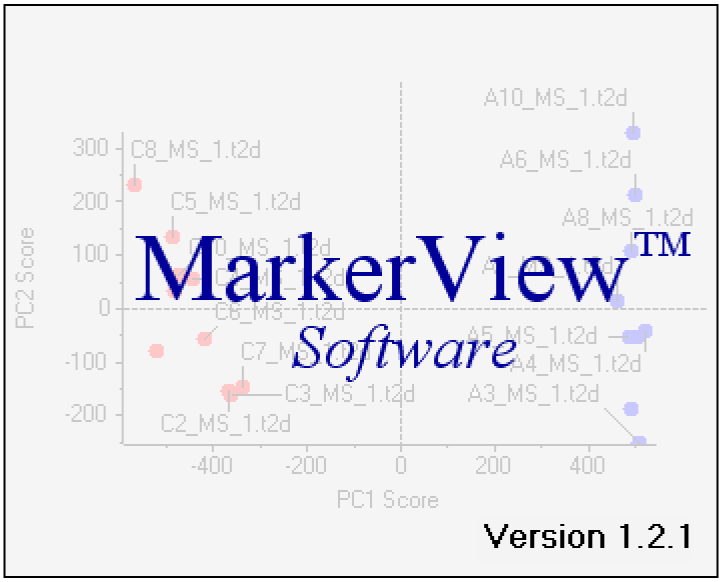
AB SCIEX



User Manual

Equipment that may be described in this document is protected under one or more patents filed in the United States, Canada, and other countries. Additional patents are pending.

Software that may be described in this document is furnished under a license agreement. It is against the law to copy, modify, or distribute the software on any medium, except as specifically allowed in the license agreement. Furthermore, the license agreement may prohibit the software from being disassembled, reverse engineered, or decompiled for any purpose.

Portions of this document may make reference to other manufacturers and/or their products, which may contain parts whose names are registered as trademarks and/or function as trademarks of their respective owners. Any such usage is intended only to designate those manufacturers' products as supplied by AB SCIEX for incorporation into its equipment and does not imply any right and/or license to use or permit others to use such manufacturers' and/or their product names as trademarks.

AB SCIEX makes no warranties or representations as to the fitness of this equipment for any particular purpose and assumes no responsibility or contingent liability, including indirect or consequential damages, for any use to which the purchaser may put the equipment described herein, or for any adverse circumstances arising therefrom.

For research use only. Not for use in diagnostic procedures.

ISO 9001
REGISTERED COMPANY

AB SCIEX
71 Four Valley Dr., Concord, Ontario, Canada. L4K 4V8.
AB SCIEX LP is ISO 9001 registered.

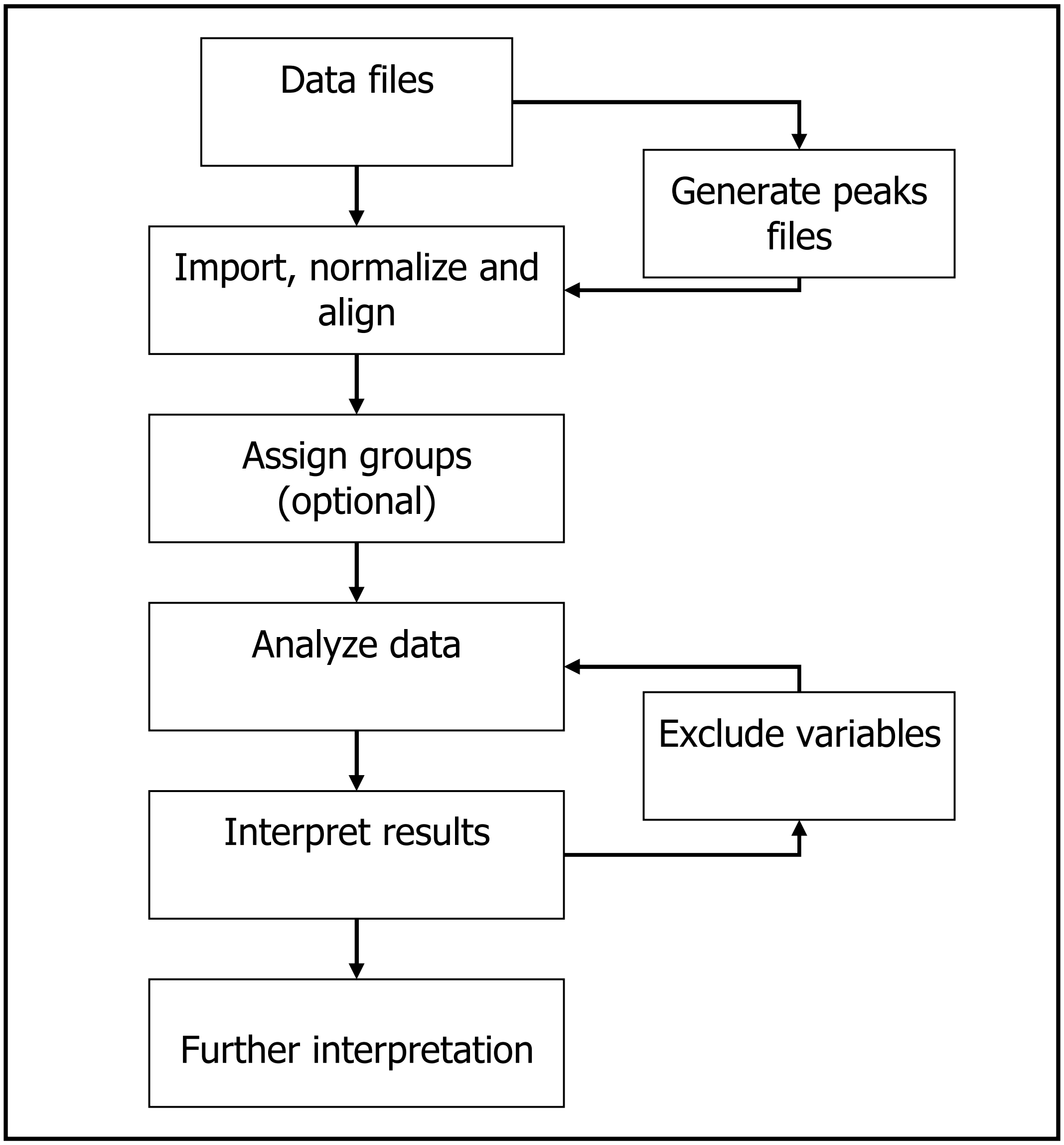
1 Table of Contents

2 Introduction and typical workflow

The MarkerView™ Software is designed to allow the data from several samples to be compared so that differences can be identified; typical applications include: metabolomics, biomarker discovery, metabolite identification, impurity profiling, etc. This manual provides an overview of some of the most common processing operations; a detailed description of the various commands, menus and dialog boxes is contained in the Reference Manual.

The program uses multivariate analysis (MVA) techniques to compare the samples and provides both supervised and unsupervised methods. Supervised methods use prior knowledge of the sample groups (for example, healthy vs. diseased) to determine the variables that distinguish the groups. In contrast, unsupervised methods allow the structure within the data to be determined and visualized. The two approaches can be combined, i.e. unsupervised methods can be used to determine the groups, and then supervised methods can be used to confirm the important variables.

A typical workflow is shown below:



MVA requires that the initial data be in the form of an array, hence the first step is importing the data to generate the array:

	Sample 1	Sample 2	Sample 3	etc…
Variable 1				
Variable 2				
etc…				

The content of a cell represents the value of the appropriate variable in the sample and can be zero if the variable was not present. The rows represent variables found in at least one sample.

It is important that the variables represent the same quantity in every sample. This is straightforward if distinct quantities have been measured (an example might be the intensity of a specific mass at a particular retention time), but care must be taken if the variables are 'found' in the data (for example centroid masses or the mass and retention time of an LCMS peak) since the same variable may be assigned slightly different values in different samples. Ensuring that the variables are correctly assigned is known as 'alignment' and is performed by the program as the data is imported.

Similarly, it is also important to allow for differences in the values of the variables due to known or expected changes in the data, for example different intensities of LCMS peaks due to differences in the amount injected or the response of the instrument. This is known as 'normalization' and is also performed during the import step.

If the data was obtained from known or suspected groups, the samples may be assigned to these groups for supervised analysis or to allow better visualization of the results. It is useful to be able to define different symbols for the groups so they may be easily recognized in subsequent plots and graphs.

Variations in the data can arise from several sources, for example:

1) Experimental variations due to changes in the instrument or experimental conditions
2) Variations that are real but not of interest, for example, male vs. female subjects, metabolites of a therapeutic agent, etc.
3) Relevant differences that reflect changes in the system being studied

During processing, the program allows variables of the first two types to be identified and excluded from further processing. The excluded variables are tracked so that they can later be examined or used for other processing. The program also allows 'interesting' variables to be saved for later interpretation.

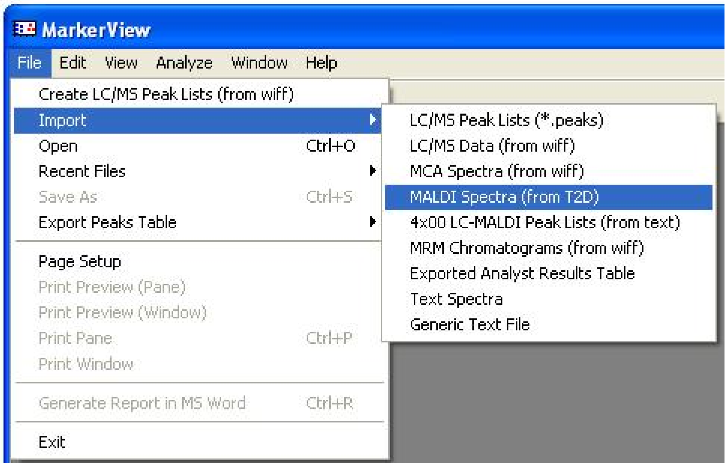
3 Supervised processing of MALDI TOF data

The data for this example consists of two sets of TOF MS spectra that were exported from the 4700 database in 'T2D' format. One set of spectra was obtained from the tryptic digest of a beta-galactosidase digest; the second set is from the same digest but spiked with a calibration standard.

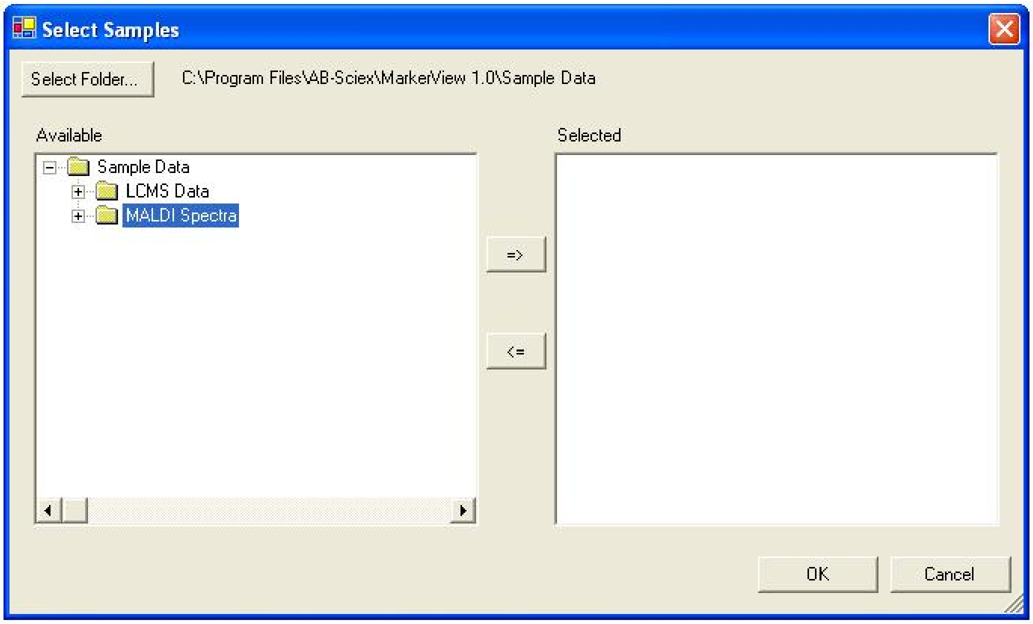
The example illustrates importing and reviewing data, and analyzing the data with a t-test.

3.1 Importing data

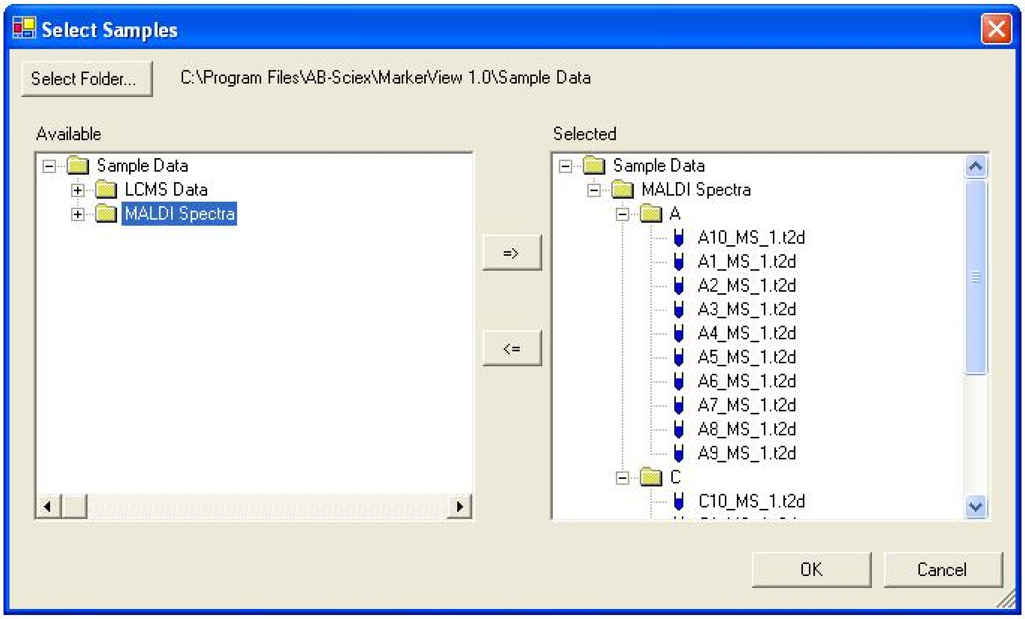
1. Select **MALDI Spectra (from T2D)** from the **File -> Import** menu.



2. In the **Select Samples** dialog box that appears, click the **Select Folder…** button and locate the folder containing the example MALDI spectra. This folder is installed to the 'AB-Sciex\MarkerView\Sample Data' subfolder of the 'Program Files' folder.


C:\Program Files\AB-Sciex\MarkerView 1.0\Sample Data

3. Select the folder 'MALDI Spectra' and click the button marked **=>**; alternatively you can drag the folder to the right side of the display marked **Selected**.



Select Samples

Select Folder... C:\Program Files\AB-Sciex\MarkerView 1.0\Sample Data

Available

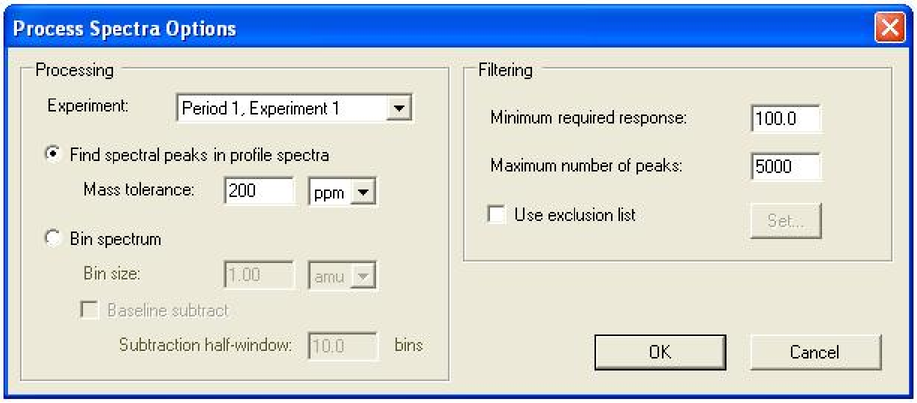
- Sample Data
  - LCMS Data
  - MALDI Spectra

Selected

- Sample Data
  - MALDI Spectra
    - A
      - A10_MS_1.t2d
      - A1_MS_1.t2d
      - A2_MS_1.t2d
      - A3_MS_1.t2d
      - A4_MS_1.t2d
      - A5_MS_1.t2d
      - A6_MS_1.t2d
      - A7_MS_1.t2d
      - A8_MS_1.t2d
      - A9_MS_1.t2d
    - C
      - C10_MS_1.t2d

4. Click **OK** to import the files on the right side of the display.

   In the **Process Spectra Options** dialog, select **Find spectral peaks in profile spectra** and enter a **Mass tolerance** of 200 ppm, set **Minimum required response** to 100, **Maximum number of peaks** to 5000 and ensure **Use exclusion list** is unchecked. Click **OK**; a dialog box will indicate the progress of the importing operation.



Process Spectra Options

Processing

Experiment: Period 1, Experiment 1

● Find spectral peaks in profile spectra

Mass tolerance: 200 ppm

○ Bin spectrum

Bin size: 1.00 amu

□ Baseline subtract

Subtraction half-window: 10.0 bins

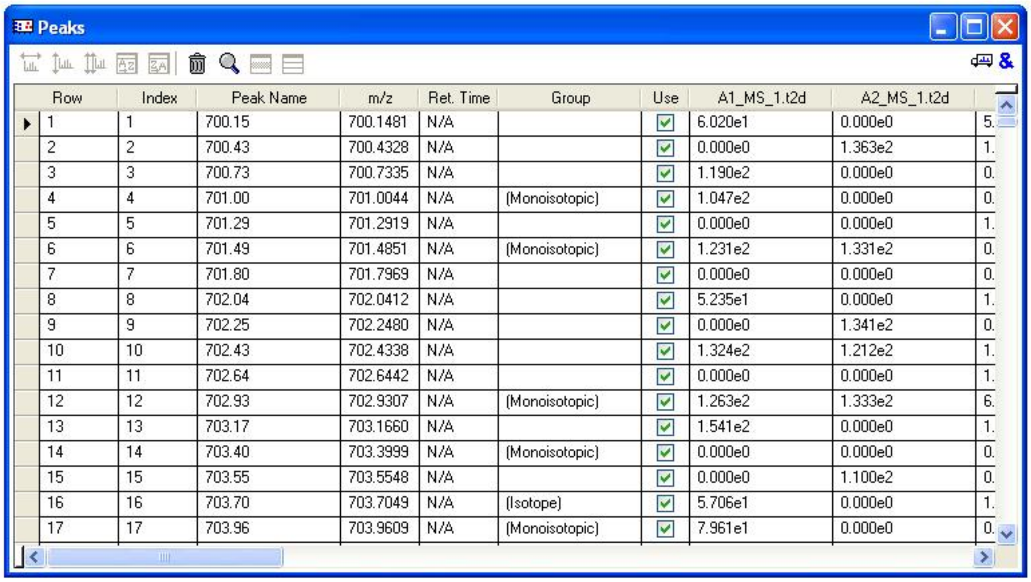
Filtering

Minimum required response: 100.0

Maximum number of peaks: 5000

□ Use exclusion list

5. When the import operation has finished, the data table will be displayed.



Row	Index	Peak Name	m/z	Ret. Time	Group	Use	A1_MS_1.t2d	A2_MS_1.t2d	
1	1	700.15	700.1481	N/A		☑	6.020e1	0.000e0	5.
2	2	700.43	700.4328	N/A		☑	0.000e0	1.363e2	1.
3	3	700.73	700.7335	N/A		☑	1.190e2	0.000e0	0.
4	4	701.00	701.0044	N/A	(Monoisotopic)	☑	1.047e2	0.000e0	0.
5	5	701.29	701.2919	N/A		☑	0.000e0	0.000e0	1.
6	6	701.49	701.4851	N/A	(Monoisotopic)	☑	1.231e2	1.331e2	0.
7	7	701.80	701.7969	N/A		☑	0.000e0	0.000e0	0.
8	8	702.04	702.0412	N/A		☑	5.235e1	0.000e0	1.
9	9	702.25	702.2480	N/A		☑	0.000e0	1.341e2	0.
10	10	702.43	702.4338	N/A		☑	1.324e2	1.212e2	1.
11	11	702.64	702.6442	N/A		☑	0.000e0	0.000e0	1.
12	12	702.93	702.9307	N/A	(Monoisotopic)	☑	1.263e2	1.333e2	6.
13	13	703.17	703.1660	N/A		☑	1.541e2	0.000e0	1.
14	14	703.40	703.3999	N/A	(Monoisotopic)	☑	0.000e0	0.000e0	0.
15	15	703.55	703.5548	N/A		☑	0.000e0	1.100e2	0.
16	16	703.70	703.7049	N/A	(Isotope)	☑	5.706e1	0.000e0	1.
17	17	703.96	703.9609	N/A	(Monoisotopic)	☑	7.961e1	0.000e0	0.

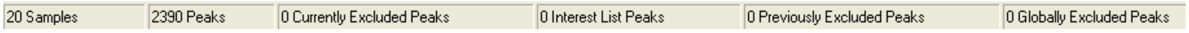
The table contains a row for each variable and a column for every sample. Variables are identified by a peak name, m/z and retention time; since these data are from mass spectra alone, there is no retention time available and the peak name is simply the m/z value.

The table also contains a column which allows the variables to be assigned to a particular group. When the application reads MS or LCMS data it attempts to determine the charge state of each variable, based on the spacing of the isotope peaks, and assigns it to one of two groups:

- Monoisotopic: charge was successfully assigned and this peak has the lowest m/z value of the isotope cluster
- Isotope: charge was assigned and this peak has a higher m/z value than the monoisotopic peak

If the group is blank then the charge state was not assigned, probably because the peak was small and no other peaks with reasonable spacing could be identified.

A status bar at the bottom of the main window indicates how many samples and peaks were read in (20 and 2390 respectively in this example).



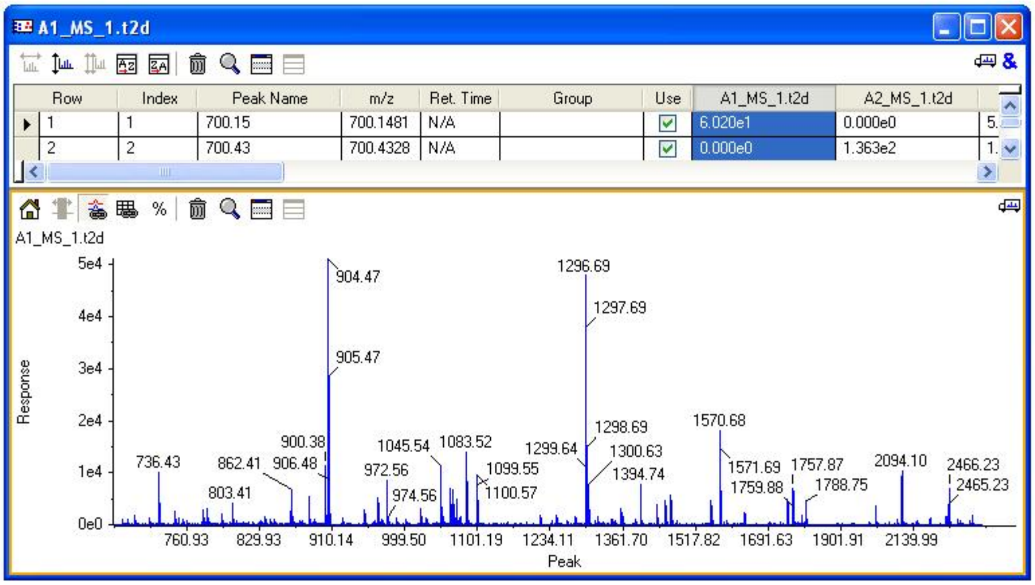
20 Samples	2390 Peaks	0 Currently Excluded Peaks	0 Interest List Peaks	0 Previously Excluded Peaks	0 Globally Excluded Peaks

The bar contains other fields that will be explained later.

3.2 Reviewing the data

You can use the controls in the toolbar at the top of the ‘Peaks’ window to graphically examine the data before performing an analysis.

1. Select any column by clicking in its title and click the **Plot Column** icon (↕). A plot of the data for that sample will appear beneath the table. Note that this is not the raw mass spectrum, but a plot of the most intense peaks found across all samples during the import process.



A1_MS_1.t2d

Row	Index	Peak Name	m/z	Ret. Time	Group	Use	A1_MS_1.t2d	A2_MS_1.t2d	
1	1	700.15	700.1481	N/A		☑	6.020e1	0.000e0	5.
2	2	700.43	700.4328	N/A		☑	0.000e0	1.363e2	1.

If you click on the **Link to Table** button (▦) in the graph header, the graph will update when different samples are selected. You can zoom or scroll the plot by dragging in the axes, as in the Analyst® Software.

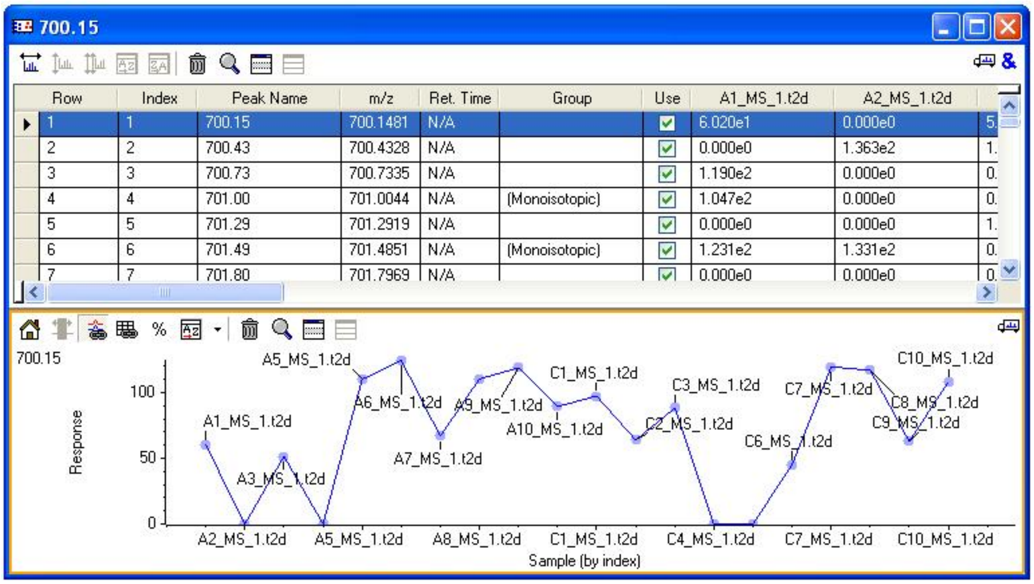
Click on the **trash can** button (🗑) in the graph header to delete the graph.

2. Select a row by clicking in the row header to the left of the row number and click the **Plot Row** icon (↔).

   The graph shows how the value of the selected variable changes for every sample in the table.

   Select the **Link to Table** button in the graph pane and then click in the table to make it active (the active pane is indicated by an orange border). Select a new row in the table and the graph will update to show the behavior of the selected variable for all samples. When the table is active, you may also use the arrow keys to change rows and quickly review the data.

   When you have finished reviewing the data, click the trash can icon to delete the graph.

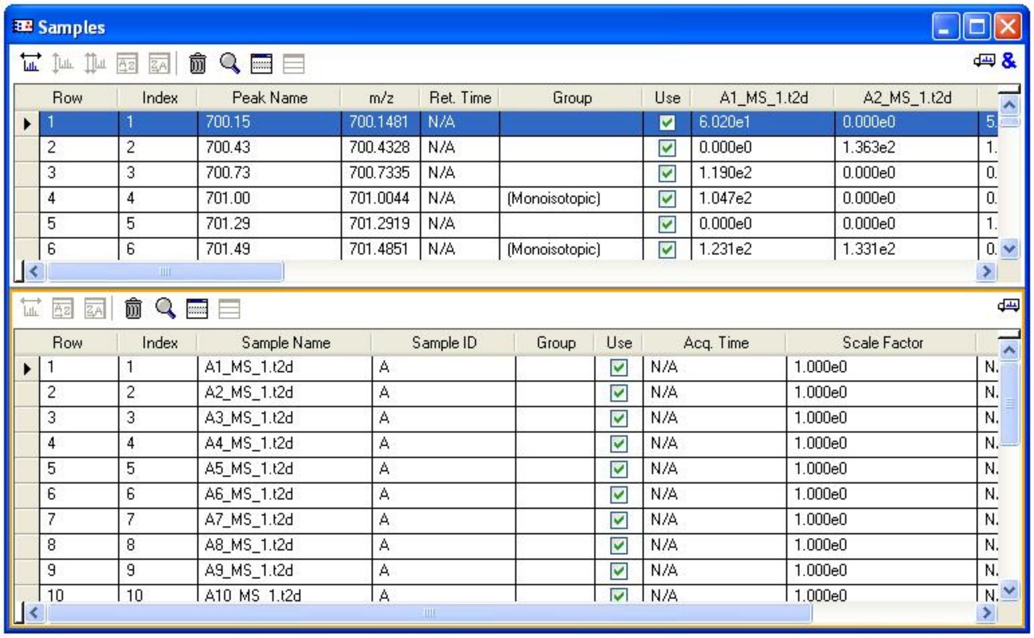


700.15

Row	Index	Peak Name	m/z	Ret. Time	Group	Use	A1_MS_1.t2d	A2_MS_1.t2d	
1	1	700.15	700.1481	N/A		☑	6.020e1	0.000e0	5.
2	2	700.43	700.4328	N/A		☑	0.000e0	1.363e2	1.
3	3	700.73	700.7335	N/A		☑	1.190e2	0.000e0	0.
4	4	701.00	701.0044	N/A	(Monoisotopic)	☑	1.047e2	0.000e0	0.
5	5	701.29	701.2919	N/A		☑	0.000e0	0.000e0	1.
6	6	701.49	701.4851	N/A	(Monoisotopic)	☑	1.231e2	1.331e2	0.
7	7	701.80	701.7969	N/A		☑	0.000e0	0.000e0	0.

3.3 Reviewing the samples and assigning groups

1. Select **Show Samples Table** from the **View** menu. The Sample table will be displayed below the Peaks table.



	Row	Index	Peak Name	m/z	Ret. Time	Group	Use	A1_MS_1.t2d	A2_MS_1.t2d	
▶	1	1	700.15	700.1481	N/A		☑	6.020e1	0.000e0	5.
	2	2	700.43	700.4328	N/A		☑	0.000e0	1.363e2	1.
	3	3	700.73	700.7335	N/A		☑	1.190e2	0.000e0	0.
	4	4	701.00	701.0044	N/A	(Monoisotopic)	☑	1.047e2	0.000e0	0.
	5	5	701.29	701.2919	N/A		☑	0.000e0	0.000e0	1.
	6	6	701.49	701.4851	N/A	(Monoisotopic)	☑	1.231e2	1.331e2	0.

	Row	Index	Sample Name	Sample ID	Group	Use	Acq. Time	Scale Factor	
▶	1	1	A1_MS_1.t2d	A		☑	N/A	1.000e0	N.
	2	2	A2_MS_1.t2d	A		☑	N/A	1.000e0	N.
	3	3	A3_MS_1.t2d	A		☑	N/A	1.000e0	N.
	4	4	A4_MS_1.t2d	A		☑	N/A	1.000e0	N.
	5	5	A5_MS_1.t2d	A		☑	N/A	1.000e0	N.
	6	6	A6_MS_1.t2d	A		☑	N/A	1.000e0	N.
	7	7	A7_MS_1.t2d	A		☑	N/A	1.000e0	N.
	8	8	A8_MS_1.t2d	A		☑	N/A	1.000e0	N.
	9	9	A9_MS_1.t2d	A		☑	N/A	1.000e0	N.
	10	10	A10_MS_1.t2d	A		☑	N/A	1.000e0	N.

The table contains a row for each sample with columns indicating if the sample is to be used in subsequent processing (**Use**), the scale factor for this sample, the associated group and other optional information.

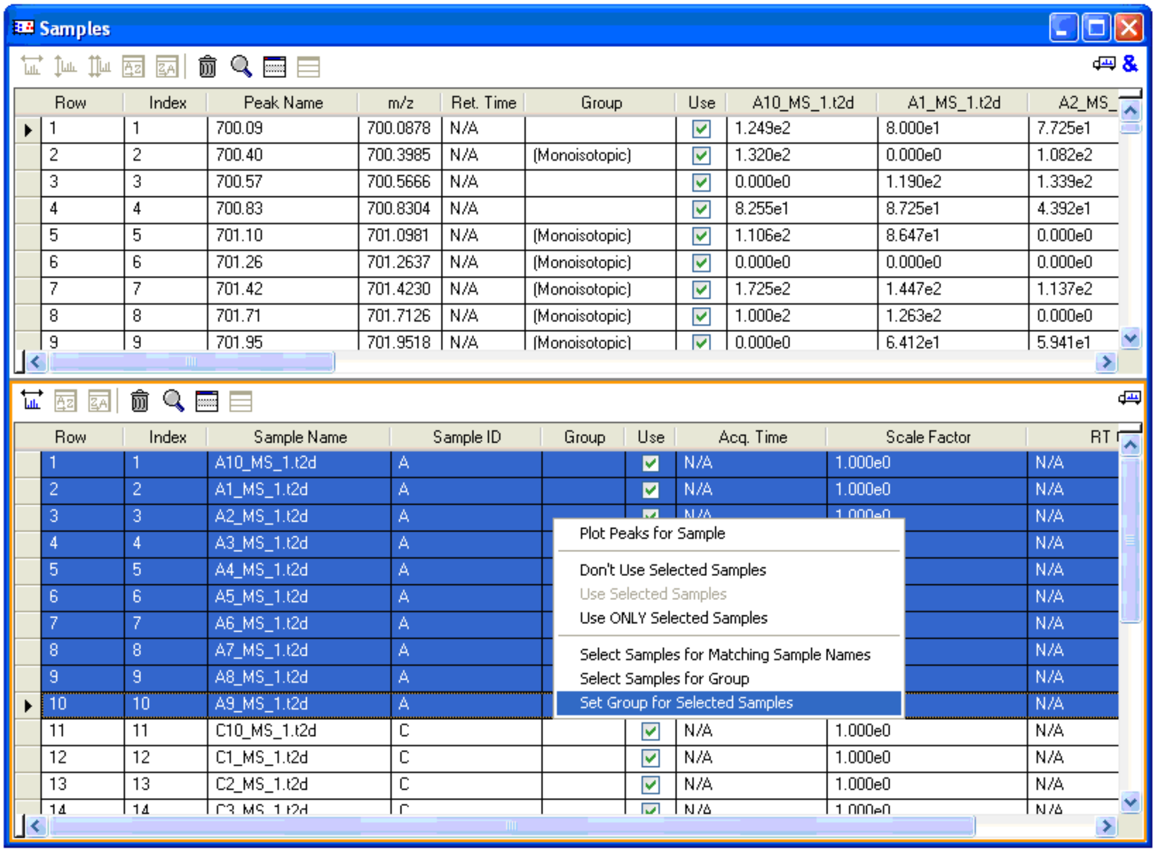
The **Sample ID** is obtained from the data file, as is the **Acquisition Time** if the data are being imported from a .wiff file.

The **Scale Factor** can be adjusted to allow for overall differences in the amount of sample used, and the **RT Correction** is used to adjust the retention time in LCMS analyses.

The **Group** information is used in supervised analyses and to select plotting symbols for the samples so that differences are more apparent.

2. Select all of the rows for samples with names starting 'A' by dragging in the row headers (to the left of the **Row** column), right-click and select **Set Group for Selected Samples**. In this case, all of the samples in group A are beta galactosidase tryptic digests with calibrant spiked at a particular level.

   (Note: if the Sample ID column contains the group information, you may quickly copy it to the Group column by clicking the column heading, hitting ctrl-C to copy the column, selecting the Group column and hitting ctrl-V).

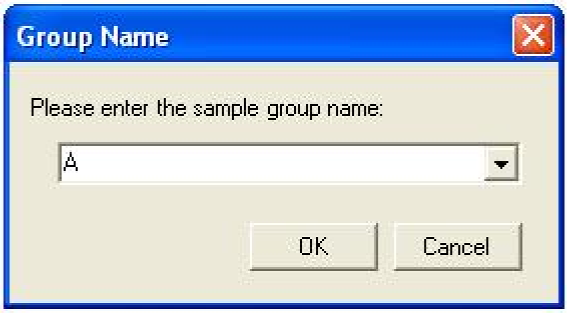


Row	Index	Peak Name	m/z	Ret. Time	Group	Use	A10_MS_1.t2d	A1_MS_1.t2d	A2_MS_
1	1	700.09	700.0878	N/A		☑	1.249e2	8.000e1	7.725e1
2	2	700.40	700.3985	N/A	(Monoisotopic)	☑	1.320e2	0.000e0	1.082e2
3	3	700.57	700.5666	N/A		☑	0.000e0	1.190e2	1.339e2
4	4	700.83	700.8304	N/A		☑	8.255e1	8.725e1	4.392e1
5	5	701.10	701.0981	N/A	(Monoisotopic)	☑	1.106e2	8.647e1	0.000e0
6	6	701.26	701.2637	N/A	(Monoisotopic)	☑	0.000e0	0.000e0	0.000e0
7	7	701.42	701.4230	N/A	(Monoisotopic)	☑	1.725e2	1.447e2	1.137e2
8	8	701.71	701.7126	N/A	(Monoisotopic)	☑	1.000e2	1.263e2	0.000e0
9	9	701.95	701.9518	N/A	(Monoisotopic)	☑	0.000e0	6.412e1	5.941e1

Row	Index	Sample Name	Sample ID	Group	Use	Acq. Time	Scale Factor	RT
1	1	A10_MS_1.t2d	A		☑	N/A	1.000e0	N/A
2	2	A1_MS_1.t2d	A		☑	N/A	1.000e0	N/A
3	3	A2_MS_1.t2d	A		☑	N/A	1.000e0	N/A
4	4	A3_MS_1.t2d	A					N/A
5	5	A4_MS_1.t2d	A					N/A
6	6	A5_MS_1.t2d	A					N/A
7	7	A6_MS_1.t2d	A					N/A
8	8	A7_MS_1.t2d	A					N/A
9	9	A8_MS_1.t2d	A					N/A
10	10	A9_MS_1.t2d	A					N/A
11	11	C10_MS_1.t2d	C		☑	N/A	1.000e0	N/A
12	12	C1_MS_1.t2d	C		☑	N/A	1.000e0	N/A
13	13	C2_MS_1.t2d	C		☑	N/A	1.000e0	N/A
14	14	C3_MS_1.t2d	C		☑	N/A	1.000e0	N/A

Plot Peaks for Sample
Don't Use Selected Samples
Use Selected Samples
Use ONLY Selected Samples
Select Samples for Matching Sample Names
Select Samples for Group
Set Group for Selected Samples

3. In the resulting **Group Name** dialog enter ‘A’ for the group name and press **OK**.



Please enter the sample group name:

A

4. Select all of the samples with names starting with ‘C’ and repeat the process, assigning ‘C’ as the group name. The samples in group C are beta galactosidase tryptic digests with calibrant spiked at a lower level than group A.

5. Click the trash can icon in the sample table to remove it from the display.

3.4 Saving the data for later retrieval

It is often useful to save the imported data so that it can be reprocessed later without having to re-import it since importing may be slow if there are many complex samples. The group information you have just entered in the previous section will also be saved with the data.

