

High-throughput Processing and Analysis of LC-MS Spectra

By Jianguo Xia (jianguox@ualberta.ca)

Last update : 02/05/2012

This tutorial shows how to process and analyze LC-MS spectra using methods provided in MetaboAnalyst. The spectra processing is performed using the **XCMS** package developed by [Smith CA, et al \(PMID: 16448051\)](#). Five steps are involved – filter and identify peaks, match peaks across samples, correct retention time, fill in missing peaks, and finally arrange peaks into a peak intensity table for statistical analysis. The test data we used is the 12 spectra (NetCDF format) that come together with XCMS package. They are from 12 mice spinal cord samples collected by LC-MS ([Saghatelian et al, PMID: 15533037](#)). Group 1- wild-type (WT) or FAAH(+/+); group 2 – knock-out (KO) or FAAH (-/-). FAAH is the abbreviation of fatty acid amide hydrolase.

Direct comparison of peak intensities without using internal standards is named *discovery metabolite profiling (DMP)* as opposed to the *selected ion monitoring (SIM)* in which the levels of specific compounds are determined using isotopic variants as internal standards. According to the paper ([Saghatelian et al, PMID: 15533037](#)), the DMP measurements were within 1.6-fold of results obtained by targeted SIM analysis, and can be used for quantitative comparisons as well as for novel biomarker discovery.


The focus of this tutorial is on spectra processing rather than statistical analysis due to small sample size. Please note, you can process spectra locally and then upload the peak list files or a peak intensity table after calibration using internal standards.

Step 1. Go to the **Data Format** page, under “Zipped file (.zip) format” option, click the [download](#) link after the “LC/GC-MS spectra (NetCDF, mzDATA, or mzXML)” option, and save the data to your local disk. (*Note: no space or special characters are allowed in either folder (group) names or spectra names.*)

Step 2. Go the MetaboAnalyst **Home** page and click “click here to start” to enter the data upload page.

Zipped files (.zip) format ([show details](#)), including :

- NMR peak lists (2 columns - chemical shift and intensity) ([download](#))
- MS peak lists (2 columns - mass and intensity) ([download](#))
- MS peak lists (3 columns - mass, retention time, and intensity) ([download](#))
- LC/GC - MS spectra (NetCDF, mzDATA, or mzXML) ([download](#))



Welcome ([click here to start](#))

MetaboAnalyst is a web-based analytical pipeline for high-throughput metabolomics studies. It offers a variety of options for metabolomic data processing, normalization, multivariate statistical analysis, as well as data annotation. The current implementation focuses on [biomarker discovery](#) and [classification](#) for the [two-group](#) discrimination problem. Its main features include:

Step 3. In the **Upload** page, go to the “Upload your data” panel. Under the “Zipped Files (.zip)” option, select the “MS spectra” option and browse to your directory of the zip file you just created, then click “Submit”. Please note, we ignore the “Pairs” option which is only required if you want to conduct paired analysis.

Zipped Files (.zip) : For WinZip 12.x, choose "Legacy compression (Zip 2.0 Compatible)"

Data type : NMR peak list MS peak list MS spectra

Data :

Pairs : (required for paired comparison)

Note: alternatively, you can directly select the #7 option in the “Try our test data” without downloading the example.

2) Try our test data : (You can download these data [here](#))

Data Type	Description
<input type="radio"/> Concentrations Tutorial Report	Urinary compound concentrations from 57 cancer patients measured by 1H NMR (unpublished data). Group 1- no weight loss; group 2 - severe weight loss
<input type="radio"/> NMR spectral bins Tutorial Report	Binned 1H NMR spectra of 50 urine samples using 0.04 ppm constant width (Psihogios NG, et al.) Group 1- control; group 2 - severe kidney disease.
<input type="radio"/> NMR peak lists	Peak lists and intensity files for 50 urine samples measured by 1H NMR (Psihogios NG, et al.). Group 1- control; group 2 - severe kidney disease.
<input type="radio"/> Concentrations (paired) Tutorial Report	Compound concentrations of 14 urine samples collected from 7 cows at two time points using 1H NMR (unpublished data). Group 1- day 1, group 2- day 4.
<input type="radio"/> MS peak intensities	LC-MS peak intensity table for 12 mice spinal cord samples (Saghatelian et al.). Group 1- wild-type; group 2 - knock-out.
<input type="radio"/> MS peak lists	Three-column LC-MS peak list files for 12 mice spinal cord samples (Saghatelian et al.). Group 1- wild-type; group 2 - knock-out.
<input checked="" type="radio"/> LC-MS spectra Tutorial Report	NetCDF spectra of 12 mice spinal cord samples collected by LC-MS (Saghatelian et al.). Group 1- wild-type; group 2 - knock-out.

