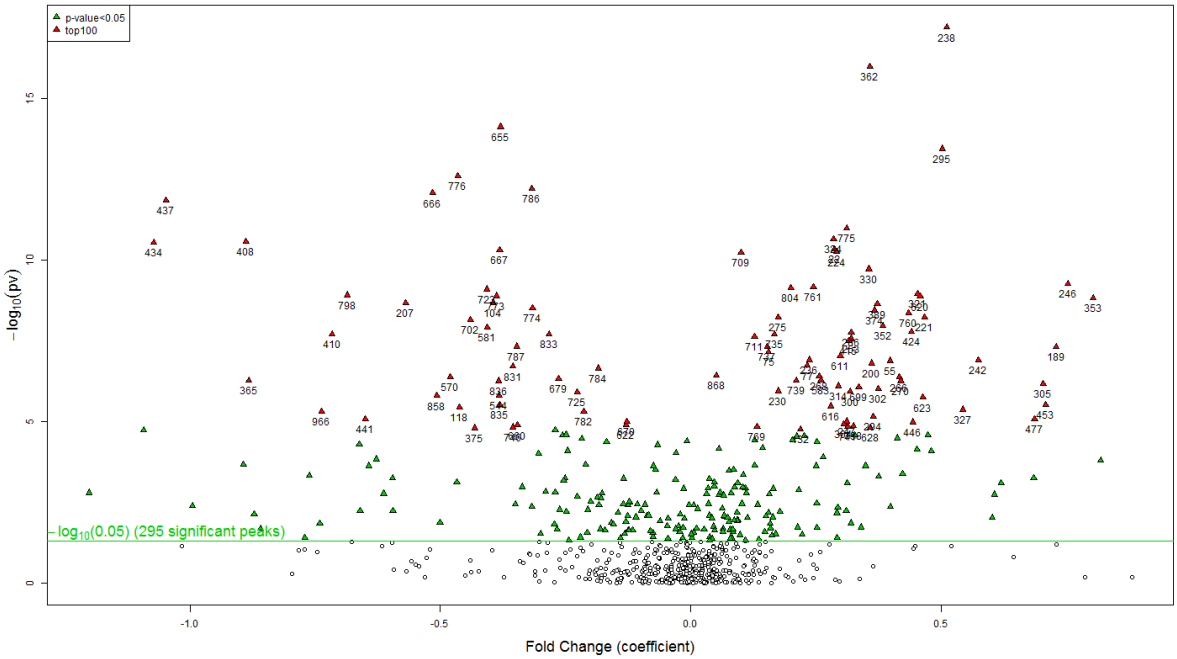


Figure 6.8.5. Volcano plot of nominal *p*-values.

To reduce the computational time of permutations, **SMART** supports parallel computing by applying the *snow* package. Users can specify the number of computation cores in the analysis (see **Figure 6.8.1**).

6.8.2 PLS/PLS-DA

To understand a quantitative relationship between multivariate responses and high-dimensional explanatory variables, SMART provides PLS regression [4,5]. PLS regression can be combined with discriminant analysis (DA) (i.e., PLS-DA) for sample classification and biomarker selection [6]. SMART provides PLS and PLS-DA by applying the *ropls* package [7] (see **Figure 6.8.6**).

Users follow the procedure to perform PLS analysis: *Click "Statistical Analysis"* ➔ *Choose "Partial Least Square (PLS/PLS-DA)"*. The first part is to provide the peak abundance data. If users would like to adjust for batch effects and covariate effects, users can perform an ANCOVA model; in the ANCOVA model, batch effects and covariates are adjusted but the factor variable (e.g., case vs. control or blood pressure (bp)) cannot be included. The residual file of each peak was exported as an input file for PLS/PLS-DA. Whatever peak abundance data or residual data, **SMART** will take a median across replicates of a subject of each peak before performing PLS/PLS-DA.

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 6.8.1**, **Table 6.8.2**, and **Table 6.8.3**).

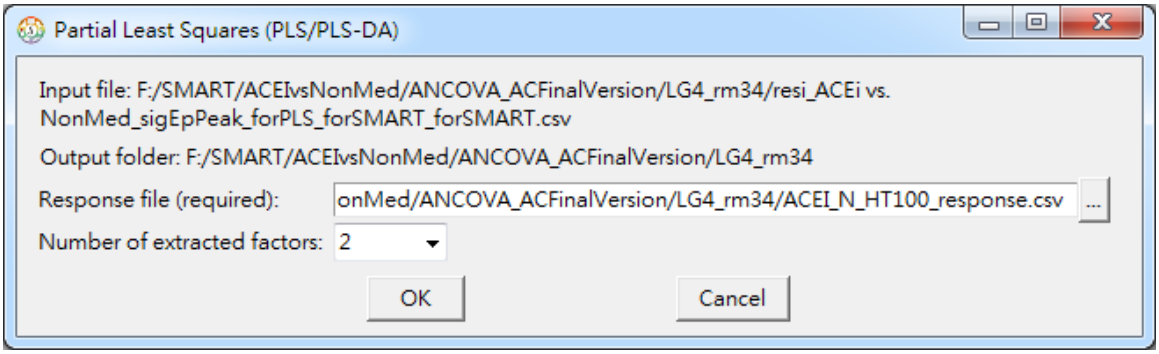


Figure 6.8.6. Interface of PLS/PLS-DA.

Table 6.8.1. 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 6.8.2. 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 6.8.3. Response variable file with more than one response variables.

SampleID	Cbp	CBMI
ACEI1	126	26.08
ACEI2	110	NA
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 6.8.7**). 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 6.8.8**). In **Figure 6.8.8**, there are 10 subjects, 5 of whom were treated with ACEi, while the remaining 5 were untreated. The discrete response variable represents the ACEi-treated group and the non-medicated group. The black ellipse shows the 95% confidence band for all 10 subjects, the blue ellipse represents the 95% confidence band for the 5 ACEi-treated subjects (2 of whom have missing peak abundance), and the blue ellipse also represents the 95% confidence band for the 5 non-medicated 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. Additionally, R2Y and Q2Y values will be generated to assess the model's performance.

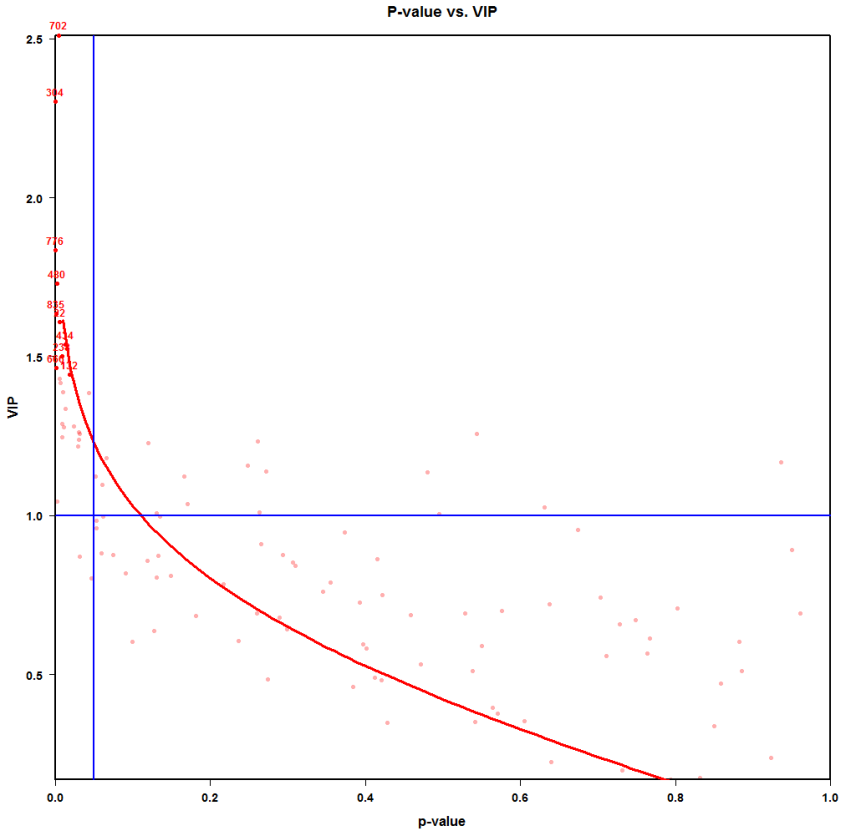


Figure 6.8.7. P-value vs. VIP plot.

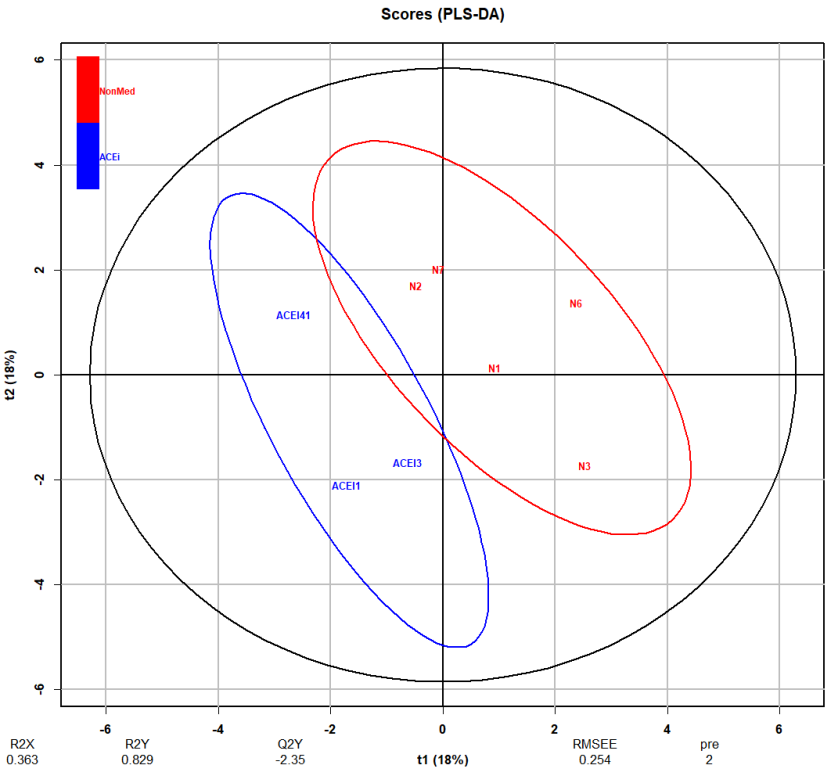


Figure 6.8.8. Score plot of the first two PLSs.