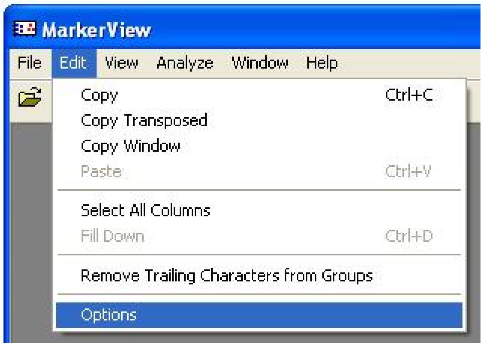
1. From the File menu select **Save As**, select a folder to save the data, enter a name and click **OK**. The data will be saved in a file with the extension ‘mrkvw’. In this example the file name is ‘Saved’.

2. To retrieve the data later, select **Open** from the **File** menu, locate the appropriate file and click **OK**.

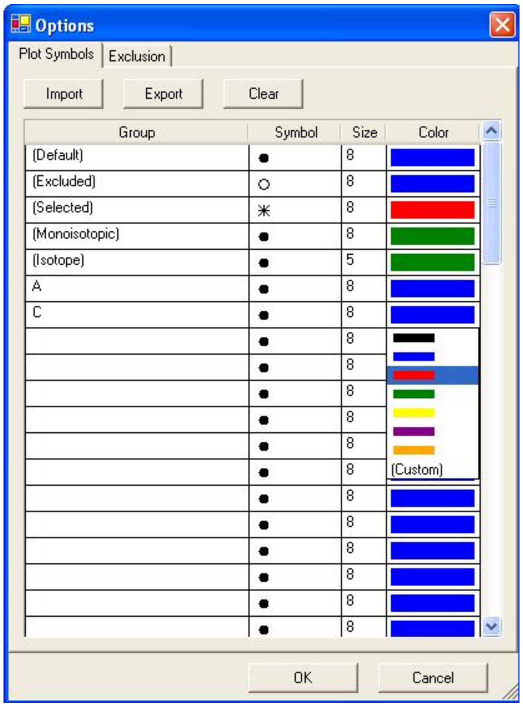
3.5 Assigning a symbol for the groups

The results are easier to visualize if a unique symbol is associated with each group.

1. From the **Edit** menu select **Options.**



In the Options dialog box, select the **Plot Symbols** tab if it is not already selected.



Group	Symbol	Size	Color
(Default)	●	8	🟦
(Excluded)	○	8	🟦
(Selected)	✳	8	🟥
(Monoisotopic)	●	8	🟩
(Isotope)	●	5	🟩
A	●	8	🟦
C	●	8	🟦
	●	8	
	●	8	
	●	8	
	●	8	
	●	8	
	●	8	
	●	8	🟦
	●	8	🟦
	●	8	🟦
	●	8	🟦
	●	8	🟦
	●	8	🟦

Select the first empty cell in the **Sample Group** column and enter 'A'. Enter 'C' in the **Sample Group** cell in the next row.

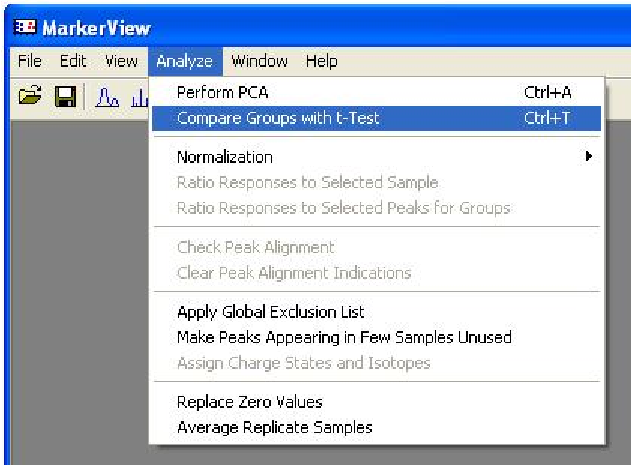
You may change the shape, size and color of the symbol by clicking in the appropriate cell and making a selection from the drop-down menu. Click in the color cell for the C row just added and select the red color. All graphs and plots that show samples will now use filled blue circles for group A and filled red circles for group C.

Note that there are five special categories – Default, Excluded, Selected, Monoisotopic and Isotope – that are used when no other symbol is defined, for excluded samples and variables, when particular samples are selected, or to indicate isotope peaks respectively. You cannot change the names of these symbols but you can edit the shape, size and color.

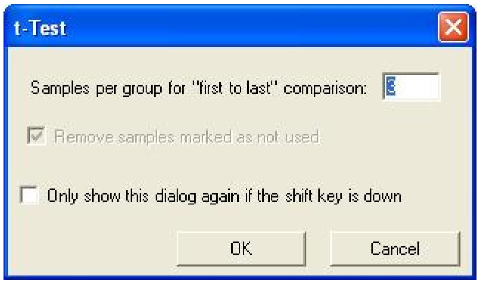
3.6 Performing a t-test

The t-test is applied to every variable in the table and determines if the mean for each group is significantly different given the standard deviation and the number of samples.

1. From the **Analyze** menu, select **Compare Groups with t-Test** or click on the t-test button in the toolbar below the menu bar (t).


MarkerView
File
Edit
View
Analyze
Window
Help
Perform PCA Ctrl+A
Compare Groups with t-Test Ctrl+T
Normalization
Ratio Responses to Selected Sample
Ratio Responses to Selected Peaks for Groups
Check Peak Alignment
Clear Peak Alignment Indications
Apply Global Exclusion List
Make Peaks Appearing in Few Samples Unused
Assign Charge States and Isotopes
Replace Zero Values
Average Replicate Samples

The following dialog appears:

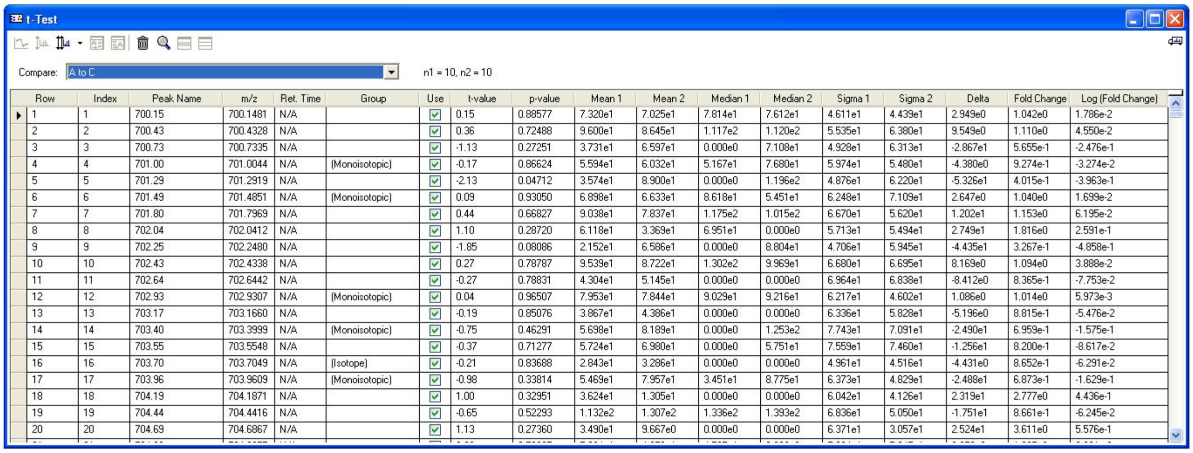

t-Test

Samples per group for "first to last" comparison: 3

☑ Remove samples marked as not used

□ Only show this dialog again if the shift key is down

2. Click **OK**.



Compare: A to C

n1 = 10, n2 = 10

Row	Index	Peak Name	m/z	Ret. Time	Group	Use	t-value	p-value	Mean 1	Mean 2	Median 1	Median 2	Sigma 1	Sigma 2	Delta	Fold Change	Log (Fold Change)
1	1	700.15	700.1481	N/A		☑	0.15	0.88577	7.320e1	7.025e1	7.814e1	7.612e1	4.611e1	4.439e1	2.949e0	1.042e0	1.786e-2
2	2	700.43	700.4328	N/A		☑	0.36	0.72488	9.600e1	8.645e1	1.117e2	1.120e2	5.535e1	6.380e1	9.549e0	1.110e0	4.550e-2
3	3	700.73	700.7335	N/A		☑	-1.13	0.27251	3.731e1	6.597e1	0.000e0	7.108e1	4.928e1	6.313e1	-2.867e1	5.655e-1	-2.476e-1
4	4	701.00	701.0044	N/A	(Monoisotopic)	☑	-0.17	0.86624	5.594e1	6.032e1	5.167e1	7.680e1	5.974e1	5.480e1	-4.380e0	9.274e-1	-3.274e-2
5	5	701.29	701.2919	N/A		☑	-2.13	0.04712	3.574e1	8.900e1	0.000e0	1.196e2	4.876e1	6.220e1	-5.326e1	4.015e-1	-3.963e-1
6	6	701.49	701.4851	N/A	(Monoisotopic)	☑	0.09	0.93050	6.898e1	6.633e1	8.618e1	5.451e1	6.248e1	7.109e1	2.647e0	1.040e0	1.699e-2
7	7	701.80	701.7969	N/A		☑	0.44	0.66827	9.038e1	7.837e1	1.175e2	1.015e2	6.670e1	5.620e1	1.202e1	1.153e0	6.195e-2
8	8	702.04	702.0412	N/A		☑	1.10	0.28720	6.118e1	3.369e1	6.951e1	0.000e0	5.713e1	5.494e1	2.749e1	1.816e0	2.591e-1
9	9	702.25	702.2480	N/A		☑	-1.85	0.08086	2.152e1	6.586e1	0.000e0	8.804e1	4.706e1	5.945e1	-4.435e1	3.267e-1	-4.858e-1
10	10	702.43	702.4338	N/A		☑	0.27	0.78787	9.539e1	8.722e1	1.302e2	9.969e1	6.680e1	6.695e1	8.169e0	1.094e0	3.888e-2
11	11	702.64	702.6442	N/A		☑	-0.27	0.78831	4.304e1	5.145e1	0.000e0	0.000e0	6.964e1	6.838e1	-8.412e0	8.365e-1	-7.753e-2
12	12	702.93	702.9307	N/A	(Monoisotopic)	☑	0.04	0.96507	7.953e1	7.844e1	9.029e1	9.216e1	6.217e1	4.602e1	1.086e0	1.014e0	5.973e-3
13	13	703.17	703.1660	N/A		☑	-0.19	0.85076	3.867e1	4.386e1	0.000e0	0.000e0	6.336e1	5.828e1	-5.196e0	8.815e-1	-5.476e-2
14	14	703.40	703.3999	N/A	(Monoisotopic)	☑	-0.75	0.46291	5.698e1	8.189e1	0.000e0	1.253e2	7.743e1	7.091e1	-2.490e1	6.959e-1	-1.575e-1
15	15	703.55	703.5548	N/A		☑	-0.37	0.71277	5.724e1	6.980e1	0.000e0	5.751e1	7.559e1	7.460e1	-1.256e1	8.200e-1	-8.617e-2
16	16	703.70	703.7049	N/A	(Isotope)	☑	-0.21	0.83688	2.843e1	3.286e1	0.000e0	0.000e0	4.961e1	4.516e1	-4.431e0	8.652e-1	-6.291e-2
17	17	703.96	703.9609	N/A	(Monoisotopic)	☑	-0.98	0.33814	5.469e1	7.957e1	3.451e1	8.775e1	6.373e1	4.829e1	-2.488e1	6.873e-1	-1.629e-1
18	18	704.19	704.1871	N/A		☑	1.00	0.32951	3.624e1	1.305e1	0.000e0	0.000e0	6.042e1	4.126e1	2.319e1	2.777e0	4.436e-1
19	19	704.44	704.4416	N/A		☑	-0.65	0.52293	1.132e2	1.307e2	1.336e2	1.393e2	6.836e1	5.050e1	-1.751e1	8.661e-1	-6.245e-2
20	20	704.69	704.6867	N/A		☑	1.13	0.27360	3.490e1	9.667e0	0.000e0	0.000e0	6.371e1	3.057e1	2.524e1	3.611e0	5.576e-1

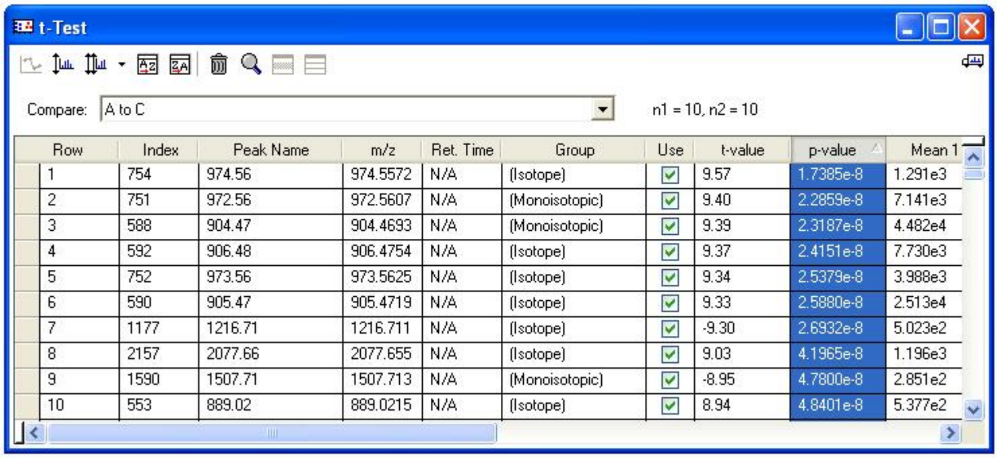
The program automatically compares all groups in pairs and each group to all of the others; the comparisons are accessed through the combo-box labeled ‘Compare’ at the top of the table. In this case there are only two groups so the comparison is selected and the resulting table displayed. The number of samples in each group (10 in both here) is also displayed.

For every variable the table displays the calculated t-value, the corresponding p-value and various metrics for both groups such as the mean (Mean 1, Mean2), the median, the difference between the means (Delta), the fold change and the log of the fold change.

The t-value is a measure of how well the variable distinguishes the two groups whereas the p-value is the probability that the delta value would occur by chance. If the value of t exceeds a calculated critical value then the variable does distinguish the groups with some confidence value; t can be positive or negative depending on the direction of the subtraction. The p-value is always positive and the smaller the value the lower the probability that this is a chance occurrence.

3.7 Reviewing the results

1. Click in the heading of the p-value column and click the **Ascending sort** button ().

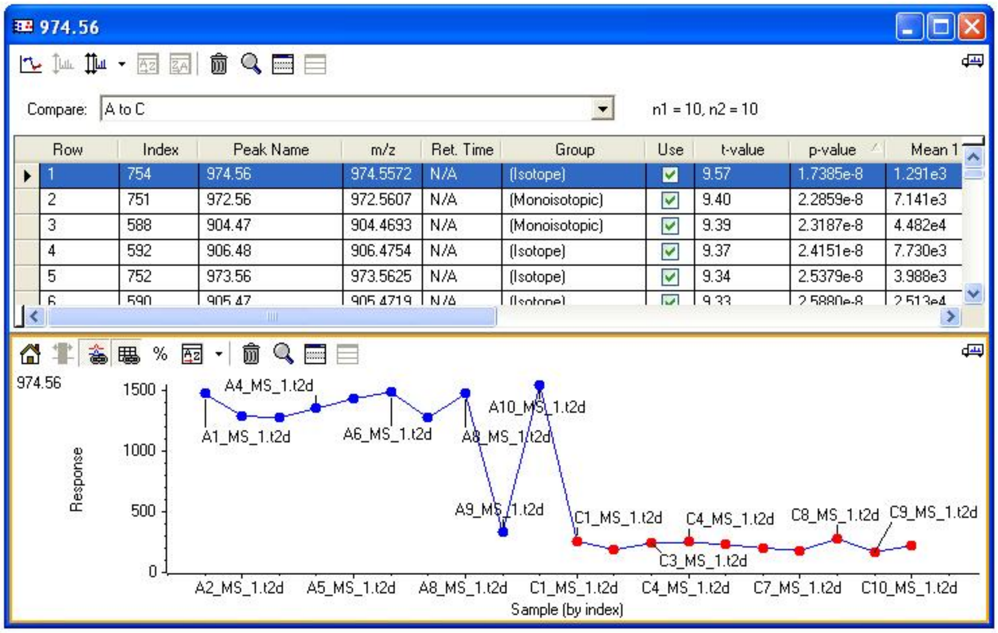


t-Test

Compare: A to C   n1 = 10, n2 = 10

Row	Index	Peak Name	m/z	Ret. Time	Group	Use	t-value	p-value	Mean 1
1	754	974.56	974.5572	N/A	(Isotope)	☑	9.57	1.7385e-8	1.291e3
2	751	972.56	972.5607	N/A	(Monoisotopic)	☑	9.40	2.2859e-8	7.141e3
3	588	904.47	904.4693	N/A	(Monoisotopic)	☑	9.39	2.3187e-8	4.482e4
4	592	906.48	906.4754	N/A	(Isotope)	☑	9.37	2.4151e-8	7.730e3
5	752	973.56	973.5625	N/A	(Isotope)	☑	9.34	2.5379e-8	3.988e3
6	590	905.47	905.4719	N/A	(Isotope)	☑	9.33	2.5880e-8	2.513e4
7	1177	1216.71	1216.711	N/A	(Isotope)	☑	-9.30	2.6932e-8	5.023e2
8	2157	2077.66	2077.655	N/A	(Isotope)	☑	9.03	4.1965e-8	1.196e3
9	1590	1507.71	1507.713	N/A	(Monoisotopic)	☑	-8.95	4.7800e-8	2.851e2
10	553	889.02	889.0215	N/A	(Isotope)	☑	8.94	4.8401e-8	5.377e2

2. Select the first row of the t-test table by clicking in the area to the left of row 1, and click the **Plot profile** button ().



974.56

Compare: A to C   n1 = 10, n2 = 10

Row	Index	Peak Name	m/z	Ret. Time	Group	Use	t-value	p-value	Mean 1
1	754	974.56	974.5572	N/A	(Isotope)	☑	9.57	1.7385e-8	1.291e3
2	751	972.56	972.5607	N/A	(Monoisotopic)	☑	9.40	2.2859e-8	7.141e3
3	588	904.47	904.4693	N/A	(Monoisotopic)	☑	9.39	2.3187e-8	4.482e4
4	592	906.48	906.4754	N/A	(Isotope)	☑	9.37	2.4151e-8	7.730e3
5	752	973.56	973.5625	N/A	(Isotope)	☑	9.34	2.5379e-8	3.988e3
6	590	905.47	905.4719	N/A	(Isotope)	☑	9.33	2.5880e-8	2.513e4

The resulting display shows how the value of the selected variable changes across all the samples (the profile). Since the data points are also labeled with the symbol defined in section 3.5, it is clear that this variable is indeed different for the two groups, and higher in group A.

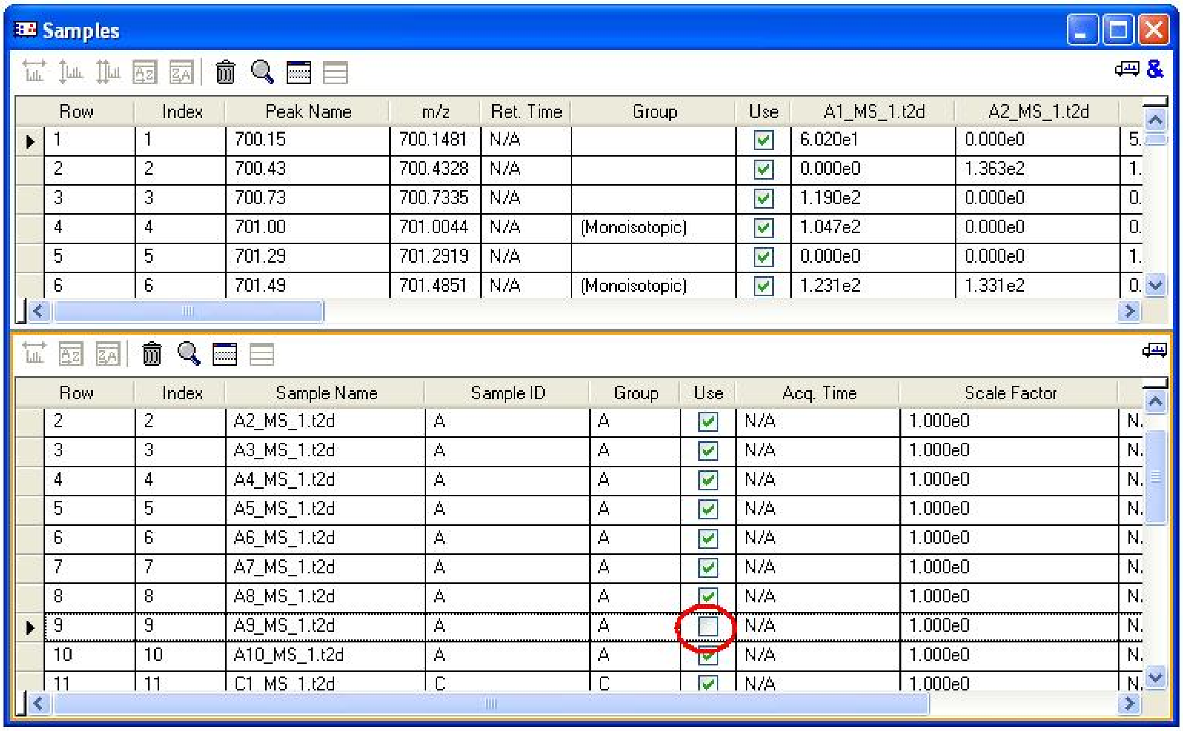
The peaks with nominal mass of 974, 1298, 1507, *etc.* with high probabilities are in fact peaks from the spiked calibration standard.

This graph is automatically locked to the table, so clicking in another row, or using the arrow keys when the table is active, will cause the graph to update to reflect the behavior of the new variable.

Note that the display reveals an anomalous sample – the first sample at the lower intensity level is labeled as an A sample (A9_MS_1.t2d) even though its behavior is more similar to group C samples. Apparently there is a problem with this sample, or the name. Removing the sample from future calculations will help to ensure that the values are correctly calculated.

3.8 Inactivating a sample

1. Close all windows except the initial Peaks window that tabulates the data.
2. From the View menu select **Show Samples Table**.
3. In the Samples table, locate the row containing sample A9_MS_1 and click the check box in the **Use** column so that it is unchecked.



	Row	Index	Peak Name	m/z	Ret. Time	Group	Use	A1_MS_1.t2d	A2_MS_1.t2d	
▶	1	1	700.15	700.1481	N/A		☑	6.020e1	0.000e0	5.
	2	2	700.43	700.4328	N/A		☑	0.000e0	1.363e2	1.
	3	3	700.73	700.7335	N/A		☑	1.190e2	0.000e0	0.
	4	4	701.00	701.0044	N/A	(Monoisotopic)	☑	1.047e2	0.000e0	0.
	5	5	701.29	701.2919	N/A		☑	0.000e0	0.000e0	1.
	6	6	701.49	701.4851	N/A	(Monoisotopic)	☑	1.231e2	1.331e2	0.

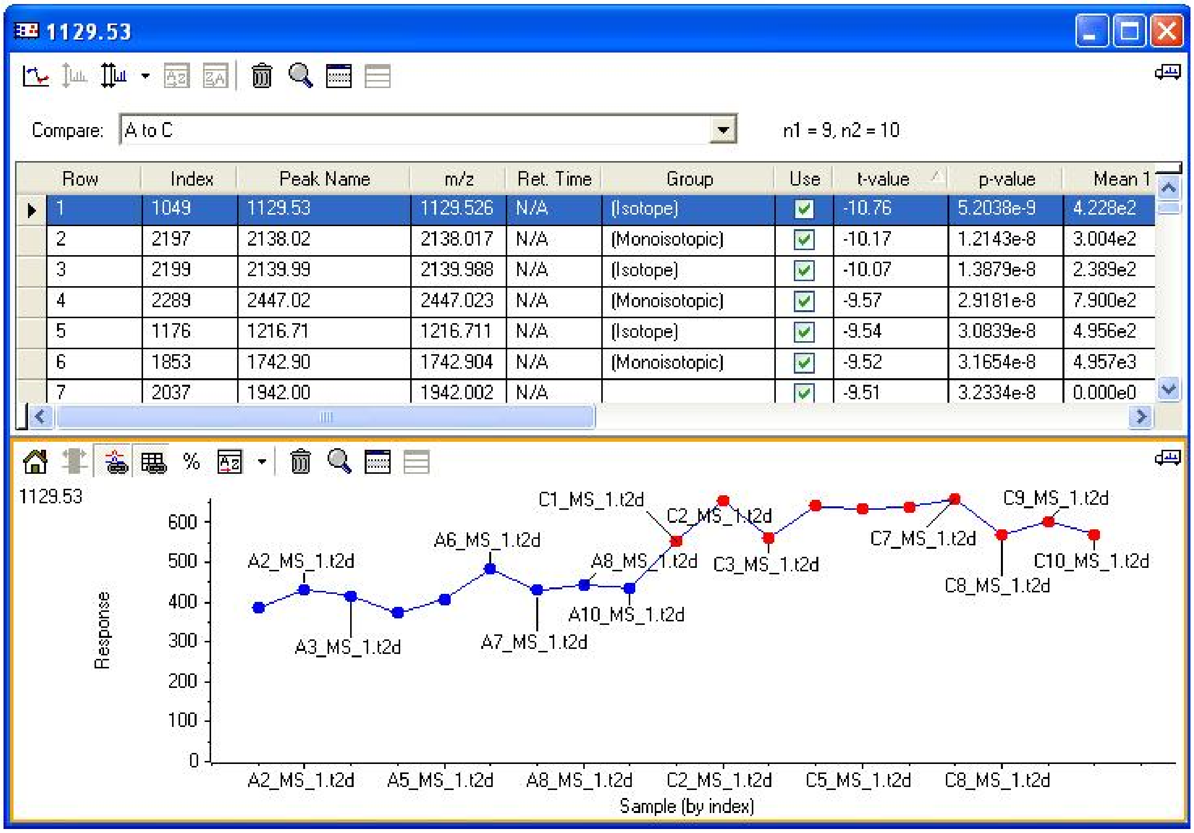
	Row	Index	Sample Name	Sample ID	Group	Use	Acq. Time	Scale Factor	
	2	2	A2_MS_1.t2d	A	A	☑	N/A	1.000e0	N.
	3	3	A3_MS_1.t2d	A	A	☑	N/A	1.000e0	N.
	4	4	A4_MS_1.t2d	A	A	☑	N/A	1.000e0	N.
	5	5	A5_MS_1.t2d	A	A	☑	N/A	1.000e0	N.
	6	6	A6_MS_1.t2d	A	A	☑	N/A	1.000e0	N.
	7	7	A7_MS_1.t2d	A	A	☑	N/A	1.000e0	N.
	8	8	A8_MS_1.t2d	A	A	☑	N/A	1.000e0	N.
▶	9	9	A9_MS_1.t2d	A	A	☐	N/A	1.000e0	N.
	10	10	A10_MS_1.t2d	A	A	☑	N/A	1.000e0	N.
	11	11	C1_MS_1.t2d	C	C	☑	N/A	1.000e0	N.

4. Repeat the t-test, ensuring that **Remove samples marked as not used** is checked, sort the results and regenerate the profile graph using the plot row icon to verify that the sample is no longer part of the display. (Zoom the graph if necessary by dragging in the horizontal axis). Sort the table by ascending p-value as before and note that the first value is now much lower.

   If **Remove samples marked as not used** is unchecked in the t-test dialog, excluded samples will not be used to calculate the t-test values but will be retained in the displays. This provides a way to classify unknown samples, i.e. compare them to known samples.

5. Sort the table by ascending t-value, select the first row and display the profile graph.

   The variables with negative t values seem to be higher in the C samples than in the A group, i.e. both groups appear to contain unique variables, not just the samples spiked with calibrant. In this case, it seems likely that this is an experimental variation – for example suppression of some peaks by the spiked compounds – so that they appear to be less intense.



1129.53

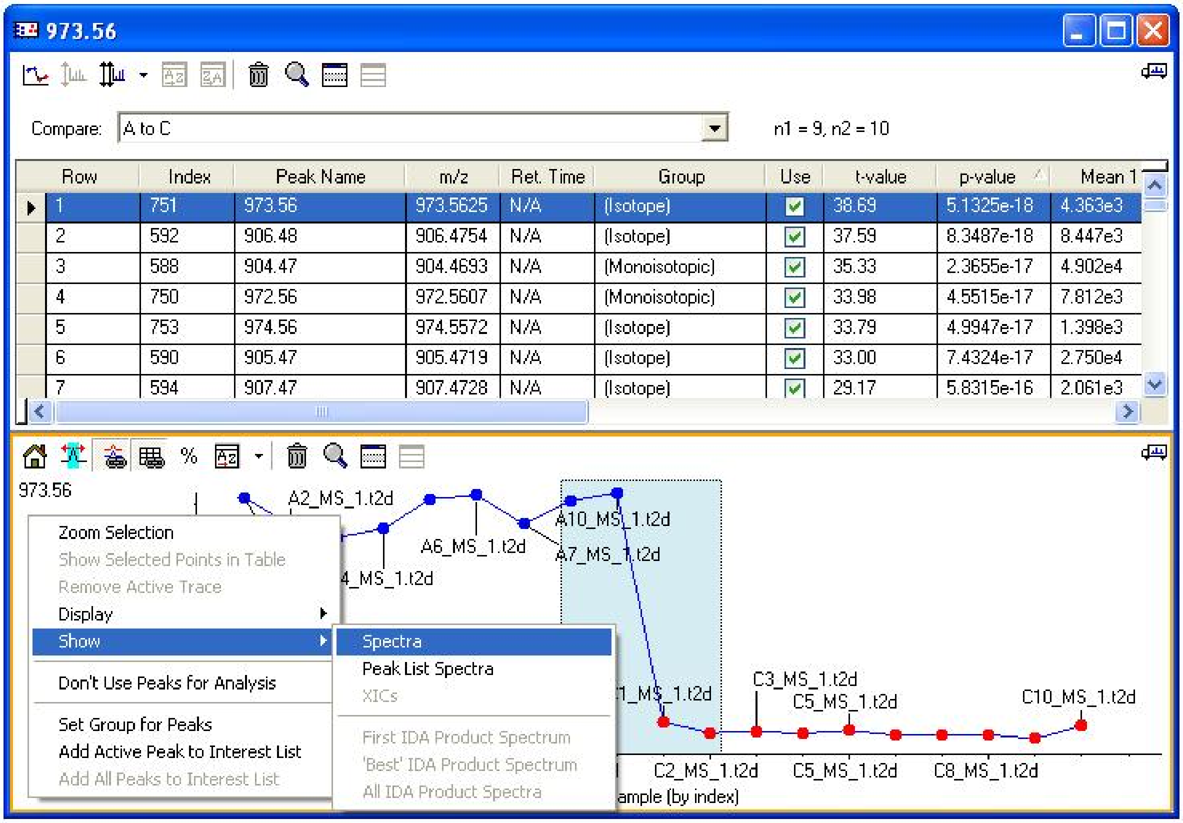
Compare: A to C

n1 = 9, n2 = 10

Row	Index	Peak Name	m/z	Ret. Time	Group	Use	t-value	p-value	Mean 1
1	1049	1129.53	1129.526	N/A	(Isotope)	☑	-10.76	5.2038e-9	4.228e2
2	2197	2138.02	2138.017	N/A	(Monoisotopic)	☑	-10.17	1.2143e-8	3.004e2
3	2199	2139.99	2139.988	N/A	(Isotope)	☑	-10.07	1.3879e-8	2.389e2
4	2289	2447.02	2447.023	N/A	(Monoisotopic)	☑	-9.57	2.9181e-8	7.900e2
5	1176	1216.71	1216.711	N/A	(Isotope)	☑	-9.54	3.0839e-8	4.956e2
6	1853	1742.90	1742.904	N/A	(Monoisotopic)	☑	-9.52	3.1654e-8	4.957e3
7	2037	1942.00	1942.002	N/A		☑	-9.51	3.2334e-8	0.000e0

3.9 Reviewing spectra

1. Sort the table in ascending p-value order, and select the first row to generate the profile graph using the plot profile icon.
2. Click and drag in the graph so that a few samples on both sides of the sharp intensity change are selected.



973.56

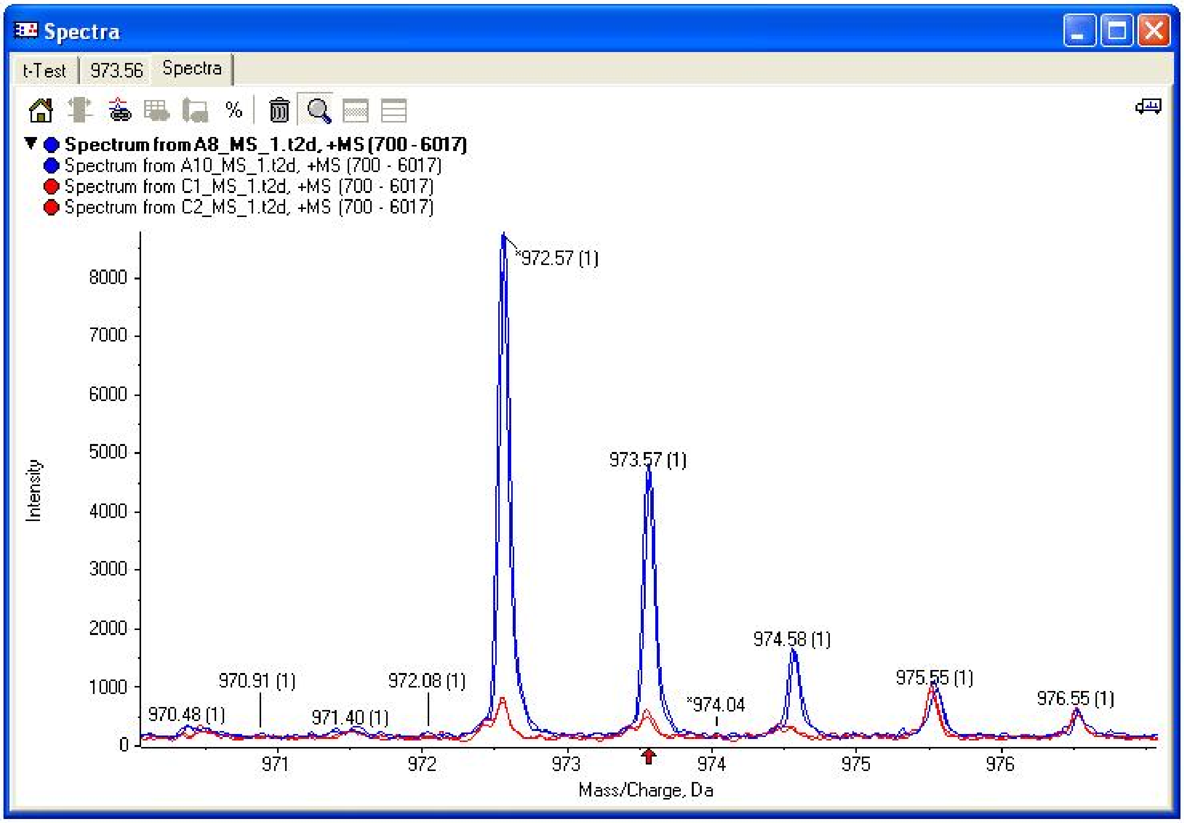
Compare: A to C

n1 = 9, n2 = 10

Row	Index	Peak Name	m/z	Ret. Time	Group	Use	t-value	p-value	Mean 1
1	751	973.56	973.5625	N/A	(Isotope)	☑	38.69	5.1325e-18	4.363e3
2	592	906.48	906.4754	N/A	(Isotope)	☑	37.59	8.3487e-18	8.447e3
3	588	904.47	904.4693	N/A	(Monoisotopic)	☑	35.33	2.3655e-17	4.902e4
4	750	972.56	972.5607	N/A	(Monoisotopic)	☑	33.98	4.5515e-17	7.812e3
5	753	974.56	974.5572	N/A	(Isotope)	☑	33.79	4.9947e-17	1.398e3
6	590	905.47	905.4719	N/A	(Isotope)	☑	33.00	7.4324e-17	2.750e4
7	594	907.47	907.4728	N/A	(Isotope)	☑	29.17	5.8315e-16	2.061e3

3. Right-click in the plot and select **Spectra** from the **Show** submenu.

   A progress bar will appear while the program locates the raw data files and extracts the spectra, followed by a graph showing the spectra.



● **Spectrum from A8_MS_1.t2d, +MS (700 - 6017)**
● Spectrum from A10_MS_1.t2d, +MS (700 - 6017)
● Spectrum from C1_MS_1.t2d, +MS (700 - 6017)
● Spectrum from C2_MS_1.t2d, +MS (700 - 6017)

The graph is zoomed so that the selected peak is centered in the display. In this case it is clear that the reported difference is real, i.e. that the peak at 972.57 is intense in samples from group A and much less intense in those from group C. For this figure the **Use Group Colors for Traces** option (from the **Display** submenu of the graph's context menu) was selected so that the group color is used for each trace, rather than a different color for each – this makes it easier to tell at a glance to which group a given spectrum belongs.

You can click on the magnifying glass icon (🔍) to make the pane containing the spectra fill the display windows. When you have finished examining the data, click the icon again to return to the normal display.

3.10 Summary

In this section you have:

- Imported a set of MALDI TOF spectra and reviewed the sample and variable data
- Reviewed the samples, assigned them to groups and created a symbol for each group
- Performed a t-test to determine how well each variable distinguishes the two groups
- Reviewed the behavior of certain variables for all samples
- Used the raw data to confirm a difference between groups
- Detected the presence of a suspicious sample and deactivated it from further calculations

These steps are the basis of all data processing in the application, and many of the operations are common regardless of the data and the type of analysis.

The next section shows how the same data can be reviewed using unsupervised techniques to confirm or identify groups, detect outliers, etc. You may also want to look at section 6.4, Selecting discriminating t-test variables, to see additional ways of determining variables that best distinguish the groups.

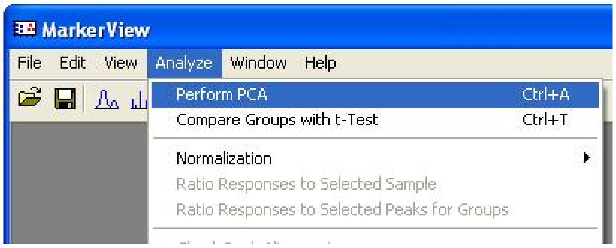
4 Reviewing the data with PCA

In this section you will learn how to review the data using an unsupervised technique – Principal Components Analysis (PCA).

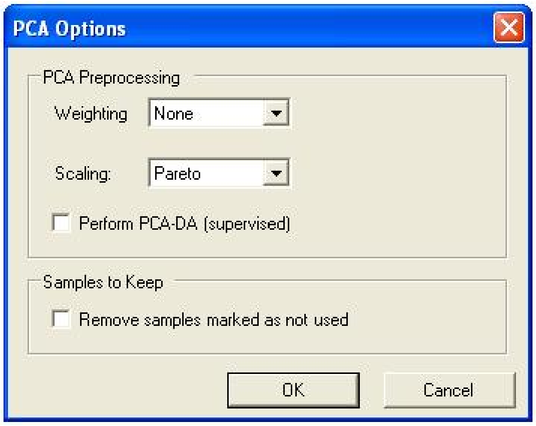
Close any open windows and then open the data table you saved in section 3.4 by selecting **File -> Open** and locating the saved data file. The data table will be displayed.