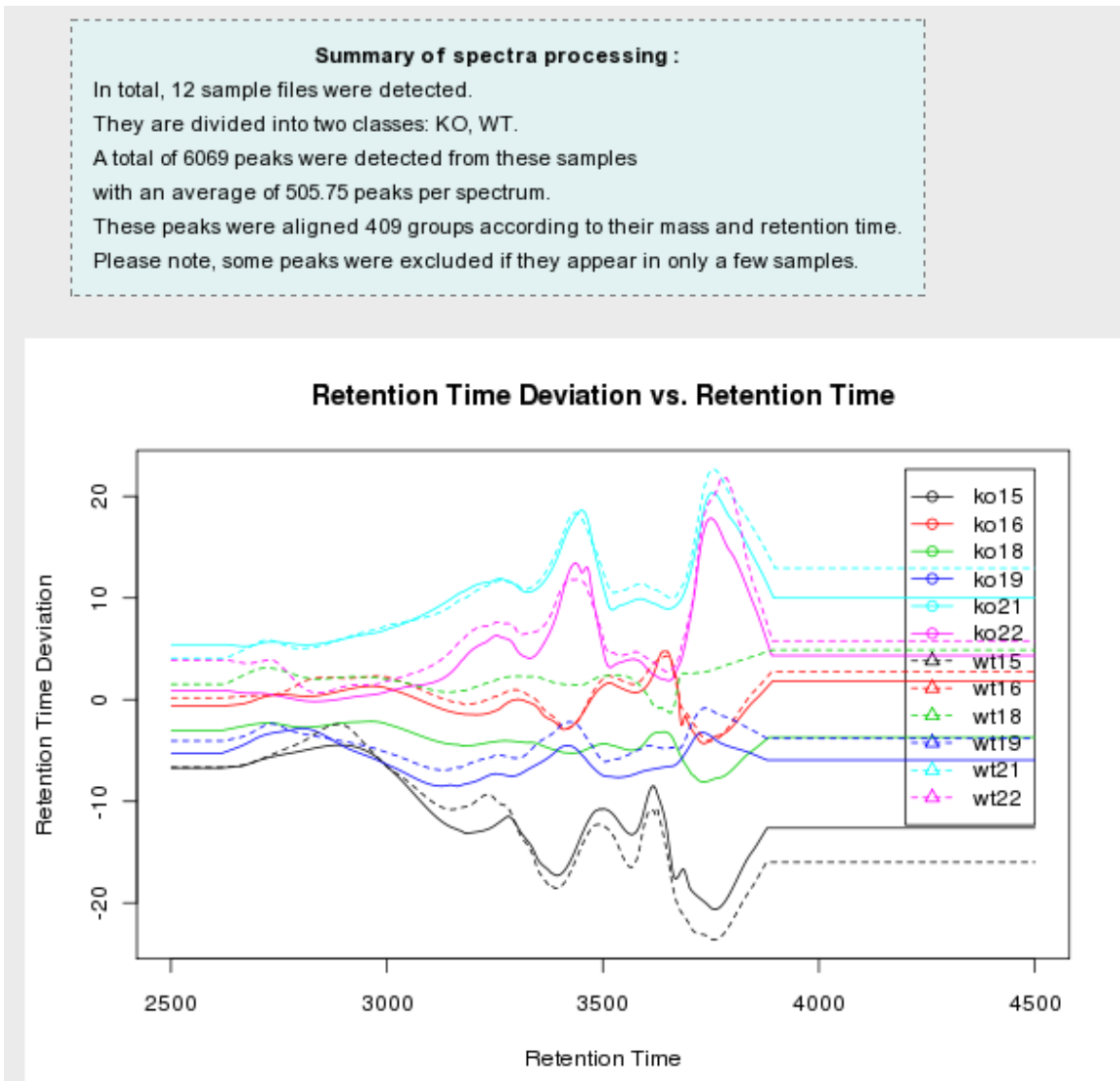
Step 4. In this step, we set the parameters for MS spectra processing. The “Full width at half maximum (fwhm)” is the most important parameter. It is used to specify a Gaussian model for peak detection and can be quite different for different chromatography. Here we use 30 (seconds) as suggested for LC-MS. Leave other parameters as default and click “Next”. Please note, the process can be very long if there are a large number of spectra uploaded.