6.8.3 Pathway analysis (Integrative Omics Pathway Analysis - IOPA)

Pathway analysis is a powerful tool that enhances our understanding of complex biological systems, supports disease research, facilitates drug development, and aids in personalized medicine. SMART integrates multi-omics data, such as combining transcriptomics with metabolomics or genomics with metabolomics, using the KEGG PATHWAY database. It incorporates pathway topology and applies analysis to human signaling pathways in KEGG via the *SPIA* package [8], and extends this to metabolomics pathways (see **Figure 6.8.9**).

Users follow the procedure to perform IOPA analysis: *Click "Statistical Analysis" ➔ Choose "Pathway Analysis (Integrative Omics Pathway Analysis)"*. The first step involves summarizing information from different omics studies. IOPA requires input data, including marker IDs, the log of fold change, and the p-value for each marker (see **Table 6.8.4**). The data should contain only three fields, in this specific order: marker IDs, the log of fold change (or beta coefficients), and the p-value for each marker. Note that gene IDs should be provided as Entrez Gene IDs, while metabolite IDs should be given as KEGG compound IDs. The log of fold change and p-value can be obtained from association analyses, such as ANCOVA. Additionally, the log of fold change can be replaced by beta coefficients.

The parameter settings in the interface include p-value and logFC for filtering markers. Finally, 'Relative accuracy' refers to the computational accuracy of Pbine. The smaller the value, the more accurate the results, but this will increase the computation time.

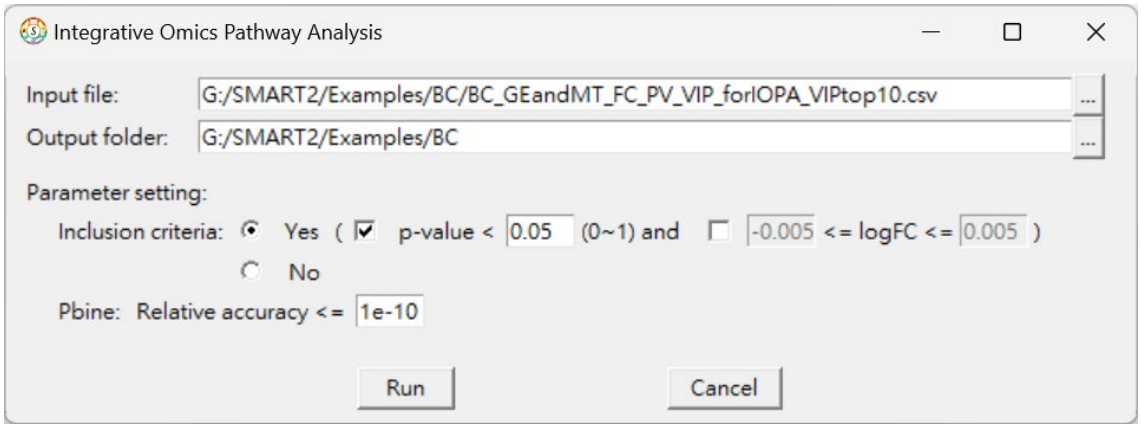


Figure 6.8.9. Interface of IOPA.

Table 6.8.4. Summarizing information file for IOPA.

ID	TUMOR_NORMAL_logFC	TUMOR_NORMAL_pv
633	0.010208589	5.34E-10
2277	-0.023064997	5.76E-16
5212	-0.002415829	5.63E-17
6854	-0.00052117	2.41E-16
8840	0.008952051	9.09E-12
22871	-0.006782005	5.76E-16
27147	-0.002096221	5.63E-17
56920	0.000324869	1.58E-16
84795	-0.003769418	1.25E-13
388335	-0.004401064	3.41E-15
C00025	0.003073349	2.82E-16
C00049	0.006398058	2.82E-16
C00148	0.001013883	9.02E-17
C00245	0.004865891	7.04E-15
C00346	0.013283466	6.93E-17
C01042	0.021779581	1.16E-15
C03626	0.003865088	6.93E-17
NA	-0.027927199	0.795677578
NA	-0.027927199	0.795677578
NA	-0.027927199	0.795677578

Finally, the statistical test results for the pathways will be output and sorted according to Pbine's p-value (see **Table 6.8.5**). Additionally, results will be visualized using Manhattan plots (see **Figure 6.8.10** and **Figure 6.8.11)** and volcano plots (see **Figure 6.8.12** and **Figure 6.8.13**).

IOPA will output both raw p-values and FDR-adjusted p-values, providing users with more detailed pathway information.

Table 6.8.5. Pathway analysis summarized table.

Name	ID	G_size	deG_n	G_pORA	C_size	deC_n	C_pORA	P_size	deP_n	pORA	pORA_Bonf	pORA_FDR	raw_eSPIA	eSPIA	eSPIA_nor	peSPIA	peSPIA_Bonf	peSPIA_FDR	pFisher	pFisher_Bonf	pFisher_FDR	pPbine	pPbine_Bonf	pPbine_FDR	Status	KEGGLINK
PI3K-Akt signaling pathway	4151	106	35	1.85E-06	0	0	1	106	35	7.85E-06	0.002056934	0.002056934	-0.696338854	-0.662305871	-0.232146951	0.001	0.255	0.0425	1.54E-07	4.04E-05	4.04E-05	4.38E-07	0.000111777	0.000111777	Inhibited	http://www.genome.jp/dbg
Calcium signaling pathway	4020	63	18	0.00389028	0	0	1	63	18	0.0075056	1	0.09364132	-0.441439162	-0.425246937	-0.35748479	5.00E-06	0.001275	0.00031875	6.79E-07	0.000177947	8.90E-05	1.71E-06	0.000437096	0.000218548	Inhibited	http://www.genome.jp/dbg
Pathways in cancer	5200	147	34	0.00497987	3	0	1	150	34	0.01725762	1	0.152266146	-0.571390567	-0.554011739	-0.185532211	5.00E-06	0.001275	0.00031875	1.49E-06	0.000390331	0.00013011	3.53E-06	0.000901132	0.000300377	Inhibited	http://www.genome.jp/dbg
Melanoma	5218	21	6	0.08175737	0	0	1	21	6	0.10347663	1	0.326637074	-0.253838699	-0.248376369	-0.600314631	5.00E-06	0.001275	0.00031875	8.01E-06	0.002097634	0.000524408	1.67E-05	0.00424622	0.001061555	Inhibited	http://www.genome.jp/dbg
Drug metabolism - cytochrome P450	982	12	8	6.86E-05	0	0	1	12	8	0.00010778	0.028238923	0.007059731	-2.987524907	-2.835152548	-0.689629269	0.012	1	0.218571429	1.88E-05	0.004933311	0.000722352	3.67E-05	0.009347822	0.001366146	Inhibited	http://www.genome.jp/dbg
EGFR tyrosine kinase inhibitor resistance	1521	22	5	0.22489265	0	0	1	22	5	0.26561789	1	0.575139557	-0.220706524	-0.217250275	-0.48048767	5.00E-06	0.001275	0.00031875	1.93E-05	0.005056466	0.000722352	3.75E-05	0.009563023	0.001366146	Inhibited	http://www.genome.jp/dbg
Gastric cancer	5226	37	14	0.00052846	0	0	1	37	14	0.00099445	0.260546882	0.033008095	-0.305626645	-0.297376399	-0.311058228	0.001	0.255	0.0425	1.47E-05	0.003861584	0.000722352	2.92E-05	0.007456461	0.001366146	Inhibited	http://www.genome.jp/dbg
Pentose phosphate pathway	30	5	3	0.02643039	6	5	0.05284472	11	8	4.17E-05	0.010935087	0.004524574	1.025991508	1.040952419	0.310509222	0.073	1	0.517083333	4.17E-05	0.010937313	0.001367164	7.65E-05	0.019498085	0.002437261	Activated	http://www.genome.jp/dbg
Focal adhesion	4510	70	21	0.00094558	0	0	1	70	21	0.00210085	0.550422188	0.045868516	-0.427565521	-0.419475605	-0.177183974	0.005	1	0.159375	0.00013092	0.034301572	0.003811286	0.00021992	0.056078542	0.006230949	Inhibited	http://www.genome.jp/dbg
Amoebiasis	5146	25	13	1.53E-05	1	0	1	26	13	5.18E-05	0.013573723	0.004524574	0.001746742	0.005077916	0.003933636	0.893	1	1	0.00050804	0.13310598	0.013310598	0.00077146	0.196721929	0.019672193	Activated	http://www.genome.jp/dbg
Breast cancer	5224	39	11	0.02413688	0	0	1	39	11	0.03625862	1	0.193872634	-0.326241792	-0.321570055	-0.411620738	0.003	0.765	0.109285714	0.00110149	0.288589981	0.026235453	0.00158065	0.40306554	0.036642322	Inhibited	http://www.genome.jp/dbg
Staphylococcus aureus infection	5150	12	7	0.00065091	0	0	1	12	7	0.00095342	0.249795879	0.033008095	0.03344143	0.03344143	0.086781981	0.275	1	0.796571429	0.00242433	0.635173708	0.052931142	0.00328713	0.838218448	0.069851537	Activated	http://www.genome.jp/dbg
Metabolism of xenobiotics by cytochrome P450	980	11	7	0.00031122	0	0	1	11	7	0.00046013	0.120554321	0.024110864	-0.186425763	-0.157108143	-0.119475935	0.699	1	1	0.00290823	0.761955221	0.05861194	0.00389259	0.992610251	0.076354635	Inhibited	http://www.genome.jp/dbg
MicroRNAs in cancer	5206	66	17	0.0148543	0	0	1	66	17	0.02600013	1	0.168425775	0.034832868	0.034832868	0.117685742	0.014	1	0.238	0.0032463	0.850529619	0.060752116	0.0043114	1	0.078529101	Activated	http://www.genome.jp/dbg
Rap1 signaling pathway	4015	76	18	0.02861294	0	0	1	76	18	0.04877708	1	0.228207042	-0.505074937	-0.482823678	-0.356757056	0.01	1	0.2125	0.00420735	1	0.073488333	0.00548668	1	0.09327351	Inhibited	http://www.genome.jp/dbg