4.1 Performing Principal Components Analysis

1. Select **Perform PCA** from the **Analyze** menu:


MarkerView
File
Edit
View
Analyze
Window
Help
Perform PCA Ctrl+A
Compare Groups with t-Test Ctrl+T
Normalization
Ratio Responses to Selected Sample
Ratio Responses to Selected Peaks for Groups

The options dialog box will appear



PCA Options

PCA Preprocessing

Weighting: None

Scaling: Pareto

☐ Perform PCA-DA (supervised)

Samples to Keep

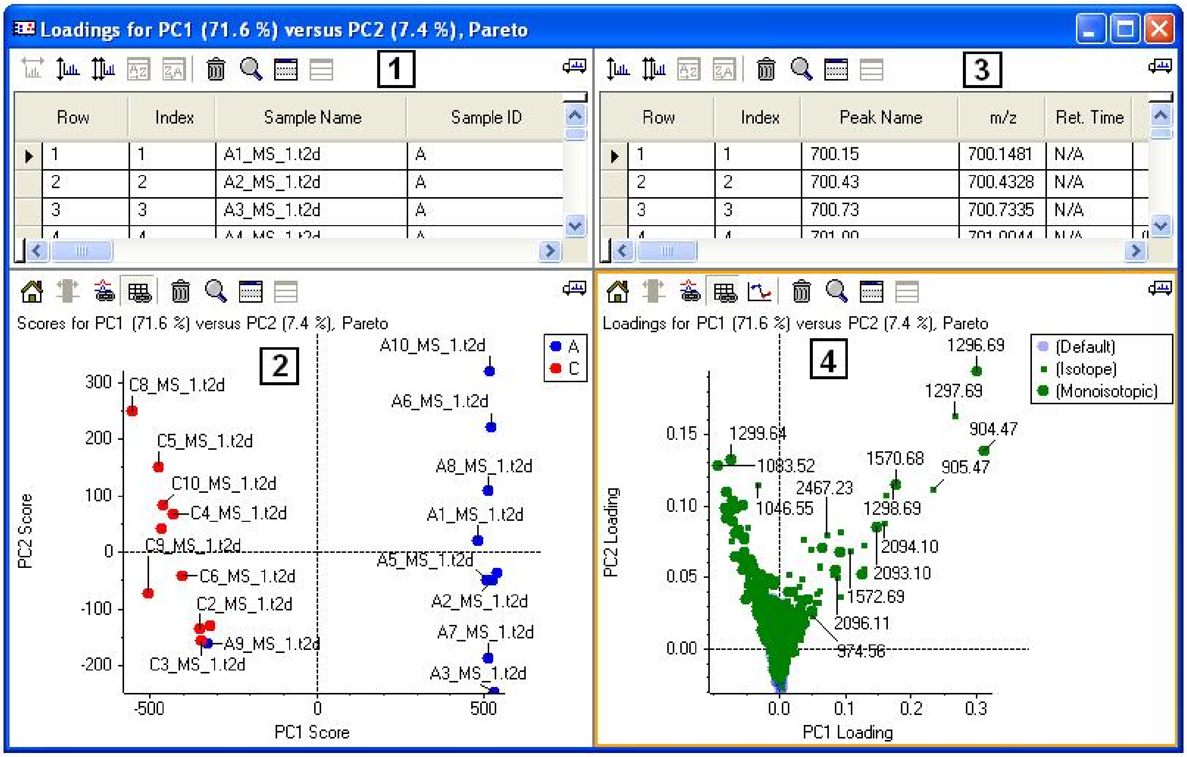
☐ Remove samples marked as not used

PCA Preprocessing determines how the data will be treated prior to the actual PCA analysis. PCA determines the variance of the data and is most affected by the largest data values; hence it is normal to scale the data so that variables have equal importance regardless of the magnitude. The most common method is known as Autoscaling and is available from the Scaling menu. Experience has shown, however, that for mass spectrometry data Pareto scaling is a good first choice; Pareto scaling reduces, but does not completely eliminate, the significance of the intensity which is appropriate for MS because larger peaks are generally more reliable and all variables are equivalent. Different scaling methods can reveal different features of the data and it is worth experimenting with these settings to observe this behavior.

2. Select **None** for the Weighting and **Pareto** for the Scaling as shown. Make sure that the **Perform PCA-DA** option is unchecked.

   PCA determines combinations of the original variables that explain the variance in the data. The first principal component (PC1) explains the greatest amount of variance; PC2 explains the next largest amount and so on. The program will stop calculating PC's when the amount of variance explained is less than 0.5% of the total variance.

3. Click **OK**. After the PC's are calculated, the following will be displayed



Row	Index	Sample Name	Sample ID
1	1	A1_MS_1.t2d	A
2	2	A2_MS_1.t2d	A
3	3	A3_MS_1.t2d	A
4	4	A4_MS_1.t2d	A

Row	Index	Peak Name	m/z	Ret. Time
1	1	700.15	700.1481	N/A
2	2	700.43	700.4328	N/A
3	3	700.73	700.7335	N/A
4	4	701.00	701.0044	N/A

4.2 Understanding the display

As mentioned above, PCA determines linear combinations (PC's) of the original variables that explain the variance in the data, i.e.:

$$PC_1 = p_1 x_1 + p_2 x_2 + p_3 x_3 \ldots$$

where the p's are called the loadings and represent the importance of the variables (x) to the PC; the larger the loading, the more important the variable. You can think of this as follows: if there are n variables originally, then every sample corresponds to a point in the n-dimensional space defined by the variables. PCA is equivalent to rotating the axes so that one – PC1 – lies along the line of maximum variance. The loadings then indicate the direction of the new axes. Each sample can be given a value on this new axis which is called the score, so we can look at the way the samples are arranged according to this new axis.

The display obtained after performing PCA consists of 4 panes as numbered in the figure above:

1. A table of the scores for each sample and each PC – the Scores Table
2. A plot of the sample scores for PC1 and PC2 – The Scores Plot
3. A table of the loadings (contributions) for each variable and each PC – the Loadings Table
4. A plot of the loadings for PC1 and PC2 – the Loadings Plot

In the scores table (1) each of the PC's has a separate column and the heading indicates the percentage of the total variance that is explained by that particular PC. In this case PC1 explains 71.6% of the variance, PC2 7.4% and PC3 2.4%. Each sample has a row showing the scores for that sample.

The scores plot (2) contains a point for each sample using the symbols assigned to the groups and defined earlier (section 3.5). Several observations can be made from this plot:

- The samples are divided into two groups along PC1 – the blue symbols (group A) have large positive PC1 scores and the red samples (group C) have large negative scores

- There is also some variation that is explained by PC2, and this seems to affect both groups in a similar manner. This variance is, however, only 7.4% of the total even though the plot visually suggests it is more significant.
- One of the blue samples (A9_MS_1.t2d) is more similar to group C (red) than it is to the other members of group A. This is the sample that was also identified as an outlier during the t-test (section 3.7) and subsequently excluded. (It is still included here because the data table was saved before the sample was excluded)

The loadings table (3) also contains a column for each PC, but in this case the rows correspond to variables and the values in the cells indicate the loading for the various PC's.

The loadings plot (4) displays a point for each variable colored according to the groups assigned as the data is imported and the symbols assigned to the default groups (see section 3.5); as illustrated, monoisotopic peaks are represented by large green circles, other isotope peaks by small green circles and unassigned peaks are blue. Coloring the variables in this way allows you to quickly determine their importance. The loadings plot has some interesting features:

- The vast majority of the points are clustered around the origin, i.e. they have small loadings and contribute little to either PC1 or PC2
- A number of variables have large positive values on PC1 and PC2. Since one group of samples (A) is separated because it has large positive PC1 scores (as shown by the clustering of the group A samples in the Scores plot), these variables with large positive PC1 loadings are responsible.
- There are also a number of variables that have negative PC1 and positive PC2 loadings; the latter may contribute to the variation of the samples in the PC2 direction since some of them have large positive PC2 scores.
- Some variables tend to lie close to straight lines that pass through the origin, for example the points labeled 1296.69, 1297.69 and 1298.69.

The behavior described in the last point arises because these points are correlated (they are all isotopic forms of the same compound as indicated by the coloring) and we used Pareto scaling which retains some of the intensity of the variable. Since the peaks are correlated they will have very similar behavior on all PC's, and thus lie on the same straight line, but the actual loading value will depend on the intensity with the largest value having the biggest loading. Hence we can say the following:

- Correlated peaks, for example isotope peaks, adducts, fragments or multiply charged variants, will have loadings such that the ratio of two PCs is the same, i.e. for a given peak:

  (PC1 loading) / (PC2 loading) = constant

  and will lie on a straight line through the origin.
- The most intense peaks will be in intensity order along this line with the most intense furthest from the origin.

So, in this example, 1296.69 is the most intense, the peak containing one ^{13}C atom is next most intense and the peak with two ^{13}C atoms is least intense.

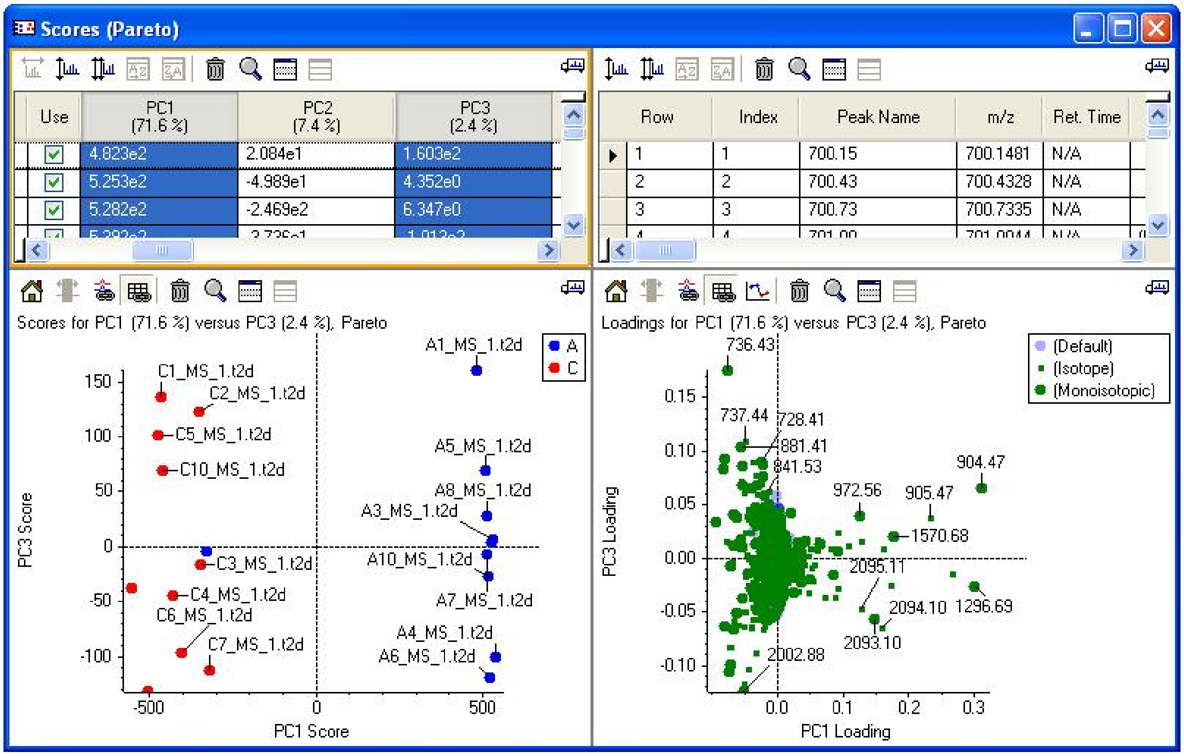
Note: in many cases correlated variables are removed before performing PCA, but here their presence helps to confirm that the observed behavior is real and not random. It also provides a way to determine peaks that are related to the same compound since these will be correlated.

4.3 Interacting with the display

This section describes some of the features of the displays and ways in which you can interact with them. The displays contain many powerful features and it is valuable to experiment with them.

4.3.1 Displaying other PC's

1. In the scores table (1) select the PC1 column by clicking its title, and then select PC3 by holding the Control key down while clicking its heading (scroll the table sideways if necessary).



Use	PC1 (71.6 %)	PC2 (7.4 %)	PC3 (2.4 %)
☑	4.823e2	2.084e1	1.603e2
☑	5.253e2	-4.989e1	4.352e0
☑	5.282e2	-2.469e2	6.347e0

Row	Index	Peak Name	m/z	Ret. Time
1	1	700.15	700.1481	N/A
2	2	700.43	700.4328	N/A
3	3	700.73	700.7335	N/A

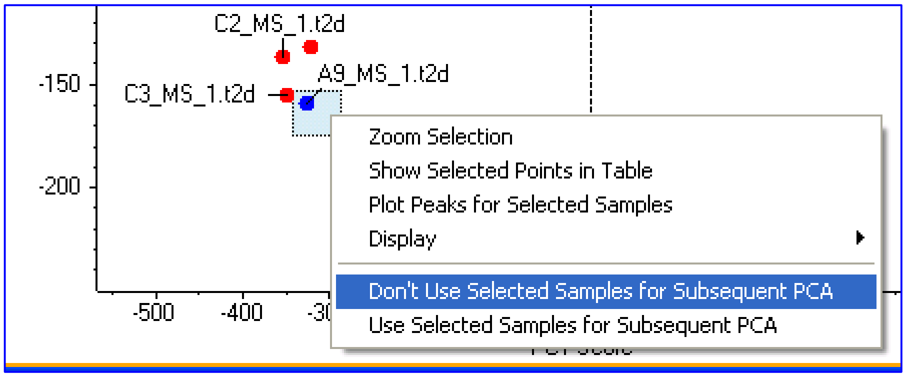
The display will update to show the scores and loadings for PC3 vs. PC1. The separation due to PC1 is maintained, but there is also some separation along PC3 and a suggestion of two groups – one with positive PC3 scores and one with negative scores – for both sample groups.

- In the scores or loadings table, click the PC1 column heading and drag so that PC2 is also selected; the original display will be restored.

4.3.2 Excluding samples

It is clear that sample A9_MS_1.t2d is unusual in some way and should be excluded from further calculations. In section 3.8 we saw how to do this by deactivating it in the sample table; here we will see how to do this from the scores plot.

1. In the scores plot, click and drag to make a selection rectangle around the abnormal sample.
2. Right-click within the selection rectangle and select **Don't Use Selected Samples for Subsequent PCA**.



C2_MS_1.t2d
A9_MS_1.t2d
C3_MS_1.t2d
-150
-200
-500
-400
-3
Zoom Selection
Show Selected Points in Table
Plot Peaks for Selected Samples
Display ▸
Don't Use Selected Samples for Subsequent PCA
Use Selected Samples for Subsequent PCA

The sample symbol is replaced with an open circle. This is the default symbol for excluded samples and variables, and can be changed as described in section 3.5 by altering the symbol for the special group (Excluded).

3. Repeat the PCA analysis by selecting **Perform PCA** from the **Analyze** menu or by clicking on the PCA button under the menu bar ().

4. In the options dialog, make sure **Remove samples marked as unused** is unchecked and click OK.

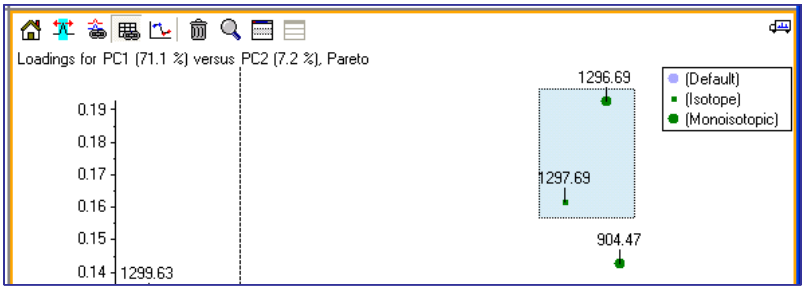
   The PC's will be recalculated and the display regenerated. Note that the excluded sample is still present in the display but is drawn using the excluded symbol and that this is reflected in the plot legend (If the legend is not displayed, right-click in the scores plot, and select **Display -> Show legend**). If **Remove samples marked as unused** had been checked the sample would not have been included in the display.

   If you save the data at this point, the resulting file will still contain the excluded sample but it will be marked as unused.

4.3.3 Selecting and displaying the behavior of variables

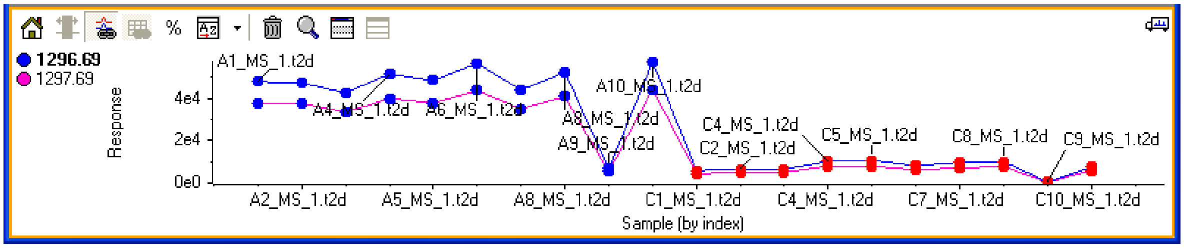
As noted previously, the variables with large, positive PC1 loadings are the ones most likely to be responsible for separating the two groups since one group has large, positive PC1 scores.

1. Draw a selection rectangle around the points representing the variables with the largest PC1 and PC2 loadings (1296.69 and 1297.69)



2. Click the **Plot profile** button ().

   This will generate a new pane in the same window showing the intensity profiles of the selected variables.



   It is clear from the display that these two variables are present in group A samples at a high level and only at a low level in group C samples. The excluded sample (A9_MS_1.t2d) also shows a low level consistent with group C.

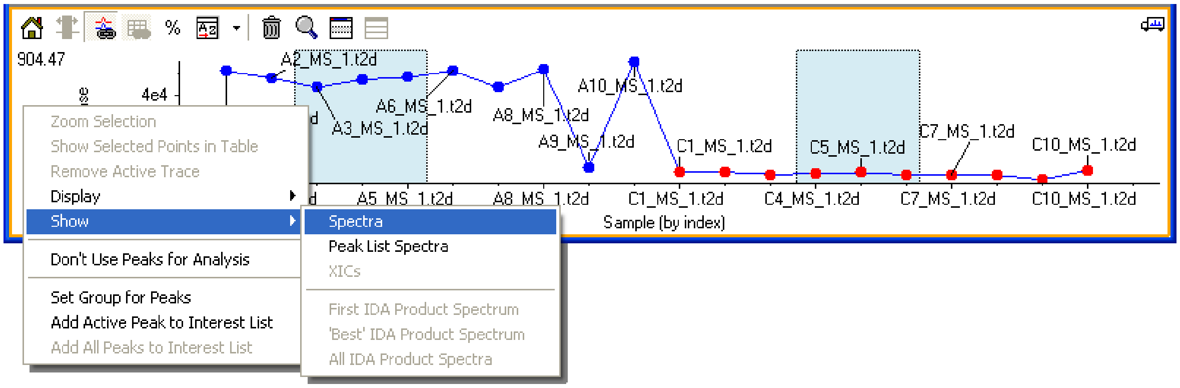
   The intensity for the peak at mass 1297.69 is lower than the peak at 1296.69 in all samples as expected for a ^{13}C isotope peak at this mass and suggested by the scores.

3. In the loadings plot, click on the point for the variable 904.47 and the display will update to show the profile of this peak.

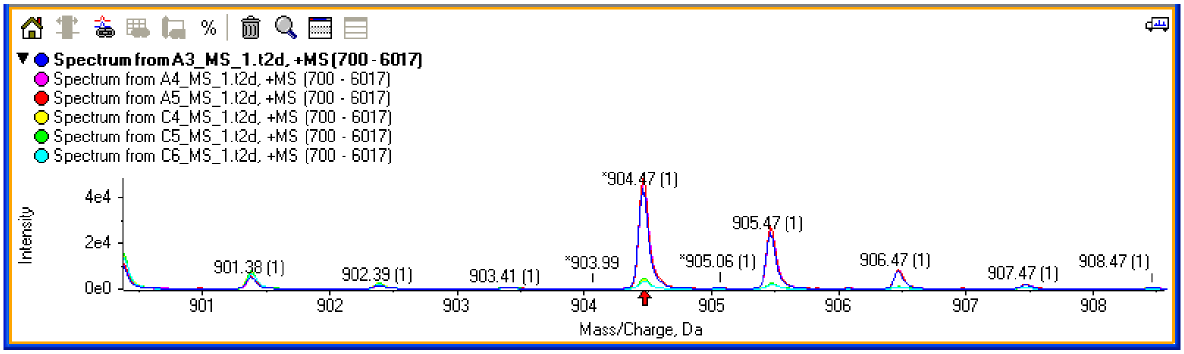
This is the default behavior when there is an active graph and a single point is selected; if you hold the shift key down when clicking on a new variable point, the profile for the new variable will be added to the existing display.

If you make a selection that includes several points you will need to click the **Plot profile** button or right-click and choose **Plot Profiles for Selected Peaks** to generate the display. By default, a new profile graph will be generated; if you hold the shift key down while generating the display the existing plot will be replaced.

4. In the Profile Plot, select a region containing a few samples from the A group, hold the shift key down and make a second selection from the C samples. Right-click in one of the selection rectangles and select **Spectra** from the **Show** submenu.



The program will locate the original data files, extract the spectra for the samples you have selected and zoom the display so that the active variable (904.47 in this case) is centered in the display. The colors are different for each sample; to color them according to the group right-click and select **Display->Use Group Colors for Traces**.


▼ ● Spectrum from A3_MS_1.t2d, +MS (700 - 6017)
● Spectrum from A4_MS_1.t2d, +MS (700 - 6017)
● Spectrum from A5_MS_1.t2d, +MS (700 - 6017)
○ Spectrum from C4_MS_1.t2d, +MS (700 - 6017)
● Spectrum from C5_MS_1.t2d, +MS (700 - 6017)
● Spectrum from C6_MS_1.t2d, +MS (700 - 6017)
Intensity
4e4
2e4
0e0
*904.47 (1)
905.47 (1)
901.38 (1)
902.39 (1)
903.41 (1)
*903.99
*905.06 (1)
906.47 (1)
907.47 (1)
908.47 (1)
901
902
903
904
905
906
907
908
Mass/Charge, Da

You may wish to enlarge the graph so that the small peaks are easier to see. This can be achieved in one of the following ways:

- Click the arrow (▼) in the top-left corner of the graph to shrink the title display to a single line reflecting the active (front and labeled) trace. The active trace can be changed by clicking on another trace in the graph and all titles redisplayed clicking the arrow again.
- Click on the top border of the pane containing the display and drag the frame upwards to enlarge the size of the pane. The cursor will change to a resizing tool when correctly positioned over the border.
- Click on the magnifying glass icon in the pane header. This will cause the display to switch to a mode where each pane is displayed on a separate tabbed page. Clicking on a different tab will change the active display, and the process can be reversed by clicking on the magnifying glass again.

5. In the loadings plot, select another variable by clicking on it. The sample graph will update but the selection rectangles will remain in place.

   Click in one of the selections and choose **Spectra** from the **Show** submenu while holding the shift key down; the spectrum graph will update to show the raw data peak for the new variable. Since these are spectral data and the spectra have already been retrieved from the files, the display will update much faster.

   If the shift key is not down a new spectrum pane will be generated using data newly retrieved from the files.

6. Remove the graph and profile plots by clicking in the **trash can** icons in each pane.

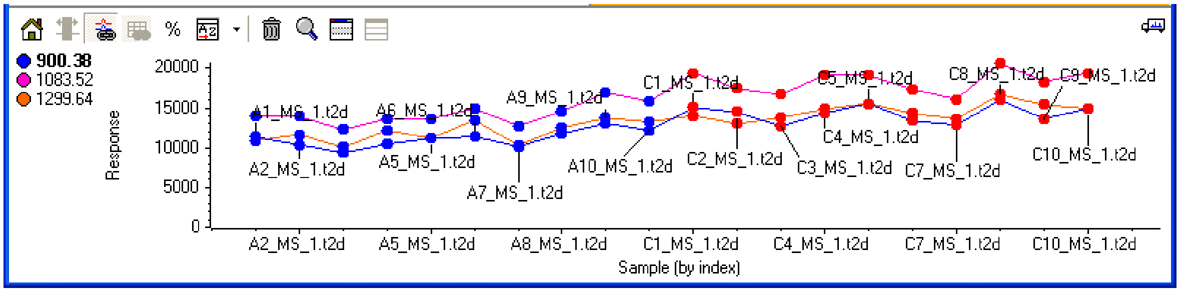
4.4 Interpreting the results

So far we have learned that the variables with large positive PC1 loadings are mainly in the group A (spiked) samples and absent, or at lower intensities, in the group C samples. But what causes some variables to have negative PC1 values? What is the source of the variation displayed by PC2? Is it significant?

Since PC1 separates the two groups, and variables with positive loadings are only in group A, it seems likely that variables with negative loadings will only be in group C (or at a lower intensity in group A). We can verify this by displaying the profiles for some of these variables:

1. In the loadings plot, select three variables with the largest positive PC2 and negative PC1 loadings (1083.52, 900.38 and 1299.64) and display the behavior of these variables by clicking the **Plot profile** button. You may need to zoom-in to do this.

   The most intense trace appears to be more intense in group C, but overall the plots suggest a gradual increase in intensity going from left to right. This is also supported by the excluded peak (open circle).



   This kind of change is quite common and can be caused by a number of gradual changes in the instrument or the samples.

   In this case the data were acquired in the order they are displayed, i.e. all group A samples were analyzed before group C, which can cause this kind of variation to appear as real differences between the two groups. To avoid this, the samples should be acquired in a random order so that members of both groups will be equally affected by experimental variation. This simple example illustrates an important point – these techniques will find differences and can be very sensitive to small changes between groups, but in order to determine 'real' biological changes of interest, the experimental system should be as closely controlled as possible.

4.5 Summary

In this section you have:

- Opened a saved data set
- Performed a Principal Components Analysis (PCA) on the data

- Understood and interacted with the display
- Excluded outliers or abnormal samples from a graph
- Displayed the profiles of variables for all samples
- Showed the raw data corresponding to the variables
- Examined the data to reveal that there are some experimental variations

These are the basic operations for using PCA and will be used frequently. The next section will apply these techniques to LCMS data and show how variables can be excluded from the calculations when they are not of interest.

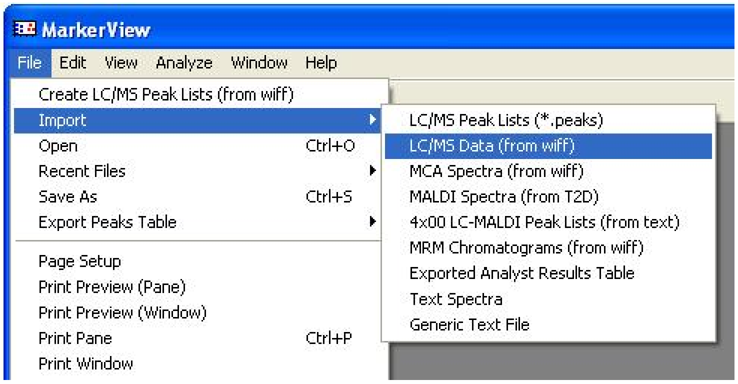
5 Unsupervised processing of LCMS data

In section 4 you learned how to process data using PCA; this section applies this technique to more complex samples resulting from the LCMS analyses of a time point study.

The data set[1] was obtained by analyzing the urine from three rats at three different time points (0 – 8, 8 – 16 and 16 – 24 hour) before and after administration of vinpocetin[2] at 10 mg/kg. Samples were analyzed by LCMS on a QStar® XL.

5.1 Importing data

1. Form the **File->Import** menu select **LC/MS Data (from wiff)**.



2. In the **Select Samples** dialog, navigate to the example data folder and drag the folder **LCMS Data** to the **Selected** side of the dialog (see section 3.1 for additional details).

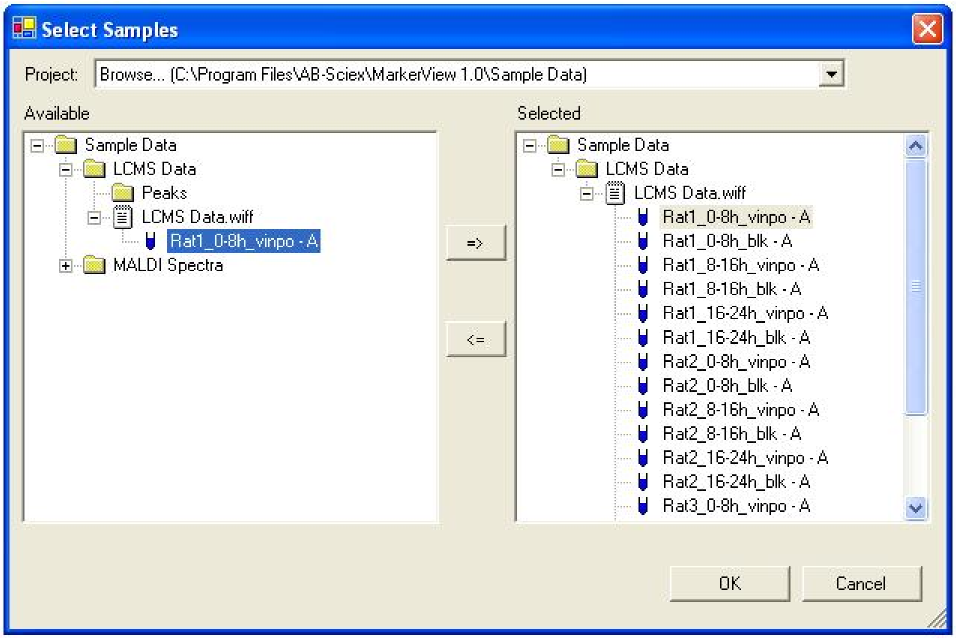
   Note that the first sample (Rat1_0-8_Vinpo_A) is included in the file list twice. After the first injection of this sample the chromatographic conditions were changed so this sample needs to be removed.

3. In the right-hand display, select the first sample and click the **<=** button to move it back to the left-hand side so it will not be imported.

   You can verify that the sample has been removed by expanding the **LCMS Data** folder and data file in the left-hand pane by clicking on the + signs.

[1] Data courtesy of Dr. Gerard Hopfgartner, University of Geneva.

[2] Vinpocetin is known as a memory enhancer; a treatment for Alzheimer's disease; a treatment for stroke; it improves circulation (especially to the brain); and it is a powerful antioxidant.


Select Samples

Project: Browse... (C:\Program Files\AB-Sciex\MarkerView 1.0\Sample Data)

Available

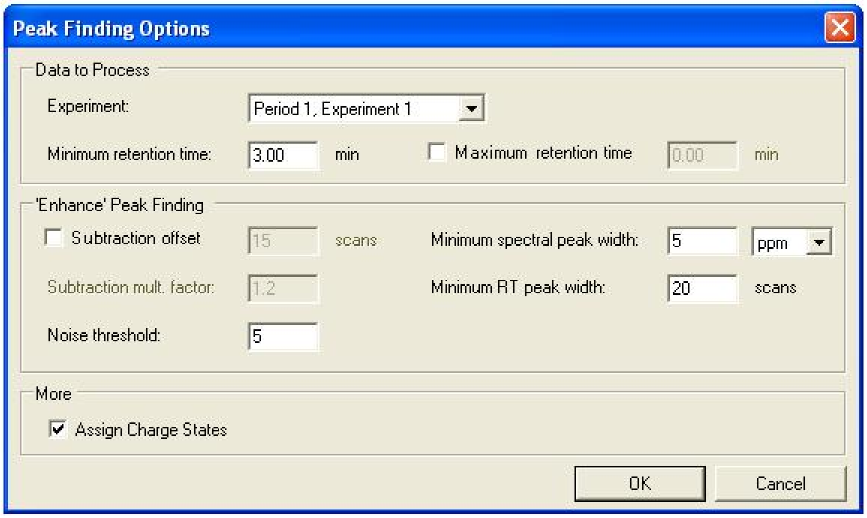
- Sample Data
  - LCMS Data
    - Peaks
    - LCMS Data.wiff
      - Rat1_0-8h_vinpo - A
  - MALDI Spectra

Selected

- Sample Data
  - LCMS Data
    - LCMS Data.wiff
      - Rat1_0-8h_vinpo - A
      - Rat1_0-8h_blk - A
      - Rat1_8-16h_vinpo - A
      - Rat1_8-16h_blk - A
      - Rat1_16-24h_vinpo - A
      - Rat1_16-24h_blk - A
      - Rat2_0-8h_vinpo - A
      - Rat2_0-8h_blk - A
      - Rat2_8-16h_vinpo - A
      - Rat2_8-16h_blk - A
      - Rat2_16-24h_vinpo - A
      - Rat2_16-24h_blk - A
      - Rat3_0-8h_vinpo - A

4. Click **OK** to begin the import process.

   Importing LCMS data occurs in two separate steps; the first step locates the peaks in the data and the second step performs the alignment and normalization.


Peak Finding Options

Data to Process

Experiment: Period 1, Experiment 1

Minimum retention time: 3.00 min

□ Maximum retention time 0.00 min

'Enhance' Peak Finding

□ Subtraction offset 15 scans

Minimum spectral peak width: 5 ppm

Subtraction mult. factor: 1.2

Minimum RT peak width: 20 scans

Noise threshold: 5

More

☑ Assign Charge States

In the **Peak Finding Options** dialog box, set the parameters as follows:

Minimum retention time to 3.00 min (to ignore the void volume)

Subtraction offset unchecked

Minimum spectral peak width 5 ppm

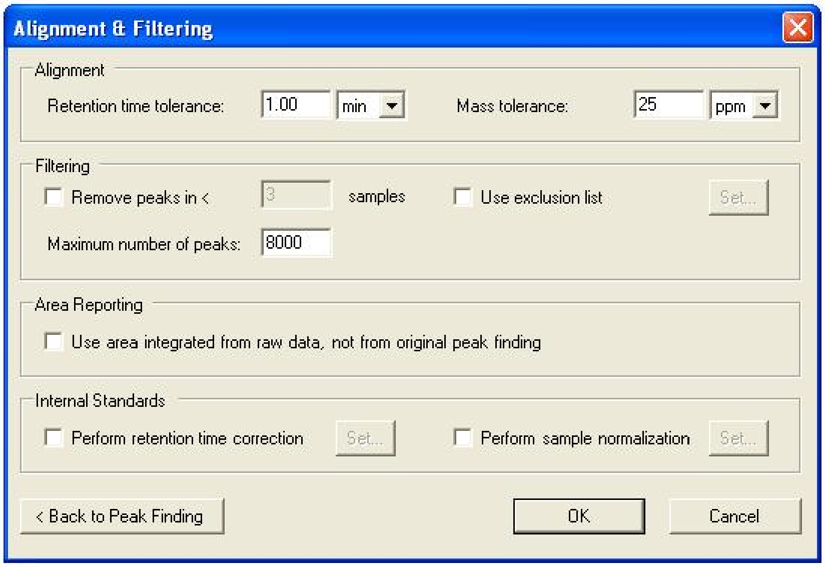
Noise threshold 5

Minimum RT peak width 20 scans

Assign Charge States checked

These settings will allow the program to find small, narrow mass peaks that may be recombined during alignment. These data were acquired using an unusually fast scan speed of 5 scans/second, so the LC peaks are wide in terms of scan numbers.

5. Click **OK**. The dialog for the second step of the import process appears.



Alignment & Filtering

Alignment

Retention time tolerance: 1.00 min

Mass tolerance: 25 ppm

Filtering

□ Remove peaks in < 3 samples

□ Use exclusion list

Maximum number of peaks: 8000

Area Reporting

□ Use area integrated from raw data, not from original peak finding

Internal Standards

□ Perform retention time correction

□ Perform sample normalization

Set the **Retention time tolerance** to 1 min. and the **Mass tolerance** to 25 ppm; peaks that are within these tolerance values, either between files or within a single file, will be aligned to the same peak.

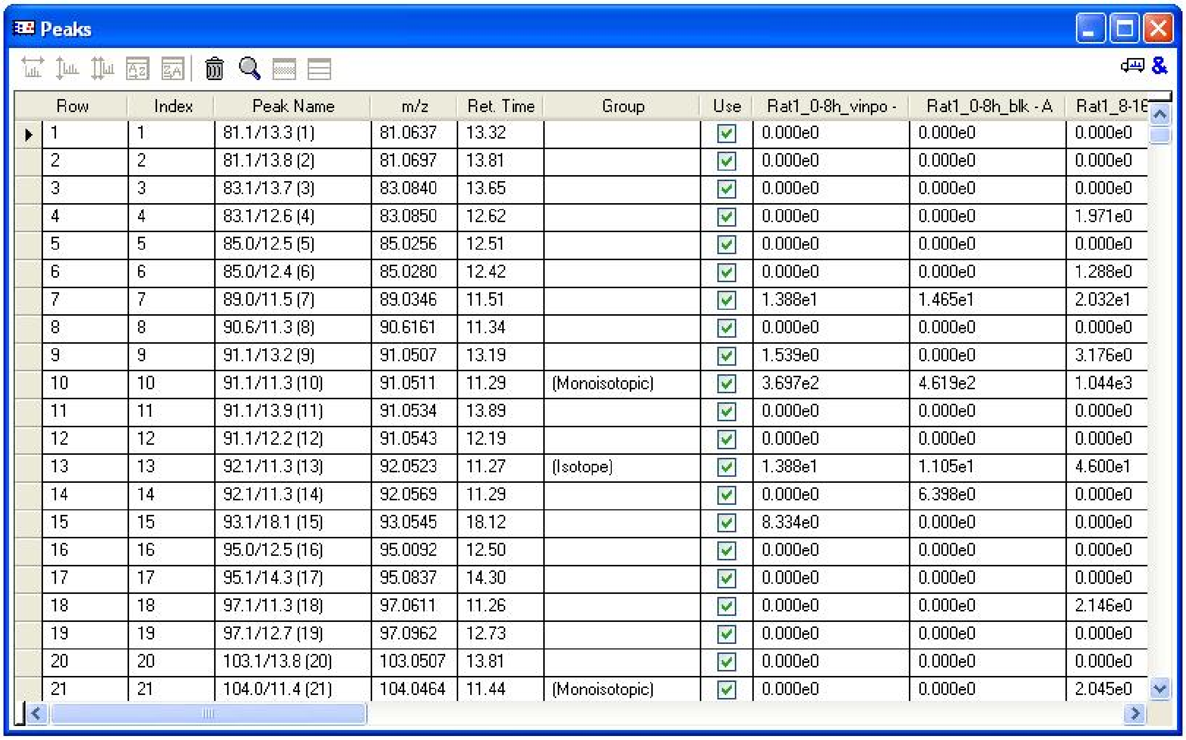
Leave the filtering parameters unchecked, and set the **Maximum number of peaks** to 8000.

Uncheck **Perform sample normalization** and **Perform retention time correction.**

6. Click **OK**. Once the import process is complete the data table will appear.

The data table is similar to the one for spectra (section 3.1), but now the retention time field is not empty and the peak name is constructed by combining the m/z value and the retention time in minutes. The name also contains an index value in brackets since it is sometimes easier to locate variables using this number.

You can review the data by selecting rows (variables) or columns (samples) and clicking on the plot column or plot row buttons at the top of the pane.