Processing parameters:

Full width at half maximum (fwhm)	<input type="text" value="30"/>	(seconds)
Profiling method	<input type="text" value="bin"/>	
Peak Intensity measure	<input type="text" value="Integrated area of original peak"/>	

Please wait, the process may take a while ...

Step 5. Peak detection, peak grouping, retention time correction, and filling of missing peak are performed sequentially. The result is summarized below. More detailed information is available in the analysis report when the analysis is complete. Click “Next” to continue.



Step 6. In this step, data sanity checks are performed with the results shown below. Click “Skip” button to go to Normalization step. Note, 110 zero values and no missing values are detected in the data. By default, these values will be replaced by half of the minimum positive values from the data since some algorithm does not work properly with zero (i.e. log transformation).

Note, missing values are represented as NA (no quotes) or empty values.

Data processing information

Checking data content ...passed

Two groups were detected based on the sample labels.

Samples are not paired.

All data values are numeric.

All data values are non-negative.

A total of 110 , (2.2 %) zero values were detected

A total of 0 , (0 %) missing values were detected

By default, these values will be replaced by a small value

Click **Skip** button if you accept the default practice

Or click **Missing value imputation** to use other methods

Missing value imputation Skip

Step 7. Now we arrive at the data normalization step. The internal data structure is now a table with each row representing a sample and each column represents a feature (peak intensities). With the data structured in this format, two types of data normalization protocols - row-wise normalization and column-wise normalization -- may be used. These are often applied sequentially to reduce systematic variance and to improve the performance for downstream statistical analysis. Row-wise normalization aims to normalize each sample (row) so that it is comparable to the other. For row-wise normalization MetaboAnalyst uses normalization to a constant sum, normalization to a reference sample (probabilistic quotient normalization), normalization to a reference feature (creatinine or an internal standard) and sample-specific normalization (dry weight or tissue volume). In contrast to row-wise normalization, column-wise normalization aims to make each feature (column) more comparable in magnitude to the other. Four widely-used methods are offered in MetaboAnalyst - log transformation, auto-scaling, Pareto scaling, and range scaling. According to the paper, the data was normalized by the amount of the tissue used to extract each sample. Therefore, we choose “Sample specific normalization” and click the link “Click here to specify” to specify tissue amount for each sample.

Row-wise normalization

None

Normalization by sum

Normalization by a reference sample <Not set> ▾

Normalization by a reference feature <Not set> ▾

Sample specific normalization (i.e. dry weight, volume) [Click here to specify](#)

Column-wise normalization

None

Log (log₂ transformation)

Autoscaling (mean-centered and divided by the standard deviation of each variable)

Pareto Scaling (mean-centered and divided by the square root of standard deviation of each variable)

Range Scaling (mean-centered and divided by the range of each variable)

Step 8. In this page, we enter the tissue amount used for the extraction of each sample. However, since we don't know this information, we will use the default values. Click the “Submit” button to go back to the Normalization page.