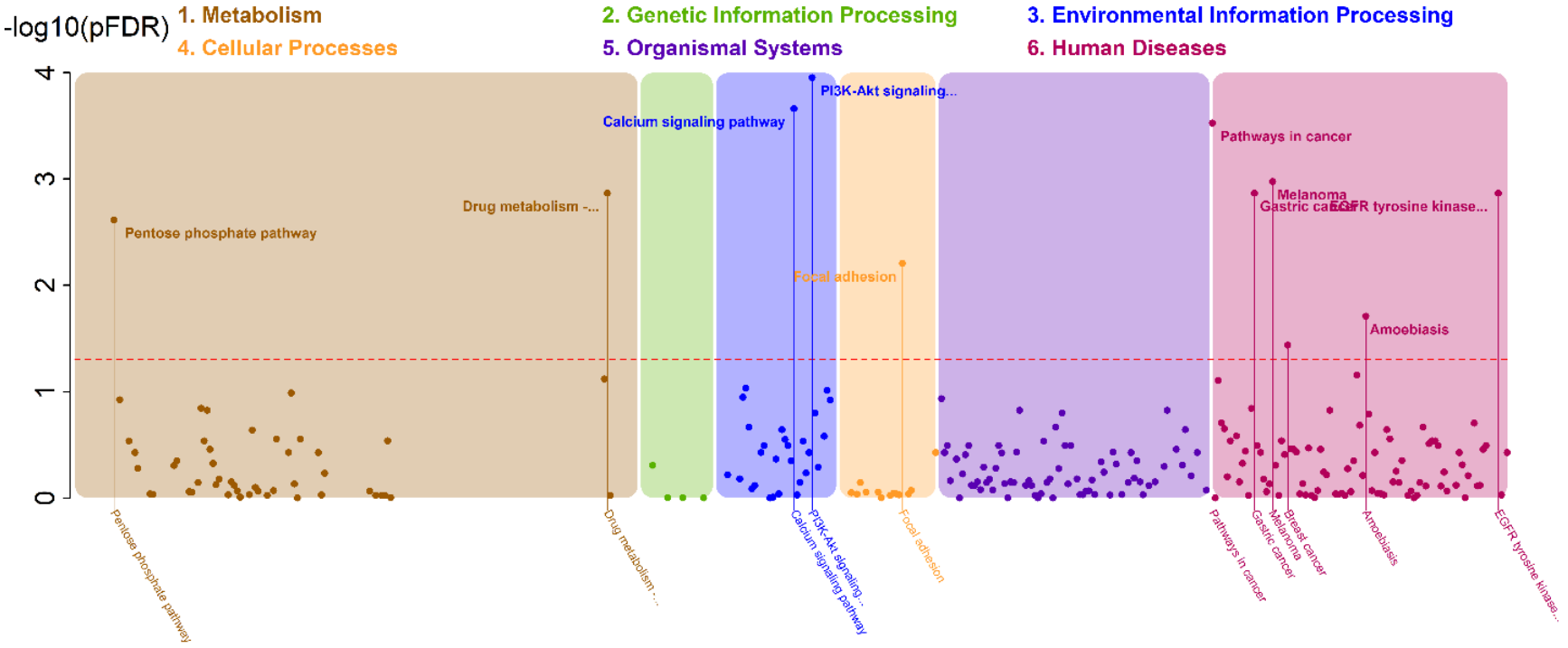


Figure 6.8.10. Manhattan plot of IOPA (pFDR of Pbine).

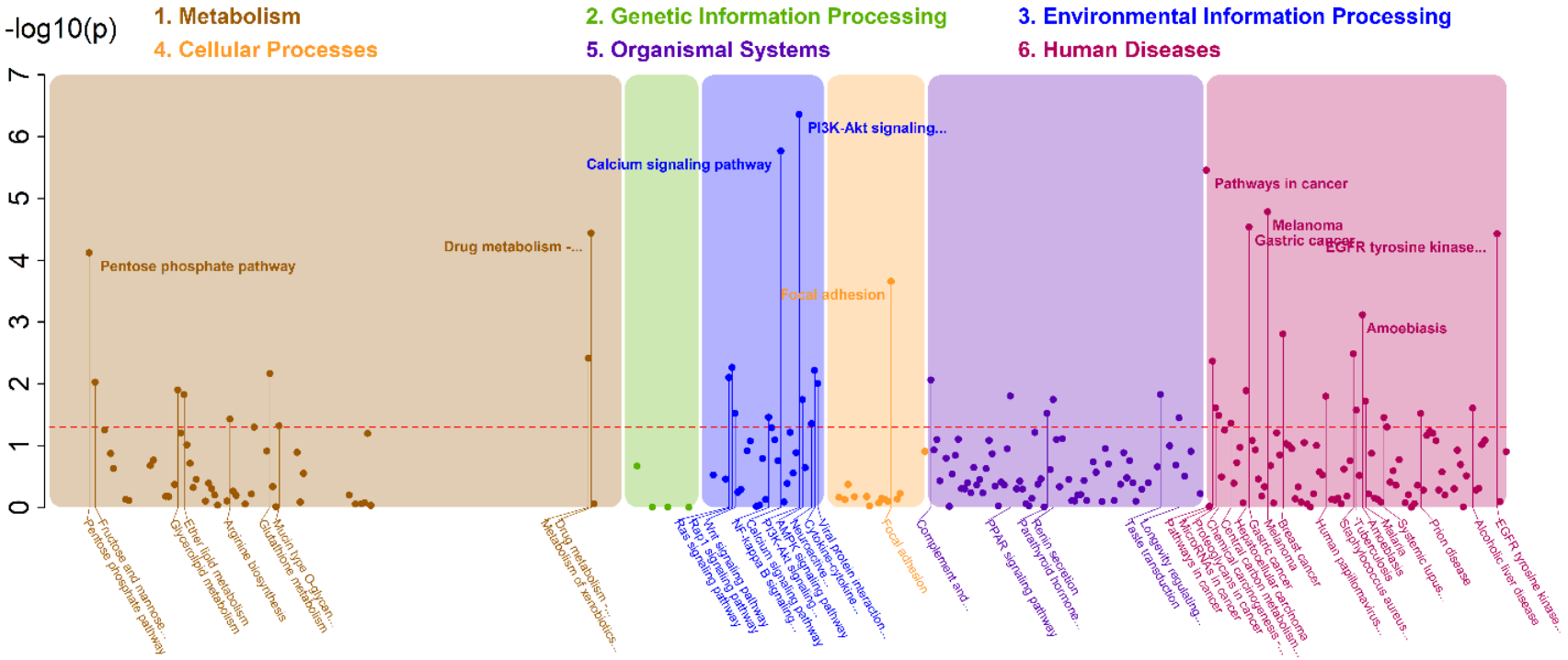


Figure 6.8.11. Manhattan plot of IOPA (p-value of Pbine).