Row	Index	Peak Name	m/z	Ret. Time	Group	Use	Rat1_0-8h_vinpo -	Rat1_0-8h_blk - A	Rat1_8-16
1	1	81.1/13.3 (1)	81.0637	13.32		☑	0.000e0	0.000e0	0.000e0
2	2	81.1/13.8 (2)	81.0697	13.81		☑	0.000e0	0.000e0	0.000e0
3	3	83.1/13.7 (3)	83.0840	13.65		☑	0.000e0	0.000e0	0.000e0
4	4	83.1/12.6 (4)	83.0850	12.62		☑	0.000e0	0.000e0	1.971e0
5	5	85.0/12.5 (5)	85.0256	12.51		☑	0.000e0	0.000e0	0.000e0
6	6	85.0/12.4 (6)	85.0280	12.42		☑	0.000e0	0.000e0	1.288e0
7	7	89.0/11.5 (7)	89.0346	11.51		☑	1.388e1	1.465e1	2.032e1
8	8	90.6/11.3 (8)	90.6161	11.34		☑	0.000e0	0.000e0	0.000e0
9	9	91.1/13.2 (9)	91.0507	13.19		☑	1.539e0	0.000e0	3.176e0
10	10	91.1/11.3 (10)	91.0511	11.29	(Monoisotopic)	☑	3.697e2	4.619e2	1.044e3
11	11	91.1/13.9 (11)	91.0534	13.89		☑	0.000e0	0.000e0	0.000e0
12	12	91.1/12.2 (12)	91.0543	12.19		☑	0.000e0	0.000e0	0.000e0
13	13	92.1/11.3 (13)	92.0523	11.27	(Isotope)	☑	1.388e1	1.105e1	4.600e1
14	14	92.1/11.3 (14)	92.0569	11.29		☑	0.000e0	6.398e0	0.000e0
15	15	93.1/18.1 (15)	93.0545	18.12		☑	8.334e0	0.000e0	0.000e0
16	16	95.0/12.5 (16)	95.0092	12.50		☑	0.000e0	0.000e0	0.000e0
17	17	95.1/14.3 (17)	95.0837	14.30		☑	0.000e0	0.000e0	0.000e0
18	18	97.1/11.3 (18)	97.0611	11.26		☑	0.000e0	0.000e0	2.146e0
19	19	97.1/12.7 (19)	97.0962	12.73		☑	0.000e0	0.000e0	0.000e0
20	20	103.1/13.8 (20)	103.0507	13.81		☑	0.000e0	0.000e0	0.000e0
21	21	104.0/11.4 (21)	104.0464	11.44	(Monoisotopic)	☑	0.000e0	0.000e0	2.045e0

5.2 Assigning groups and symbols

It is convenient to assign groups and symbols so that the pre- and post-dose samples can be distinguished, as well as the different time points. We will assign the groups and symbols according to the following table:

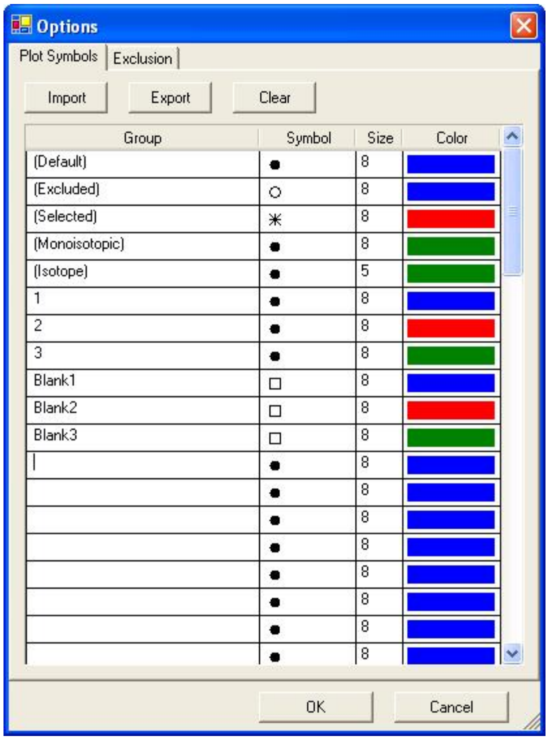
Time point (hr.)	Post-dose group	Sample symbol	Pre-dose group	Blank symbol
0 – 8	1	Closed blue circle	blank1	Open blue square
8 – 16	2	Closed red circle	blank2	Open red square
16 – 24	3	Closed green circle	blank3	Open green square

So that the time points are distinguished by color and the pre- and post-dose by shape.

We will start by assigning the symbols first; this saves some typing since the groups can then be assigned by selecting them from a menu.

7. Select **Options** from the **Edit** menu. You can either manually fill-in the point symbols for the six groups as shown in the figure below or import them from a file included with the program.

   To import symbols from the example file, click the **Import** button and navigate to the 'AB-Sciex\MarkerView\Sample Data\LCMS Data' subfolder of the 'Program Files' folder in the resulting 'Open' dialog. Select the **LCMS Plot Symbols.ptsym** file and then click **OK** to close the **Options** dialog.



Options

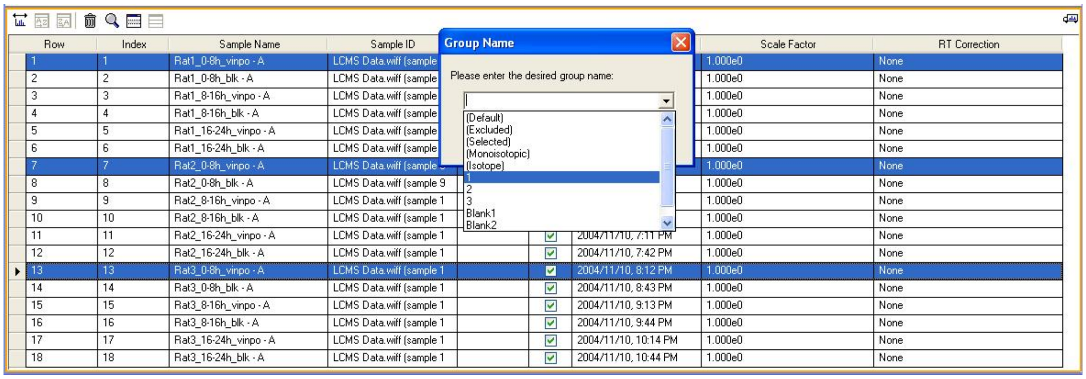
Plot Symbols | Exclusion

Import | Export | Clear

Group	Symbol	Size	Color
(Default)	●	8	🟦
(Excluded)	○	8	🟦
(Selected)	✳	8	🟥
(Monoisotopic)	●	8	🟩
(Isotope)	●	5	🟩
1	●	8	🟦
2	●	8	🟥
3	●	8	🟩
Blank1	□	8	🟦
Blank2	□	8	🟥
Blank3	□	8	🟩
	●	8	🟦
	●	8	🟦
	●	8	🟦
	●	8	🟦
	●	8	🟦
	●	8	🟦
	●	8	🟦
	●	8	🟦

OK | Cancel

8. From the **View** menu, select **Show Samples Table.**
9. In the sample table select the row containing the sample **Rat1_0-8h_vinpo – A** which corresponds to the 0 – 8 hour post-dose sample from Rat1.
10. Hold the control key down and select the rows for **Rat2_0-8h_vinpo – A** and **Rat3_0-8h_vinpo – A.**
11. Right-click in the table and select **Set Group for Selected Samples.**
12. In the **Group Name** dialog, click on the pop-up menu, select '1' and click **OK.**



Row	Index	Sample Name	Sample ID				Scale Factor	RT Correction
1	1	Rat1_0-8h_vinpo - A	LCMS Data.wiff (sample				1.000e0	None
2	2	Rat1_0-8h_blk - A	LCMS Data.wiff (sample				1.000e0	None
3	3	Rat1_8-16h_vinpo - A	LCMS Data.wiff (sample				1.000e0	None
4	4	Rat1_8-16h_blk - A	LCMS Data.wiff (sample				1.000e0	None
5	5	Rat1_16-24h_vinpo - A	LCMS Data.wiff (sample				1.000e0	None
6	6	Rat1_16-24h_blk - A	LCMS Data.wiff (sample				1.000e0	None
7	7	Rat2_0-8h_vinpo - A	LCMS Data.wiff (sample				1.000e0	None
8	8	Rat2_0-8h_blk - A	LCMS Data.wiff (sample 9				1.000e0	None
9	9	Rat2_8-16h_vinpo - A	LCMS Data.wiff (sample 1				1.000e0	None
10	10	Rat2_8-16h_blk - A	LCMS Data.wiff (sample 1				1.000e0	None
11	11	Rat2_16-24h_vinpo - A	LCMS Data.wiff (sample 1		☑	2004/11/10, 7:11 PM	1.000e0	None
12	12	Rat2_16-24h_blk - A	LCMS Data.wiff (sample 1		☑	2004/11/10, 7:42 PM	1.000e0	None
13	13	Rat3_0-8h_vinpo - A	LCMS Data.wiff (sample 1		☑	2004/11/10, 8:12 PM	1.000e0	None
14	14	Rat3_0-8h_blk - A	LCMS Data.wiff (sample 1		☑	2004/11/10, 8:43 PM	1.000e0	None
15	15	Rat3_8-16h_vinpo - A	LCMS Data.wiff (sample 1		☑	2004/11/10, 9:13 PM	1.000e0	None
16	16	Rat3_8-16h_blk - A	LCMS Data.wiff (sample 1		☑	2004/11/10, 9:44 PM	1.000e0	None
17	17	Rat3_16-24h_vinpo - A	LCMS Data.wiff (sample 1		☑	2004/11/10, 10:14 PM	1.000e0	None
18	18	Rat3_16-24h_blk - A	LCMS Data.wiff (sample 1		☑	2004/11/10, 10:44 PM	1.000e0	None

Group Name

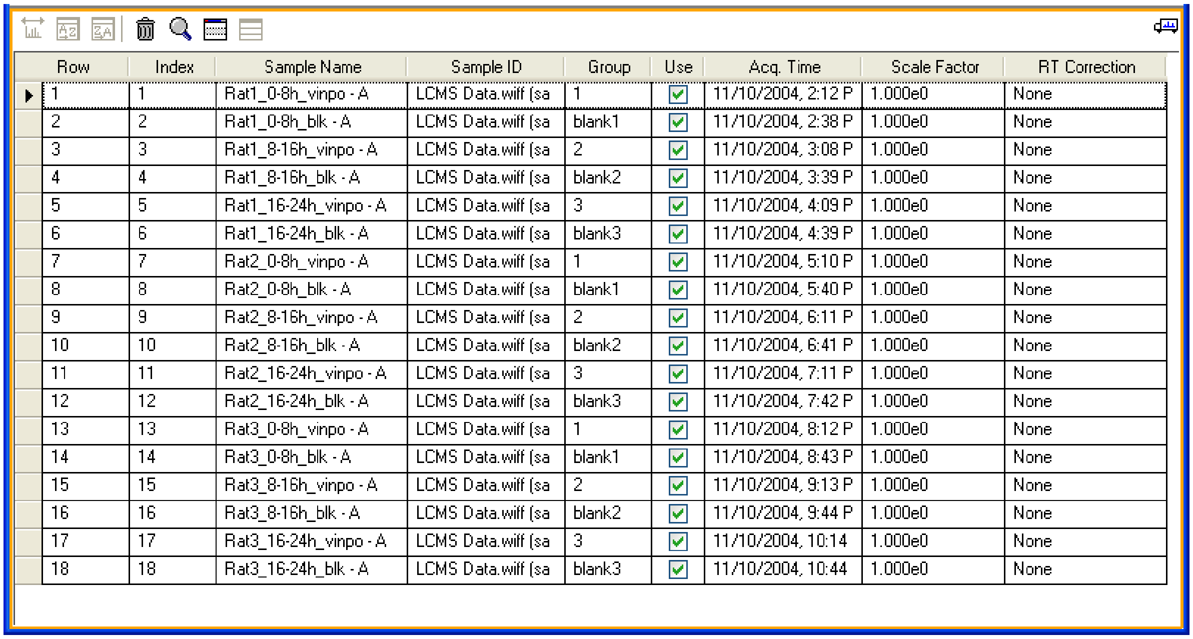
Please enter the desired group name:

13. Repeat the process assigning the groups as follows:

Samples	Group
Rat1_8-16h_vinpo – A Rat2_8-16h_vinpo – A Rat3_8-16h_vinpo – A	2
Rat1_16-24h_vinpo – A Rat2_16-24h_vinpo – A	

Rat3_16-24h_vinpo – A	3
Rat1_0-8h_blk – A Rat2_0-8h_blk – A Rat3_0-8h_blk – A	Blank1
Rat1_8-16h_ blk – A Rat2_8-16h_ blk – A Rat3_8-16h_ blk – A	Blank2
Rat1_16-24h_ blk – A Rat2_16-24h_ blk – A Rat3_16-24h_ blk – A	Blank3

The finished sample table will look like:



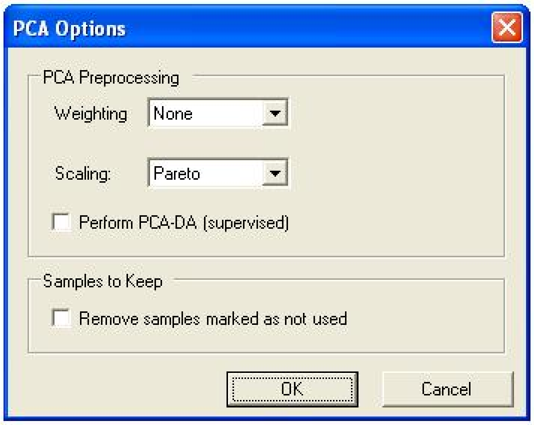
Row	Index	Sample Name	Sample ID	Group	Use	Acq. Time	Scale Factor	RT Correction
1	1	Rat1_0-8h_vinpo - A	LCMS Data.wiff (sa	1	☑	11/10/2004, 2:12 P	1.000e0	None
2	2	Rat1_0-8h_blk - A	LCMS Data.wiff (sa	blank1	☑	11/10/2004, 2:38 P	1.000e0	None
3	3	Rat1_8-16h_vinpo - A	LCMS Data.wiff (sa	2	☑	11/10/2004, 3:08 P	1.000e0	None
4	4	Rat1_8-16h_blk - A	LCMS Data.wiff (sa	blank2	☑	11/10/2004, 3:39 P	1.000e0	None
5	5	Rat1_16-24h_vinpo - A	LCMS Data.wiff (sa	3	☑	11/10/2004, 4:09 P	1.000e0	None
6	6	Rat1_16-24h_blk - A	LCMS Data.wiff (sa	blank3	☑	11/10/2004, 4:39 P	1.000e0	None
7	7	Rat2_0-8h_vinpo - A	LCMS Data.wiff (sa	1	☑	11/10/2004, 5:10 P	1.000e0	None
8	8	Rat2_0-8h_blk - A	LCMS Data.wiff (sa	blank1	☑	11/10/2004, 5:40 P	1.000e0	None
9	9	Rat2_8-16h_vinpo - A	LCMS Data.wiff (sa	2	☑	11/10/2004, 6:11 P	1.000e0	None
10	10	Rat2_8-16h_blk - A	LCMS Data.wiff (sa	blank2	☑	11/10/2004, 6:41 P	1.000e0	None
11	11	Rat2_16-24h_vinpo - A	LCMS Data.wiff (sa	3	☑	11/10/2004, 7:11 P	1.000e0	None
12	12	Rat2_16-24h_blk - A	LCMS Data.wiff (sa	blank3	☑	11/10/2004, 7:42 P	1.000e0	None
13	13	Rat3_0-8h_vinpo - A	LCMS Data.wiff (sa	1	☑	11/10/2004, 8:12 P	1.000e0	None
14	14	Rat3_0-8h_blk - A	LCMS Data.wiff (sa	blank1	☑	11/10/2004, 8:43 P	1.000e0	None
15	15	Rat3_8-16h_vinpo - A	LCMS Data.wiff (sa	2	☑	11/10/2004, 9:13 P	1.000e0	None
16	16	Rat3_8-16h_blk - A	LCMS Data.wiff (sa	blank2	☑	11/10/2004, 9:44 P	1.000e0	None
17	17	Rat3_16-24h_vinpo - A	LCMS Data.wiff (sa	3	☑	11/10/2004, 10:14	1.000e0	None
18	18	Rat3_16-24h_blk - A	LCMS Data.wiff (sa	blank3	☑	11/10/2004, 10:44	1.000e0	None

14. Select **Save As** from the **File** menu and save the imported data in the **LCMS Data** folder with the name **LCMS Saved,** overwriting the file if it already exists.

    This will save the imported data and the assigned groups so they can be easily retrieved in future.

5.3 Performing PCA and interpreting the results

1. Click the **trash can** icon in the sample table pane to close it
2. Click the PCA button or select **Perform PCA** from the **Analyze** menu.
3. In the **PCA** dialog, select **None** for the **Weighting** and **Pareto** for the **Scaling** and click **OK.**



PCA Options

PCA Preprocessing

Weighting: None

Scaling: Pareto

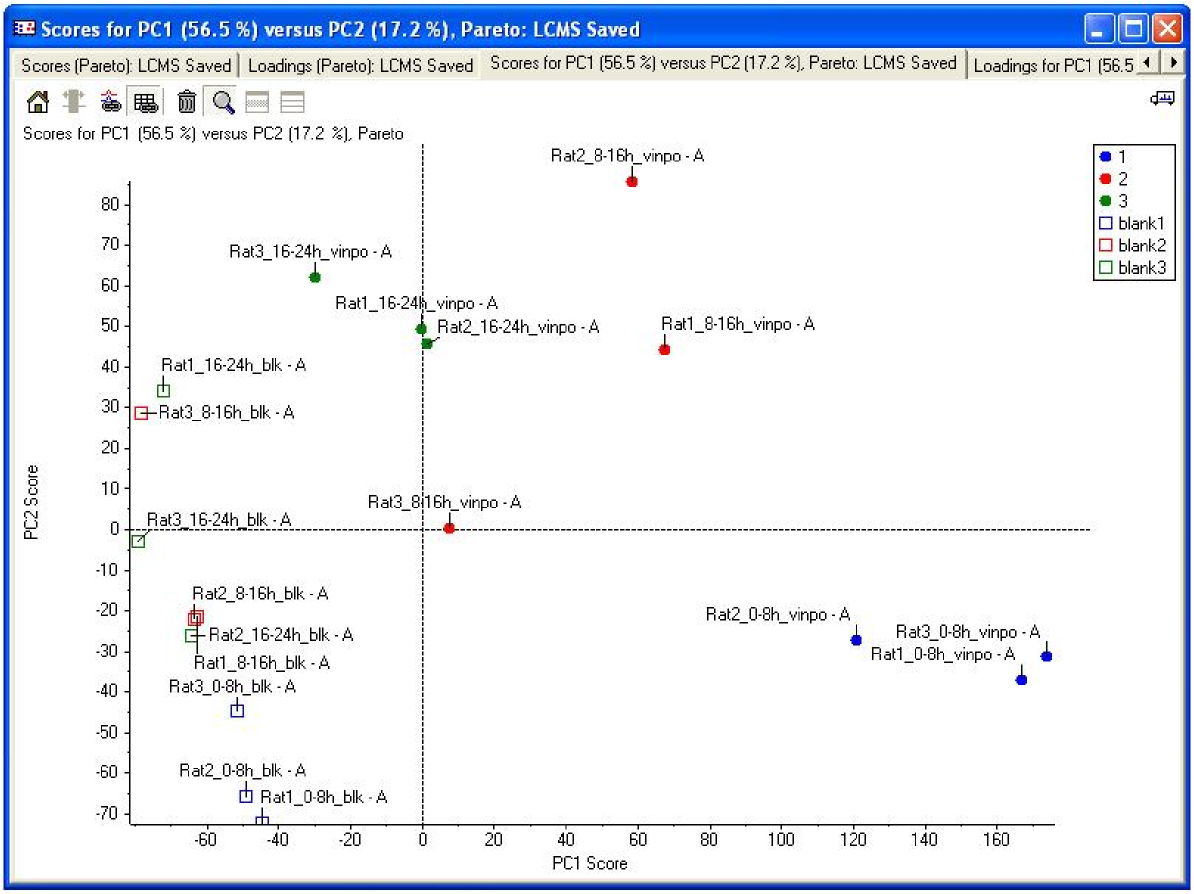
□ Perform PCA-DA (supervised)

Samples to Keep

□ Remove samples marked as not used

The resulting display will show the scores and loading in both tabular and graphical form as described in detail section 4.2.

4. Click the magnifying glass button (🔍) in the scores plot so that it is easier to examine.

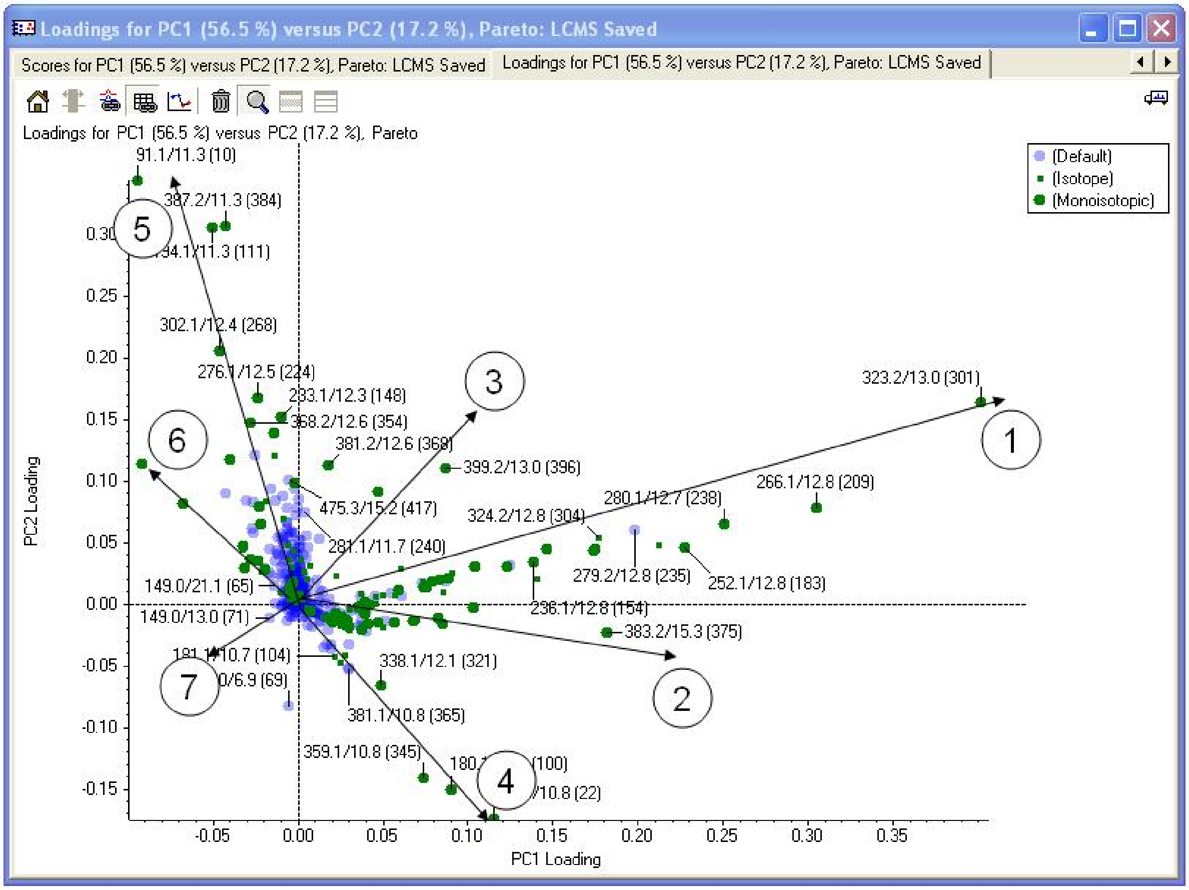


Scores for PC1 (56.5 %) versus PC2 (17.2 %), Pareto: LCMS Saved

It is clear that PC1 (ca. 56.5% of the variance) separates the pre-dose samples (open squares) from the post-dose (closed circles), with the 0 – 8 hour samples having the highest PC1 scores, the 8 – 16 having the next highest and the 16 – 24 hour samples being closest to the pre-dose. This suggests that the biggest change occurs in the first 8 hours and that the magnitude of the change lessens over time.

PC2 (17.2%) appears to separate the samples according to the sampling interval with the 0 – 8 samples (pre- and post-dose) having the most negative values and the other time points being less well separated. This suggests that there is a diurnal variation in the samples that is unaffected by administration of vinpocetin.

Click on the tab to display the loadings plot (PC1 Loading versus PC2 Loading).



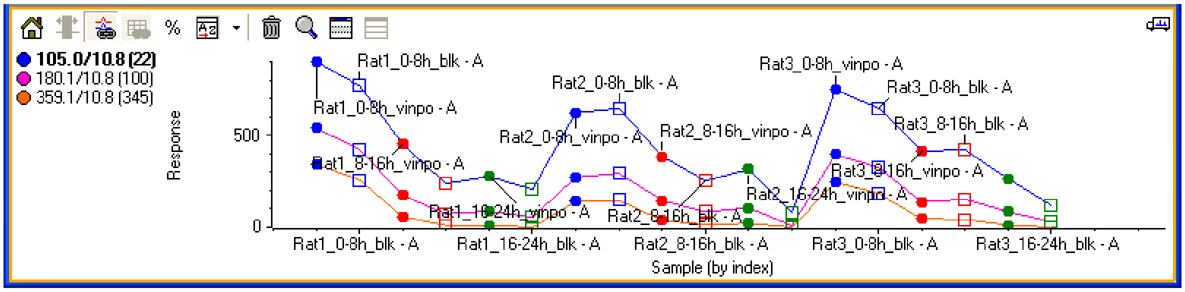
As explained in section 4.2, Pareto scaling causes correlated variables to lie on straight lines that pass through the origin. Examination of the loadings plot indicates the presence of several families of correlated variables such as those shown above.

The families marked 1 and 2 have the highest positive PC1 loadings and will contribute most to the separation of the post- and pre-dose samples, although families 3 and 4 may also have an affect. The variables indicated by the line marked 7 may also contribute to this difference but in the opposite sense – if 1 and 2 correspond to variables present in the post-dose but not the pre-dose, variables in family 7 will be predominantly in the pre-dose samples.

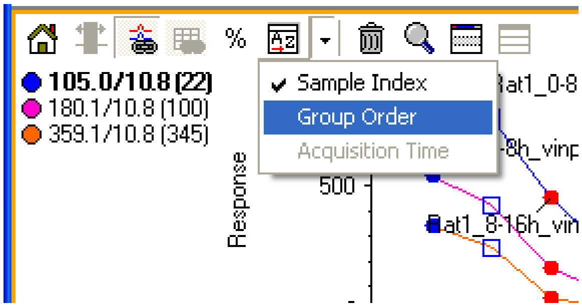
The variables in family 4 seem most likely to be in the 0 – 8 hour samples (pre- and post-dose) since these had the largest negative PC2 loadings, whereas 5 and 6 will be more prominent in the 8 – 16 and 16- 24 hour samples.

We will start by exploring the diurnal variation.

5. Click the magnifying glass to return to the multi-pane display, select the variables that are furthest away in the direction of arrow 4 and click the **Plot Profile** button.



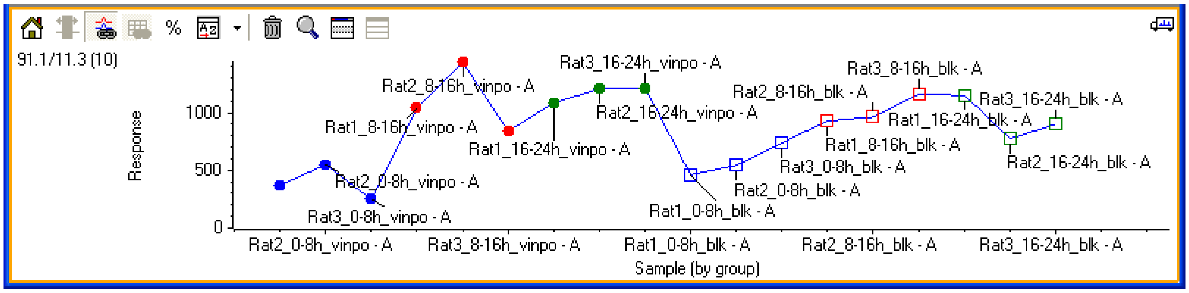
6. In the toolbar of the new graph, click on the downwards pointing arrow adjacent to the **Sort Order** button and make sure **Group Order** is selected. The data will be drawn in group order where the groups are sorted alphanumerically, i.e. in this case the order is: 1, 2, 3, blank1, blank2, blank3.



The profile graph shows that the selected ions do behave as expected, i.e. they are more intense in the 0 – 8 hour samples than the 8 – 16 and 16 – 24, and comparable in the pre- and post-dose samples.

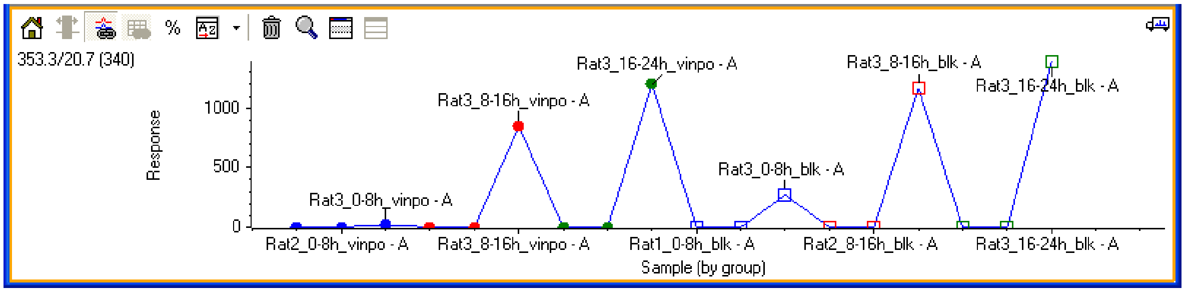
Click on other variables in the direction of family 4 to update the profile display, and note that they all have similar behavior although the peaks get smaller, and the noise higher, closer to the origin.

7. Click on a variable that is furthest from the origin in the direction of arrow 5, e.g. the variable 91.1/11.3.



Note that the variables in this direction show the opposite behavior to those in direction 4, i.e. they are lowest in the 0 – 8 samples. In this case there may also be some difference between the pre- and post-does samples.

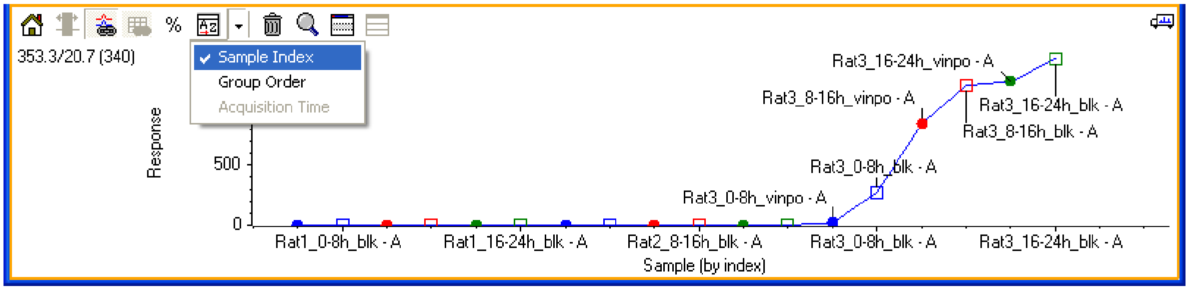
8. Click on the variable furthest from the origin in direction 6 (353.3/20.7).



At first sight, this variable appears to be present only in the rat 3 samples, and two of the rat 3 blanks have the largest negative PC1 scores, but there may be other explanations for this behavior, for example a systematic variation.

To check this it is useful to switch the display so that the samples are arranged in index or acquisition order.

9. In the **Sort Order** pull down list select **Sample Index**.

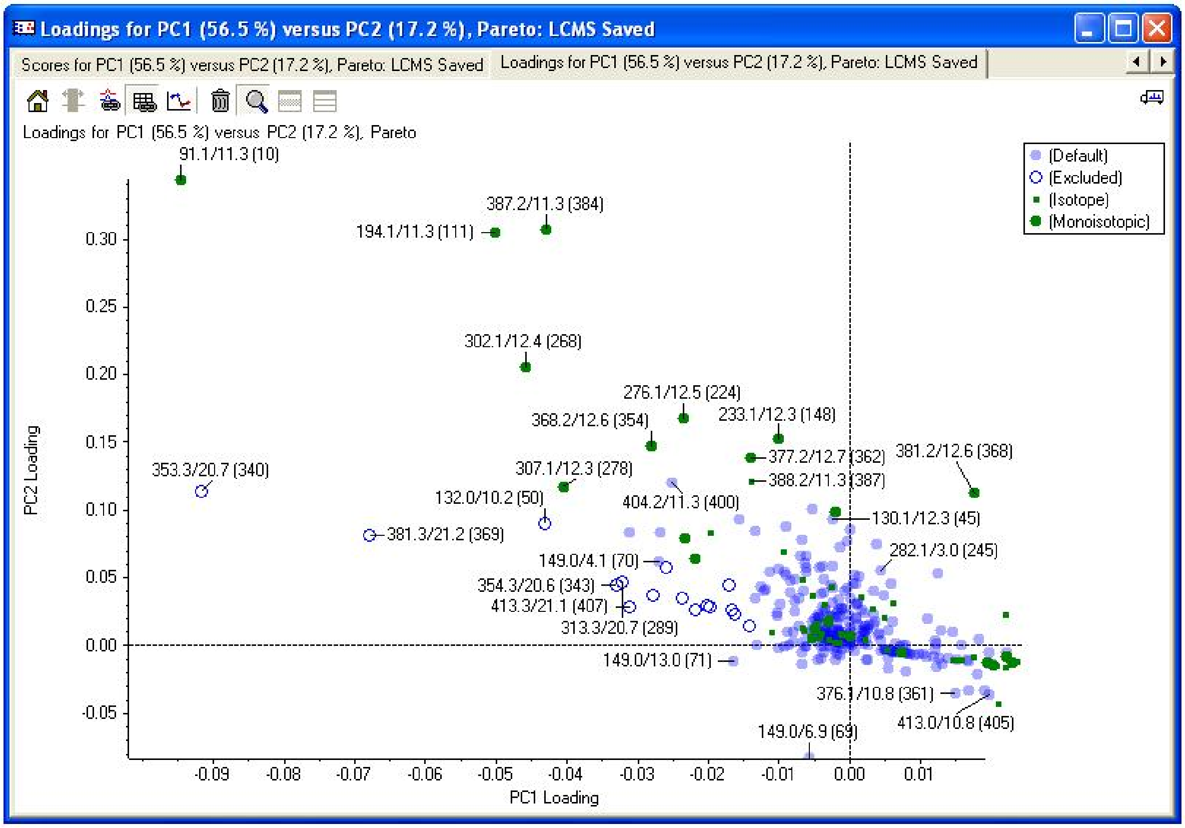


The graph illustrates that the samples were run in order, i.e. rat 1 followed by rat 2 and rat 3, and that this variable appears to be a contaminant that occurs later in the analyses and is only present in rat 3.

10. Hide the two tables and the scores plot by clicking in the **Hide pane** button (▤) in each of these panes. This results in a display that consists of the loadings plot and the profile graph making it easier to select variables.

11. Search for variables that have similar behavior by clicking on symbols in the direction of family 6, but closer to the origin. To make this easier, you may need to zoom the graph either by dragging in the axes or by selecting a rectangle in the graph, right-clicking and selecting **Zoom Selection.**

12. As the variables are encountered draw a small rectangle around them, right-click in the rectangle and select **Don’t use Selected Peaks for Subsequent PCA.** The symbols will be replaced by the symbol for excluded points (by default an open blue circle). The finished display will resemble the one below.



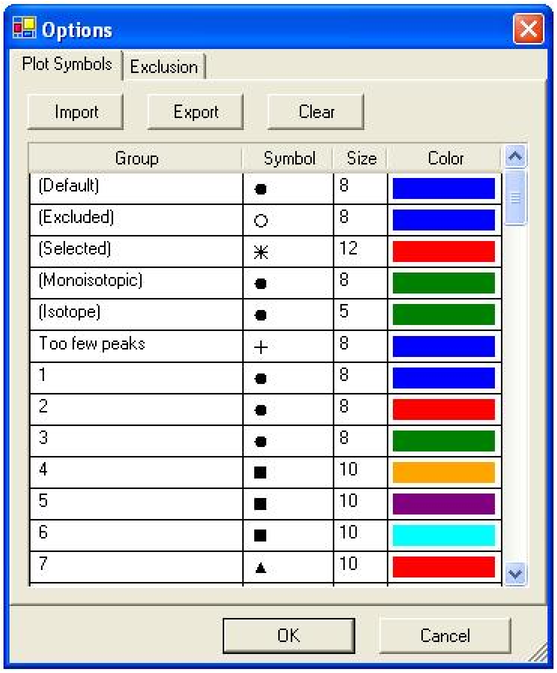
Repeat the PCA analysis by clicking the PCA button in the toolbar at the top of the window. The resulting scores and loadings plots are very similar to those obtained earlier, the most obvious difference being that the variance explained by PC1 has increased to ca. 61% (the exact value will depend on the variables you excluded), and the 16-24 hour post-dose samples seem to be grouped slightly more tightly.

5.3.1 Principal Component Variable Grouping Utility

The MarkerView™ software includes a utility which allows variables to be grouped in an automated way to facilitate data interpretation.

Follow these steps:

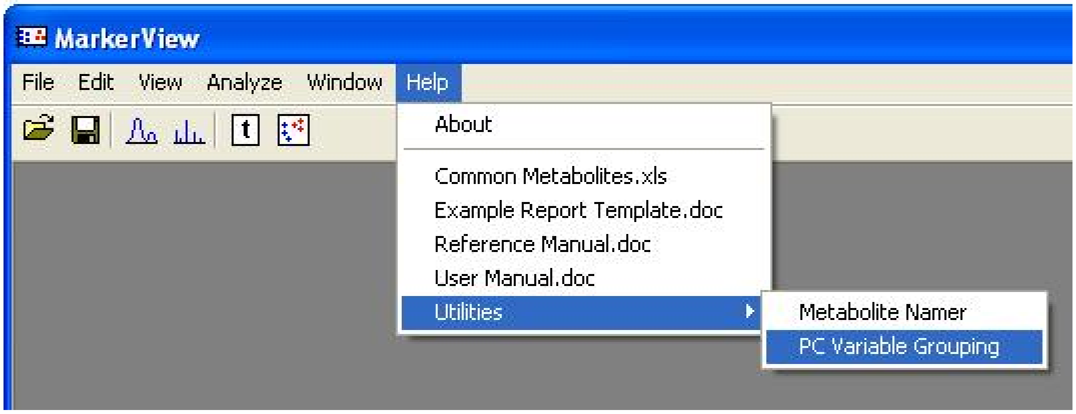
1. First, select **Options** from the **Edit** menu and define plot symbols for groups numbered '1' through '7' as shown in the figure below. Note that you do not need to use identical symbols to those shown here provided that you can distinguish these groups.



Group	Symbol	Size	Color
(Default)	●	8	🟦
(Excluded)	○	8	🟦
(Selected)	✳	12	🟥
(Monoisotopic)	●	8	🟩
(Isotope)	●	5	🟩
Too few peaks	+	8	🟦
1	●	8	🟦
2	●	8	🟥
3	●	8	🟩
4	■	10	🟧
5	■	10	🟪
6	■	10	🟦
7	▲	10	🟥

2. Return the display to the state shown in step (4) above by closing the current window or by activating the previous window. Ensure that the Loadings Plot is active.