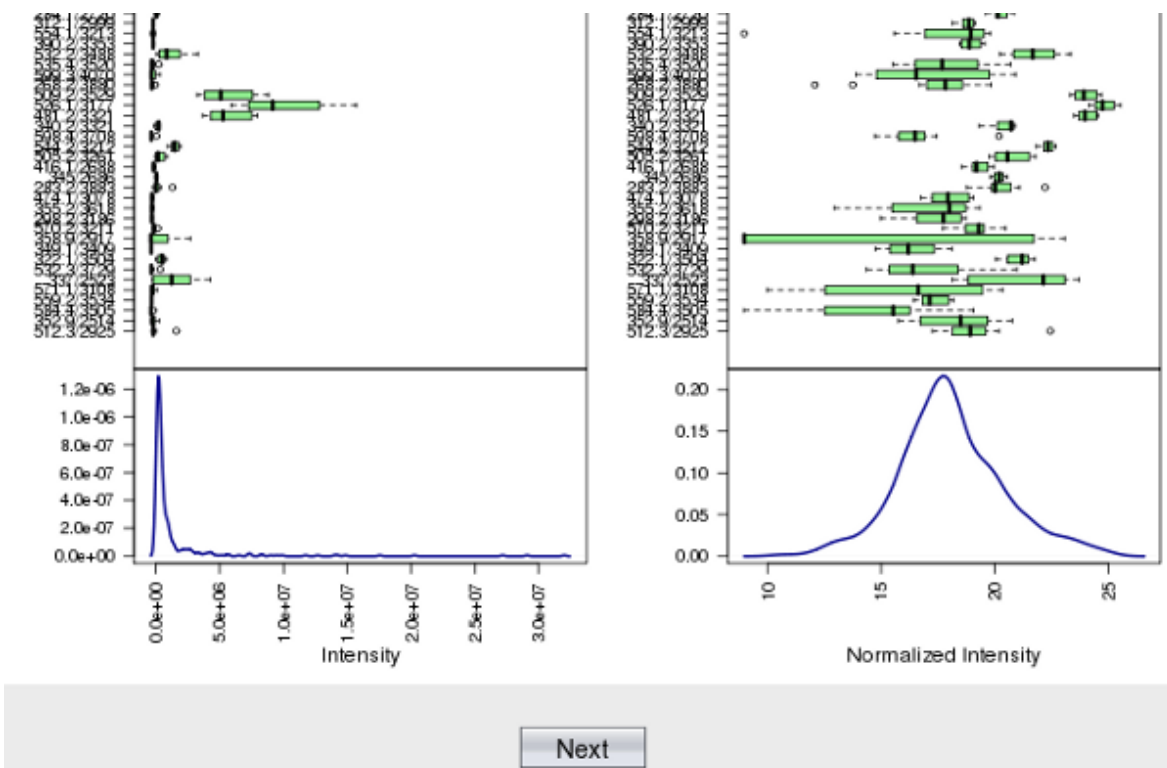
Sample specific normalization

Note: normalization factor refer to the sample specific information (i.e. dry weight, volume). Each sample will be normalized by its own normalization factor.

Sample ID	Normalization factor
ko15	1.0
ko16	1.0
ko18	1.0
ko19	1.0
ko21	1.0
ko22	1.0
wt15	1.0
wt16	1.0
wt18	1.0
wt19	1.0
wt21	1.0
wt22	1.0

The radio button becomes unselected after you go back, make sure the “Sample specific normalization” option is re-selected! Choose “Log normalization” or “None” for column-wise normalization since we are primarily interested in fold changes between the two groups (this is the analysis used on the paper). Click the “Process” button at the bottom to continue.

The normalization result is shown below. On the left is a plot (box-whisker plot on top, linear distribution plot on the bottom) of the data prior to normalization. On the right is a plot (box-whisker plot on top, linear distribution plot on the bottom) of the data after normalization. As can be seen by comparing the linear concentration curve on the left (which has an exponential decay character to it) to the log-transformed curve on the right (which looks reasonably Gaussian), the normalization procedures makes the peak intensity data reasonably “normal”. You can also try other normalization approaches and compare their results. Note, the Click “Next” button to continue