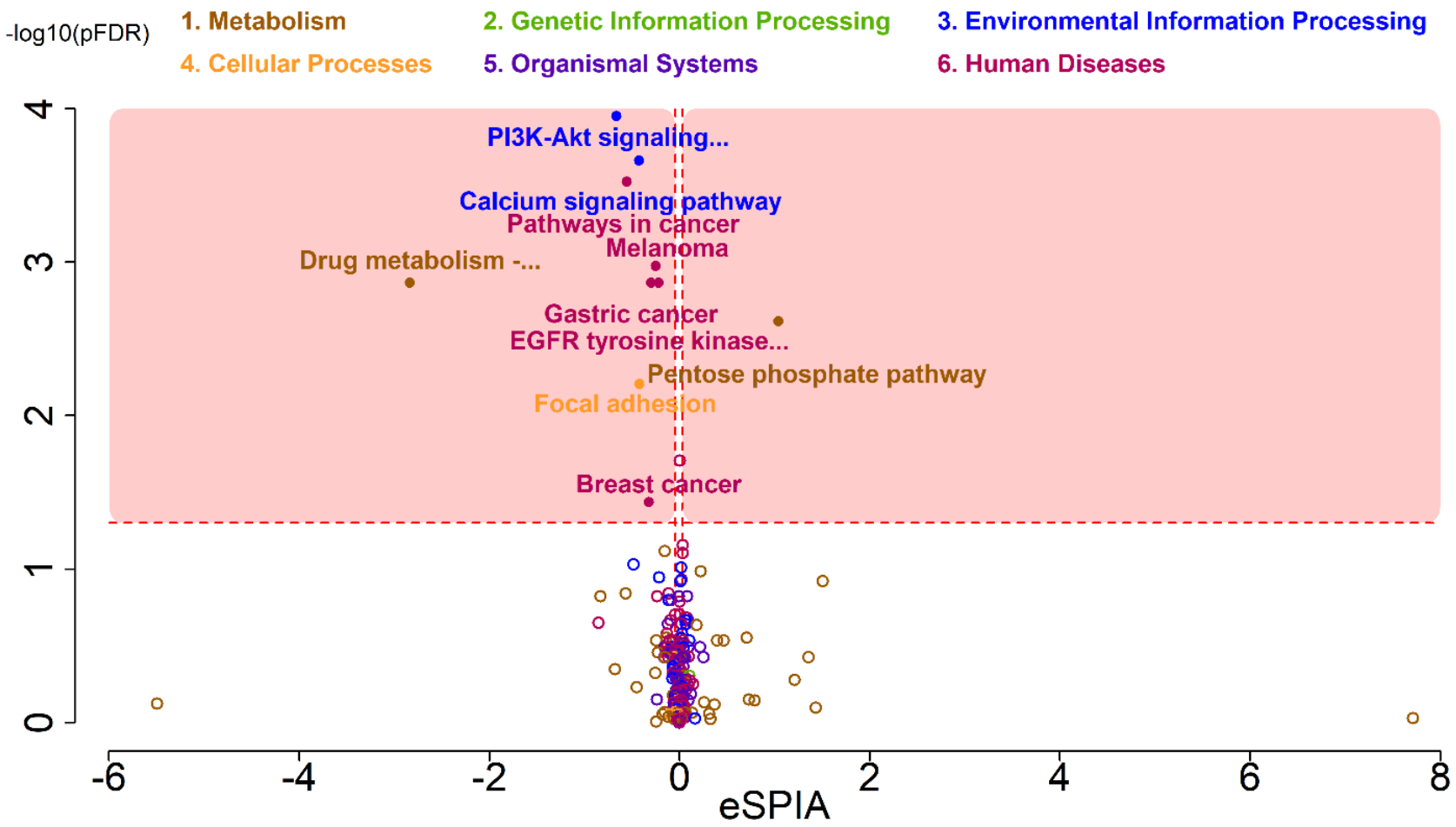


Figure 6.8.12. Volcano plot of IOPA (eSPIA).

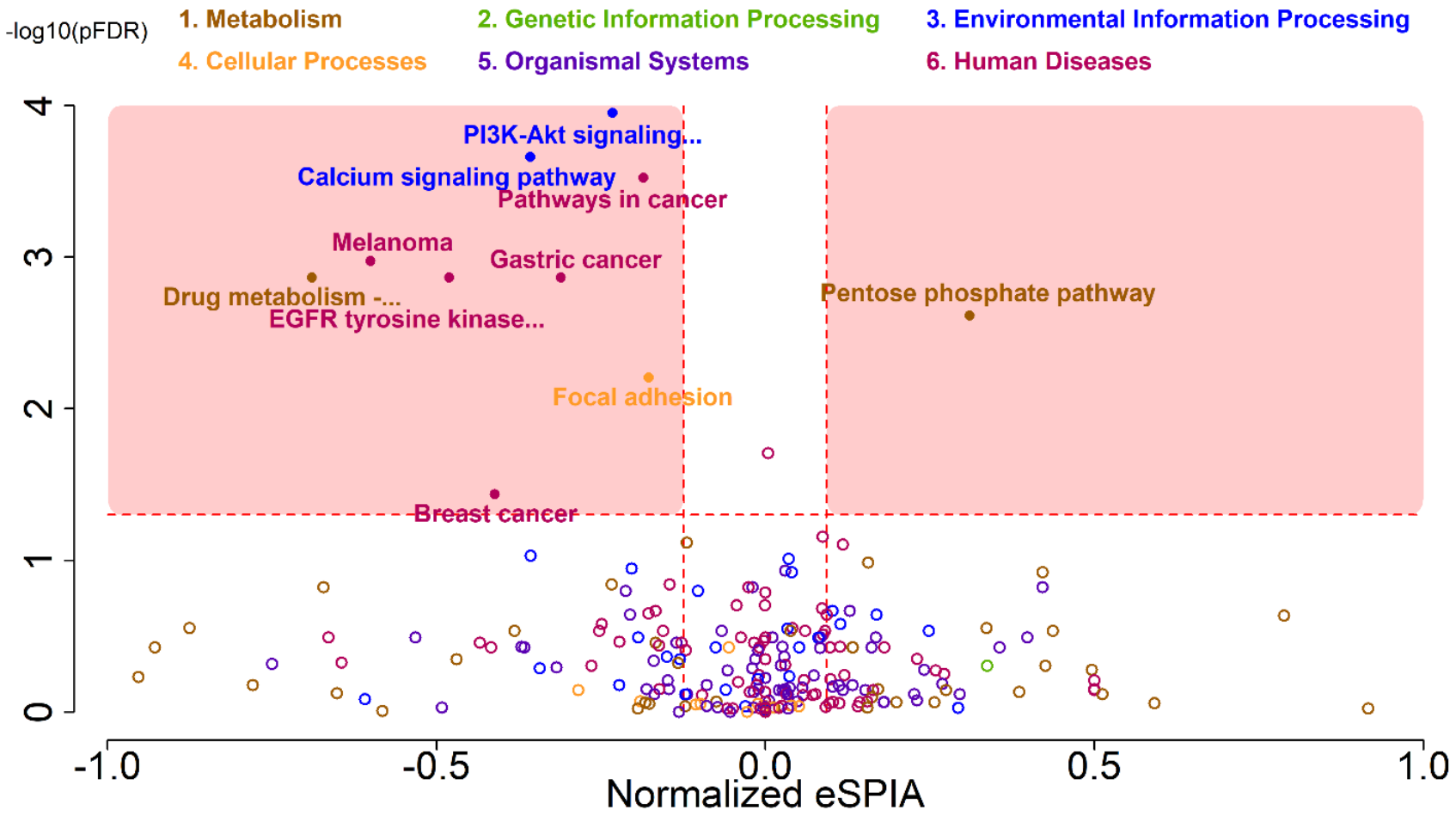


Figure 6.8.13. Volcano plot of IOPA (normalized eSPIA).

IOPA will analyze all KEGG pathways, but it is possible that few or no significant pathway results will be found. This may be due to the limited number of markers remaining after applying the fold change and p-value filters. It is recommended to consider relaxing the filtering criteria.

Tips in the instructions

- ID preparation:
    - Tips: If only gene symbols are available, Entrez Gene IDs are required for IOPA analysis. Given the gene symbols, Entrez Gene IDs can be queried using SYNGO (https://www.syngoportal.org/convert) or g:Profiler (https://biit.cs.ut.ee/gprofiler/convert). Metabolite KEGG IDs can be obtained through SMART's peak identification module or via MetaboAnalyst (https://www.metaboanalyst.ca/MetaboAnalyst/upload/ConvertView.xhtml).

- Combine the p-values from ORA and eSPIA using the Pbine method, taking into account the correlation r of their p-values.
    - Tips: Be aware that the correlation r between the p-values from the two pathway tests should not be too close to 1。

- If the combined p-value from Pbine is not significant after FDR correction, the IOPA analysis will indicate: “IOPA's analysis shows no significant pathways, likely due to the small number of markers after filtering. Consider relaxing the inclusion criteria.”
    - Tips: Users can re-evaluate the appropriate p-value cut-off or adjust the fold change range to be more lenient.

- If no pair of significant p-values is found, IOPA will display: “No significant pathway identified by either ORA or eSPIA.”
    - Tips: Users should verify if the number of markers in the input data is too small to correspond with the KEGG database. Additionally, they should check if these markers are not significant in the association test (e.g., ANCOVA).

6.9 Post analysis

6.9.1 Peak identification

For peak identification, **SMART** offers a batch search interface to query m/z values from HMDB and MassBank. Metabolite data from HMDB in XML format were downloaded from http://www.hmdb.ca/downloads, and metabolite data from MassBank in JSON format were downloaded from https://massbank.eu/MassBank/.

Users follow the procedure to perform peak identification: *Click "Post Analysis"* ➔ *Choose "Peak Identification"*. In the interface of peak identification (see **Figure 6.9.1**), users can load a csv file that the first three columns are peak index, m/z, and RT (e.g., extracted from 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.

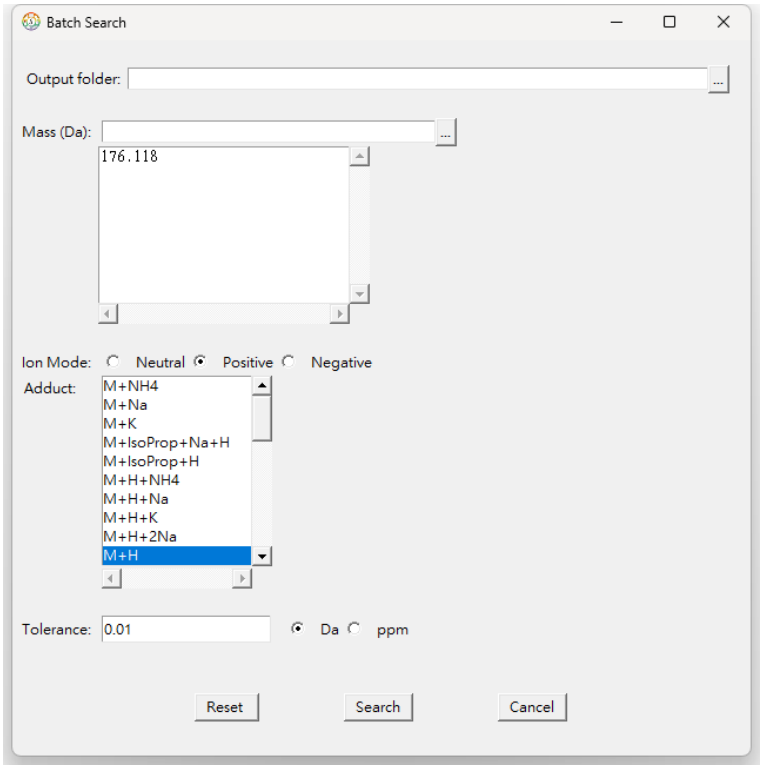


Figure 6.9.1. Interface of peak identification module.

After *clicking "Search"*, **SMART** shows a table including "Target" (i.e., the m/z we queried), "Adduct", "Database", "Database ID", "Name", Chemical Formula", "Mass" (i.e., M: the monoisotopic molecular weight), KEGG ID, and "Delta" (see **Table 6.9.1**). Also, the peak identification table is exported as a csv file.