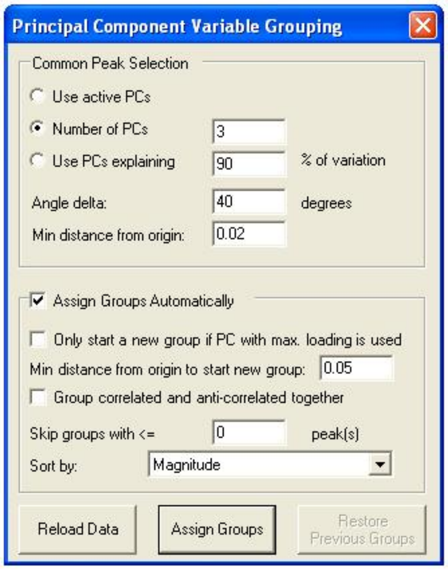
3. Select the **PC Variable Grouping** menu item from the **Utilities** sub-menu of the **Help** menu as shown below.



The window shown below is presented.

4. Fill-in the parameters as shown in the figure. In particular set the **Number of PCs** to '3' and de-select the **Only start a new group if PC with max. loading is used** checkbox.

5. Click the **Assign Groups** button and close the window by clicking in its close box.



Principal Component Variable Grouping

Common Peak Selection

○ Use active PCs

● Number of PCs: 3

○ Use PCs explaining: 90 % of variation

Angle delta: 40 degrees

Min distance from origin: 0.02

☑ Assign Groups Automatically

☐ Only start a new group if PC with max. loading is used

Min distance from origin to start new group: 0.05

☐ Group correlated and anti-correlated together

Skip groups with <= 0 peak(s)

Sort by: Magnitude

The loadings plot should appear as shown below. The variables have been automatically assigned to one of six groups (in addition to the 'Default' group for certain very small variables). These groups roughly correspond to the numbered groups discussed in step (4) of the previous section (section 5.3).

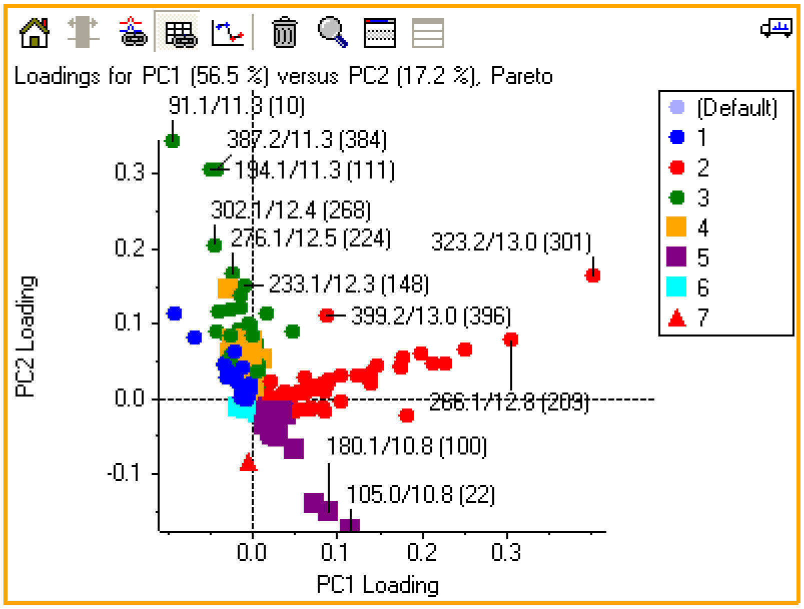
One reason that the grouping is not identical to the visually identified groups of the previous section is because the automatic groups were assigned using information from the first *three* principal components (as selected in the figure above), whereas the visual grouping was based on only the two visible components. This is an important point since it allows a two-dimensional display to be colored in such a way that additional variation not otherwise visible can be seen.

For a detailed discussion of this tool see the MarkerView™ Reference Manual. The concepts underlying the grouping itself are discussed in the following paper:

> *Dimensionality Reduction and Visualization in Principal Component Analysis*
> Anal. Chem., 2008, **80** (13), pp 4933-4944

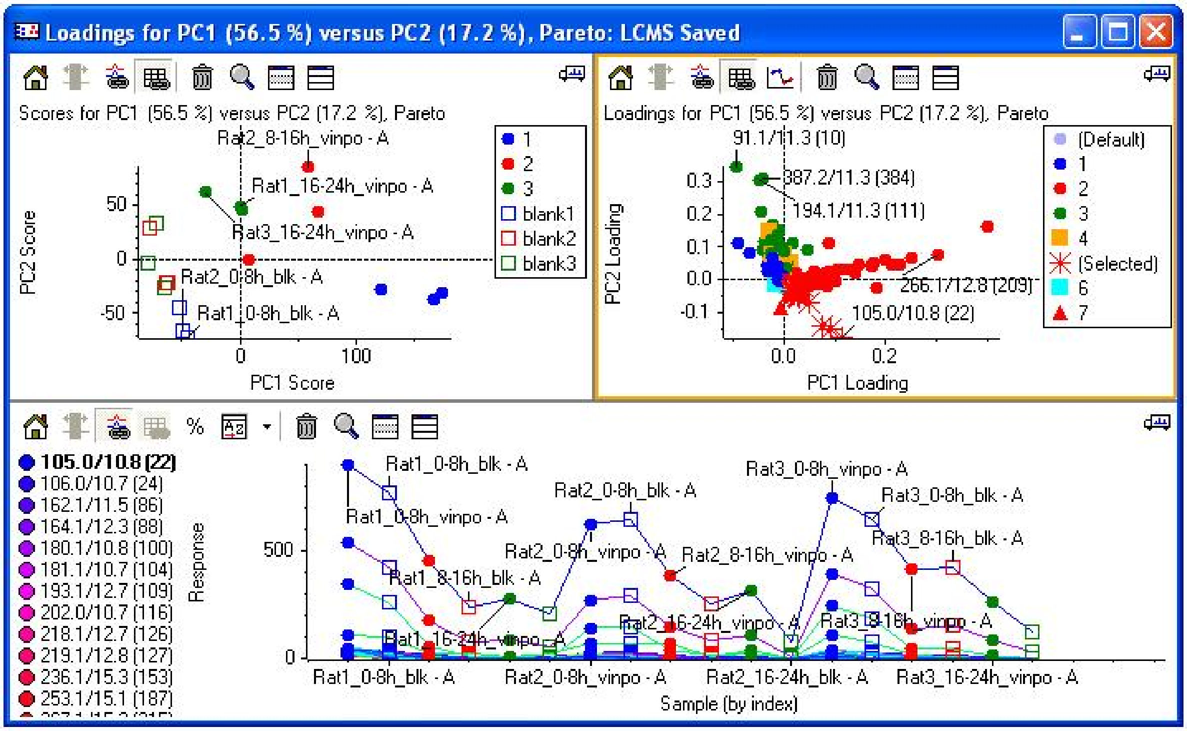
Which is available for download as a pdf file from:

> http://pubs.acs.org/doi/abs/10.1021/ac800110w



6. Use the magnifying glass tool (🔍) so that the Scores Plot is also visible.
7. Select *any* variable in the Loadings Plot (by drawing a selection box around it) and click the **Plot Profile** button to generate a Profile Plot to display that (arbitrary) variable.
8. Click on the color spot to the immediate left of the text for group '5' in the Loadings Plot (you can also double-click the '5' text itself). The display should appear as shown below.

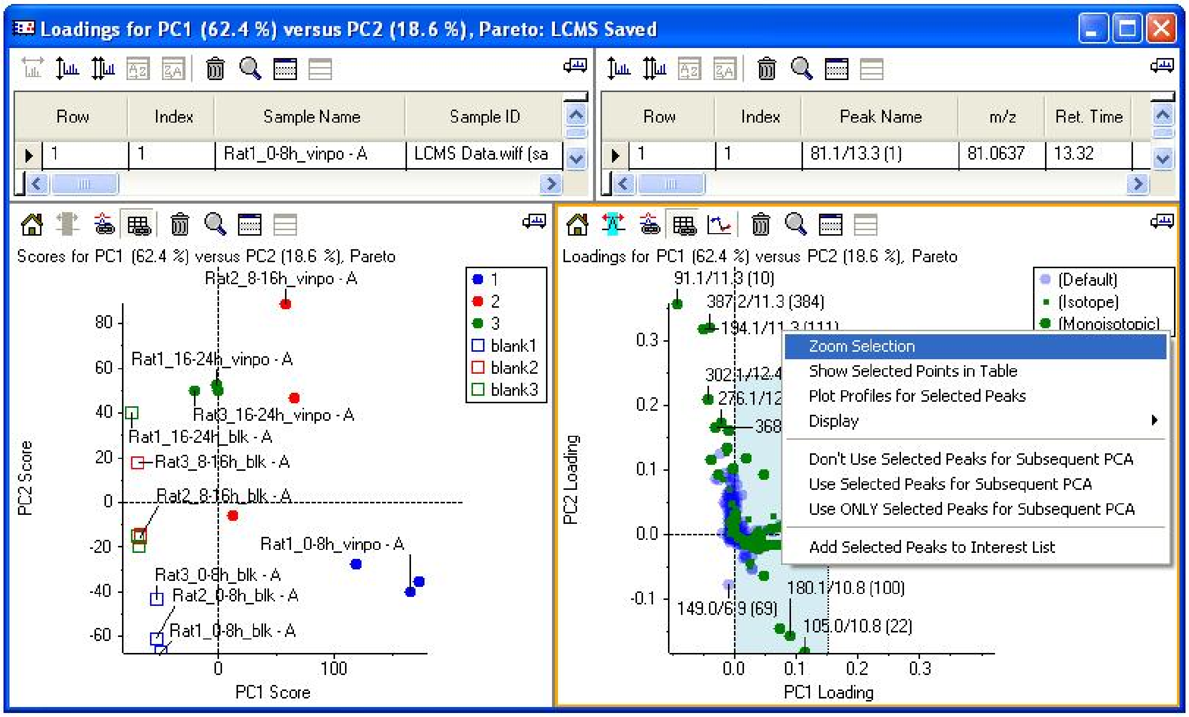
   The Profile Plot will update so that all variables assigned to group 5 are overlaid. This is a very similar display to that shown in step (5) for the previous section (section 5.3). The main difference is that traces for a larger number of variables are overlaid since all group members are used, rather than the subset which as chosen in the manual case.



5.4 Working with the excluded and interest lists

The variables we have looked at so far seem to show 1) a diurnal variation, possibly somewhat suppressed in the post-dose samples and 2) systematic variations that may be due to a contaminant. We will now explore the variables with positive PC1 loadings that we believe are due to metabolites of vinpocetin.

1. From the display which is the result of performing PCA with some variables excluded (you may need to regenerate this plot if you closed it in the previous section), select a region of the loadings plot containing variables in families 1, 2, 3, and 7 (earlier figure) and extending to a PC1 loading of about 0.15, right-click and select **Zoom Selection**.



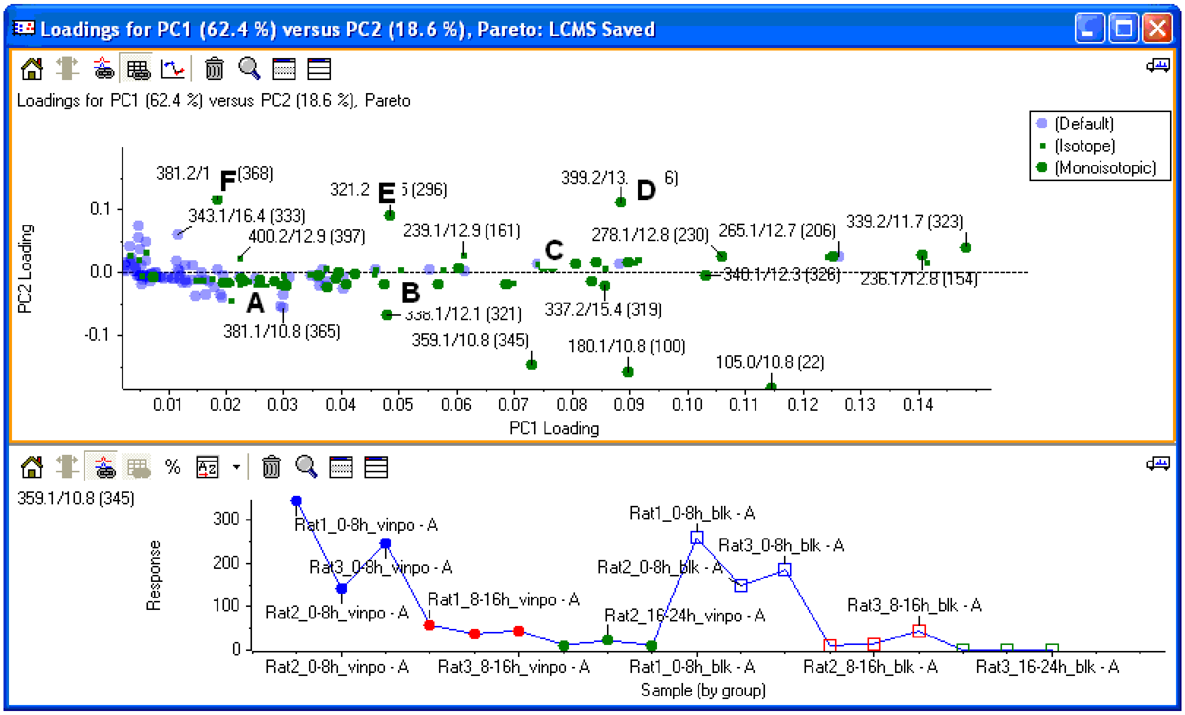
Row	Index	Sample Name	Sample ID
1	1	Rat1_0-8h_vinpo - A	LCMS Data.wiff (sa

Row	Index	Peak Name	m/z	Ret. Time
1	1	81.1/13.3 (1)	81.0637	13.32

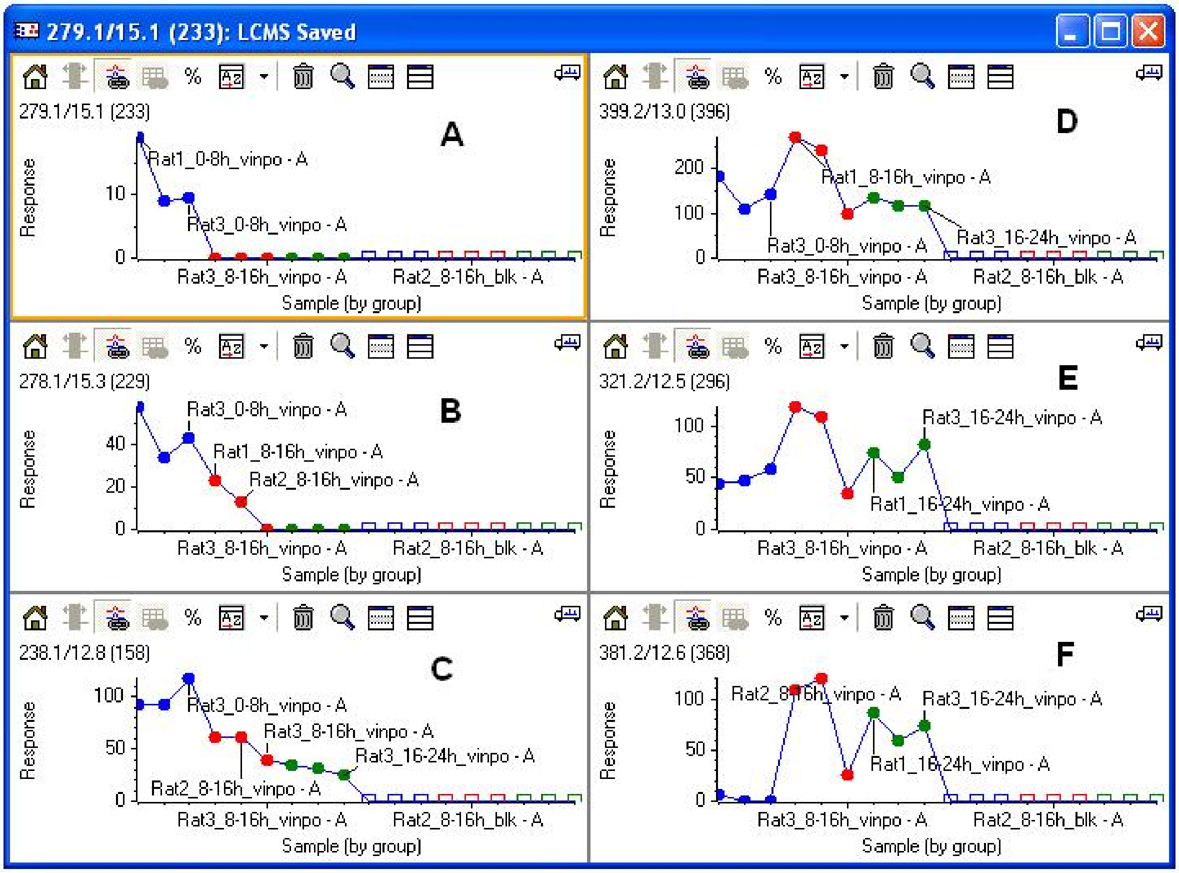
2. Hide all panes except the loadings plot by clicking the **Hide pane** button, select one of the variables with the largest negative PC2 loading (e.g. 359.1/10.8) and generate the profile plot.

3. Change the sorting to **Group Order.**

   As we've seen before, these are the variables that demonstrate the diurnal variation. Some other 'families' of variables are also apparent and are marked (A – F) in the figure below.



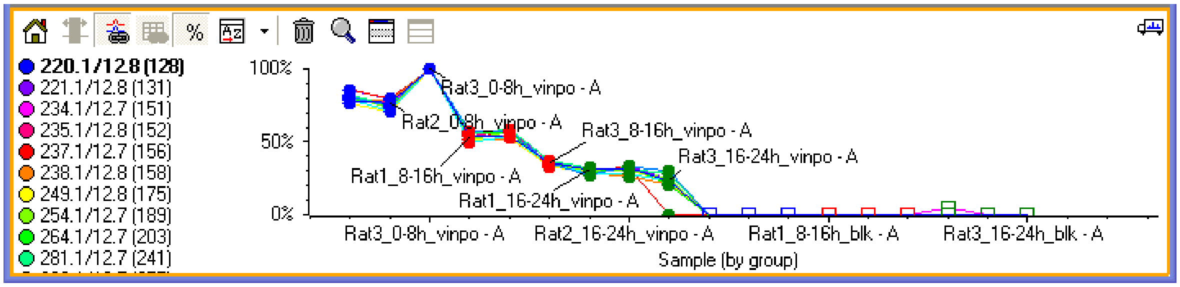
4. As you click on other variable in the loadings plot, the profile graph will update to display the selected variable. Check that **Group Order** is selected for sorting and explore the behavior of other variables, such as those indicated with circles in the above figure; note that the intensity pattern changes as you move counter-clockwise as shown below. (This figure was generated by drawing selection rectangles around the variables, rather than clicking on them, and plotting the profiles. The panes were arranged by dragging the moving truck icon).



The different families illustrate the different kinetics for different metabolites. Those lying along line A occur only in the 0 – 8 hour samples, while the relative amounts in the 8 – 16 and 16 – 24 hour samples increase in going from A to D. For E the 8 – 16 hour intensity is greater than the 0 – 8, and greater still in panel F.

Thus the radial lines correspond to different variables that illustrate the different temporal behavior of the metabolites. Vinpocetin fragments easily so many of the correlated ions are fragments formed in the orifice. A good way to check the correlation is to generate profile plots and use a relative, rather than absolute, y-axis.

5. Delete any profile graphs, click on the **Home** button (🏠) in the loadings plot to restore the full view, select some of the variables with the largest loadings in family C and generate the profile plot. Click the **%** button in the toolbar.

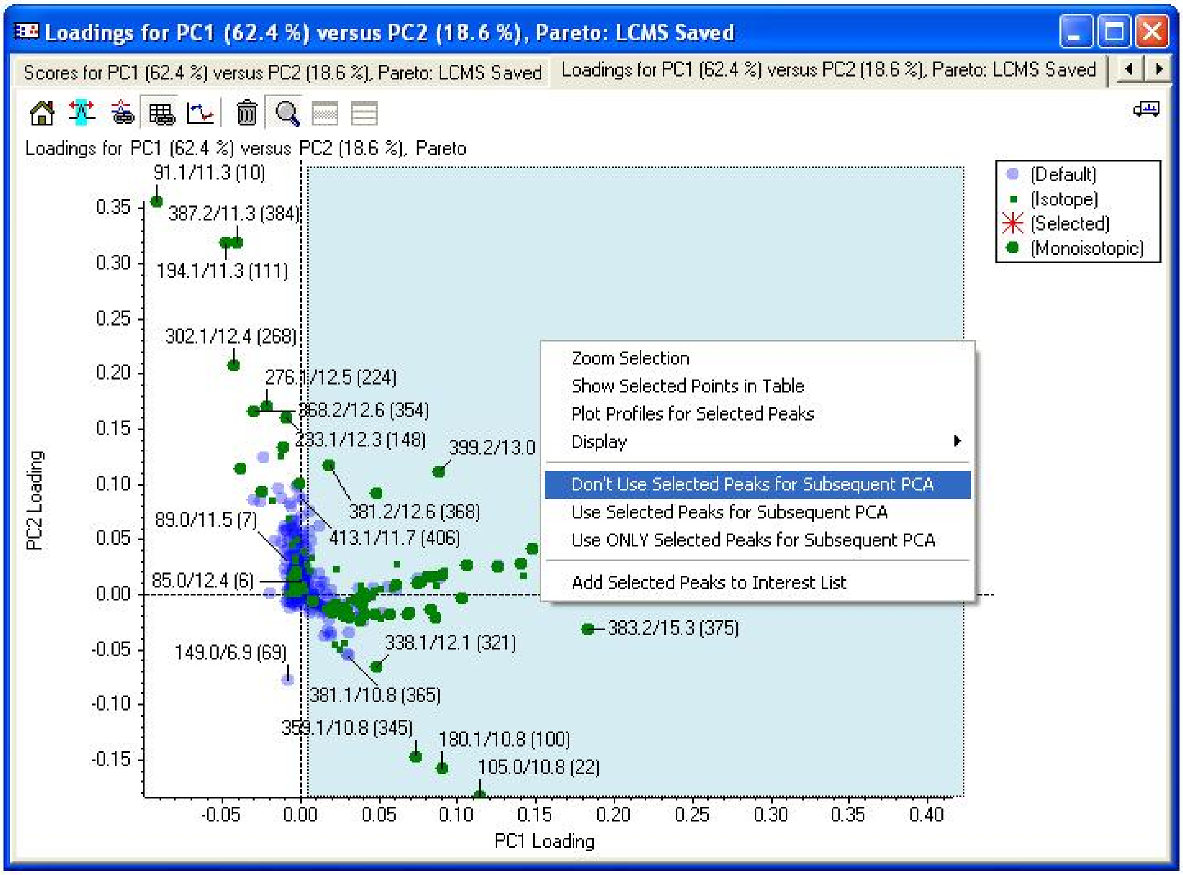


The similarity of the graphs shows that they have similar behavior in the different samples, i.e. they are well correlated (as would be expected if they are all related).

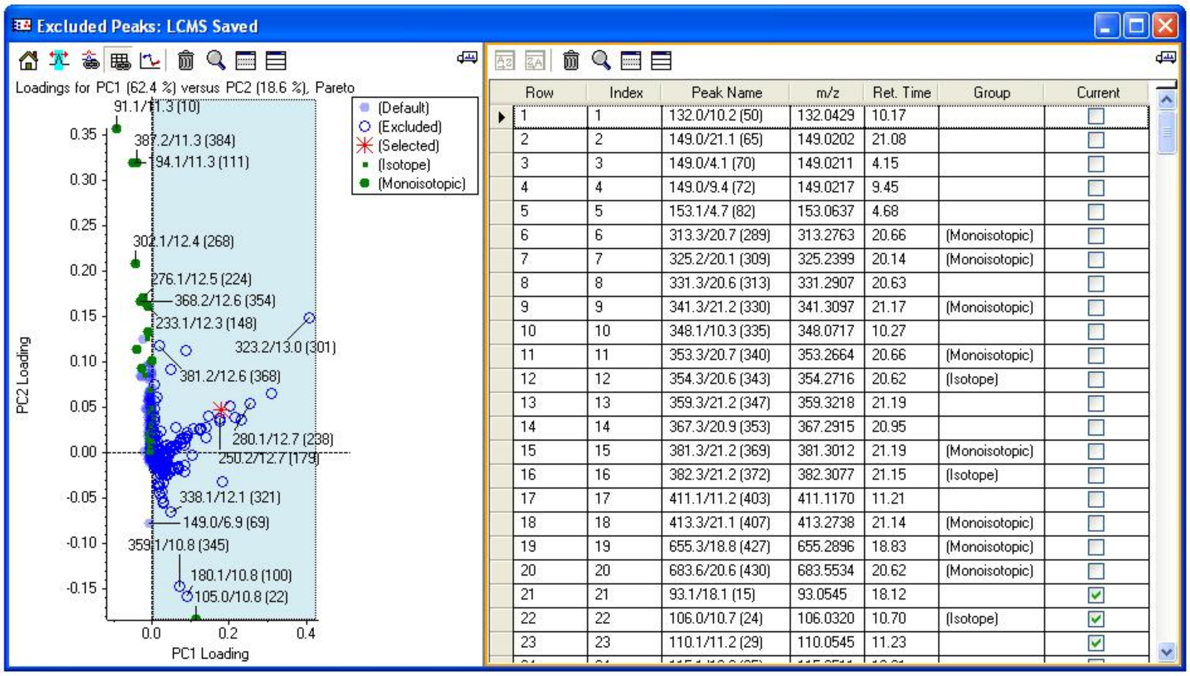
In many cases we are interested in changes in the endogenous metabolites, rather than the xenobiotic metabolites arising from the dosed compound, so we need to exclude the latter from the display; these appear to be variables that have PC1 loadings greater than ca. 0.05.

6. Switch the profile display back to using an absolute scale by clicking the **%** button again.

7. Draw a selection rectangle that includes all of the variables with PC1 loading values greater than ca. 0.005. The simplest way to do this is to start to the right, and slightly above, the point with the largest PC1 loading and drag towards the origin. Right-click and select **Don’t Use Selected Points for Subsequent PCA**. Note that the variables are now drawn with the excluded symbol (and open blue circle by default).


Loadings for PC1 (62.4 %) versus PC2 (18.6 %), Pareto: LCMS Saved
Scores for PC1 (62.4 %) versus PC2 (18.6 %), Pareto: LCMS Saved
Loadings for PC1 (62.4 %) versus PC2 (18.6 %), Pareto: LCMS Saved
Loadings for PC1 (62.4 %) versus PC2 (18.6 %), Pareto
91.1/11.3 (10)
387.2/11.3 (384)
194.1/11.3 (111)
302.1/12.4 (268)
276.1/12.5 (224)
368.2/12.6 (354)
233.1/12.3 (148)
399.2/13.0
381.2/12.6 (368)
413.1/11.7 (406)
89.0/11.5 (7)
85.0/12.4 (6)
383.2/15.3 (375)
338.1/12.1 (321)
149.0/6.9 (69)
381.1/10.8 (365)
359.1/10.8 (345)
180.1/10.8 (100)
105.0/10.8 (22)
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
PC2 Loading
-0.05
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
PC1 Loading
(Default)
(Isotope)
(Selected)
(Monoisotopic)
Zoom Selection
Show Selected Points in Table
Plot Profiles for Selected Peaks
Display
Don't Use Selected Peaks for Subsequent PCA
Use Selected Peaks for Subsequent PCA
Use ONLY Selected Peaks for Subsequent PCA
Add Selected Peaks to Interest List

8. Select **Show Excluded Peaks** from the **View** menu and use the **Truck** icon to drag the resulting list so that it is alongside the loadings plot. (When you are dragging the list pane, the edge of the loadings plot pane will turn red to indicate where it will be drawn. Release the mouse button when the right edge of the plot is red and the list will be drawn in the correct position.)



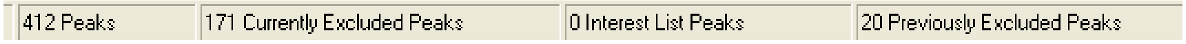
Excluded Peaks: LCMS Saved

Row	Index	Peak Name	m/z	Ret. Time	Group	Current
1	1	132.0/10.2 (50)	132.0429	10.17		☐
2	2	149.0/21.1 (65)	149.0202	21.08		☐
3	3	149.0/4.1 (70)	149.0211	4.15		☐
4	4	149.0/9.4 (72)	149.0217	9.45		☐
5	5	153.1/4.7 (82)	153.0637	4.68		☐
6	6	313.3/20.7 (289)	313.2763	20.66	(Monoisotopic)	☐
7	7	325.2/20.1 (309)	325.2399	20.14	(Monoisotopic)	☐
8	8	331.3/20.6 (313)	331.2907	20.63		☐
9	9	341.3/21.2 (330)	341.3097	21.17	(Monoisotopic)	☐
10	10	348.1/10.3 (335)	348.0717	10.27		☐
11	11	353.3/20.7 (340)	353.2664	20.66	(Monoisotopic)	☐
12	12	354.3/20.6 (343)	354.2716	20.62	(Isotope)	☐
13	13	359.3/21.2 (347)	359.3218	21.19		☐
14	14	367.3/20.9 (353)	367.2915	20.95		☐
15	15	381.3/21.2 (369)	381.3012	21.19	(Monoisotopic)	☐
16	16	382.3/21.2 (372)	382.3077	21.15	(Isotope)	☐
17	17	411.1/11.2 (403)	411.1170	11.21		☐
18	18	413.3/21.1 (407)	413.2738	21.14	(Monoisotopic)	☐
19	19	655.3/18.8 (427)	655.2896	18.83	(Monoisotopic)	☐
20	20	683.6/20.6 (430)	683.5534	20.62	(Monoisotopic)	☐
21	21	93.1/18.1 (15)	93.0545	18.12		☑
22	22	106.0/10.7 (24)	106.0320	10.70	(Isotope)	☑
23	23	110.1/11.2 (29)	110.0545	11.23		☑

Note that the list contains a **Current** column which is checked for some variables and not for others.

Each PCA plot maintains a list of the variables that were excluded when the display was generated (these do not have a check mark in the **Current** column) and a list of the peaks that are excluded in the display but were in use when the display was generated (these have a check mark). In the figure above, the first 20 variables were excluded before the display was generated (i.e. steps 12 and 13 in section 5.3 above), the rest correspond to the ones selected after these plots were generated.

This is also reflected in the status bar at the bottom of the main window



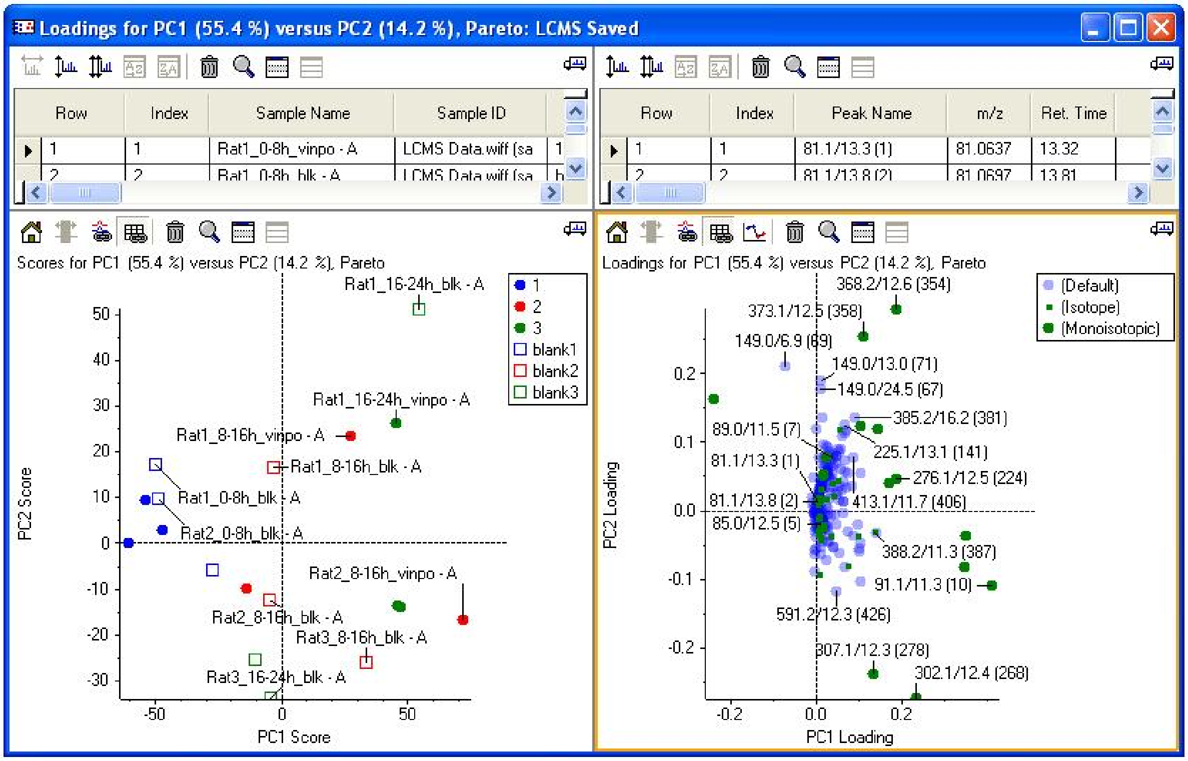
412 Peaks | 171 Currently Excluded Peaks | 0 Interest List Peaks | 20 Previously Excluded Peaks

which, in this case, indicates that 20 variables were previously excluded and 171 have been selected and excluded.

The **Excluded Peaks** list behaves as a normal table. You may sort on any column by clicking on the column heading and then of one of the two sort buttons. You may select one or more columns by dragging in the column headings and these can then be copied to the clipboard, by typing ctrl-C or selecting **Copy** from the **Edit** menu, and pasted into another program such as Excel.

9. Perform another PCA analysis and display the excluded peak list; verify that the new list has no checkmarks indicating that all of the listed variables were excluded before generating the display.

The analysis display will resemble the one shown below, *however* depending on exactly which variables you have excluded, the display map flip about the PC2 axis..



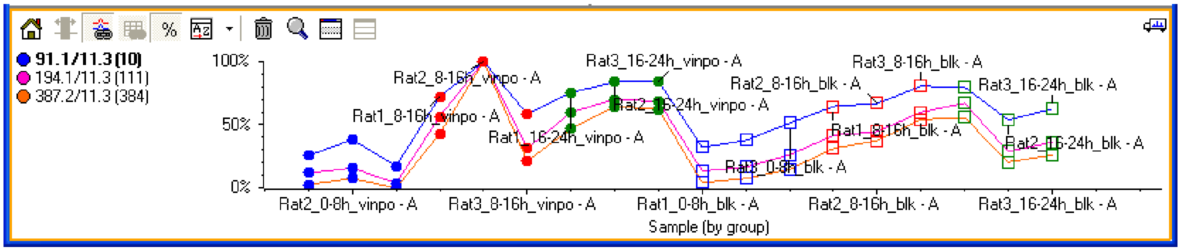
Loadings for PC1 (55.4 %) versus PC2 (14.2 %), Pareto: LCMS Saved

Row	Index	Sample Name	Sample ID	
1	1	Rat1_0-8h_vinpo - A	LCMS Data.wiff (sa	1
2	2	Rat1_0-8h_blk - A	LCMS Data.wiff (sa	h

Row	Index	Peak Name	m/z	Ret. Time
1	1	81.1/13.3 (1)	81.0637	13.32
2	2	81.1/13.8 (2)	81.0697	13.81

The 0 – 8 hour samples (pre- and post-dose) appear to be separated from the rest of the samples with negative PC1 scores, and the majority of the remaining post-dose samples have high positive PC1 scores. This suggests that the variables with high, positive PC1 loadings will be more intense in the remaining (8-16 and 16-24) post-dose samples.

10. Select the variables with the highest positive PC1 loadings and plot the profiles.

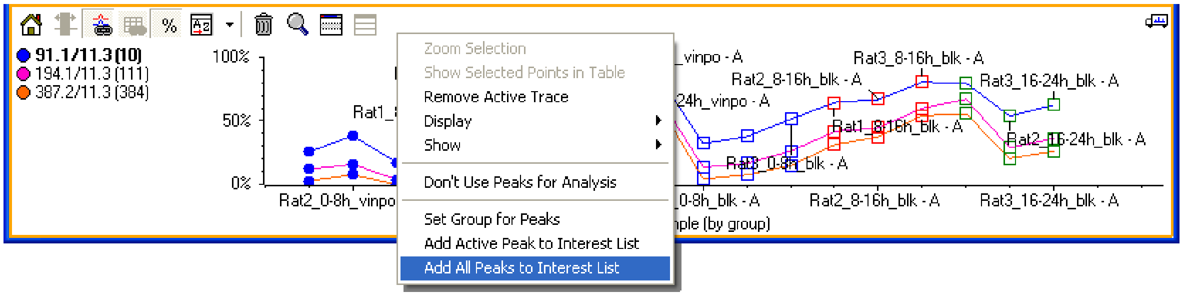


These variables appear in most samples and the behavior is modified in those obtained post-dose, being somewhat higher in the 8 – 16 and 16 – 24 hour samples and, perhaps, lower in the 0 – 8 hour samples. Since they all have the same retention time (11.3 min.) they are likely related: m/z 387 is probably a dimer (2M + H^+) of the ion at 194 (MH^+) and 91.1 a fragment.

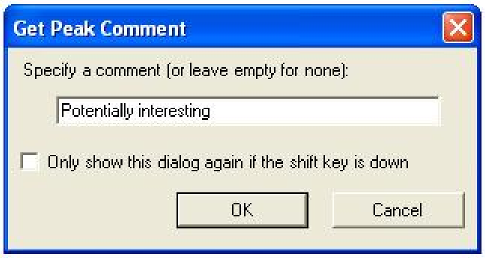
These may be variables that we want to process further, so we will transfer them to the interest list.

11. The profile graph's context menu allows some flexibility in editing the variables displayed and adding them to the interest list. If you click on a data point in any trace, that trace will be made active, i.e. it will appear at the top of the variable list at the left of the display and will be labeled. You may remove it by selecting **Remove Active Trace** or add it to the list by selecting **Add Active Peak to Interest List**. Removing traces in this way is useful if you have accidentally displayed a variable that is not relevant, perhaps because its profile shows no variation.

    Right-click in the profile graph and select **Add All Peaks to Interest List** in the context menu.



A dialog box will appear so that you may enter a comment; when you click **OK** the variables and the comment will be added to the interest list.



Get Peak Comment

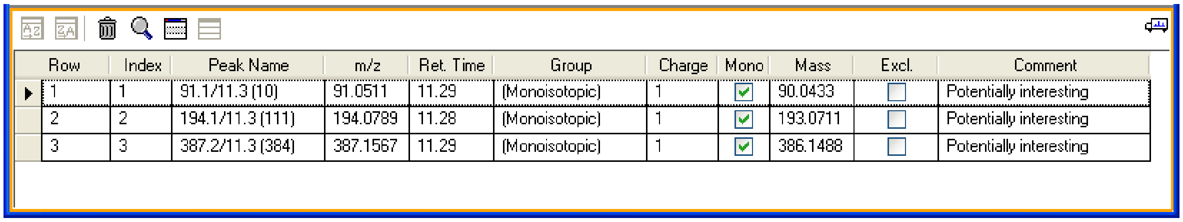
Specify a comment (or leave empty for none):

Potentially interesting

☐ Only show this dialog again if the shift key is down

12. From the **View** menu select **Show Interest List.**

You may manipulate the interest list (sort, copy, etc.) as with other tables. Unlike the exclusion list, there is only one interest list.