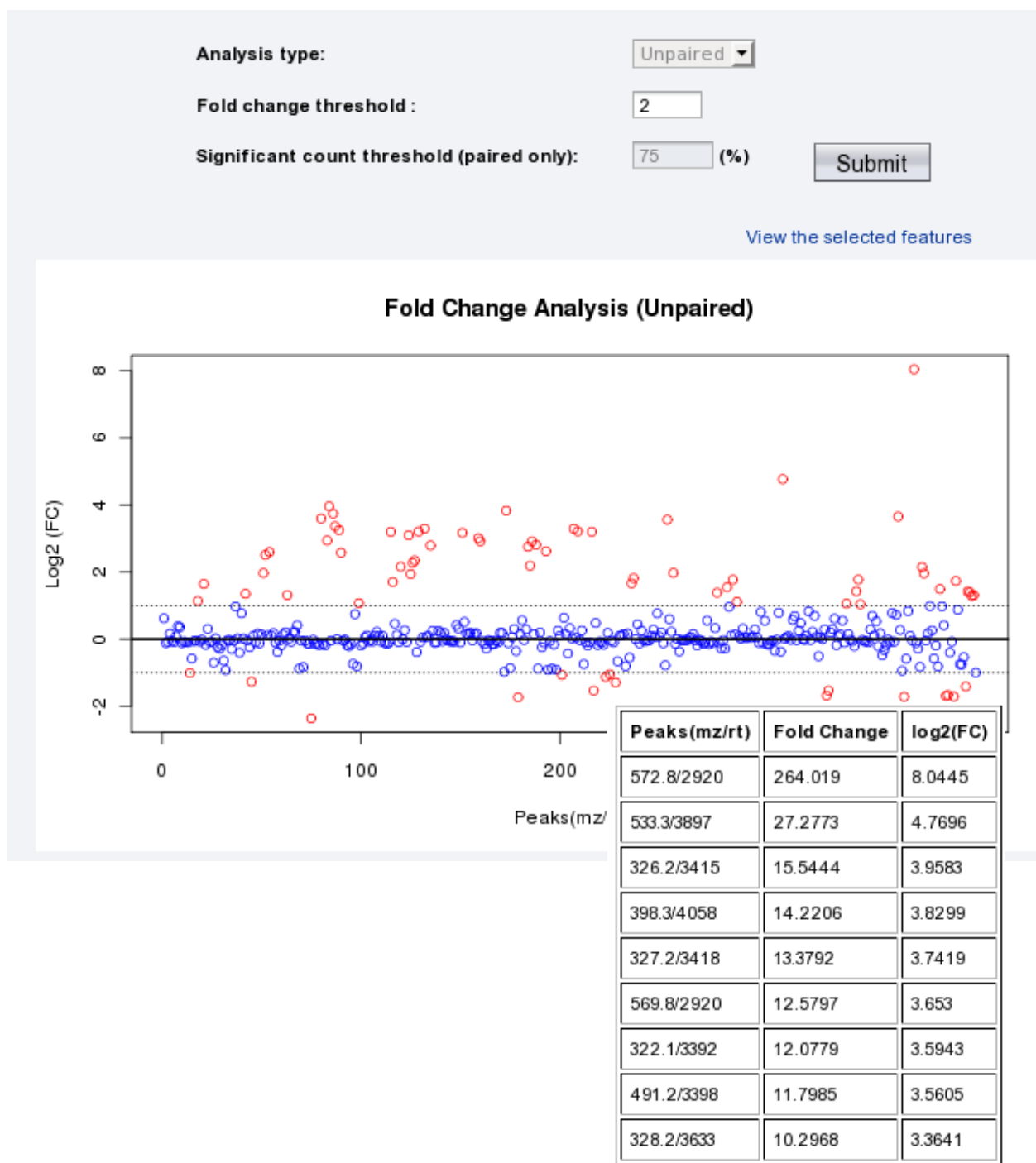


Step 9. After we finish data processing and normalization, the data is suitable for different statistical analysis. There are many methods available in MetaboAnalyst for identification of features that are significantly different between two groups. However, given the small sample size, only the Univariate analysis will be performed here.

The screenshot displays the MetaboAnalyst web interface. On the left, a vertical 'Steps' menu is visible, with '4. Analyze' selected and expanded to show various analysis options: Univariate, PCA, PLSDA, SAM, EBAM, Tree & heatmap, Kmean & SOM, RandomForest, R-SVM, 5. Peak Search, 6. Pathway Mapping, 7. Download, and Log Out. The main content area is titled 'Select an analysis path to explore :'. Below this title, several analysis categories are listed with their respective sub-methods:

- Univariate Analysis**
 - [Fold Change, t-Tests, and Volcano plot](#)
- Chemometrics**
 - [Principal Component Analysis \(PCA\)](#)
 - [Partial-Least Square - Discriminant Analysis \(PLS-DA\)](#)
- Significant Feature Identification**
 - [Significance Analysis of Microarray \(SAM\)](#)
 - [Empirical Bayesian Analysis of Microarray \(EBAM\)](#)
- Cluster Analysis**
 - [Hierarchical Clustering - Dendrogram and Heatmap](#)
 - [Partitional Clustering - K-Means and Self Organizing Map \(SOM\)](#)
- Classification & Feature Selection**
 - [Random Forest](#)
 - [Support Vector Machine \(SVM\)](#)

Step 8. Click the “Univariate” link on the navigation panel. Many features are above the the default fold change threshold. Note, the fold changes are log2 transformed so that up-regulated and down-regulated features will be plotted symmetrically on the graph (i.e. 2 fold change will be the same distance to the baseline (0) as 0.5, since $\log_2(2) = 1$, $\log_2(0.5) = -1$). Click “view selected features” for a table view. A subset of the table is shown below.