Table 6.9.1. Peak identification table.

Target	Adduct	DB	DB_ID	Name	Chem.Formu	Mass	KEGG_ID	abs.delta
370.1643	M+H	HMDB	HMDB0014038	N-Deisopropyl-fluvastatin	C21H20FNO4	369.1376		0.01936
370.1643	M+H	HMDB	HMDB0015288	Trimetrexate	C19H23N5O3	369.1801	C11154	0.023093
370.1643	M+H	HMDB	HMDB0015633	Amisulpride	C17H27N3O4S	369.1722		0.01523
370.1643	M+H	HMDB	HMDB0029360	Isorheagenine	C20H19NO6	369.1212		0.035759
370.1643	M+H	HMDB	HMDB0029401	Ochratoxin B	C20H19NO6	369.1212		0.035759
370.1643	M+H	HMDB	HMDB0031947	Niazicinin	C17H23NO8	369.1424		0.01463
286.1431	M+H	HMDB	HMDB0002171	Glycylprolylhydroxyproline	C12H19N3O5	285.1325		0.003399
286.1431	M+H	HMDB	HMDB0014440	Morphine	C17H19NO3	285.1365	C01516	0.000624
286.1431	M+H	HMDB	HMDB0014472	Hydromorphone	C17H19NO3	285.1365	C07042	0.000624
286.1431	M+H	HMDB	HMDB0015692	Isothipendyl	C16H19N3S	285.13		0.005901
286.1431	M+H	HMDB	HMDB0029377	Piperine	C17H19NO3	285.1365	C03882	0.000624
286.1431	M+H	HMDB	HMDB0030256	Erysopine	C17H19NO3	285.1365		0.000624
286.1431	M+H	HMDB	HMDB0032961	Secodemethylclausenamide	C17H19NO3	285.1365		0.000624

6.9.2 Concentration calibration

For concentration calibration, **SMART** provides two components: calibration curve construction and concentration calculation. First, the intensities of known standards at different concentrations are used to construct a calibration curve, which is then used to estimate the concentration of unknown samples. We used concentrations of 200, 300, 400, 500, and 700 ppb as standards to construct the calibration curve, with 600 ppb as the unknown test sample. Both the standards and the test sample were analyzed using targeted peak analysis (see **Section 6.3.2**), and their results are presented in the peak abundance table.

- **Calibration curve construction**

Users should follow these steps to perform calibration curve construction: *Click "Post Analysis" ➔ Choose "Concentration Calibration" ➔ Choose "Calibration Curve Construction"*. In the calibration curve construction interface (see **Figure 6.9.2**), the first step is to import the peak abundance data. This data can be obtained from targeted peak analysis (see **Section 6.3.2**) base on different known standards.

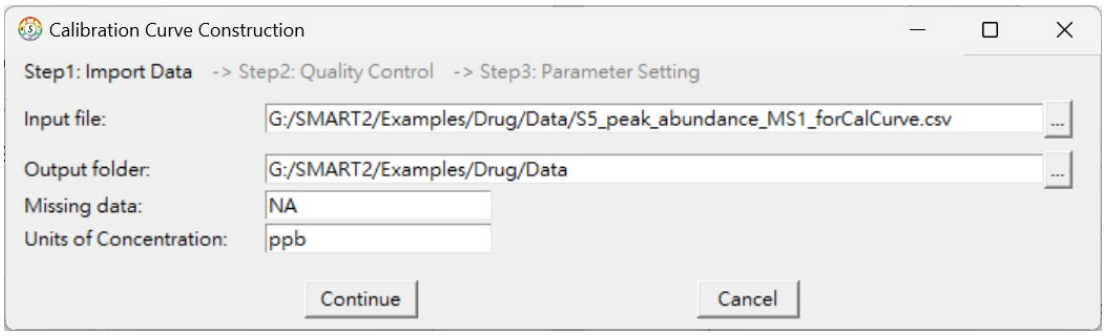


Figure 6.9.2. Interface of calibration curve construction (step 1: import data).

After *clicking "Continue"*, the data fields will be displayed. Please note that a name field is required (see **Figure 6.9.3**). You can select the standard samples for constructing the calibration curve (see **Figure 6.9.4**).

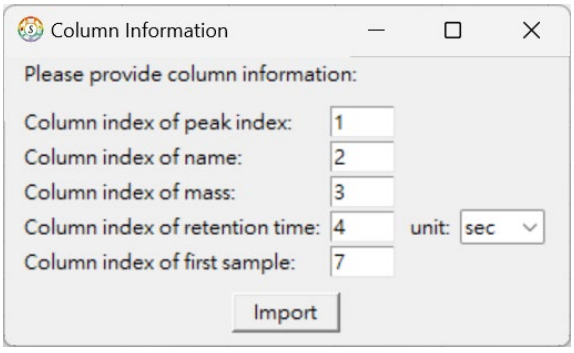


Figure 6.9.3. Setting up the column information for the dataset.

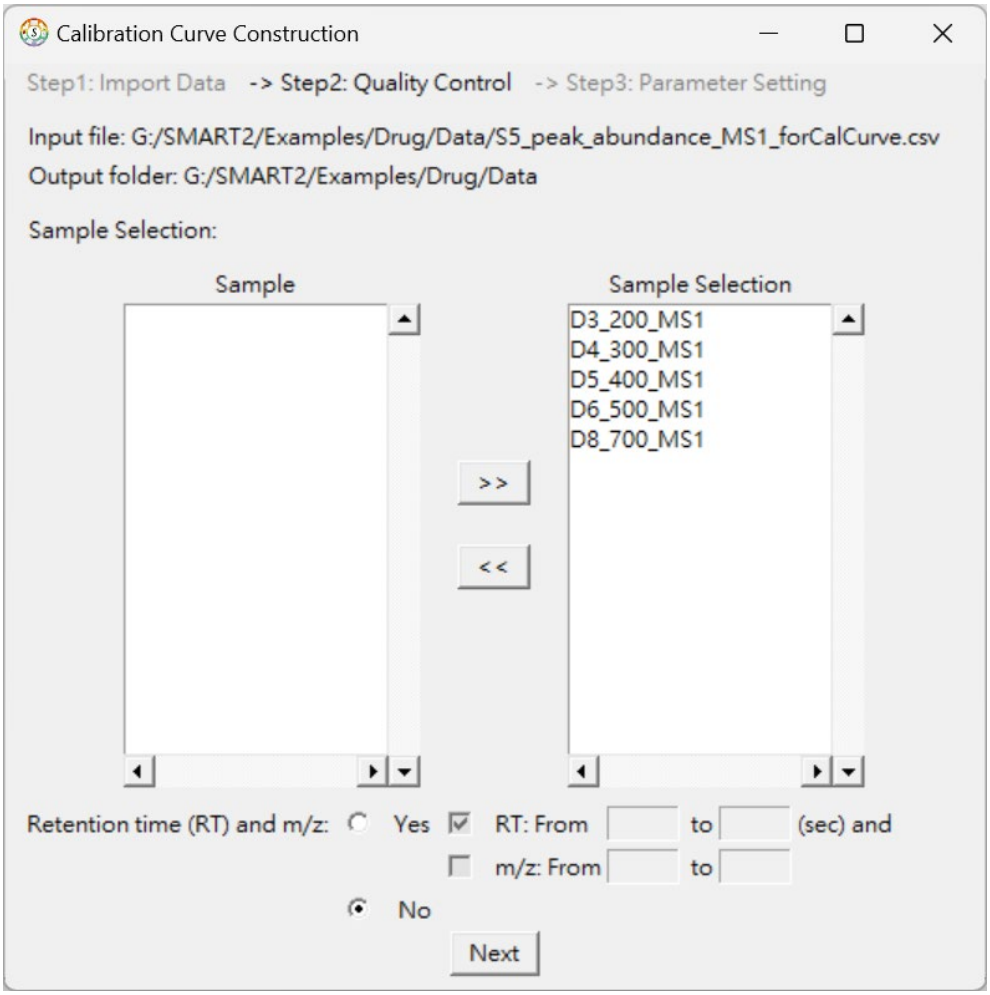


Figure 6.9.4. Interface of calibration curve construction (step 2: quality control).

Next, *click 'Next'* to select the method for calculating the calibration curve and detecting outliers (see **Figure 6.9.5**). **SMART** supports linear and quadratic regression models for building the calibration curve and allows for the application of weighting factor of 1, 1/x, or $1/x^2$. **SMART** also provides Cook's distance (Cook's D), confidence intervals (CI), and discrepancies from the true concentration (Bias) to help identify outliers. The quality of the model is assessed using either the R^2 or the adjusted R^2.