Row	Index	Peak Name	m/z	Ret. Time	Group	Charge	Mono	Mass	Excl.	Comment
1	1	91.1/11.3 (10)	91.0511	11.29	(Monoisotopic)	1	☑	90.0433	☐	Potentially interesting
2	2	194.1/11.3 (111)	194.0789	11.28	(Monoisotopic)	1	☑	193.0711	☐	Potentially interesting
3	3	387.2/11.3 (384)	387.1567	11.29	(Monoisotopic)	1	☑	386.1488	☐	Potentially interesting

The interest list contains other peak metrics such as the assigned variable group, charge state, calculated mass (cf. m/z) etc. The calculated mass is obtained from m/z and charge assuming that protons are gained in positive mode and lost in negative mode; isotope peaks have their own mass not that of the monoisotopic peak.

Since displays are not removed as new ones are generated, it is possible to 'back up' to an earlier display and continue exploring the data. Close the current window and the previous window, including the selection region used to exclude variables will be revealed. Right-click in the selection rectangle and select **Use Selected Peaks for Subsequent PCA** to restore those variables.

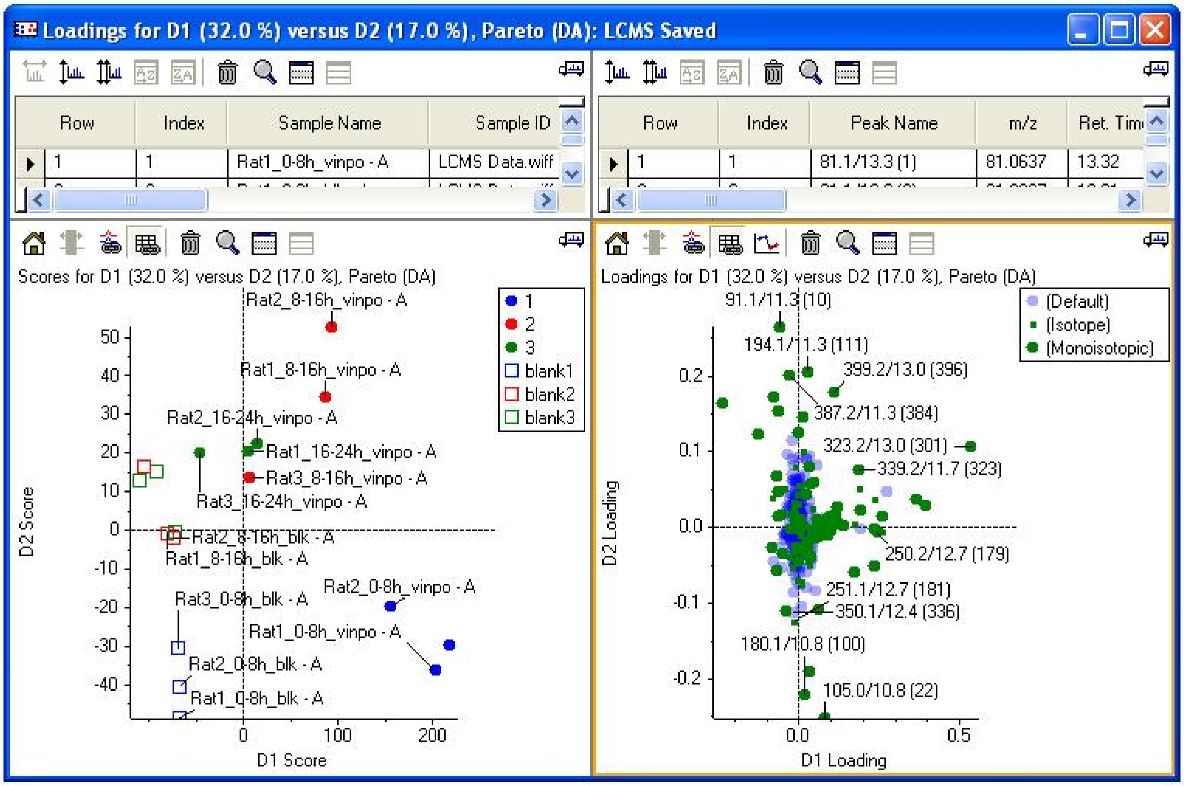
If you save the data, probably with a different file name, the excluded samples and variables are remembered so that the exclusion process does not need to be repeated.

5.5 Using Principal Components Analysis – Discriminant Analysis (PCA-DA)

Discriminant analysis (DA), like the t-test, is a supervised method that is used to find differences between known groups. The MarkerView™ Software allows DA to be combined with PCA by clicking on the **Perform PCA-DA (supervised)** checkbox in the PCA Options dialog box (see section 5.3, Performing PCA and interpreting the results).

When this box is checked, the software first performs PCA as normal using the weighting and scaling parameters specified, which reduces the dimensionality of the data by generating a few PC's that are combinations of the original variables. The PC's are then combined with the group information to find combinations that maximize the variance between groups while minimizing the variance within groups. This can often dramatically enhance the appearance of the separation as shown by the scores plot; the results are interpreted as before.

1. Close all open windows and open the data table that was saved in section 5.2.
2. Perform PCA with no weighting and Pareto scaling but click on the **Perform PCA-DA (supervised)** checkbox to select it. The result is shown below.



Loadings for D1 (32.0 %) versus D2 (17.0 %), Pareto (DA): LCMS Saved

Row	Index	Sample Name	Sample ID
1	1	Rat1_0-8h_vinpo - A	LCMS Data.wiff

Row	Index	Peak Name	m/z	Ret. Tim
1	1	81.1/13.3 (1)	81.0637	13.32

Note that the labeling is now shown as D1, D2, etc. in order to distinguish this type of analysis from normal PCA, and that only five discriminants are needed.

In this particular example the grouping in the scores plot does not change greatly (compare the figure above with the scores and loadings plots in section 5.3). However members of the individual groups are closer together and the separation between the 0 – 8 hr. samples and all others is enhanced.

The loadings plot has changed to reflect the new processing but is interpreted as before.

By constructing artificial groups, PCA-DA can be used to determine and exclude variables that correspond to changes that are not relevant to your study, for example the diurnal changes that result in the 0 – 8 hr samples being separated from the others.

5.6 Summary

In this section you have learned how to

- Import LCMS data and perform sample alignment
- Assign multiple groups and symbols to allow better visualization of the results
- Perform a PCA analysis and interpret the results
- Detect and exclude variables that appear to arise from a systematic experimental variation

- Detect and exclude variables that appear to be xenobiotic metabolites (a careful examination would require more detailed knowledge of the compound as well as its metabolic and fragmentation behavior).
- Review the excluded peaks and copy them for further processing
- Add selected variables to an interest list for additional processing
- 'Back up' to an earlier state and continue processing
- Use PCA-DA to enhance the separation of known groups

These sections have described the most common operations; more advanced topics are covered in the following sections and more details on the various parameters, dialogs, etc. can be found in the reference manual.

6 Miscellaneous

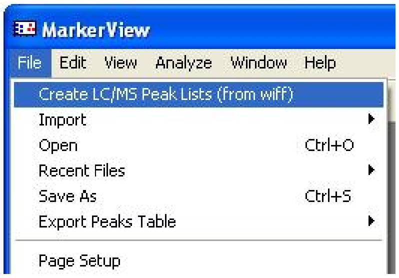
This section describes some of the many additional features of the MarkerView™ Software. It assumes that you have worked through the rest of this manual so only new material is described in detail.

6.1 Generating and importing Peaks files

When you are working with large, complex LCMS data sets, the process of importing, aligning and normalizing the data may be slow. The program allows you to divide this into two separate steps so peak finding, which is the slowest part, need only be performed once and you can experiment more easily with the alignment and normalization parameters. In addition, both steps have separate minimum intensity parameters so you can use a very low threshold to find the peaks initially and later reject small peaks that may be due to noise.

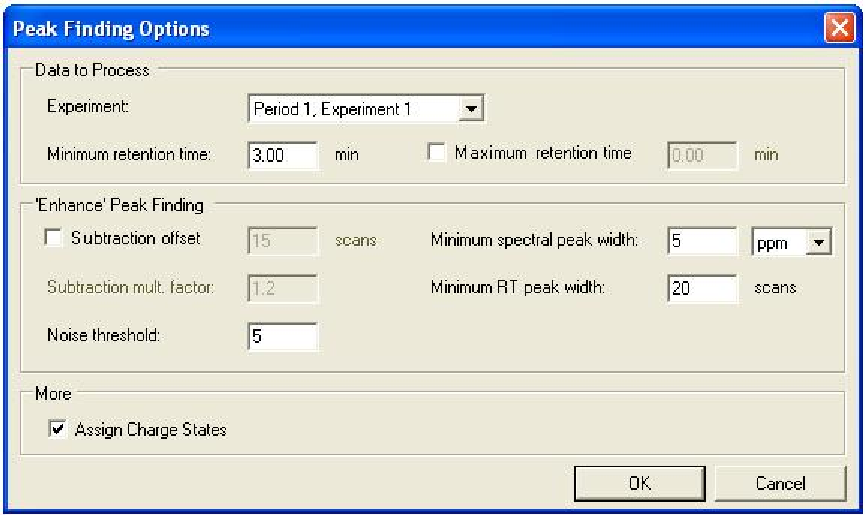
6.1.1 Generating peak list files

1. Select **Create LC/MS Peaks Lists (from wiff)** from the **File** menu.



   A dialog box will explain the purpose of the command. Click **OK** to dismiss it.

2. In the **Select Samples** dialog select the LCMS data files, removing the first, as described in section 5.1 and click **OK.**
3. The program will ask for a folder to receive the peak list files. In the **Browse For Folder** dialog, locate a convenient folder (for example the original **LCMS Data** folder), click the **Make New Folder** button and change the name of the new folder to **LCMS Peaks.** Click **OK.**
4. In the **Peak Finding Options** dialog, fill in the parameters as shown below.


Peak Finding Options

Data to Process

Experiment: Period 1, Experiment 1

Minimum retention time: 3.00 min

☐ Maximum retention time 0.00 min

'Enhance' Peak Finding

☐ Subtraction offset 15 scans

Minimum spectral peak width: 5 ppm

Subtraction mult. factor: 1.2

Minimum RT peak width: 20 scans

Noise threshold: 5

More

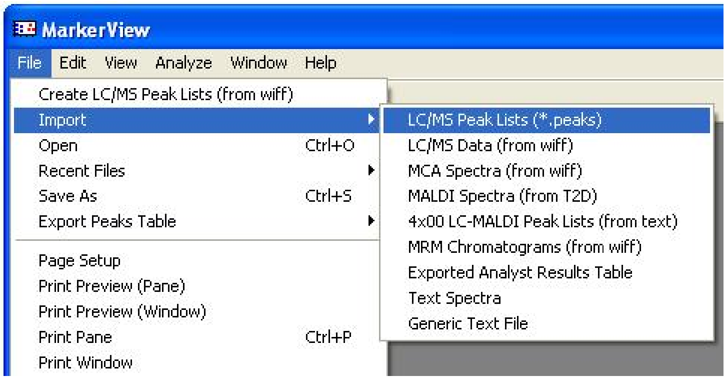
☑ Assign Charge States

5. Click **OK.** The files will be processed individually and a peak list file generated for each.

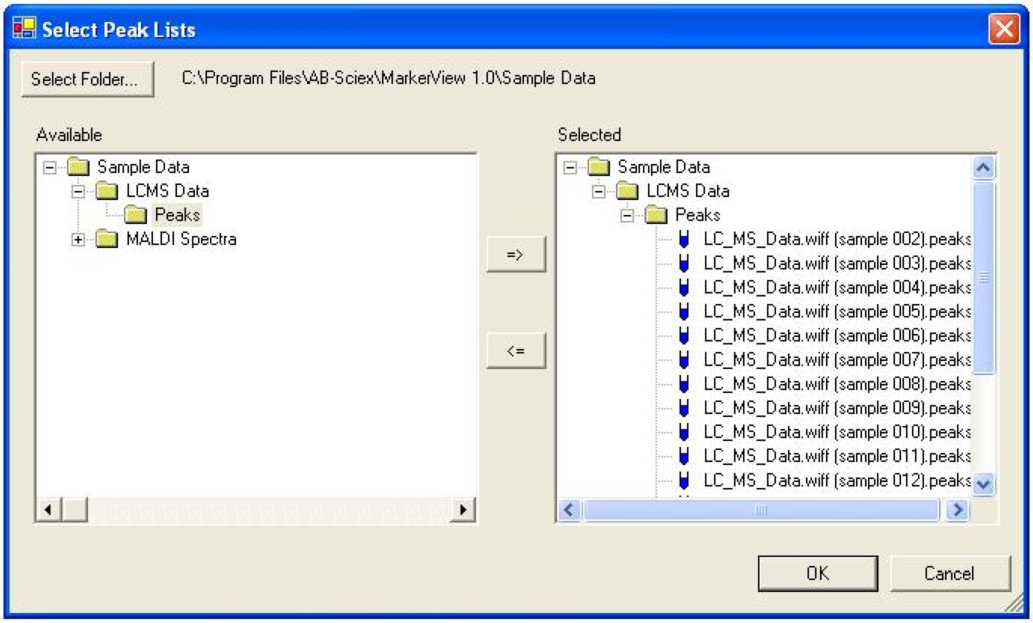
*Note: the folder named 'Peaks' already contains the peak lists for these samples so you can skip the last step if you wish by clicking **Cancel** instead of **OK**.*

6.1.2 Importing peak list files

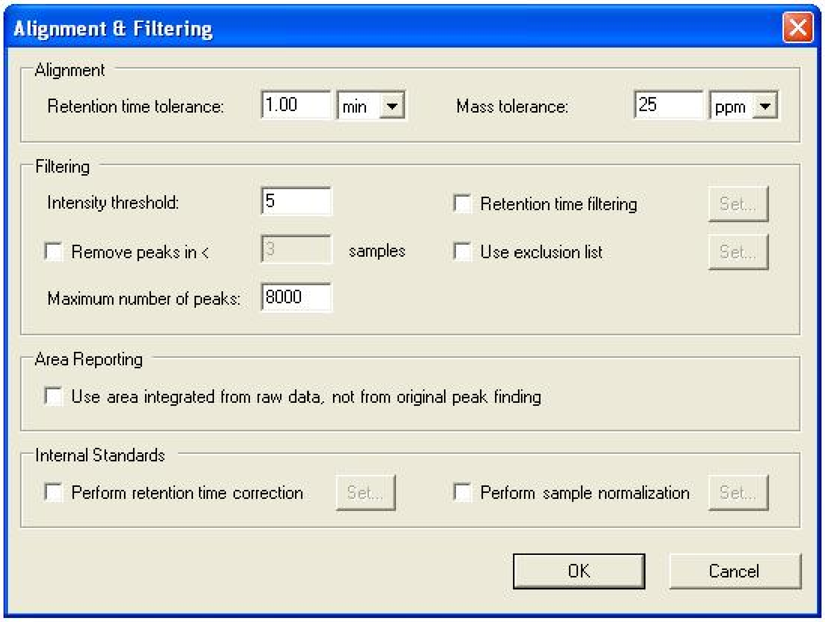
1. From the **File** menu select **LC/MS Peak Lists (*.peaks)**.



2. Locate the **Peaks** folder in the **MarkerView\Example Data\LCMS data** (or the folder you created in section 6.1.1) and drag it to the right side of the display. Click **OK.**


Select Peak Lists
Select Folder...
C:\Program Files\AB-Sciex\MarkerView 1.0\Sample Data
Available
Sample Data
LCMS Data
Peaks
MALDI Spectra
=>
<=
Selected
Sample Data
LCMS Data
Peaks
LC_MS_Data.wiff (sample 002).peaks
LC_MS_Data.wiff (sample 003).peaks
LC_MS_Data.wiff (sample 004).peaks
LC_MS_Data.wiff (sample 005).peaks
LC_MS_Data.wiff (sample 006).peaks
LC_MS_Data.wiff (sample 007).peaks
LC_MS_Data.wiff (sample 008).peaks
LC_MS_Data.wiff (sample 009).peaks
LC_MS_Data.wiff (sample 010).peaks
LC_MS_Data.wiff (sample 011).peaks
LC_MS_Data.wiff (sample 012).peaks
OK
Cancel

3. The dialog box that appears resembles that seen in section 5.1 but has some additional parameters to control the way the data is filtered. Fill in the fields as shown below and click **OK.**



Alignment

Retention time tolerance: 1.00 min

Mass tolerance: 25 ppm

Filtering

Intensity threshold: 5

□ Retention time filtering Set...

□ Remove peaks in < 3 samples

□ Use exclusion list Set...

Maximum number of peaks: 8000

Area Reporting

□ Use area integrated from raw data, not from original peak finding

Internal Standards

□ Perform retention time correction Set...

□ Perform sample normalization Set...

When the import process is complete the sample table (as in section 5.1) will appear.

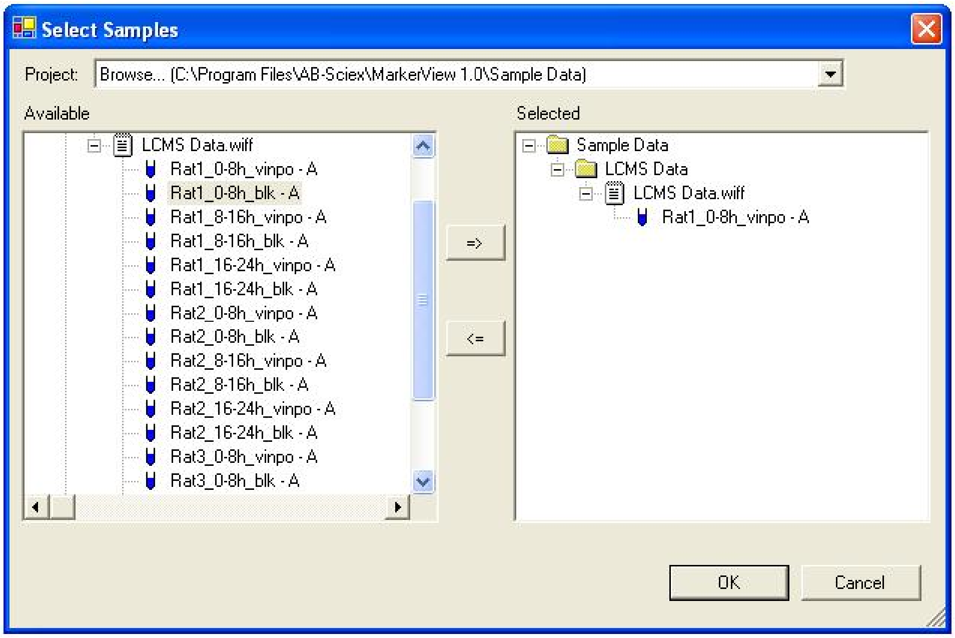
Since importing peak lists is much faster than importing from the original data files, you may want to experiment with the different parameters and observe the effect on the PCA displays. Particularly important are the alignment parameters since these determine if peaks that are close in m/z and or retention time will be combined or not. The intensity and minimum retention time parameters can also have a significant effect, but will have no effect if set to values that are less than those used to import the peaks initially.

6.2 Reviewing peak finder performance

Peak finding is a critical part of the program and it is important to set the parameters correctly to generate the best results. This is invariably a compromise since including small noise peaks will add no value to the calculations and may confuse the displays, while small real peaks may be critical to the separation desired.

A good way to evaluate the peak finder is to import a small range of the data from a single sample and observe the behavior using chromatograms and contour plots as described in this section.

1. From the **File** menu select **Import -> LC/MS Data (from wiff)**.
2. Locate the **LCMS data** folder, expand the **LCMS Data.wiff** file by clicking on the '+' sign adjacent to the file name, drag the *second* sample to the right side of the display (**Selected**), and click **OK**.



Select Samples

Project: Browse... (C:\Program Files\AB-Sciex\MarkerView 1.0\Sample Data)

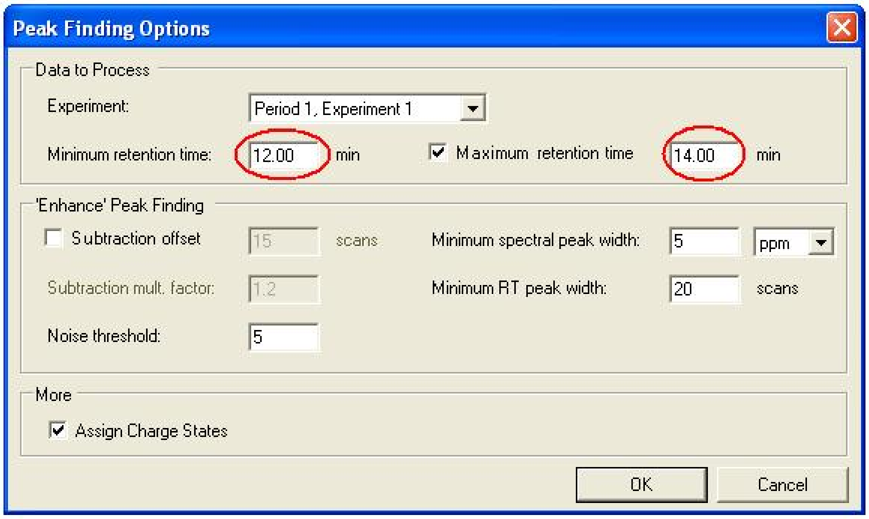
Available

- LCMS Data.wiff
  - Rat1_0-8h_vinpo - A
  - Rat1_0-8h_blk - A
  - Rat1_8-16h_vinpo - A
  - Rat1_8-16h_blk - A
  - Rat1_16-24h_vinpo - A
  - Rat1_16-24h_blk - A
  - Rat2_0-8h_vinpo - A
  - Rat2_0-8h_blk - A
  - Rat2_8-16h_vinpo - A
  - Rat2_8-16h_blk - A
  - Rat2_16-24h_vinpo - A
  - Rat2_16-24h_blk - A
  - Rat3_0-8h_vinpo - A
  - Rat3_0-8h_blk - A

Selected

- Sample Data
  - LCMS Data
    - LCMS Data.wiff
      - Rat1_0-8h_vinpo - A

3. Set the **Minimum retention time** to 12 min, click to check the **Maximum retention time** check box and enter 14 min. Set the other parameters as shown below (these are the same settings as used in section 5.1) and click **OK**.



Peak Finding Options

Data to Process

Experiment: Period 1, Experiment 1

Minimum retention time: 12.00 min ☑ Maximum retention time 14.00 min

'Enhance' Peak Finding

☐ Subtraction offset 15 scans

Subtraction mult. factor: 1.2

Noise threshold: 5

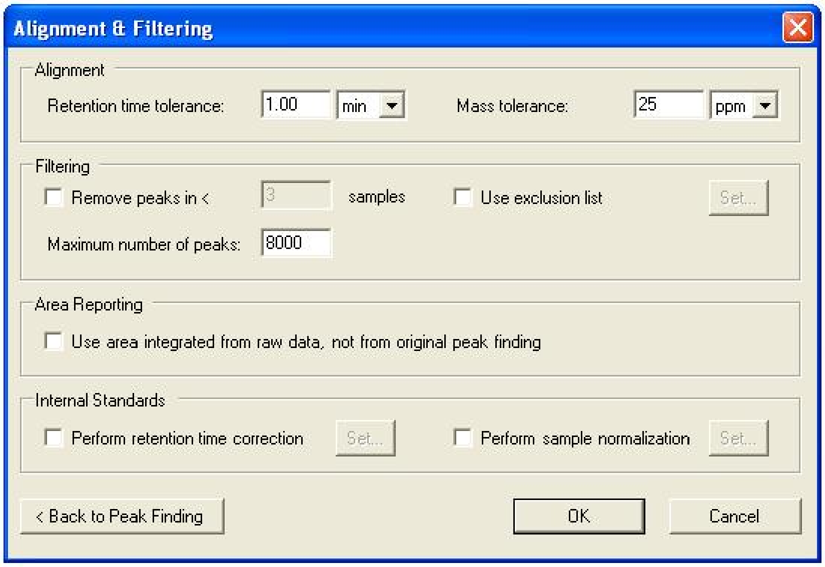
Minimum spectral peak width: 5 ppm

Minimum RT peak width: 20 scans

More

☑ Assign Charge States

4. The next dialog box allows you to set the alignment and filtering parameters. While the purpose of alignment is mainly to ensure that peaks in separate files with similar m/z and retention time values are assigned to the same variable, it is also applied to the peaks within one sample. Set the **Retention time tolerance** to 1 min. and the **Mass tolerance** to 25 ppm. Click **OK.**



Alignment

Retention time tolerance: 1.00 min

Mass tolerance: 25 ppm

Filtering

☐ Remove peaks in < 3 samples

☐ Use exclusion list

Maximum number of peaks: 8000

Area Reporting

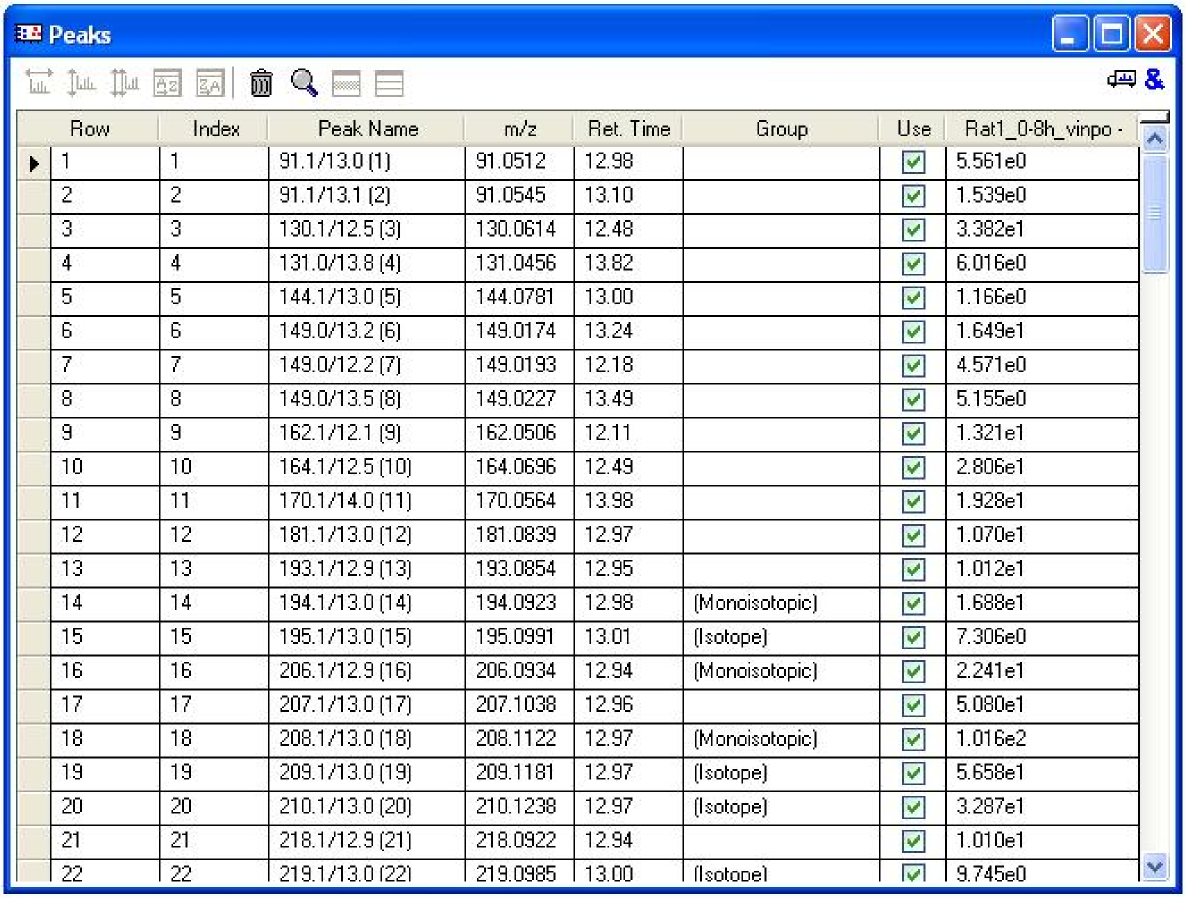
☐ Use area integrated from raw data, not from original peak finding

Internal Standards

☐ Perform retention time correction

☐ Perform sample normalization

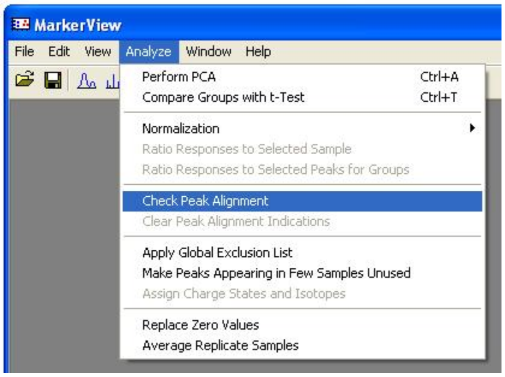
When the import process is complete, a data table with a single sample column will be generated. With the parameters given the table will contain 88 rows (peaks).

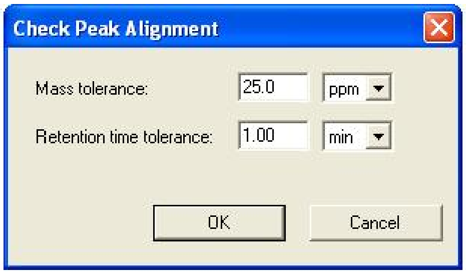


Row	Index	Peak Name	m/z	Ret. Time	Group	Use	Rat1_0-8h_vinpo -
1	1	91.1/13.0 (1)	91.0512	12.98		☑	5.561e0
2	2	91.1/13.1 (2)	91.0545	13.10		☑	1.539e0
3	3	130.1/12.5 (3)	130.0614	12.48		☑	3.382e1
4	4	131.0/13.8 (4)	131.0456	13.82		☑	6.016e0
5	5	144.1/13.0 (5)	144.0781	13.00		☑	1.166e0
6	6	149.0/13.2 (6)	149.0174	13.24		☑	1.649e1
7	7	149.0/12.2 (7)	149.0193	12.18		☑	4.571e0
8	8	149.0/13.5 (8)	149.0227	13.49		☑	5.155e0
9	9	162.1/12.1 (9)	162.0506	12.11		☑	1.321e1
10	10	164.1/12.5 (10)	164.0696	12.49		☑	2.806e1
11	11	170.1/14.0 (11)	170.0564	13.98		☑	1.928e1
12	12	181.1/13.0 (12)	181.0839	12.97		☑	1.070e1
13	13	193.1/12.9 (13)	193.0854	12.95		☑	1.012e1
14	14	194.1/13.0 (14)	194.0923	12.98	(Monoisotopic)	☑	1.688e1
15	15	195.1/13.0 (15)	195.0991	13.01	(Isotope)	☑	7.306e0
16	16	206.1/12.9 (16)	206.0934	12.94	(Monoisotopic)	☑	2.241e1
17	17	207.1/13.0 (17)	207.1038	12.96		☑	5.080e1
18	18	208.1/13.0 (18)	208.1122	12.97	(Monoisotopic)	☑	1.016e2
19	19	209.1/13.0 (19)	209.1181	12.97	(Isotope)	☑	5.658e1
20	20	210.1/13.0 (20)	210.1238	12.97	(Isotope)	☑	3.287e1
21	21	218.1/12.9 (21)	218.0922	12.94		☑	1.010e1
22	22	219.1/13.0 (22)	219.0985	13.00	(Isotope)	☑	9.745e0

5. Close the table and re-import the data using a **Retention time tolerance** of 0.5 min. and a **Mass tolerance** of 10 ppm. The resulting table will contain 104 rows indicating that there are several peaks that are very close and were merged in the first operation.

6. In order to see these peaks, select **Check Peak Alignment** from the **Analyze** menu, enter a **Mass tolerance** of 25 ppm and a **Retention time tolerance** of 1 min. and click **OK**.


MarkerView
File
Edit
View
Analyze
Window
Help
Perform PCA Ctrl+A
Compare Groups with t-Test Ctrl+T
Normalization
Ratio Responses to Selected Sample
Ratio Responses to Selected Peaks for Groups
Check Peak Alignment
Clear Peak Alignment Indications
Apply Global Exclusion List
Make Peaks Appearing in Few Samples Unused
Assign Charge States and Isotopes
Replace Zero Values
Average Replicate Samples


Check Peak Alignment

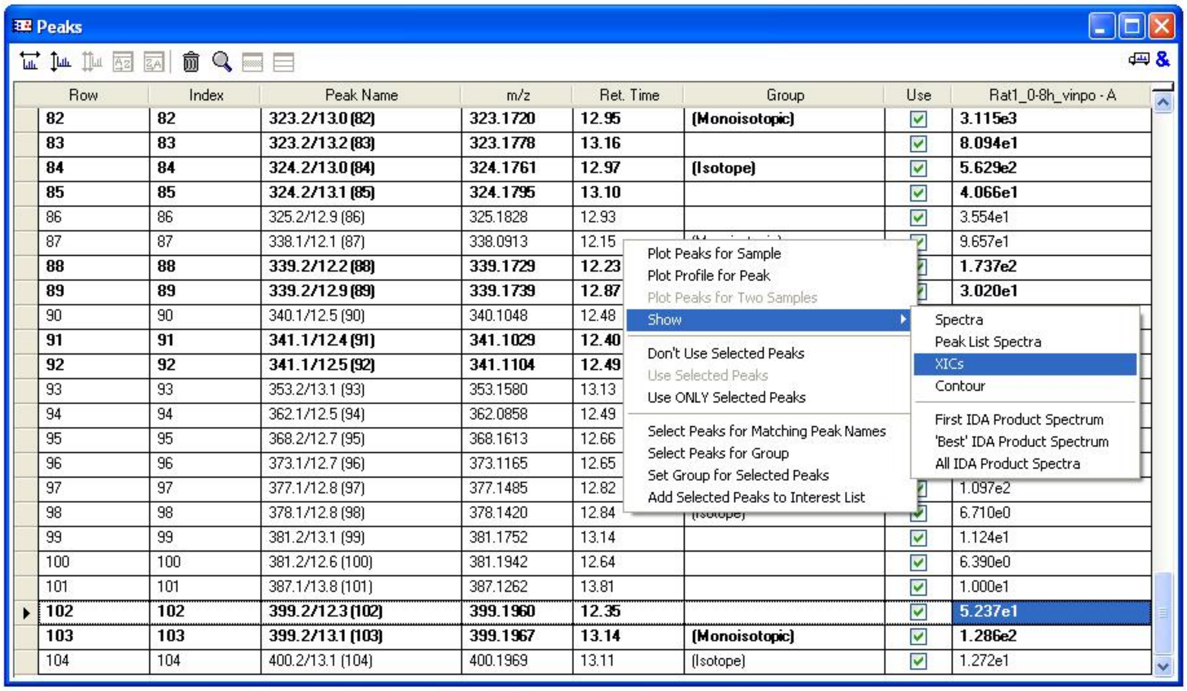
Mass tolerance: 25.0 ppm

Retention time tolerance: 1.00 min

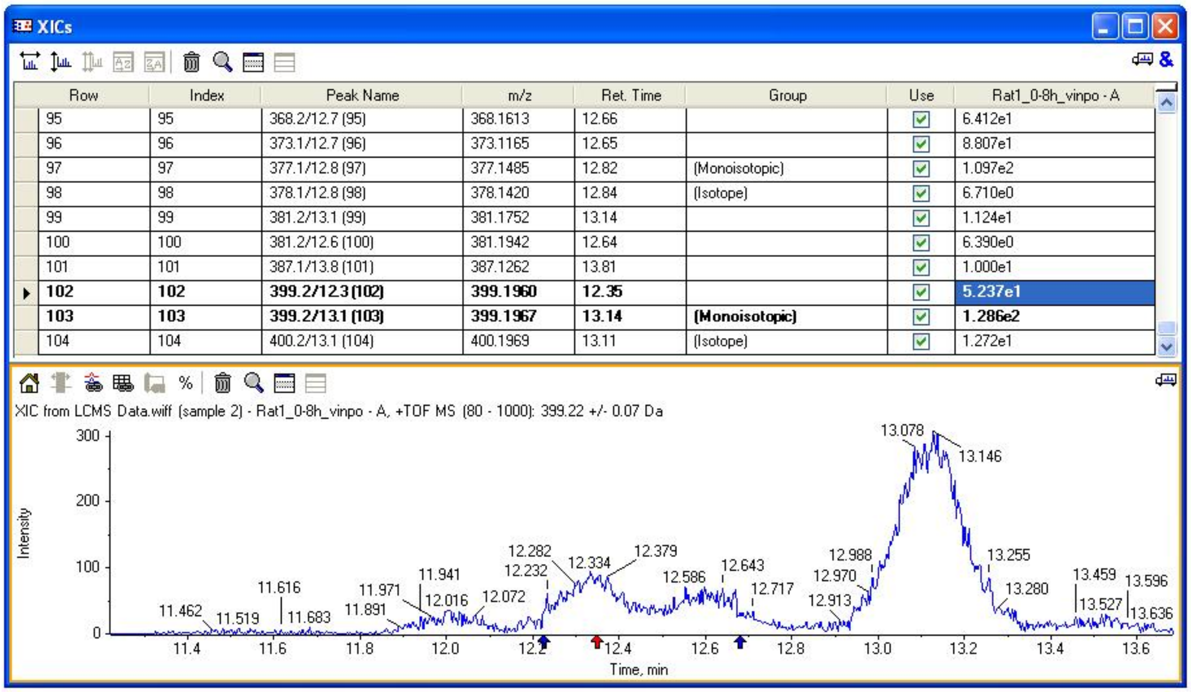
Rows in the table that are within these tolerance values will be highlighted in bold so you can locate them and determine if they are separate peaks or not.

7. Scroll the table so that the rows containing the variables with m/z 399.2 are visible. In this case the m/z values are very similar but the retention times are different by 0.79 min (47 sec.).