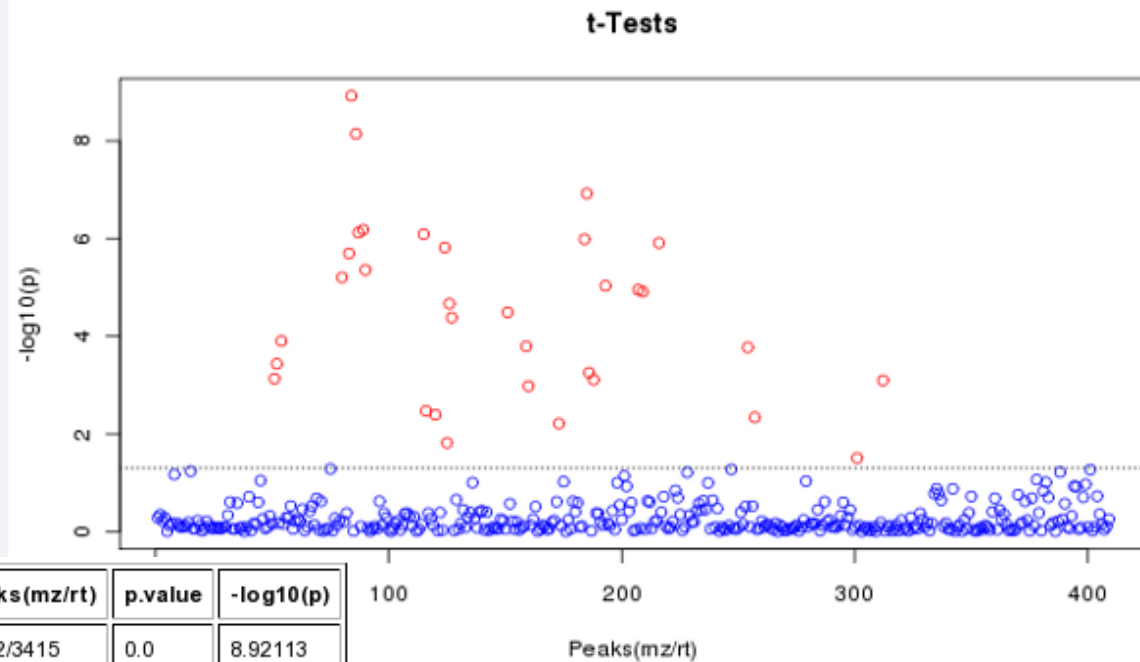
Step 9. Click the “t-Tests” tab and you will see the following result with default p-value 0.1. Again, click “View the selected features” for a detailed table view. A subset of the table is shown below.

Analysis type :

Group variance :