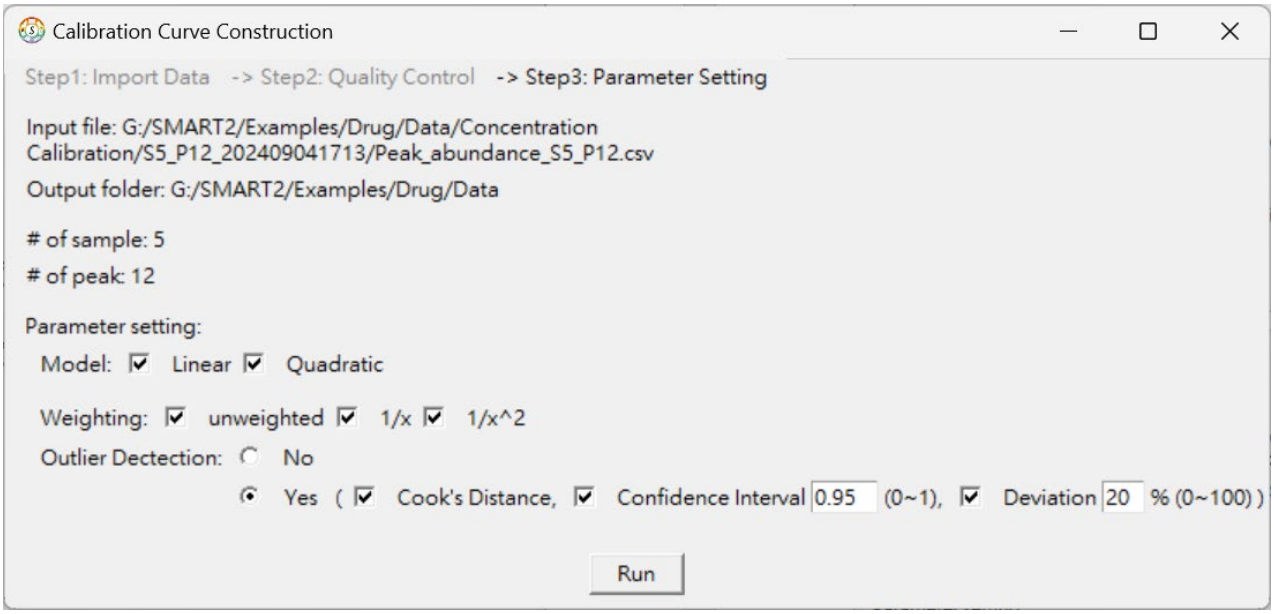


Figure 6.9.5. Interface of calibration curve construction (step 3: parameter setting).

Calibration curves can be constructed simultaneously for different compounds. For each compound, a calibration curve plot for the standards at various concentrations (e.g., PID1_Heroin_mz370_lm.tiff) will be generated (see **Figure 6.9.6**). Additionally, results for all calibration curves will be provided in a file (e.g., Calibration-curve_info.csv) (see **Table 6.9.2**), along with a concentration table for all compounds (e.g., Peak_abundance_S10_P12.csv) (see **Table 6.9.3**).

Tips in the instructions

Constructing a calibration curve involves designing standard concentration experiments based on expert knowledge and selecting the appropriate calibration curve model according to the characteristics of the standards.

- In some cases, the calibration curve cannot be computed. For instance, when there are too many outliers, the model parameters may fail to be estimated, resulting in the inability to generate the calibration curve (see **Figure 6.9.6**). If the calibration curve cannot be generated, the experimental conditions for the standards should be re-examined.

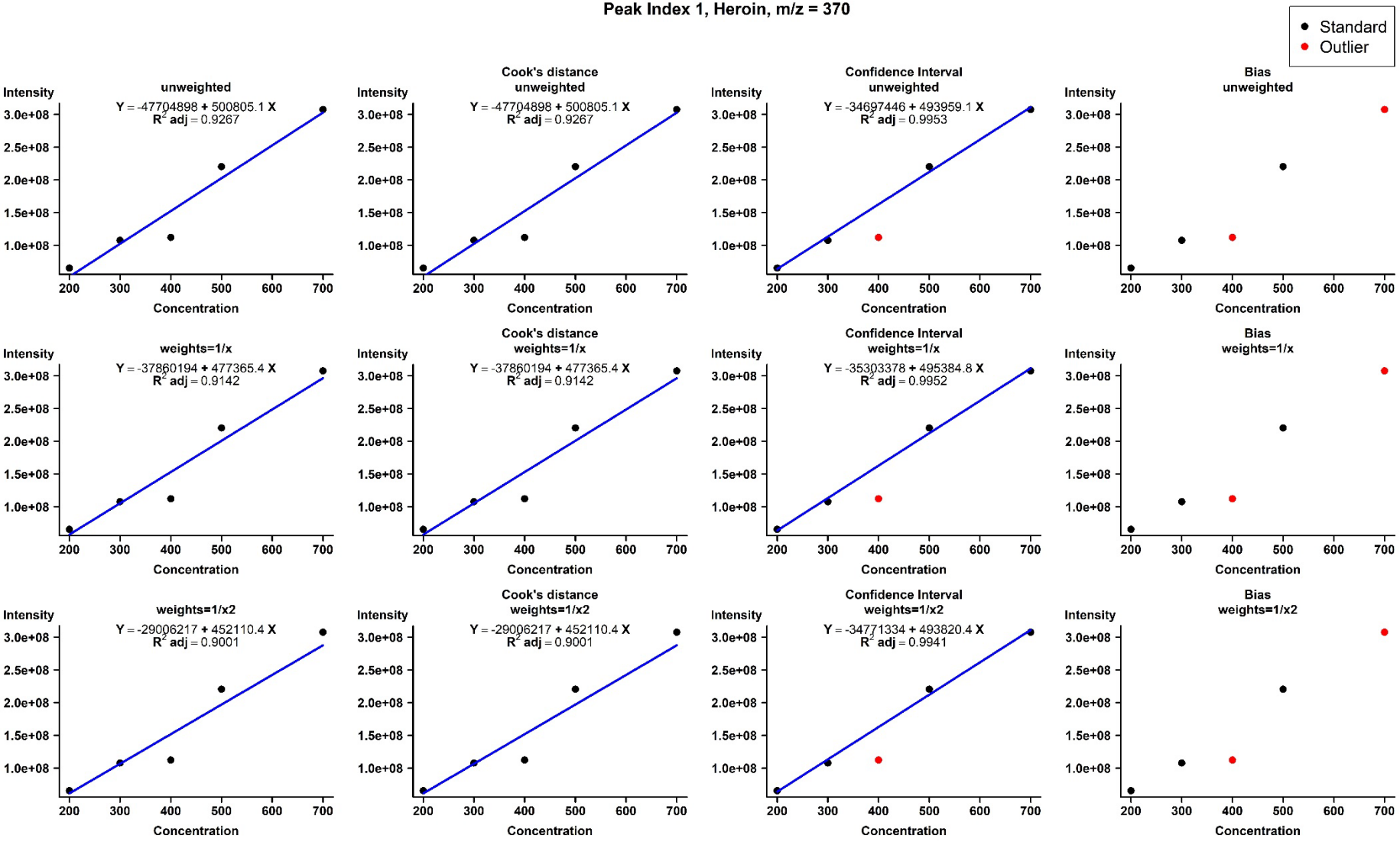


Figure 6.9.6. Calibration curve plots of Heroin.

Table 6.9.2. Calibration curve information.

Peak_Index	Name	mz	Ret_time.sec	model	weights	Original	concentration	theta1	theta2	theta3	R2	adj.R2	Cook's Distance	concentration	theta1	theta2	theta3	R2	adj.R2	Confidence Interval	concentration	theta1	theta2	theta3	R2	adj.R2	Bias	Bias_na	concentration	theta1	theta2	theta3	R2	adj.R2
1	Heroin	370	685.2	linear	unweighted		200;300;400;500;700	-4.8E+07	500805.1		0.945048	0.926731		200;300;400;500;700	-4.8E+07	500805.1		0.945048	0.926731		200;300;500;700	-3.5E+07	493959.1		0.996895	0.995342		700						
1	Heroin	370	685.2	linear	1/x		200;300;400;500;700	-3.8E+07	477365.4		0.935687	0.91425		200;300;400;500;700	-3.8E+07	477365.4		0.935687	0.91425		200;300;500;700	-3.5E+07	495384.8		0.996833	0.99525		700						
1	Heroin	370	685.2	linear	1/x2		200;300;400;500;700	-2.9E+07	452110.4		0.925085	0.900114		200;300;400;500;700	-2.9E+07	452110.4		0.925085	0.900114		200;300;500;700	-3.5E+07	493820.4		0.996057	0.994086		700						
1	Heroin	370	685.2	quadratic	unweighted		200;300;400;500;700	7647495	214235	315.5683	0.954354	0.908708		200;300;400;500;700	7647495	214235	315.5683	0.954354	0.908708		200;300;400;500;700	7647495	214235	315.5683	0.954354	0.908708			200;300;500;700	-5E+07	578348.7	-93.4144	0.997547	0.99264
1	Heroin	370	685.2	quadratic	1/x		200;300;400;500;700	18940558	155768.6	379.951	0.951281	0.902561		200;300;400;500;700	18940558	155768.6	379.951	0.951281	0.902561		200;300;400;500;700	18940558	155768.6	379.951	0.951281	0.902561			200;300;500;700	-4E+07	523130.2	-32.2908	0.996914	0.990743
1	Heroin	370	685.2	quadratic	1/x2		200;300;400;500;700	24333748	125233.2	416.027	0.945776	0.891552		200;300;400;500;700	24333748	125233.2	416.027	0.945776	0.891552		200;300;400;500;700	24333748	125233.2	416.027	0.945776	0.891552		200						
2	Morphine	286	147	linear	unweighted		200;300;400;500;700	-2.6E+07	602625.6		0.943209	0.924279		200;300;400;500;700	-2.6E+07	602625.6		0.943209	0.924279		200;300;500;700	-1.1E+07	594310.2		0.995848	0.993772		700						
2	Morphine	286	147	linear	1/x		200;300;400;500;700	-1.4E+07	573710.3		0.933302	0.91107		200;300;400;500;700	-1.4E+07	573710.3		0.933302	0.91107		200;300;500;700	-1.1E+07	595601.3		0.995606	0.99341		700						
2	Morphine	286	147	linear	1/x2		200;300;400;500;700	-3042166	541973.7		0.921835	0.89578		200;300;400;500;700	-3042166	541973.7		0.921835	0.89578		200;300;400;500;700	-3042166	541973.7		0.921835	0.89578		700						
2	Morphine	286	147	quadratic	unweighted		200;300;400;500;700	40525293	256590.9	381.0501	0.952561	0.905123		200;300;400;500;700	40525293	256590.9	381.0501	0.952561	0.905123		200;300;400;500;700	40525293	256590.9	381.0501	0.952561	0.905123			200;300;500;700	-3E+07	699576.8	-116.524	0.996548	0.989644
2	Morphine	286	147	quadratic	1/x		200;300;400;500;700	55900849	176988.6	468.7074	0.949689	0.899378		200;300;400;500;700	55900849	176988.6	468.7074	0.949689	0.899378		200;300;400;500;700	55900849	176988.6	468.7074	0.949689	0.899378			200;300;500;700	-1.5E+07	620726.4	-29.2412	0.995652	0.986957
2	Morphine	286	147	quadratic	1/x2		200;300;400;500;700	63987641	131202.5	522.8015	0.944492	0.888984		200;300;400;500;700	63987641	131202.5	522.8015	0.944492	0.888984		200;300;400;500;700	63987641	131202.5	522.8015	0.944492	0.888984		200						

Table 6.9.3. Concentration table.