8. Select one of the cells in the only sample column, right-click and select **Show XICs**.


Peaks

Row	Index	Peak Name	m/z	Ret. Time	Group	Use	Rat1_0-8h_vinpo - A
82	**82**	**323.2/13.0 (82)**	**323.1720**	**12.95**	**(Monoisotopic)**	☑	**3.115e3**
83	**83**	**323.2/13.2 (83)**	**323.1778**	**13.16**		☑	**8.094e1**
84	**84**	**324.2/13.0 (84)**	**324.1761**	**12.97**	**(Isotope)**	☑	**5.629e2**
85	**85**	**324.2/13.1 (85)**	**324.1795**	**13.10**		☑	**4.066e1**
86	86	325.2/12.9 (86)	325.1828	12.93		☑	3.554e1
87	87	338.1/12.1 (87)	338.0913	12.15		☑	9.657e1
88	**88**	**339.2/12.2 (88)**	**339.1729**	**12.23**		☑	**1.737e2**
89	**89**	**339.2/12.9 (89)**	**339.1739**	**12.87**		☑	**3.020e1**
90	90	340.1/12.5 (90)	340.1048	12.48			
91	**91**	**341.1/12.4 (91)**	**341.1029**	**12.40**			
92	**92**	**341.1/12.5 (92)**	**341.1104**	**12.49**			
93	93	353.2/13.1 (93)	353.1580	13.13			
94	94	362.1/12.5 (94)	362.0858	12.49			
95	95	368.2/12.7 (95)	368.1613	12.66			
96	96	373.1/12.7 (96)	373.1165	12.65			
97	97	377.1/12.8 (97)	377.1485	12.82		☑	1.097e2
98	98	378.1/12.8 (98)	378.1420	12.84	(Isotope)	☑	6.710e0
99	99	381.2/13.1 (99)	381.1752	13.14		☑	1.124e1
100	100	381.2/12.6 (100)	381.1942	12.64		☑	6.390e0
101	101	387.1/13.8 (101)	387.1262	13.81		☑	1.000e1
102	**102**	**399.2/12.3 (102)**	**399.1960**	**12.35**		☑	**5.237e1**
103	**103**	**399.2/13.1 (103)**	**399.1967**	**13.14**	**(Monoisotopic)**	☑	**1.286e2**
104	104	400.2/13.1 (104)	400.1969	13.11	(Isotope)	☑	1.272e1


XICs

	Row	Index	Peak Name	m/z	Ret. Time	Group	Use	Rat1_0-8h_vinpo - A
	95	95	368.2/12.7 (95)	368.1613	12.66		☑	6.412e1
	96	96	373.1/12.7 (96)	373.1165	12.65		☑	8.807e1
	97	97	377.1/12.8 (97)	377.1485	12.82	(Monoisotopic)	☑	1.097e2
	98	98	378.1/12.8 (98)	378.1420	12.84	(Isotope)	☑	6.710e0
	99	99	381.2/13.1 (99)	381.1752	13.14		☑	1.124e1
	100	100	381.2/12.6 (100)	381.1942	12.64		☑	6.390e0
	101	101	387.1/13.8 (101)	387.1262	13.81		☑	1.000e1
▶	**102**	**102**	**399.2/12.3 (102)**	**399.1960**	**12.35**		☑	**5.237e1**
	103	**103**	**399.2/13.1 (103)**	**399.1967**	**13.14**	**(Monoisotopic)**	☑	**1.286e2**
	104	104	400.2/13.1 (104)	400.1969	13.11	(Isotope)	☑	1.272e1

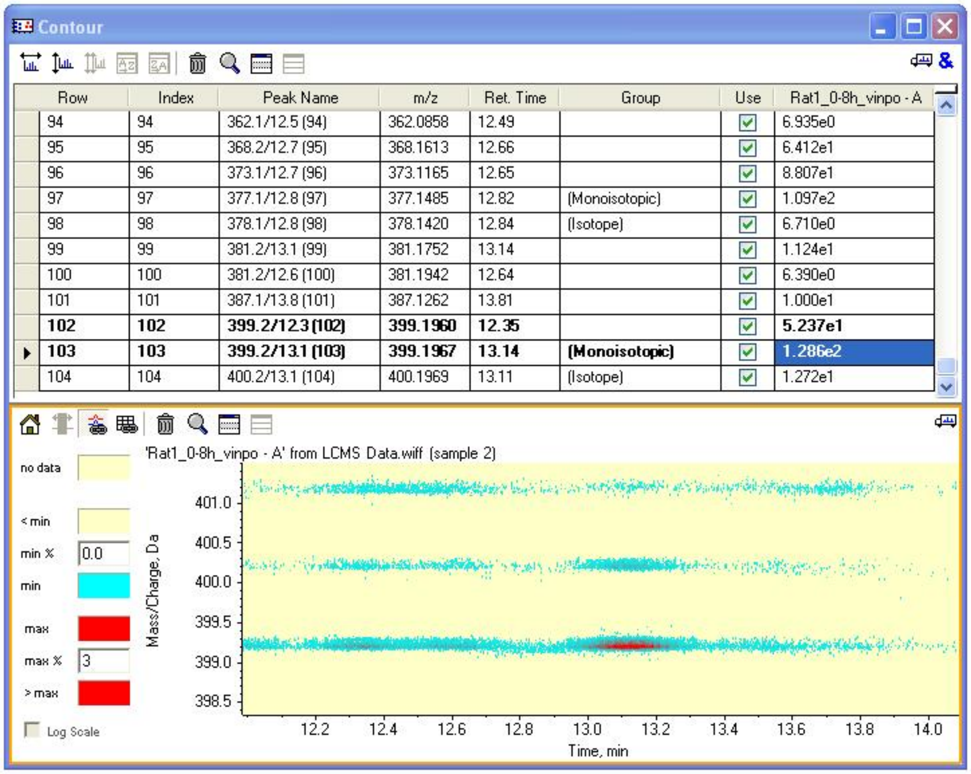
XIC from LCMS Data.wiff (sample 2) - Rat1_0-8h_vinpo - A, +TOF MS (80 - 1000): 399.22 +/- 0.07 Da

The system will generate the extracted ion chromatogram (XIC) for a small mass window around the selected m/z value; the region between the blue arrows in the x-axis indicates the range for the peak selected. In this case it is clear that not only are the peaks at 12.35 and 13.1 min. correct, but there is an additional peak at ca. 12.6 min. that was merged with the peak at 12.35 min.!

9. Click the **Link to Table** button at the top of the chromatogram pane and select another row in the variable table, e.g. one of the rows for the peaks with m/z 381.2, and the chromatogram will update to show the behavior of this variable.

   While it may be possible to find parameters that separate closely eluting peaks (in this case a retention time tolerance of 0.1 min. will allow the peak at 12.6 to be retained), this may not be wise when there are several samples since small retention time shifts between the runs may cause different peaks to be aligned. In complex samples it is definitely an advantage to introduce an internal standard to allow the retention times to be corrected so smaller tolerances can be used.

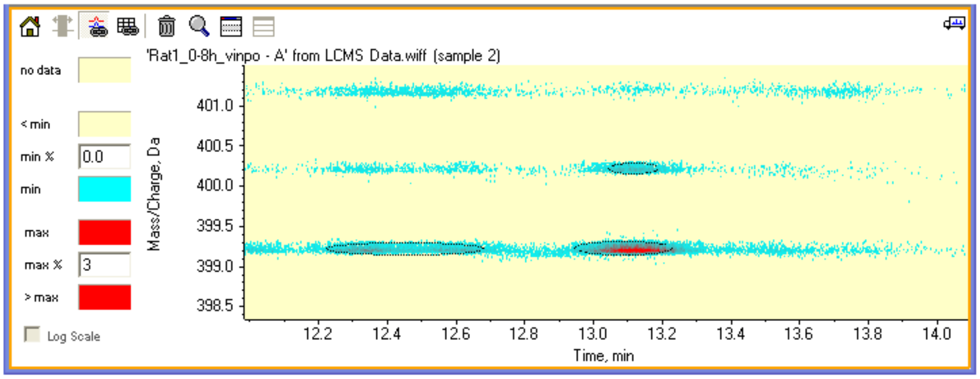
10. Click the **trash can** icon to delete the chromatogram window, right-click in any sample column cell and select **Show -> Contour.** Drag in the x axis to select the 12 to 14 min range (as imported) and in the y axis to select a region roughly 5 amu wide around m/z 400. If the color selection tools are not visible, right-click in the contour and select **Show Color Selection Tools;** set the **max% value** to 3 – this will change the way color and intensity are mapped so that the smaller peaks are more visible.



Contour

	Row	Index	Peak Name	m/z	Ret. Time	Group	Use	Rat1_0-8h_vinpo - A
	94	94	362.1/12.5 (94)	362.0858	12.49		☑	6.935e0
	95	95	368.2/12.7 (95)	368.1613	12.66		☑	6.412e1
	96	96	373.1/12.7 (96)	373.1165	12.65		☑	8.807e1
	97	97	377.1/12.8 (97)	377.1485	12.82	(Monoisotopic)	☑	1.097e2
	98	98	378.1/12.8 (98)	378.1420	12.84	(Isotope)	☑	6.710e0
	99	99	381.2/13.1 (99)	381.1752	13.14		☑	1.124e1
	100	100	381.2/12.6 (100)	381.1942	12.64		☑	6.390e0
	101	101	387.1/13.8 (101)	387.1262	13.81		☑	1.000e1
	102	**102**	**399.2/12.3 (102)**	**399.1960**	**12.35**		☑	**5.237e1**
▶	**103**	**103**	**399.2/13.1 (103)**	**399.1967**	**13.14**	**(Monoisotopic)**	☑	**1.286e2**
	104	104	400.2/13.1 (104)	400.1969	13.11	(Isotope)	☑	1.272e1

11. Right-click in the contour plot and select **Show Peak regions for All Peaks**. Ellipses will be drawn around the areas where peaks were located and the extent will indicate the time duration and m/z width of the peaks found. (The same command is available when more than one peak has been imported, but in this case the ellipses will indicate the combined extent of the peaks in all samples).



The display shows that for m/z 399.2 the peaks at 12.35 and 12.6 min. were found as a single peak (the ellipse covers both) and the peak at 13.1 min. was also found. For m/z 400.2 only the peak at 13.1 min. was found and there appear to be several other small peaks in the area that were not found.

12. Right-click in the contour and select **Show Tooltips**. As you move the cursor in the contour plot a tool tip will appear indicating the m/z, retention time and intensity (z) of the point under the cursor. In this case, the intensity of the peaks in this area do not exceed 5, the value that we initially used as a threshold when importing the data.

In addition to the intensity, there are a number of other reasons why small peaks may be rejected by the peak finder, for example:

- The mass peak does not appear in enough contiguous scans (less than the **Minimum RT peak width** defined when importing the data).
- The m/z width of the peak is less than the **Minimum spectral peak width**.
- If **Subtraction offset** was checked, for any given peak there may be another peak ahead of it by the offset value used; when subtracted this may cause the target peak to be less than the specified intensity thresholds.

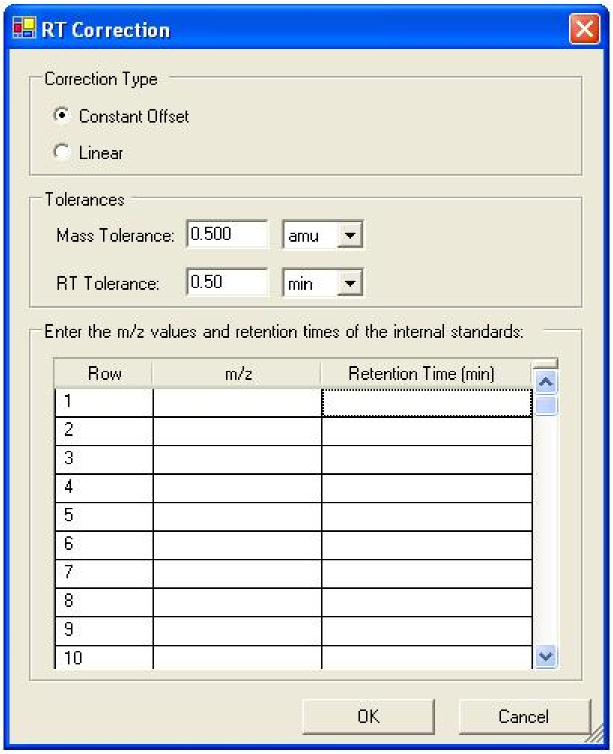
The operation of the peak finder is described in detail in the reference manual and you are encouraged to experiment with the parameters and observe the results using the tools and approach described here.

6.3 Aligning, normalizing and filtering data

6.3.1 Aligning and normalizing

As indicated in the previous section, aligning peaks is essential for best performance.

If you have added one or more internal standards to the samples, you may specify these in the **RT Correction** and **Normalization** sub-dialogs of the **Alignment & Filtering** dialog box. The tolerances in these dialogs refer to the windows used to locate the internal standards and are typically wider than the values used to actually align the data.



RT Correction

Correction Type

● Constant Offset

○ Linear

Tolerances

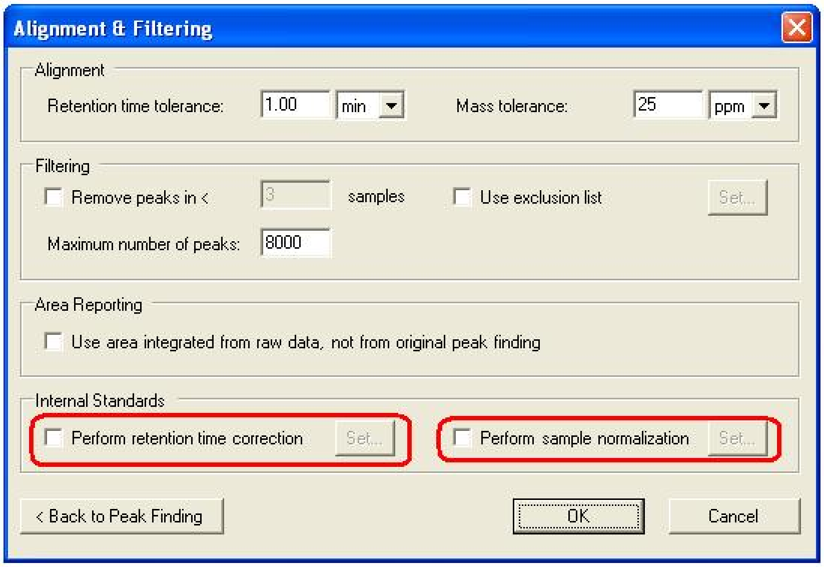
Mass Tolerance: 0.500 amu

RT Tolerance: 0.50 min

Enter the m/z values and retention times of the internal standards:

Row	m/z	Retention Time (min)
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

When importing data you select **Perform retention time correction** and/or **Perform sample normalization** and the data will be aligned and normalized as it is being read.


Alignment & Filtering

Alignment

Retention time tolerance: 1.00 min

Mass tolerance: 25 ppm

Filtering

□ Remove peaks in < 3 samples

□ Use exclusion list

Maximum number of peaks: 8000

Area Reporting

□ Use area integrated from raw data, not from original peak finding

Internal Standards

□ Perform retention time correction

□ Perform sample normalization

The alignment process is described in more detail in the reference manual, but if using more than one retention time standard it is best to have them well separated and use **Linear offset**. With this mode the program will calculate the offset as a function of retention time; standards that are close in time can cause the slope of this function to be incorrect.

While alignment can only be performed as the data is imported, normalization (with or without internal standards) can be performed on an existing data table. If you have used internal standards, you can normalize the data by selecting **Normalize LC/MS Using Internal Standards** from the **Normalization** sub-menu of the **Analyze** menu.

If you have not used internal standards you can still normalize the data, but this should be done carefully since there is no real way to ensure that the selected peak(s) should indeed be constant for all samples. The following describes the process for the vinpocetin data used in earlier sections.

1. Open the LCMS data file you saved in section 5.2, step 8.

   Here are some tips for picking peaks to use to normalize in this way

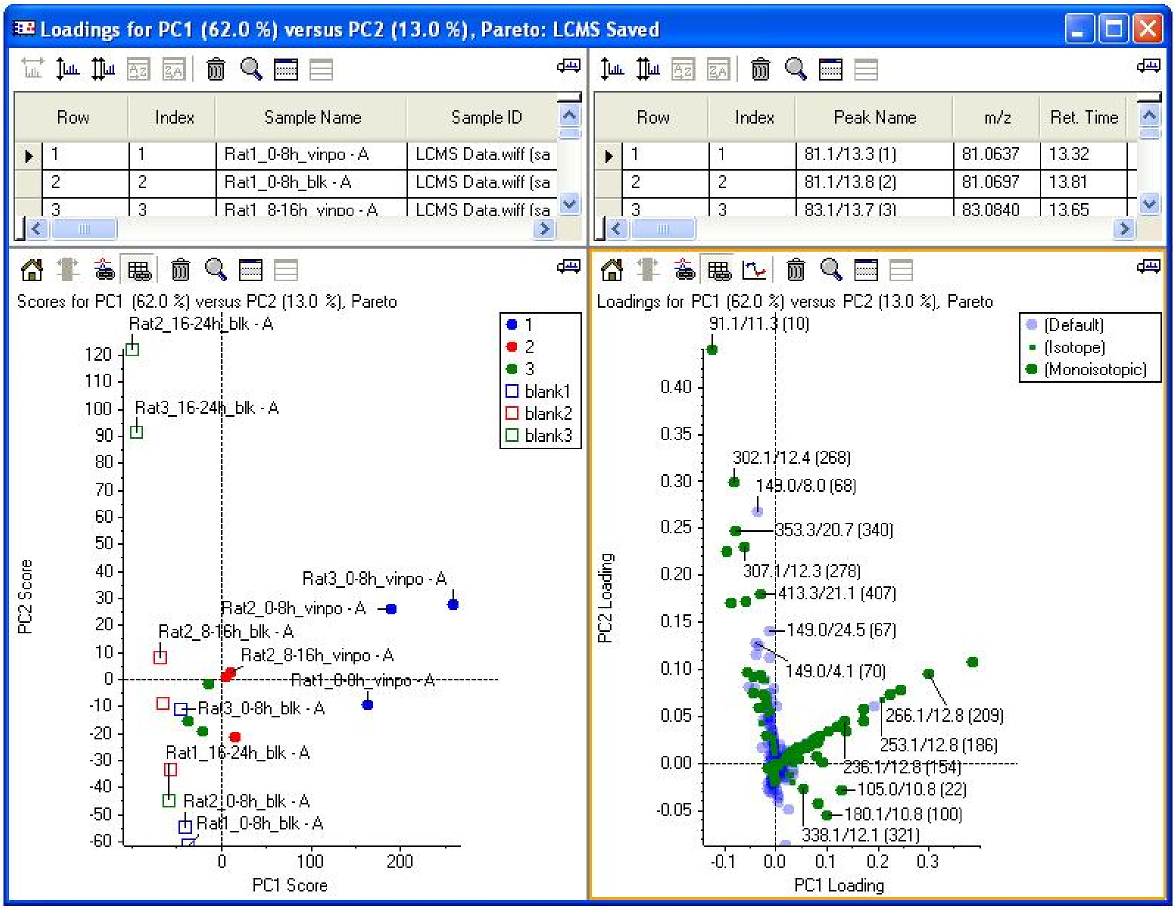
   - The peak should appear in every sample and preferably be a single peak (i.e. have no close isomers that may be picked incorrectly)
   - The intensity should not be very small (noise) or very large (possibly saturated)
   - There should be no, or little, dependence on the group
   - There should be no systematic variation (click the **Sort Order** button and select **Sample Index** to look for this)

   Examination of the data shows that the peak at m/z 384.1 and 10.5 min. appears in all samples and although it may have some group dependence, this is not large and we will assume it is suitable.

2. Select the row containing this peak and plot its profile using the **Plot row** button. Verify that there is no systematic variation and the group dependence is relatively low.

3. Make sure the data table is the active pane and the 384.1/10.5 row is selected and select **Normalize Using Selected Peaks** from the **Normalization** submenu of the **Analyze** menu. A new data table will be generated containing the now normalized values.

4. Select **Show Samples Table** from the **View** menu and note that the **Scale Factor** column now contains a value for each sample. Ideally these values will all be close to one, indicating that the peaks used for the normalization were of comparable intensity in all samples. If any of the values seems abnormally large you should check that the reference peak is present in that sample and has been selected correctly. You may need to adjust the tolerances in the **Normalization** dialog.
5. Perform a PCA analysis.



Loadings for PC1 (62.0 %) versus PC2 (13.0 %), Pareto: LCMS Saved

Row	Index	Sample Name	Sample ID
1	1	Rat1_0-8h_vinpo - A	LCMS Data.wiff (sa
2	2	Rat1_0-8h_blk - A	LCMS Data.wiff (sa
3	3	Rat1_8-16h_vinpo - A	LCMS Data.wiff (sa

Row	Index	Peak Name	m/z	Ret. Time
1	1	81.1/13.3 (1)	81.0637	13.32
2	2	81.1/13.8 (2)	81.0697	13.81
3	3	83.1/13.7 (3)	83.0840	13.65

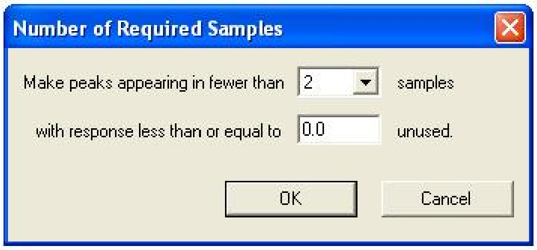
Explore the data using the techniques and tools described in section 5 and confirm that while the scores and loadings plots look different, and the amount of variance explained by the principal components is also different, the conclusions drawn earlier still apply.

6.3.2 Filtering data

In many case the data will contain variables that are suspect (e.g. too small), artifacts (e.g. arising from contamination) or not wanted in the analysis. This section briefly describes some methods of identifying and removing such peaks.

1. Close all open windows, re-open the saved LCMS data and perform a PCA analysis.
2. One useful way to filter data is to exclude variables that do not appear in a certain minimum number of samples. This is particularly relevant in data such as this since there are three samples for each time point and dose, so variables appearing in just one are likely noise, individual variation or misaligned.

   Select **Make Peaks Appearing in Few Samples Unused** from the **Analyze** menu, and select 2 in the combo box in the resulting dialog.



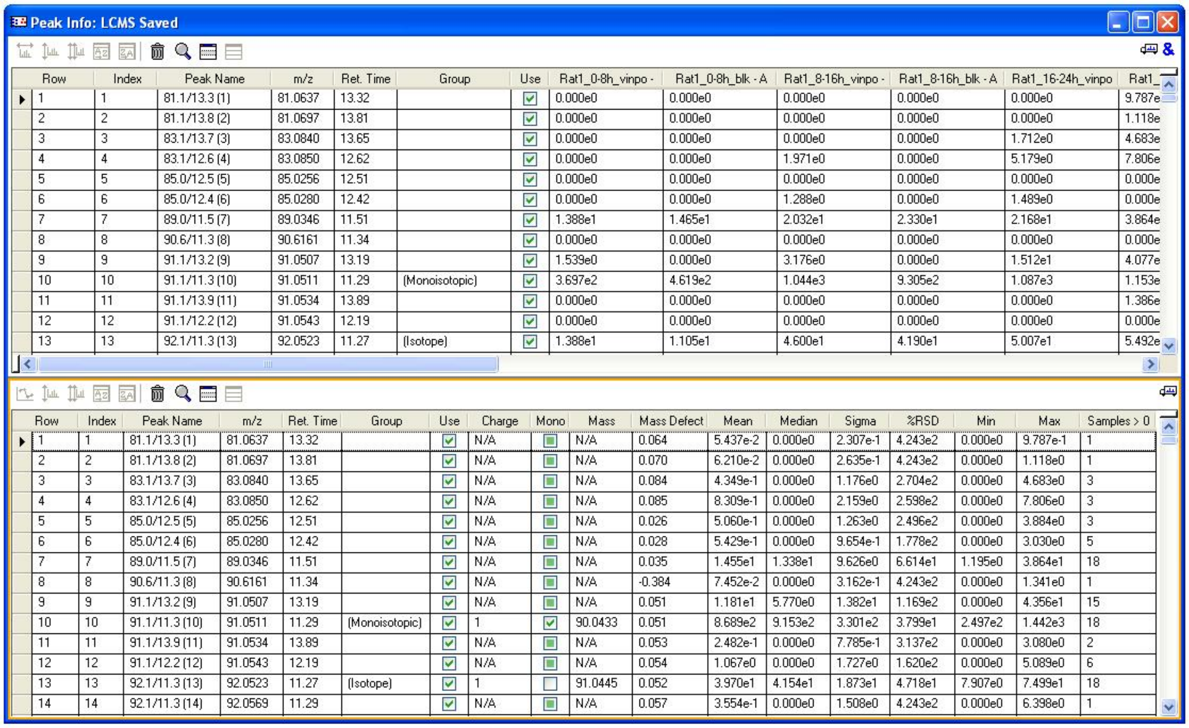
Number of Required Samples

Make peaks appearing in fewer than 2 samples

with response less than or equal to 0.0 unused.

In the loadings plot a number of variables close to the origin will now be drawn as open circles to show they have been excluded (you may need to zoom to see this). Since these are small peaks a new PCA analysis will show little change.

3. Close all windows except the peaks table and select **Show Peak Info** from the **View** menu; the peak info table appears in the lower part of the window:



Peak Info: LCMS Saved

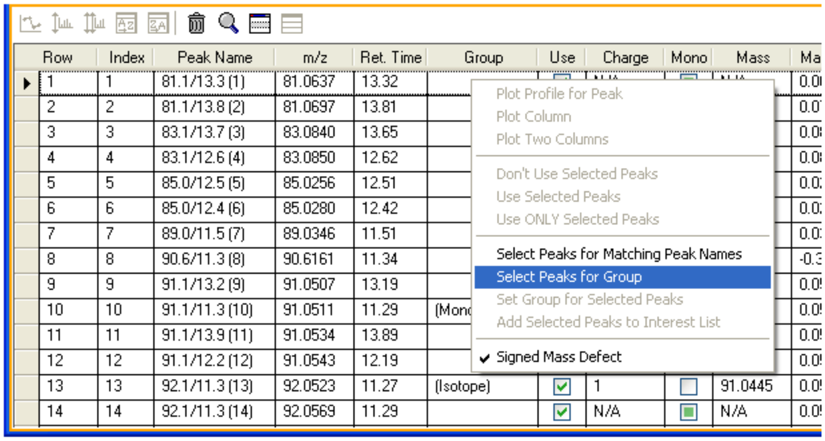
Row	Index	Peak Name	m/z	Ret. Time	Group	Use	Rat1_0-8h_vinpo -	Rat1_0-8h_blk - A	Rat1_8-16h_vinpo -	Rat1_8-16h_blk - A	Rat1_16-24h_vinpo	Rat1_
1	1	81.1/13.3 (1)	81.0637	13.32		☑	0.000e0	0.000e0	0.000e0	0.000e0	0.000e0	9.787e
2	2	81.1/13.8 (2)	81.0697	13.81		☑	0.000e0	0.000e0	0.000e0	0.000e0	0.000e0	1.118e
3	3	83.1/13.7 (3)	83.0840	13.65		☑	0.000e0	0.000e0	0.000e0	0.000e0	1.712e0	4.683e
4	4	83.1/12.6 (4)	83.0850	12.62		☑	0.000e0	0.000e0	1.971e0	0.000e0	5.179e0	7.806e
5	5	85.0/12.5 (5)	85.0256	12.51		☑	0.000e0	0.000e0	0.000e0	0.000e0	0.000e0	0.000e
6	6	85.0/12.4 (6)	85.0280	12.42		☑	0.000e0	0.000e0	1.288e0	0.000e0	1.489e0	0.000e
7	7	89.0/11.5 (7)	89.0346	11.51		☑	1.388e1	1.465e1	2.032e1	2.330e1	2.168e1	3.864e
8	8	90.6/11.3 (8)	90.6161	11.34		☑	0.000e0	0.000e0	0.000e0	0.000e0	0.000e0	0.000e
9	9	91.1/13.2 (9)	91.0507	13.19		☑	1.539e0	0.000e0	3.176e0	0.000e0	1.512e1	4.077e
10	10	91.1/11.3 (10)	91.0511	11.29	(Monoisotopic)	☑	3.697e2	4.619e2	1.044e3	9.305e2	1.087e3	1.153e
11	11	91.1/13.9 (11)	91.0534	13.89		☑	0.000e0	0.000e0	0.000e0	0.000e0	0.000e0	1.386e
12	12	91.1/12.2 (12)	91.0543	12.19		☑	0.000e0	0.000e0	0.000e0	0.000e0	0.000e0	0.000e
13	13	92.1/11.3 (13)	92.0523	11.27	(Isotope)	☑	1.388e1	1.105e1	4.600e1	4.190e1	5.007e1	5.492e

Row	Index	Peak Name	m/z	Ret. Time	Group	Use	Charge	Mono	Mass	Mass Defect	Mean	Median	Sigma	%RSD	Min	Max	Samples > 0
1	1	81.1/13.3 (1)	81.0637	13.32		☑	N/A	▣	N/A	0.064	5.437e-2	0.000e0	2.307e-1	4.243e2	0.000e0	9.787e-1	1
2	2	81.1/13.8 (2)	81.0697	13.81		☑	N/A	▣	N/A	0.070	6.210e-2	0.000e0	2.635e-1	4.243e2	0.000e0	1.118e0	1
3	3	83.1/13.7 (3)	83.0840	13.65		☑	N/A	▣	N/A	0.084	4.349e-1	0.000e0	1.176e0	2.704e2	0.000e0	4.683e0	3
4	4	83.1/12.6 (4)	83.0850	12.62		☑	N/A	▣	N/A	0.085	8.309e-1	0.000e0	2.159e0	2.598e2	0.000e0	7.806e0	3
5	5	85.0/12.5 (5)	85.0256	12.51		☑	N/A	▣	N/A	0.026	5.060e-1	0.000e0	1.263e0	2.496e2	0.000e0	3.884e0	3
6	6	85.0/12.4 (6)	85.0280	12.42		☑	N/A	▣	N/A	0.028	5.429e-1	0.000e0	9.654e-1	1.778e2	0.000e0	3.030e0	5
7	7	89.0/11.5 (7)	89.0346	11.51		☑	N/A	▣	N/A	0.035	1.455e1	1.338e1	9.626e0	6.614e1	1.195e0	3.864e1	18
8	8	90.6/11.3 (8)	90.6161	11.34		☑	N/A	▣	N/A	-0.384	7.452e-2	0.000e0	3.162e-1	4.243e2	0.000e0	1.341e0	1
9	9	91.1/13.2 (9)	91.0507	13.19		☑	N/A	▣	N/A	0.051	1.181e1	5.770e0	1.382e1	1.169e2	0.000e0	4.356e1	15
10	10	91.1/11.3 (10)	91.0511	11.29	(Monoisotopic)	☑	1	☑	90.0433	0.051	8.689e2	9.153e2	3.301e2	3.799e1	2.497e2	1.442e3	18
11	11	91.1/13.9 (11)	91.0534	13.89		☑	N/A	▣	N/A	0.053	2.482e-1	0.000e0	7.785e-1	3.137e2	0.000e0	3.080e0	2
12	12	91.1/12.2 (12)	91.0543	12.19		☑	N/A	▣	N/A	0.054	1.067e0	0.000e0	1.727e0	1.620e2	0.000e0	5.089e0	6
13	13	92.1/11.3 (13)	92.0523	11.27	(Isotope)	☑	1	☐	91.0445	0.052	3.970e1	4.154e1	1.873e1	4.718e1	7.907e0	7.499e1	18
14	14	92.1/11.3 (14)	92.0569	11.29		☑	N/A	▣	N/A	0.057	3.554e-1	0.000e0	1.508e0	4.243e2	0.000e0	6.398e0	1

This table contains detailed metrics for each of the variables in the data (these are explained in the Reference Manual) and can be used to filter the variables.

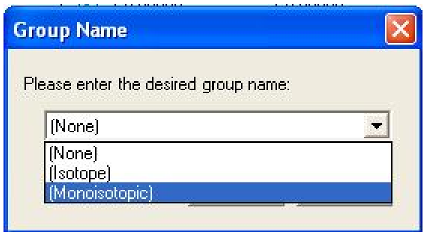
4. Another way to exclude peaks appearing in only a few samples is as follows:
   - select the **Samples >** column
   - sort in ascending order
   - select the unwanted variables
   - right click and select **Don’t Use Selected Peaks**

5. You can also select and use any variable groups(s). For example, to use only the monoisotopic peaks (this can also be performed directly from the peaks table) right click in the table and chose **Select Peaks For Group**.



Row	Index	Peak Name	m/z	Ret. Time	Group	Use	Charge	Mono	Mass	Ma
1	1	81.1/13.3 (1)	81.0637	13.32						0.0
2	2	81.1/13.8 (2)	81.0697	13.81						0.0
3	3	83.1/13.7 (3)	83.0840	13.65						0.0
4	4	83.1/12.6 (4)	83.0850	12.62						0.0
5	5	85.0/12.5 (5)	85.0256	12.51						0.0
6	6	85.0/12.4 (6)	85.0280	12.42						0.0
7	7	89.0/11.5 (7)	89.0346	11.51						0.0
8	8	90.6/11.3 (8)	90.6161	11.34						-0.3
9	9	91.1/13.2 (9)	91.0507	13.19						0.0
10	10	91.1/11.3 (10)	91.0511	11.29	(Mon					0.0
11	11	91.1/13.9 (11)	91.0534	13.89						0.0
12	12	91.1/12.2 (12)	91.0543	12.19						0.0
13	13	92.1/11.3 (13)	92.0523	11.27	(Isotope)	☑	1	☐	91.0445	0.0
14	14	92.1/11.3 (14)	92.0569	11.29		☑	N/A	■	N/A	0.0

6. The resulting dialog shows all the assigned variable groups; select **(Monoisotopic)** and click **OK**.

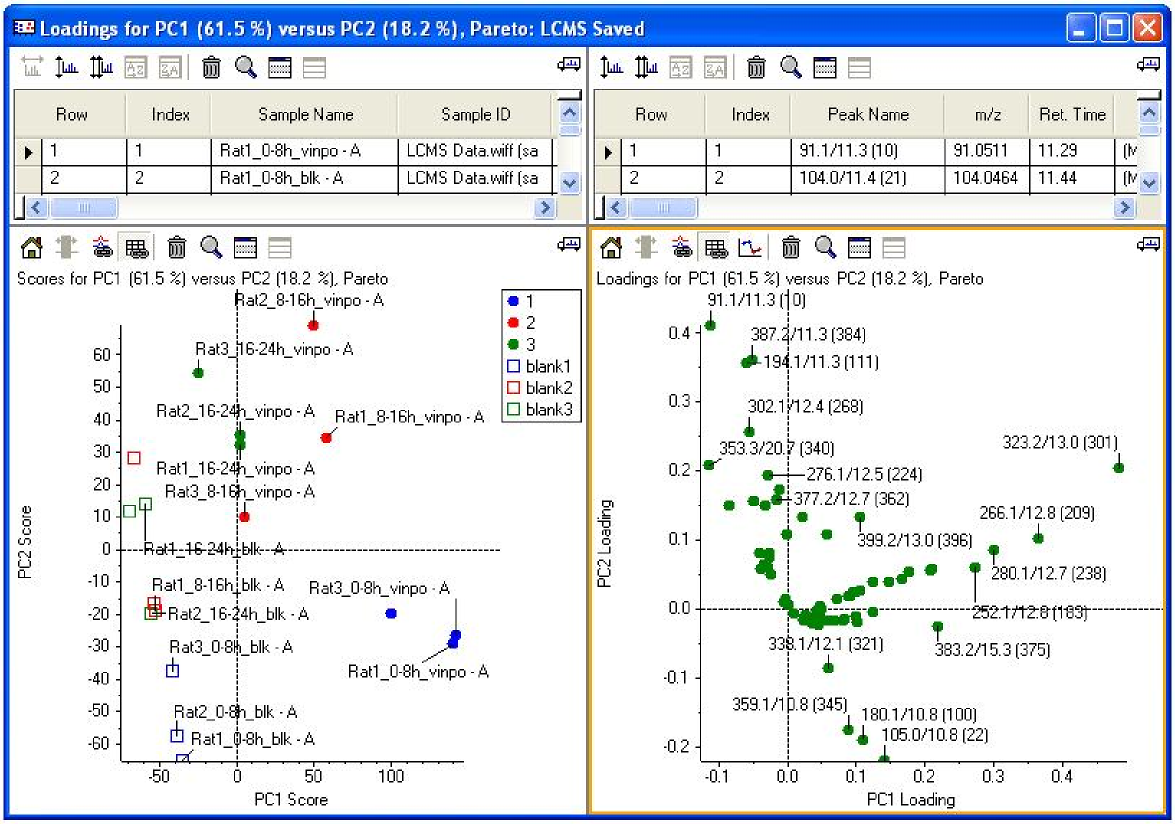


Group Name

Please enter the desired group name:

(None)

7. This automatically selects all the peaks assigned to the Monoisotopic group; right-click in the table, select **Use ONLY Selected Peaks** and perform a PCA analysis. The resulting display is similar to those obtained earlier but with a much simplified loadings plot since there are far fewer variables:



Loadings for PC1 (61.5 %) versus PC2 (18.2 %), Pareto: LCMS Saved

Row	Index	Sample Name	Sample ID
1	1	Rat1_0-8h_vinpo - A	LCMS Data.wiff (sa
2	2	Rat1_0-8h_blk - A	LCMS Data.wiff (sa

Row	Index	Peak Name	m/z	Ret. Time	
1	1	91.1/11.3 (10)	91.0511	11.29	(M
2	2	104.0/11.4 (21)	104.0464	11.44	(M

8. The **Peak Info** table also contains a column for mass defect – the difference between the measured m/z and the nearest integer. Simple metabolic changes made to xenobiotics tend to shift the m/z value without substantially altering the defect, so looking for compounds with similar mass defects to the parent drug can help identify metabolites. The table can be used to filter compounds based on their mass defects.

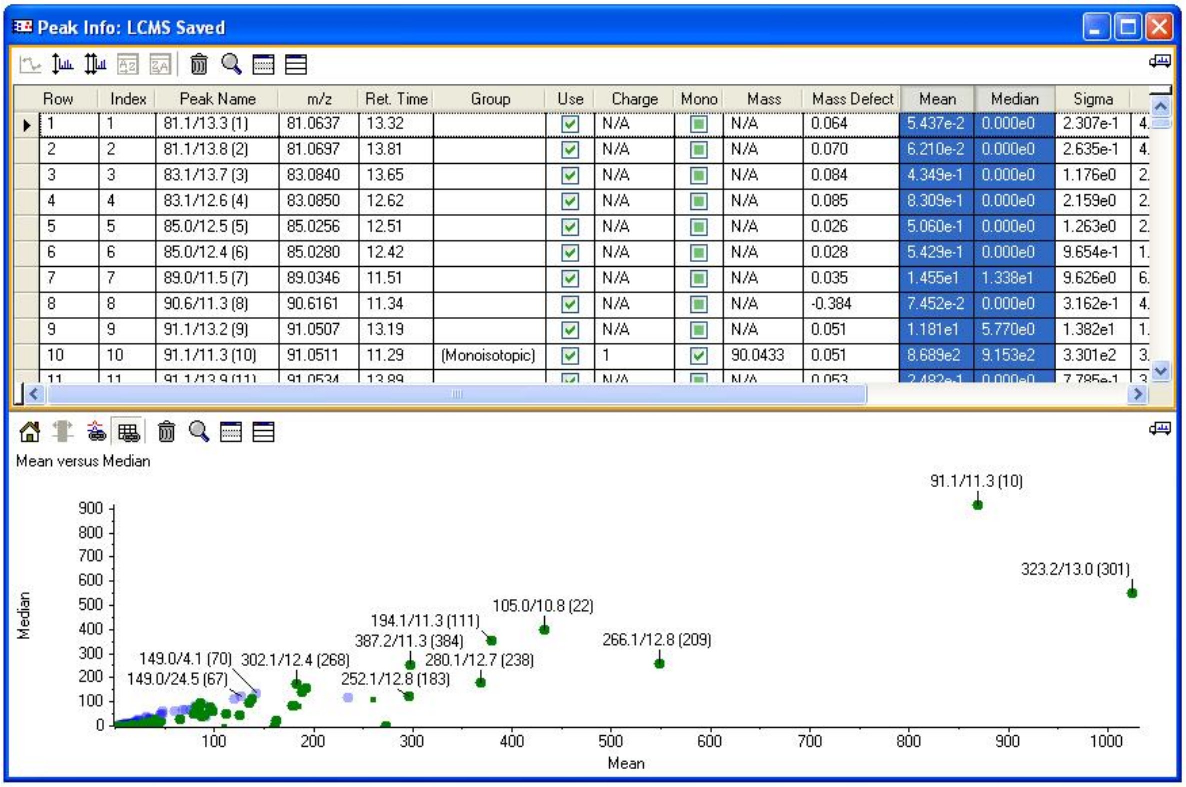
   Mass defect can be expressed in two ways:

   - Relative to the nearest integer. In this case some values will have negative values relative to a higher integer, e.g. 300.8 would have a defect of -0.2.
   - Relative to the lower integer. In this case the defects are always positive, i.e. 300.8 has a defect of 0.8

   To change between the two modes, right click in the **Peak Info** table and click **Signed Mass Defect**.

9. The Peak Info table also allows columns to be plotted, either individually or one may be plotted against another. This can help visualize characteristics of the data or select particular variables, for example, in choosing variables to use for normalization it might be appropriate to select variables with relatively high values (mean or median) that are relatively constant (low standard deviation); plotting sigma against mean can help select such variables.