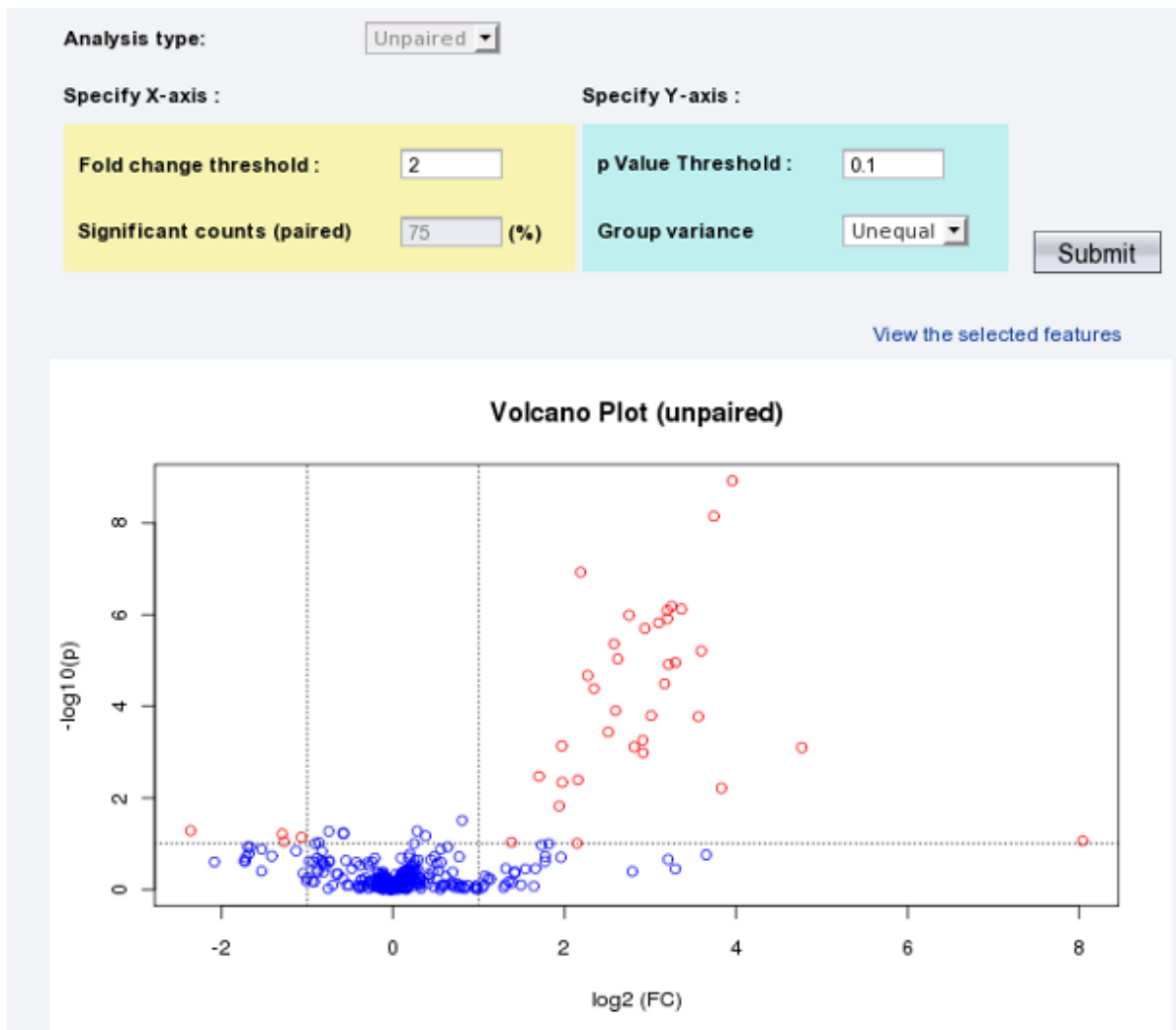
Threshold : p value <

[View the selected features](#)



Peaks(mz/rt)	p.value	-log10(p)
326.2/3415	0.0	8.92113
327.2/3418	0.0	8.14622
411.2/3937	0.0	6.9265
329.2/3630	0.0	6.18045
328.2/3633	0.0	6.12014
348.2/3288	0.0	6.08935
410.3/3937	0.0	5.98747
449.1/3290	0.0	5.91087
354.2/3618	0.0	5.81689
324.2/3276	0.0	5.69781

Step 10. Click the “Volcano plot” tab to see the result image from volcano analysis. Volcano plot combines fold change analysis and t-tests in each dimension. Each analysis can be adjusted individually. The further away its position from (0,0), the more significant the corresponding feature. Note, both x and y-axis are on log scale.



Click the “View the selected features” link to see the details of these important features. A subset of the table is shown below.

Peaks(mz/rt)	FC	log2(FC)	p.value	-log10(p)
326.2/3415	15.544	3.958	0.0	8.921
327.2/3418	13.379	3.742	0.0	8.146
411.2/3937	4.562	2.19	0.0	6.926
329.2/3630	9.518	3.251	0.0	6.18
328.2/3633	10.297	3.364	0.0	6.12
348.2/3288	9.185	3.199	0.0	6.089
410.3/3937	6.745	2.754	0.0	5.987
449.1/3290	9.197	3.201	0.0	5.911
354.2/3618	8.557	3.097	0.0	5.817
324.2/3276	7.671	2.939	0.0	5.698
330.2/3631	5.966	2.577	0.0	5.36
322.1/3392	12.078	3.594	0.0	5.206

Step 11. Now we show how to identify MS peaks using the build-in peak search tools. Click the “Peak Search” link on the navigation panel, then click the “MS search” tab. Let's try the first 3 peaks from the volcano plots, enter the **mz** value of each peak and click “Search” with default parameters. The figure below shows the search result for the third peak.