Peak_Index	Name	mz	Ret_time.sec	D3_200_MS1	D4_300_MS1	D5_400_MS1	D6_500_MS1	D8_700_MS1
1	Heroin	370	685.2	65461544.76	107776084.2	112225595.9	220292579.3	307410481.3
2	Morphine	286	147	110735239.3	159373768.1	165676372.9	297485459.8	400678174.4
3	Cocaine	304	702	279877977.6	454131140.2	478788584.5	973838986.6	1323450213
4	Thebaine	312	678.6	57057964.48	112046429.8	113939447.1	288690780.8	440569940.3
5	delta9-THC	315	1513.8	3017275.381	4419164.538	4255491.315	6385888.112	9572913.568
6	Amphetamine	136	422.4	36745807.48	37802677.71	38425017.72	38130786.73	37695091.02
7	MA	150	495.6	184181822.8	247915491.6	257212578.3	368418622.8	443134868.9
8	MDMA	194	539.4	173095172.4	331593012	310502864.9	695515436.4	956921664.8
9	Love Drug	180	506.4	5215888.396	7552810.671	8060482.148	14862760.19	21387477.76
10	Ketamine	238	600	304202322	513432835.7	538497108	1074832066	1413000680
11	FM2	314	1044	83573478.22	128436996.5	131222983.1	239439990.9	310006664.3
12	Nimetazepam	296	1041	91247723.91	144161146.3	147600911.2	256815496.3	345056745.4

- **Concentration calculation**

Users should follow these steps to perform concentration calculation of an unknown sample: *Click "Post Analysis" ➔ Choose "Concentration Calibration" ➔ Choose "Concentration Calculation"*. In the concentration calculation interface (see **Figure 6.9.7**), the first step is to import the peak abundance data for the unknown samples that you wish to estimate. This data can be obtained through targeted peak analysis (see **Section 6.3.2**).

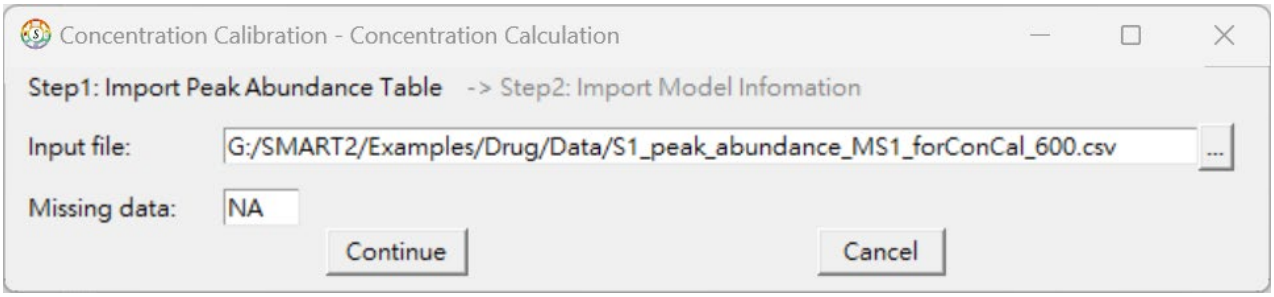


Figure 6.9.7. Interface of concentration calculation (step 1: import data).

After *clicking "Continue"*, the data fields will be displayed. Please note that a name field is required (see **Figure 6.9.3**). The next interface allows users to input the model information for the calibration curve (see **Figure 6.9.8**). Users can either select the calibration curve model generated by SMART or manually enter their own model information. The model information files must include parameter values estimated by various models. These estimates can be based on different methods, such as model variations, weighting schemes, or outlier detection techniques. The key requirement is the accuracy of the parameter values. The format should follow either the **SMART** output guidelines (see **Table 6.9.2**) or an RData file.

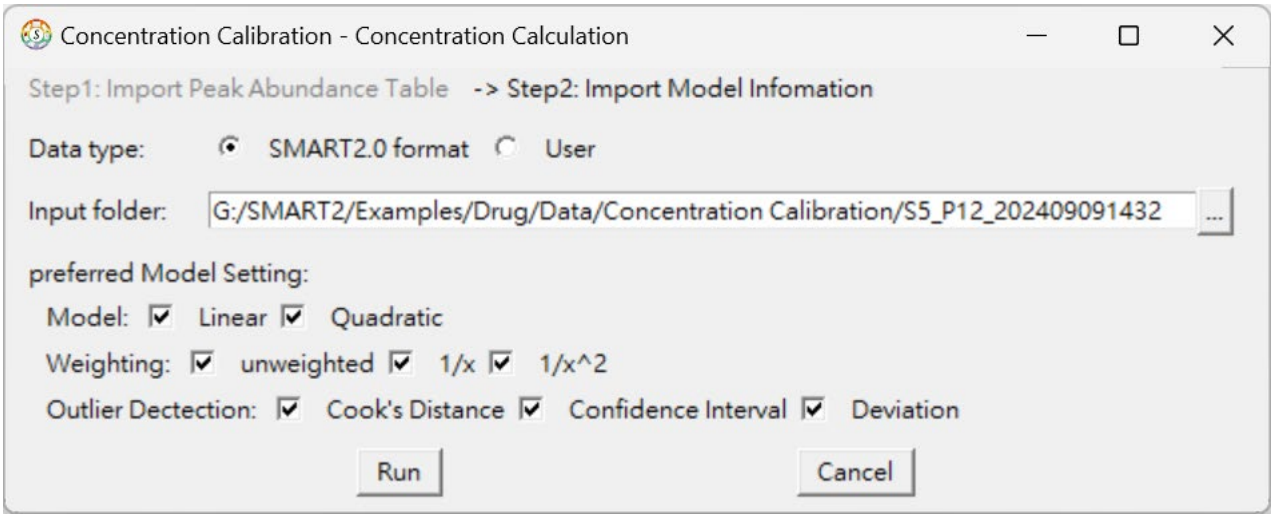


Figure 6.9.8. Interface of concentration calculation (step 2: import model information).

Use the previously constructed calibration curve to calculate the compound concentration of the unknown sample. **SMART** will output the model parameters of the optimal calibration curve (see **Table 6.9.3**), the best concentration estimation results (see **Table 6.9.4**), and the corresponding figure (see **Figure 6.9.9**).

Table 6.9.3. The best Calibration curve for each compound.

Peak_Index	Name	mz	Ret_time.sec	model	weights	outlier_detection	theta1	theta2	theta3
1	Heroin	370	685.2	linear	unweighted	Confidence Interval	-3.5E+07	493959.1	
2	Morphine	286	147	linear	unweighted	Confidence Interval	-1.1E+07	594310.2	
3	Cocaine	304	702	linear	weights=1/x	Confidence Interval	-1.6E+08	2164556	
4	Thebaine	312	678.6	linear	unweighted	Confidence Interval	-1.1E+08	786201.3	
5	delta9-THC	315	1513.8	linear	weights=1/x	Confidence Interval	517699	12543.79	
6	Amphetamine	136	422.4	quadratic	weights=1/x2		33612688	19840.16	-20.3188
7	MA	150	495.6	linear	unweighted		78432140	527953.7	
8	MDMA	194	539.4	linear	weights=1/x	Confidence Interval	-1.5E+08	1616052	
9	Love Drug	180	506.4	linear	unweighted	Confidence Interval	-1802212	33075.17	
10	Ketamine	238	600	linear	weights=1/x	Confidence Interval	-1.6E+08	2321723	
11	FM2	314	1044	linear	weights=1/x	Confidence Interval	-1E+07	471871	
12	Nimetazepam	296	1041	linear	weights=1/x	Confidence Interval	-1.1E+07	519278.4	

Table 6.9.4. The best concentration estimation for each compound.

Peak_Index	Name	mz	Ret_time.sec	Test1_600_MS1
1	Heroin	370	685.2	512.2577
2	Morphine	286	147	530.2785
3	Cocaine	304	702	534.6121
4	Thebaine	312	678.6	526.4279
5	delta9-THC	315	1513.8	600.662
6	Amphetamine	136	422.4	NA
7	MA	150	495.6	NA
8	MDMA	194	539.4	538.2119
9	Love Drug	180	506.4	537.7085
10	Ketamine	238	600	555.6778
11	FM2	314	1044	535.9544
12	Nimetazepam	296	1041	536.8149