10. In the peak Info table select the Mean column and drag to include the Median column; click the two-way plot icon to get the following display (the sample table has been hidden for clarity):



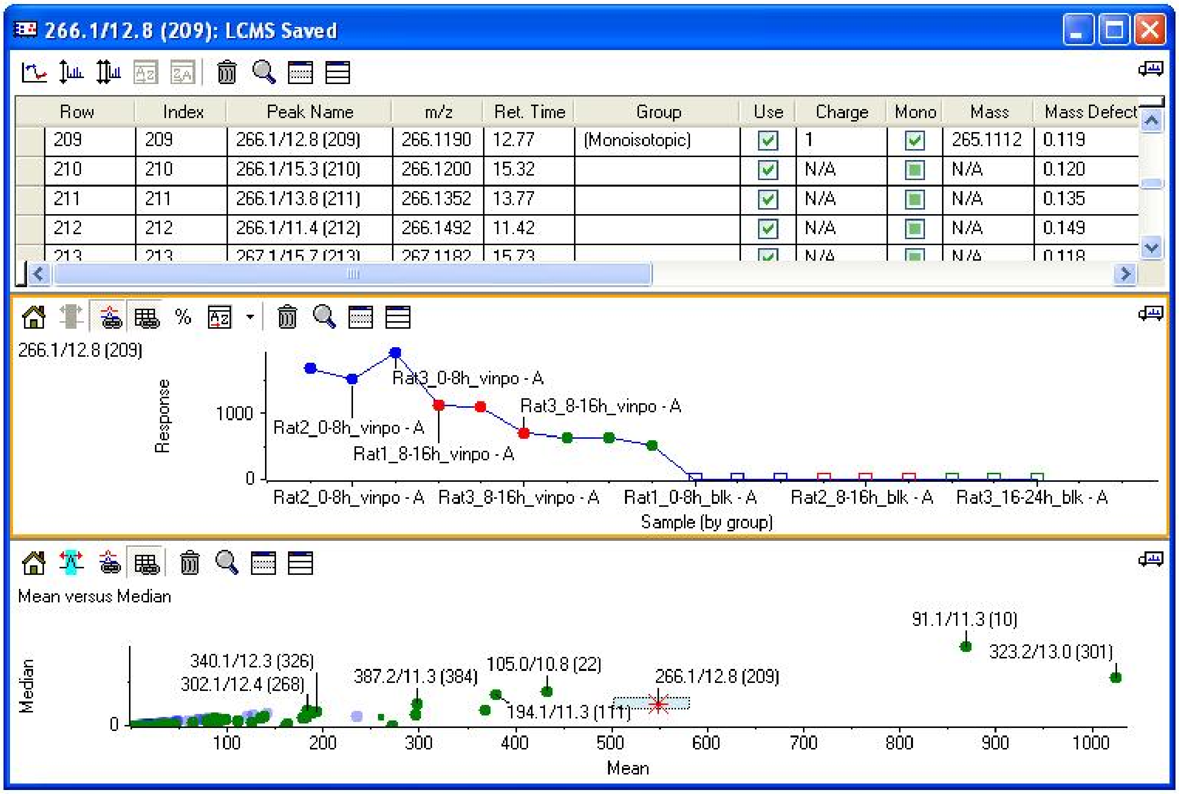
Peak Info: LCMS Saved

Row	Index	Peak Name	m/z	Ret. Time	Group	Use	Charge	Mono	Mass	Mass Defect	Mean	Median	Sigma	
1	1	81.1/13.3 (1)	81.0637	13.32		☑	N/A	▣	N/A	0.064	5.437e-2	0.000e0	2.307e-1	4.
2	2	81.1/13.8 (2)	81.0697	13.81		☑	N/A	▣	N/A	0.070	6.210e-2	0.000e0	2.635e-1	4.
3	3	83.1/13.7 (3)	83.0840	13.65		☑	N/A	▣	N/A	0.084	4.349e-1	0.000e0	1.176e0	2.
4	4	83.1/12.6 (4)	83.0850	12.62		☑	N/A	▣	N/A	0.085	8.309e-1	0.000e0	2.159e0	2.
5	5	85.0/12.5 (5)	85.0256	12.51		☑	N/A	▣	N/A	0.026	5.060e-1	0.000e0	1.263e0	2.
6	6	85.0/12.4 (6)	85.0280	12.42		☑	N/A	▣	N/A	0.028	5.429e-1	0.000e0	9.654e-1	1.
7	7	89.0/11.5 (7)	89.0346	11.51		☑	N/A	▣	N/A	0.035	1.455e1	1.338e1	9.626e0	6.
8	8	90.6/11.3 (8)	90.6161	11.34		☑	N/A	▣	N/A	-0.384	7.452e-2	0.000e0	3.162e-1	4.
9	9	91.1/13.2 (9)	91.0507	13.19		☑	N/A	▣	N/A	0.051	1.181e1	5.770e0	1.382e1	1.
10	10	91.1/11.3 (10)	91.0511	11.29	(Monoisotopic)	☑	1	☑	90.0433	0.051	8.689e2	9.153e2	3.301e2	3.
11	11	91.1/13.9 (11)	91.0534	13.89		☑	N/A	▣	N/A	0.053	2.482e-1	0.000e0	7.785e-1	3

Mean versus Median

If the variables are normally distributed we expect the mean and the median to be identical, i.e. the plot should be a straight line with a slope of one. While there are many variables that meet this condition, there are also several that have a lower median than expected (zero in some cases). This arises when the data is not normally distributed, for example there may be two groups with the variable absent (zero) in one group; in this case, depending on the number of samples in each group, the median may be zero while the mean is not. In any case these are likely to be interesting variables.

11. Make a selection rectangle around the variable 266.1/12.8, right-click and select **Show Selected Points In Table**, and then click the **Plot Profile** tool in the table's toolbar. In the profile plot click the **Sort Order** tool to get the following display



266.1/12.8 (209): LCMS Saved

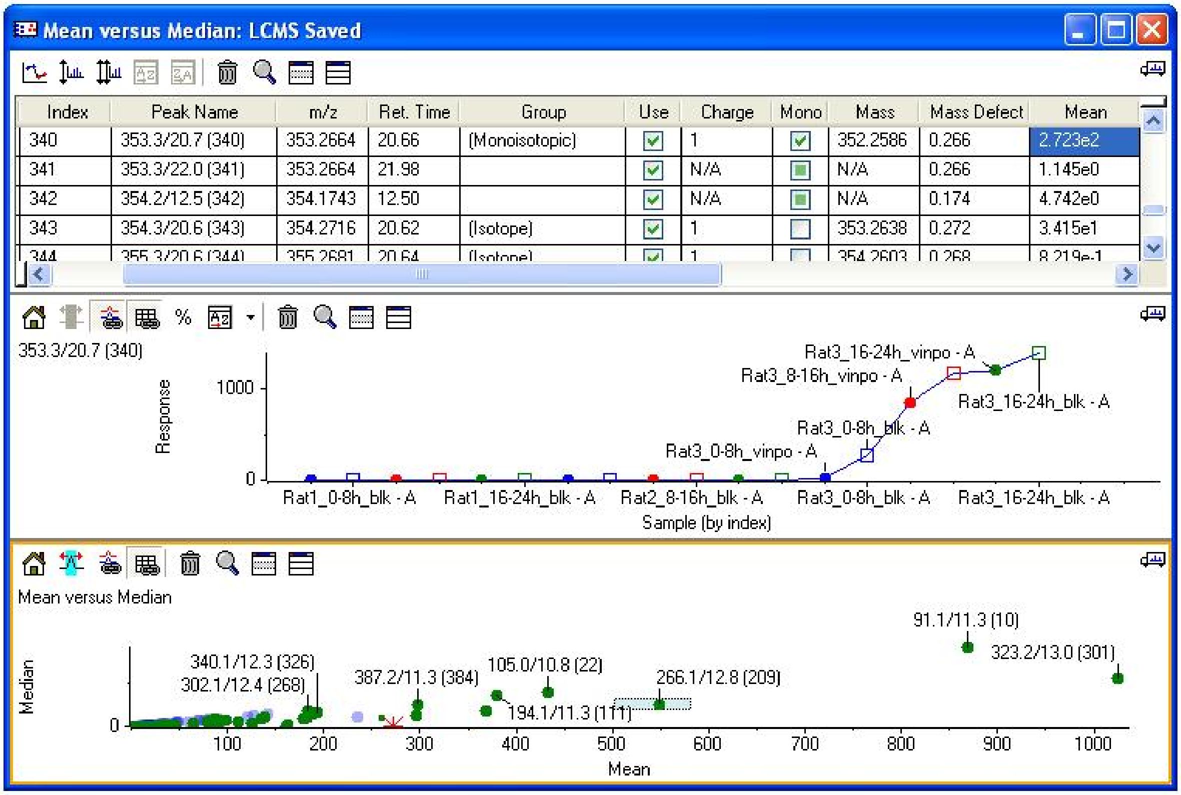
Row	Index	Peak Name	m/z	Ret. Time	Group	Use	Charge	Mono	Mass	Mass Defect
209	209	266.1/12.8 (209)	266.1190	12.77	(Monoisotopic)	☑	1	☑	265.1112	0.119
210	210	266.1/15.3 (210)	266.1200	15.32		☑	N/A	■	N/A	0.120
211	211	266.1/13.8 (211)	266.1352	13.77		☑	N/A	■	N/A	0.135
212	212	266.1/11.4 (212)	266.1492	11.42		☑	N/A	■	N/A	0.149
213	213	267.1/15.7 (213)	267.1182	15.73		☑	N/A	■	N/A	0.118

This is clearly a drug metabolite and the number of samples in which the variable is zero is the same as the number where it is non-zero and the overall number of samples is even. Hence the median will be the average of zero and the smallest non-zero value while the mean will be the average of all samples; in this case the latter is higher. If the variable was zero in more samples (for example because it is metabolized quickly) then the median would be zero.

12. Select one of the points with a large mean but close to the median = 0 axis. Since the table and both plots are linked, the variable will be selected in the table and the profile plot generated. Click the **Sort Order** button to get the display shown below.

    In this case the variable corresponds to some contamination that appears late in the run and is therefore most obvious in **Sample Index** order. Because the number of zero values is greater than the number of non-zero values, the median is zero while the mean is still positive.

    Zooming the display to better view the points that are close to the median = 0 axis and clicking on variables, quickly reveals variables that belong in the above classes. If the sort order is left as **Sample Index** the contamination peaks are very obvious and can be quickly excluded.



Mean versus Median: LCMS Saved

Index	Peak Name	m/z	Ret. Time	Group	Use	Charge	Mono	Mass	Mass Defect	Mean
340	353.3/20.7 (340)	353.2664	20.66	(Monoisotopic)	☑	1	☑	352.2586	0.266	2.723e2
341	353.3/22.0 (341)	353.2664	21.98		☑	N/A	■	N/A	0.266	1.145e0
342	354.2/12.5 (342)	354.1743	12.50		☑	N/A	■	N/A	0.174	4.742e0
343	354.3/20.6 (343)	354.2716	20.62	(Isotope)	☑	1	□	353.2638	0.272	3.415e1
344	355.3/20.6 (344)	355.2681	20.64	(Isotope)	☑	1	□	354.2603	0.268	8.219e-1

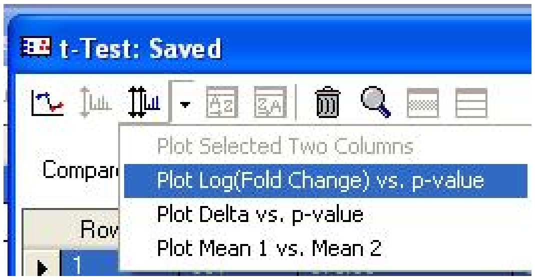
6.4 Selecting discriminating t-test variables

The metric columns and plotting capabilities associated with the t-test table provide a number of ways to asses the quality of variables and to select those that best differentiate groups.

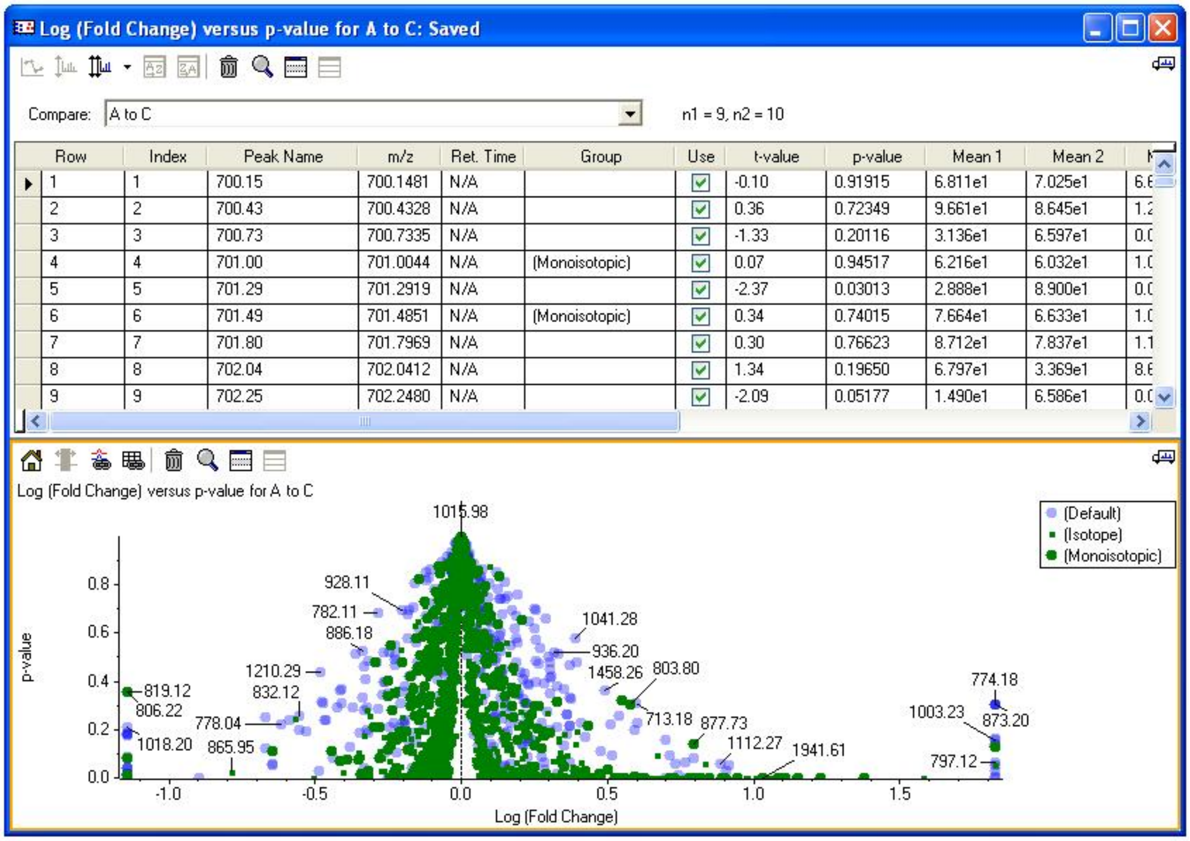
1. Open the MALDI data that was saved in section 3.4, make sure that the anomalous sample A9_MS_1.t2d is not used and perform a t-test.

   The plot two columns button (⇕ ▾) allows you to select any two columns and plot one against the other, but it also contains a combo-box that is accessed via the small downward pointing arrow and provides quick access to some pre-defined plots.

2. Click the small arrow and select 'Plot Log(Fold Change) vs. p-value' from the context menu:



   This generates a plot that is similar to the one shown below



Log (Fold Change) versus p-value for A to C: Saved

Compare: A to C

n1 = 9, n2 = 10

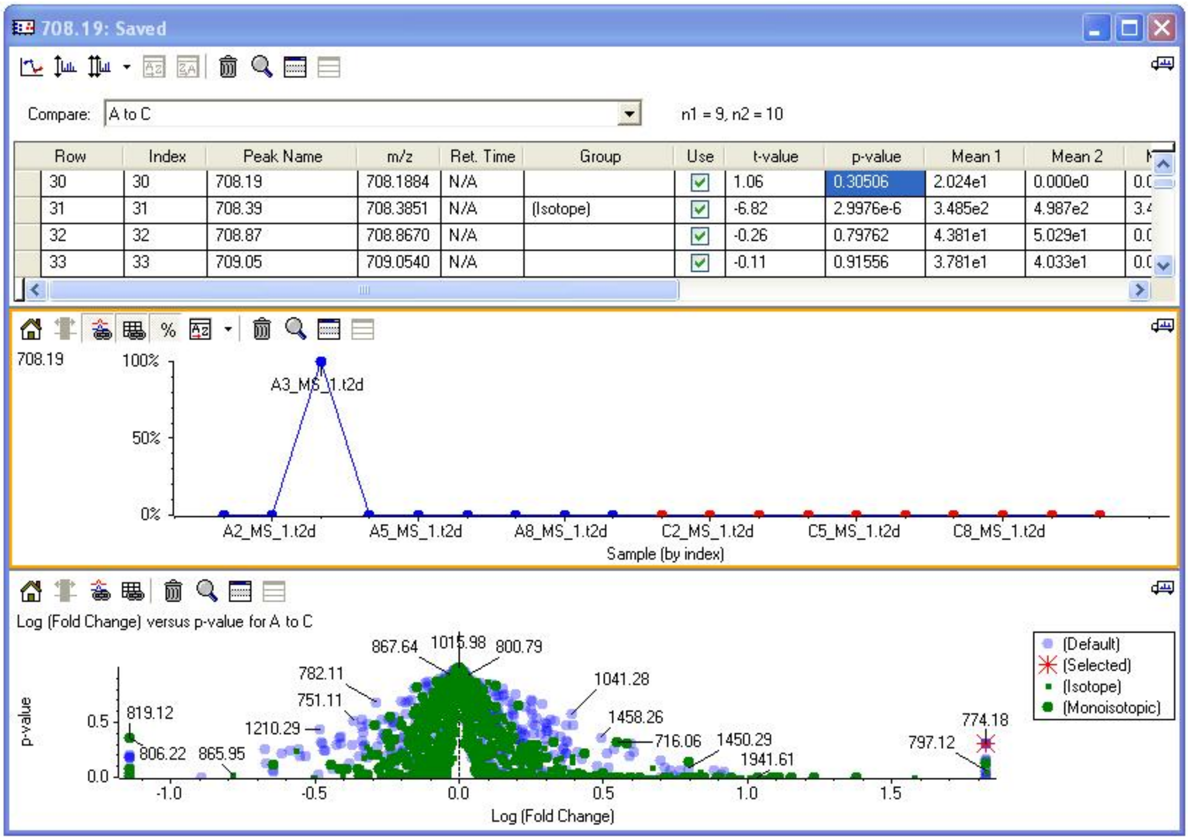
Row	Index	Peak Name	m/z	Ret. Time	Group	Use	t-value	p-value	Mean 1	Mean 2	M
1	1	700.15	700.1481	N/A		☑	-0.10	0.91915	6.811e1	7.025e1	6.6
2	2	700.43	700.4328	N/A		☑	0.36	0.72349	9.661e1	8.645e1	1.2
3	3	700.73	700.7335	N/A		☑	-1.33	0.20116	3.136e1	6.597e1	0.0
4	4	701.00	701.0044	N/A	(Monoisotopic)	☑	0.07	0.94517	6.216e1	6.032e1	1.0
5	5	701.29	701.2919	N/A		☑	-2.37	0.03013	2.888e1	8.900e1	0.0
6	6	701.49	701.4851	N/A	(Monoisotopic)	☑	0.34	0.74015	7.664e1	6.633e1	1.0
7	7	701.80	701.7969	N/A		☑	0.30	0.76623	8.712e1	7.837e1	1.1
8	8	702.04	702.0412	N/A		☑	1.34	0.19650	6.797e1	3.369e1	8.6
9	9	702.25	702.2480	N/A		☑	-2.09	0.05177	1.490e1	6.586e1	0.0

Log (Fold Change) versus p-value for A to C

Here the x-axis is the log of the fold change (the ratio of the means of the two groups) and the y-axis is the p-value. Variables that appear in one group but not in the other, i.e. that have an infinite fold change, are drawn slightly beyond the real values (819.12 on the left and 774.18 on the right for example).

Since small p-values indicate variables that distinguish the groups well, the most significant are those that have low p-values but high fold changes – those that have high p-values and low fold changes are not useful. If the variables are colored according to their isotopic status, you can select the monoisotopic peaks, or ignore those that are unassigned.

3. Click on a variable with a large positive change to select it in the t-test results table and, in that table, click the **Plot Profile** button to get the following:



708.19: Saved

Compare: A to C   n1 = 9, n2 = 10

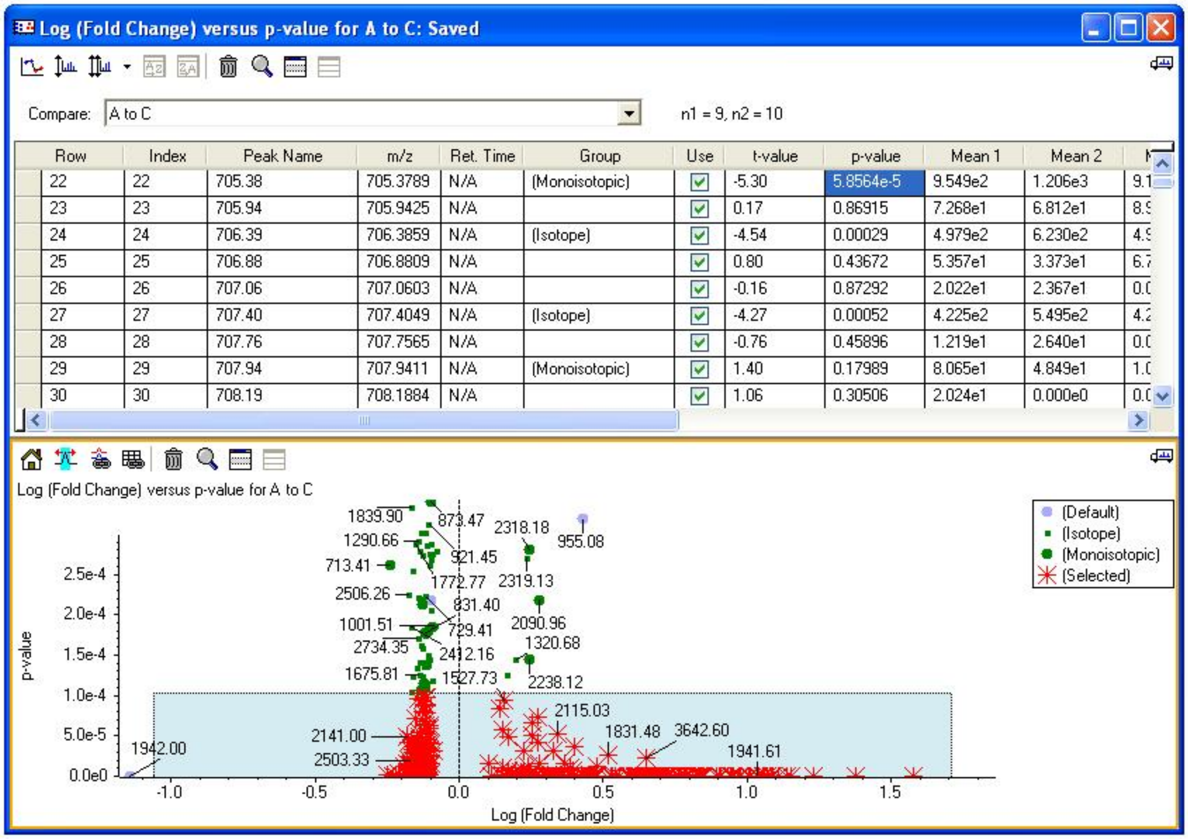
Row	Index	Peak Name	m/z	Ret. Time	Group	Use	t-value	p-value	Mean 1	Mean 2	M
30	30	708.19	708.1884	N/A		☑	1.06	0.30506	2.024e1	0.000e0	0.0
31	31	708.39	708.3851	N/A	(Isotope)	☑	-6.82	2.9976e-6	3.485e2	4.987e2	3.4
32	32	708.87	708.8670	N/A		☑	-0.26	0.79762	4.381e1	5.029e1	0.0
33	33	709.05	709.0540	N/A		☑	-0.11	0.91556	3.781e1	4.033e1	0.0

In this particular case (708.19 Da) the variable clearly represents only noise. Since the peak was not detected for the 'C' samples the fold change was reported as infinite.

The plots and the table are linked so the profile plot will update as you select different variables in the lower display.

By zooming the vertical axis of the p-value vs. log(fold change) display you can quickly select the variables that provide the greatest discrimination between the two groups.

4. Delete the profile plot and zoom the p-value axis. Select all points with a p-value less than 1e-4 as shown below, right-click in the graph and select **Show Points In Table**.

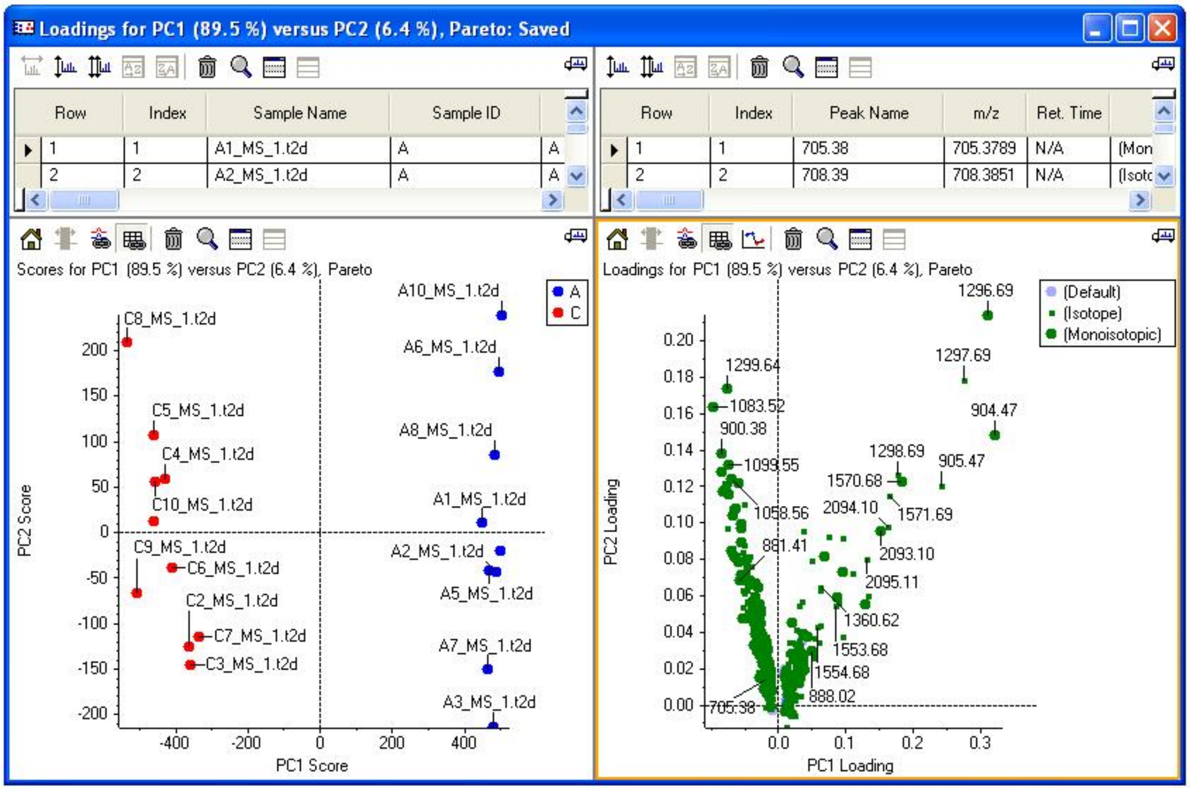


Log (Fold Change) versus p-value for A to C: Saved

Compare: A to C   n1 = 9, n2 = 10

Row	Index	Peak Name	m/z	Ret. Time	Group	Use	t-value	p-value	Mean 1	Mean 2	M
22	22	705.38	705.3789	N/A	(Monoisotopic)	☑	-5.30	5.8564e-5	9.549e2	1.206e3	9.1
23	23	705.94	705.9425	N/A		☑	0.17	0.86915	7.268e1	6.812e1	8.9
24	24	706.39	706.3859	N/A	(Isotope)	☑	-4.54	0.00029	4.979e2	6.230e2	4.9
25	25	706.88	706.8809	N/A		☑	0.80	0.43672	5.357e1	3.373e1	6.7
26	26	707.06	707.0603	N/A		☑	-0.16	0.87292	2.022e1	2.367e1	0.0
27	27	707.40	707.4049	N/A	(Isotope)	☑	-4.27	0.00052	4.225e2	5.495e2	4.2
28	28	707.76	707.7565	N/A		☑	-0.76	0.45896	1.219e1	2.640e1	0.0
29	29	707.94	707.9411	N/A	(Monoisotopic)	☑	1.40	0.17989	8.065e1	4.849e1	1.0
30	30	708.19	708.1884	N/A		☑	1.06	0.30506	2.024e1	0.000e0	0.0

5. Right-click in the table (without making any other selection), select **Use ONLY Selected Peaks** and perform a PCA analysis.



Loadings for PC1 (89.5 %) versus PC2 (6.4 %), Pareto: Saved

Row	Index	Sample Name	Sample ID	
1	1	A1_MS_1.t2d	A	A
2	2	A2_MS_1.t2d	A	A

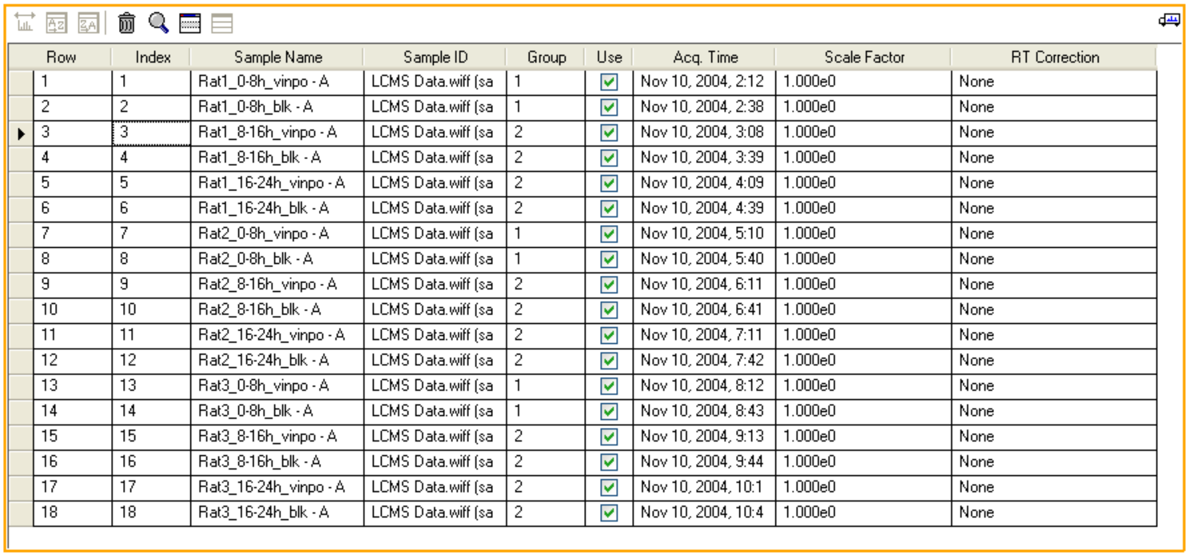
Row	Index	Peak Name	m/z	Ret. Time	
1	1	705.38	705.3789	N/A	(Mon
2	2	708.39	708.3851	N/A	(Isotc

Since the small variables with minimal separation power have been removed, the distinction between the groups – and the variables responsible – is now very clear.

6.5 Combining t-test and PCA

In some cases the t-test can be used to remove, or select, variables before PCA is performed. For example, in the LCMS data we have noticed that there is a significant diurnal variation and we may wish to remove the variables that segregate the 0 – 8 hour sample (pre- and post-does) from all the other samples. One way to do this is to create one group for the 0 – 8 hour samples and a second group for all the other samples and use the t-test to determine the distinguishing variables.

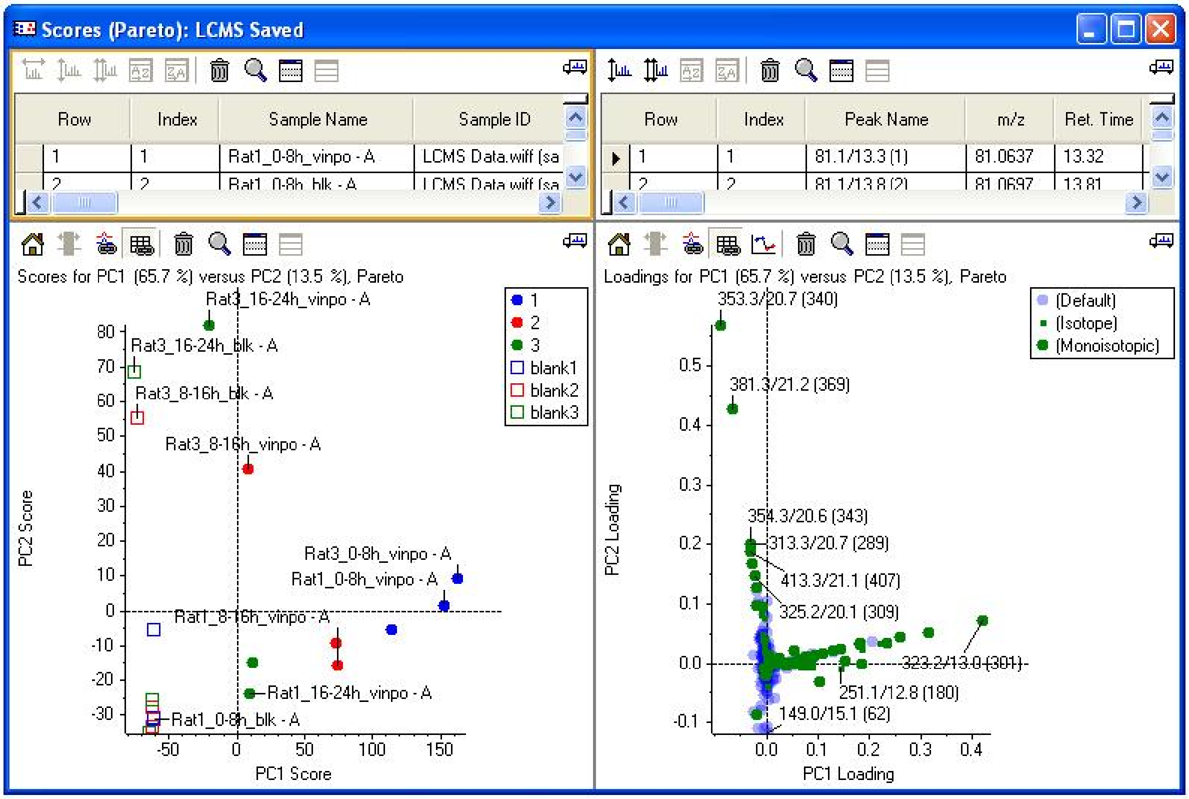
1. Open the data table saved in section 5.2 as **LCMS Saved** and show the sample table by selecting **View -> Show Samples Table**.
2. Click the heading of the **Group** column and select **Edit -> Copy** or type ctrl-C. This will copy the settings for this column to the clipboard so we can retrieve them later.
3. Change all the 0 – 8 hour samples to be group 1 and the rest to group 2



	Row	Index	Sample Name	Sample ID	Group	Use	Acq. Time	Scale Factor	RT Correction
	1	1	Rat1_0-8h_vinpo - A	LCMS Data.wiff (sa	1	☑	Nov 10, 2004, 2:12	1.000e0	None
	2	2	Rat1_0-8h_blk - A	LCMS Data.wiff (sa	1	☑	Nov 10, 2004, 2:38	1.000e0	None
▶	3	3	Rat1_8-16h_vinpo - A	LCMS Data.wiff (sa	2	☑	Nov 10, 2004, 3:08	1.000e0	None
	4	4	Rat1_8-16h_blk - A	LCMS Data.wiff (sa	2	☑	Nov 10, 2004, 3:39	1.000e0	None
	5	5	Rat1_16-24h_vinpo - A	LCMS Data.wiff (sa	2	☑	Nov 10, 2004, 4:09	1.000e0	None
	6	6	Rat1_16-24h_blk - A	LCMS Data.wiff (sa	2	☑	Nov 10, 2004, 4:39	1.000e0	None
	7	7	Rat2_0-8h_vinpo - A	LCMS Data.wiff (sa	1	☑	Nov 10, 2004, 5:10	1.000e0	None
	8	8	Rat2_0-8h_blk - A	LCMS Data.wiff (sa	1	☑	Nov 10, 2004, 5:40	1.000e0	None
	9	9	Rat2_8-16h_vinpo - A	LCMS Data.wiff (sa	2	☑	Nov 10, 2004, 6:11	1.000e0	None
	10	10	Rat2_8-16h_blk - A	LCMS Data.wiff (sa	2	☑	Nov 10, 2004, 6:41	1.000e0	None
	11	11	Rat2_16-24h_vinpo - A	LCMS Data.wiff (sa	2	☑	Nov 10, 2004, 7:11	1.000e0	None
	12	12	Rat2_16-24h_blk - A	LCMS Data.wiff (sa	2	☑	Nov 10, 2004, 7:42	1.000e0	None
	13	13	Rat3_0-8h_vinpo - A	LCMS Data.wiff (sa	1	☑	Nov 10, 2004, 8:12	1.000e0	None
	14	14	Rat3_0-8h_blk - A	LCMS Data.wiff (sa	1	☑	Nov 10, 2004, 8:43	1.000e0	None
	15	15	Rat3_8-16h_vinpo - A	LCMS Data.wiff (sa	2	☑	Nov 10, 2004, 9:13	1.000e0	None
	16	16	Rat3_8-16h_blk - A	LCMS Data.wiff (sa	2	☑	Nov 10, 2004, 9:44	1.000e0	None
	17	17	Rat3_16-24h_vinpo - A	LCMS Data.wiff (sa	2	☑	Nov 10, 2004, 10:1	1.000e0	None
	18	18	Rat3_16-24h_blk - A	LCMS Data.wiff (sa	2	☑	Nov 10, 2004, 10:4	1.000e0	None

4. Perform a t-test; the resulting table will contain one comparison indicating the variables that distinguish group 1 from group 2. The variables that best differentiate these two groups are likely those arising from the diurnal variation.

   Sort the p-value column in ascending order, select the row with the lowest p-value and click the **Plot Profile** button to review the profile of this variable. Use the arrow keys to review some of the other top variables and notice that they are larger in one group than the other.
5. In the t-test result table select the variables with the lowest p-values, for example less than 0.001, right-click and select **Don’t Use Selected Peaks.**
6. Display the sample table, select the **Group** column and type ctrl-v. This will restore the original group assignments. Perform a PCA analysis.



Row	Index	Sample Name	Sample ID
1	1	Rat1_0-8h_vinpo - A	LCMS Data.wiff (sa
2	2	Rat1_0-8h_blk - A	LCMS Data.wiff (sa

Row	Index	Peak Name	m/z	Ret. Time
1	1	81.1/13.3 (1)	81.0637	13.32
2	2	81.1/13.8 (2)	81.0697	13.81

Note that the scores and loading plots have changed and that the samples from rat 3 are now well separated in the positive PC2 direction. The corresponding variables with large positive PC2 loadings are from the contamination that we noted earlier and can easily be excluded.