Paste MS peak list (in one column as shown) :

411.2

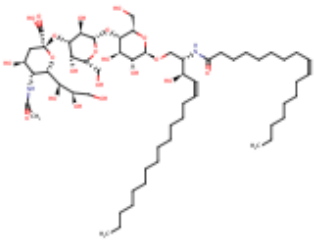
Database type :

Mass tolerance (\pm) :

[View all library hits](#)

Common Name	Molecular Weight	Adduct	HMDB Link
Ganglioside GM3 (d18:1/9Z-18:1)	1164.71	M+3Na	HMDB04843
Sialyl-Lewis X	820.3	M+2H	HMDB06565
3-Sialyl Lewis	820.3	M+2H	HMDB06579
3'-Sialyl-3-fucosyllactose	779.27	M+ACN+2H	HMDB06606
1-Palmitoyl lysophosphatidic acid	410.24	M+H	HMDB00327
Risperidone	410.21	M+H	HMDB05020

Click the top **HMDB link** to get more details. The screen shot below shows the MetaboCard for top hit for “411.2”

Chemical Formula	C ₅₈ H ₁₀₄ N ₂ O ₂₁										
Chemical Structure											
Chemical Taxonomy	<table border="1"> <tr> <td data-bbox="695 856 1201 898">Kingdom</td> <td data-bbox="706 898 1201 940"> <ul style="list-style-type: none"> Organic </td> </tr> <tr> <td data-bbox="695 940 1201 982">Class</td> <td data-bbox="706 982 1201 1024"> <ul style="list-style-type: none"> Glycolipids </td> </tr> <tr> <td data-bbox="695 1024 1201 1066">Family</td> <td data-bbox="706 1066 1201 1108"> <ul style="list-style-type: none"> Mammalian Metabolite </td> </tr> <tr> <td data-bbox="695 1108 1201 1150">Species</td> <td data-bbox="706 1150 1201 1392"> <ul style="list-style-type: none"> acetal primary alcohol secondary alcohol 1,2-diol carboxylic acid secondary carboxylic acid amide alkene heterocyclic compound </td> </tr> <tr> <td data-bbox="695 1392 1201 1434">Biofunction</td> <td data-bbox="706 1434 1201 1476"> <ul style="list-style-type: none"> Membrane component </td> </tr> </table>	Kingdom	<ul style="list-style-type: none"> Organic 	Class	<ul style="list-style-type: none"> Glycolipids 	Family	<ul style="list-style-type: none"> Mammalian Metabolite 	Species	<ul style="list-style-type: none"> acetal primary alcohol secondary alcohol 1,2-diol carboxylic acid secondary carboxylic acid amide alkene heterocyclic compound 	Biofunction	<ul style="list-style-type: none"> Membrane component
Kingdom	<ul style="list-style-type: none"> Organic 										
Class	<ul style="list-style-type: none"> Glycolipids 										
Family	<ul style="list-style-type: none"> Mammalian Metabolite 										
Species	<ul style="list-style-type: none"> acetal primary alcohol secondary alcohol 1,2-diol carboxylic acid secondary carboxylic acid amide alkene heterocyclic compound 										
Biofunction	<ul style="list-style-type: none"> Membrane component 										

Step 12. Now, assume we have finished the analysis. Click the “Download” link on the left panel. A detailed analysis report will be generated (MetaboAnalystReport.pdf) containing introductions and results for every steps we have performed. Now, you can directly click and download the “Download.zip” file which includes all the processed data, images, and the PDF report. Alternatively, you can ask MetaboAnalyst to send you the result via email by entering your email address.

Steps

- 1. Upload
- 2. Process
- 3. Normalize
- 4. Analyze
 - Univariate
 - PCA
 - PLSDA
 - SAM
 - EBAM
 - Tree & heatmap
 - Kmean & SCM
 - RandomForest
 - R-SVM
- 5. Peak Search
- 6. Pathway Mapping
- 7. Download
- Log Out

Download

The "Download.zip" contains all the files inside your home directory. You can directly download the result or let MetaAnalyst to send you via email.

Email address :

Files in your home directory
Download.zip
MetaboAnalystReport.pdf
conc-norm.png
data_normalized.csv
data_original.csv
data_processed.csv
retcor-ms.png
Rhistory.R
univar_fc.png
univar_t.png
univar_volcano.png

-----End of tutorial-----