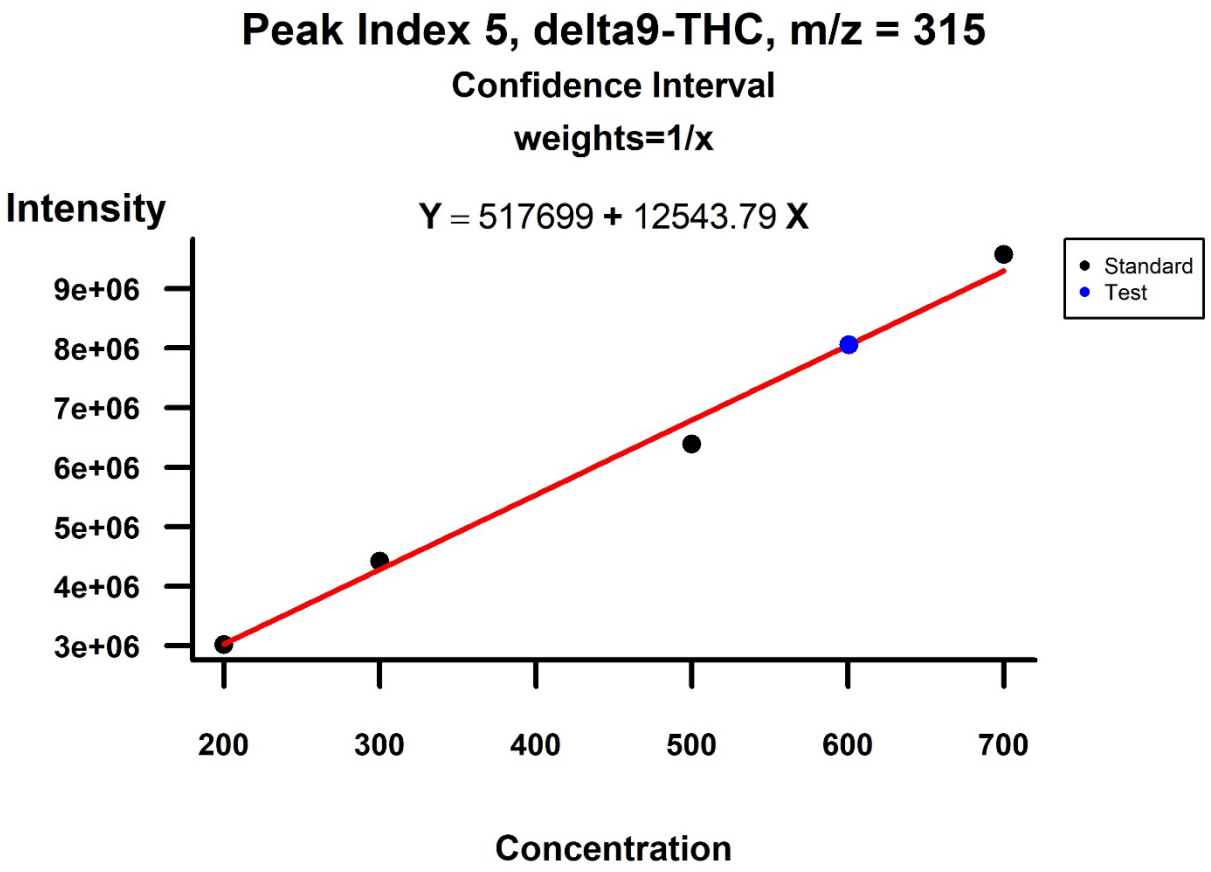


Figure 6.9.9. Concentration calculation of delta9-THC, with 600 ppb as the unknown sample.

Tips in the instructions

To estimate the compound concentration of an unknown sample, a calibration curve is typically constructed using standards of varying concentrations, based on the assumption of a linear relationship. Consequently, a linear model is frequently employed for concentration estimation. It is crucial to include a sufficient number of concentration points when constructing the calibration curve, and unknown samples should be tested with replicates.

If the concentration estimation for unknown samples is inaccurate—particularly if there is significant variation among replicates—first ensure that the experimental conditions match those of the standards. Additionally, examine the peak analysis spectrum for large variations and confirm that the peak abundance estimation is correct.

7. Examples

We offer three datasets to demonstrate **SMART**, available for download from both the **SMART** website at http://www.stat.sinica.edu.tw/hsinchou/metabolomics/SMART.htm and on GitHub at https://github.com/YuJenL/SMART.

7.1 Antihypertensive pharmacometabolomics study (HT)

We provides a real example from our antihypertensive pharmacometabolomics study [9]. In this example, there are 10 young-onset hypertensive patients from the Academia Sinica Multi-Center Young-Onset Hypertension Study [10]. 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.

- HT data for Data Import, Data Visualization, untargeted Peak Analysis, Data Preprocessing, Quality Control, Batch Effect Analysis, ANCOVA:
    - Raw spectrum files (Rawfile_10HT.7z): .raw file of 5 nonmedication and 5 ACEi medication patients.
    - Covariate file (ACEI_N_HT10_cov.csv): Age, gender, and BMI of 10 HT patients.
    - Factor file (ACEI_N_HT10_factor.csv): Contains information indicating whether subjects belong to the medication or non-medication group.
- HT data for PLS-DA:
    - Residual data (HT10_resi_ACEi vs. NonMed_FDRSig.csv): This dataset is analyzed using ANCOVA, with adjustments for residuals based on age and gender. Only peaks with significant FDR are retained.
    - Response file (HT10_ACEI_N_response.csv): Contains information on whether subjects belong to the medication or non-medication group.

7.2 Breast cancer study (BC)

This breast cancer research included both gene expression and metabolomics experiments [11,12]. Users can download the complete gene expression and metabolomics data from the *IntLIM* R package on GitHub (https://github.com/mathelab/IntLIM). Additionally, we have provided a smaller subset of BC data for PLS-DA and IOPA analysis.

- BC data for PLS-DA:
  - Residual data (BC_resi_TUMOR vs. NORMAL_FDRSig.csv): Combine the gene expression and metabolite data, apply ANCOVA to adjust for age and gender, and then retain only the markers with significant FDR from the resulting residuals. This represents a subset of the data.
  - Response file (BC_response.csv): Information on whether subjects belong to the breast cancer group or the control group.
- BC data for IOPA:
  - Multi-omics data (BC_GEandMT_FC_PV_VIP1.0_forIOPA.csv): After performing ANCOVA and PLS-DA analysis on the BC data, markers with significant FDR and VIP > 1.0 are selected for further IOPA analysis.

7.3 Narcotics study (Drug)

The narcotics study, designed by the Genome Research Center of Academia Sinica, includes MS1 and MS2 data for 12 drugs prepared at 11 concentrations ranging from 50 to 1000 ppb. We provide a sample subset of this data for reference.

- Drug data for Peak Analysis (targeted):
  - MS1 and MS2 data (forPeakAnalysis(Targeted).zip): It contains the original mzXML files of MS1 and MS2 for 12 mixed drugs at a concentration of 200 ppb, as well as the accurate m/z information for these 12 drugs.
- Drug data for Concentration Calibration (forConcentrationCalibration.zip): The peak abundance table, analyzed from targeted peak analysis, contains MS1 information for 12 narcotics at six concentrations: 200 ppb, 300 ppb, 400 ppb, 500 ppb, 600 ppb, and 700 ppb.
  - For calibration curve construction (S5_peak_abundance_MS1_forCalCurve_23457.csv): The peak abundance table contains MS1 information for 12 narcotics at five concentrations—200 ppb, 300 ppb, 400 ppb, 500 ppb, and 700 ppb—which are used as standards to construct the calibration curve.
  - For concentration calculation (S1_peak_abundance_MS1_forConCal_600.csv): The peak abundance table includes MS1 information for 12 narcotics at a single concentration of 600 ppb, which is used as an unknown sample for concentration estimation.

8. References

(1) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. *Anal Chem* **2006**, *78*, 779-787.

(2) Kuhl, C.; Tautenhahn, R.; Bottcher, C.; Larson, T. R.; Neumann, S. *Anal Chem* **2012**, *84*, 283-289.

(3) Durbin, B. P.; Hardin, J. S.; Hawkins, D. M.; Rocke, D. M. *Bioinformatics* **2002**, *18 Suppl 1*, S105-110.

(4) Nicholson, J. K.; Lindon, J. C. *Nature* **2008**, *455*, 1054-1056.

(5) Wold, S.; Sjostrom, M.; Eriksson, L. *Chemometr Intell Lab* **2001**, *58*, 109-130.

(6) Szymanska, E.; Saccenti, E.; Smilde, A. K.; Westerhuis, J. A. *Metabolomics* **2012**, *8*, S3-S16.

(7) Thevenot, E. A.; Roux, A.; Xu, Y.; Ezan, E.; Junot, C. *Journal of Proteome Research* **2015**, *14*, 3322-3335.

(8) Tarca, A. L.; Draghici, S.; Khatri, P.; Hassan, S. S.; Mittal, P.; Kim, J. S.; Kim, C. J.; Kusanovic, J. P.; Romero, R. *Bioinformatics* **2009**, *25*, 75-82.

(9) Liang, Y. J.; Chiang, K. M.; Xiu, L. L.; Chung, C. M.; Lo, C. J.; Shiao, M. S.; Cheng, M. L.; Kuo, C. C.; Yang, H. C.; Pan, W. H. *Comput Struct Biotechnol J* **2022**, *20*, 6458-6466.

(10) Yang, H. C.; Liang, Y. J.; Wu, Y. L.; Chung, C. M.; Chiang, K. M.; Ho, H. Y.; Ting, C. T.; Lin, T. H.; Sheu, S. H.; Tsai, W. C.; Chen, J. H.; Leu, H. B.; Yin, W. H.; Chiu, T. Y.; Chen, C. I.; Fann, C. S.; Wu, J. Y.; Lin, T. N.; Lin, S. J.; Chen, Y. T., et al. *PLoS One* **2009**, *4*, e5459.

(11) Terunuma, A.; Putluri, N.; Mishra, P.; Mathe, E. A.; Dorsey, T. H.; Yi, M.; Wallace, T. A.; Issaq, H. J.; Zhou, M.; Killian, J. K.; Stevenson, H. S.; Karoly, E. D.; Chan, K.; Samanta, S.; Prieto, D.; Hsu, T. Y. T.; Kurley, S. J.; Putluri, V.; Sonavane, R.; Edelman, D. C., et al. *J Clin Invest* **2014**, *124*, 398-412.

(12) Siddiqui, J. K.; Baskin, E.; Liu, M. R.; Cantemir-Stone, C. Z.; Zhang, B. F.; Bonneville, R.; McElroy, J. P.; Coombes, K. R.; Mathe, E. A. *Bmc Bioinformatics* **2018**, *19*.

9. News & updates

- 27 September 2024
  - ✓ **SMART** version 2.0 released.
    - Add S/N to QC module for peak filtering.
    - 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 new module for peak identification.
    - Add a new module for concentration calibration.