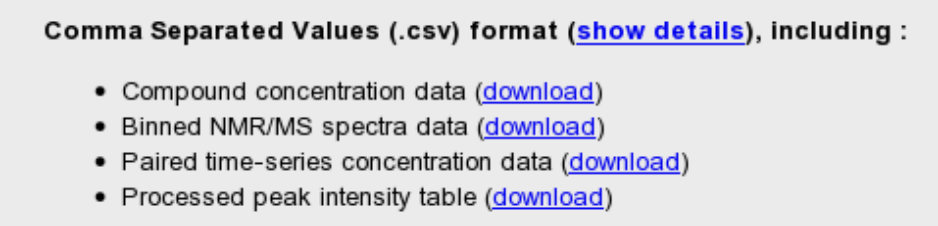
Identify Significant Features Using Concentration Data

By Jianguo Xia (jianguox@ualberta.ca)

Last update: 4/15/2009

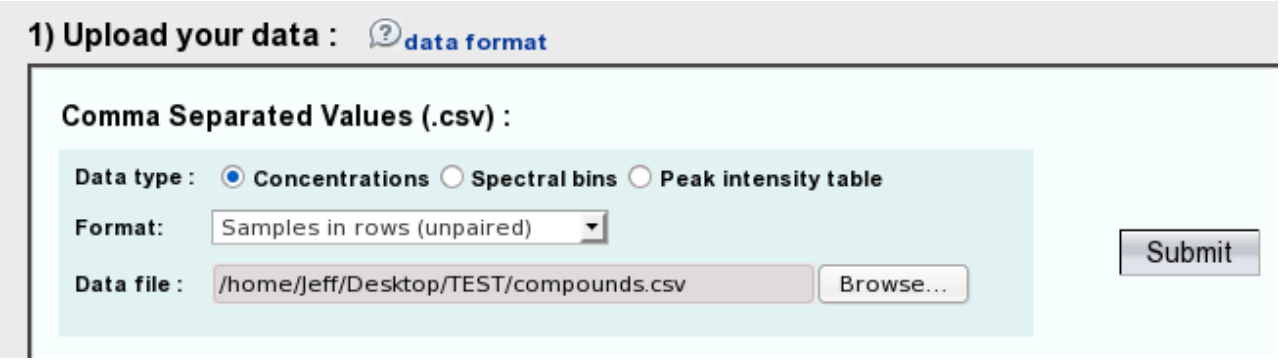
This tutorial shows how to identify significant features using methods provided in MetaboAnalyst. The example used is compound concentration data obtained by targeted (i.e. quantitative) metabolic profiling of ^{1}H NMR spectra of urine samples collected from 57 cancer patients. There are two groups of patients – Cachexia (Y) refers to the group with significant skeletal muscle loss; Cachexia (N) refers to the group with no obvious skeletal muscle loss. Cachexia is defined as the loss of weight, muscle atrophy, fatigue, weakness and significant loss of appetite in someone who is not actively trying to lose weight. Cachexia is often seen in end-stage cancer, and in that context is called "cancer cachexia". The exact mechanism behind cachexia is poorly understood, but there is probably a role for inflammatory cytokines, such as tumor necrosis factor-alpha (TNF-α) - which is also nicknamed cachexin, Interferon gamma (IFNγ), and Interleukin 6 (IL-6). The goal in this tutorial is to identify metabolites that are significantly different between these two groups of cancer patients (cachexic vs. non-cachexic). These metabolites could serve as potential early-stage biomarkers for detecting cachexia and for exploring its underlying metabolic basis.

Step 1. Go to the **Data Formats** page (found by clicking on the **Data Formats** hyperlink on MetaboAnalyst's home page), click the download link after the "Compound concentration data" option to download the compressed zip file. Unzip the file and save as "compounds.csv".

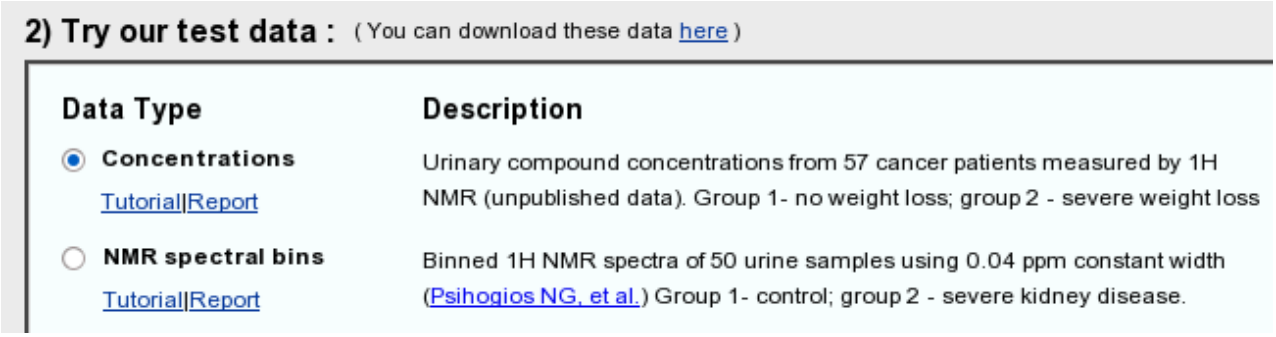


Step 2. Go the MetaboAnalyst **Home** page and click "click here to start" to enter the data upload page.

Step 3. In the **Upload** page, go to the "Upload your data" panel, select the options as indicated below and click "Submit"

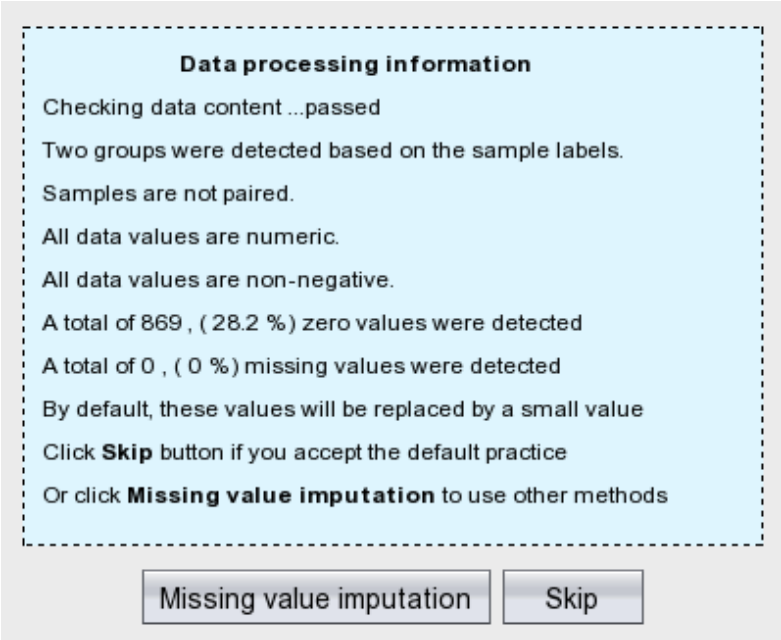


***Note**: Alternatively, you can directly select the first option in the "Try our test data" without downloading the example.*



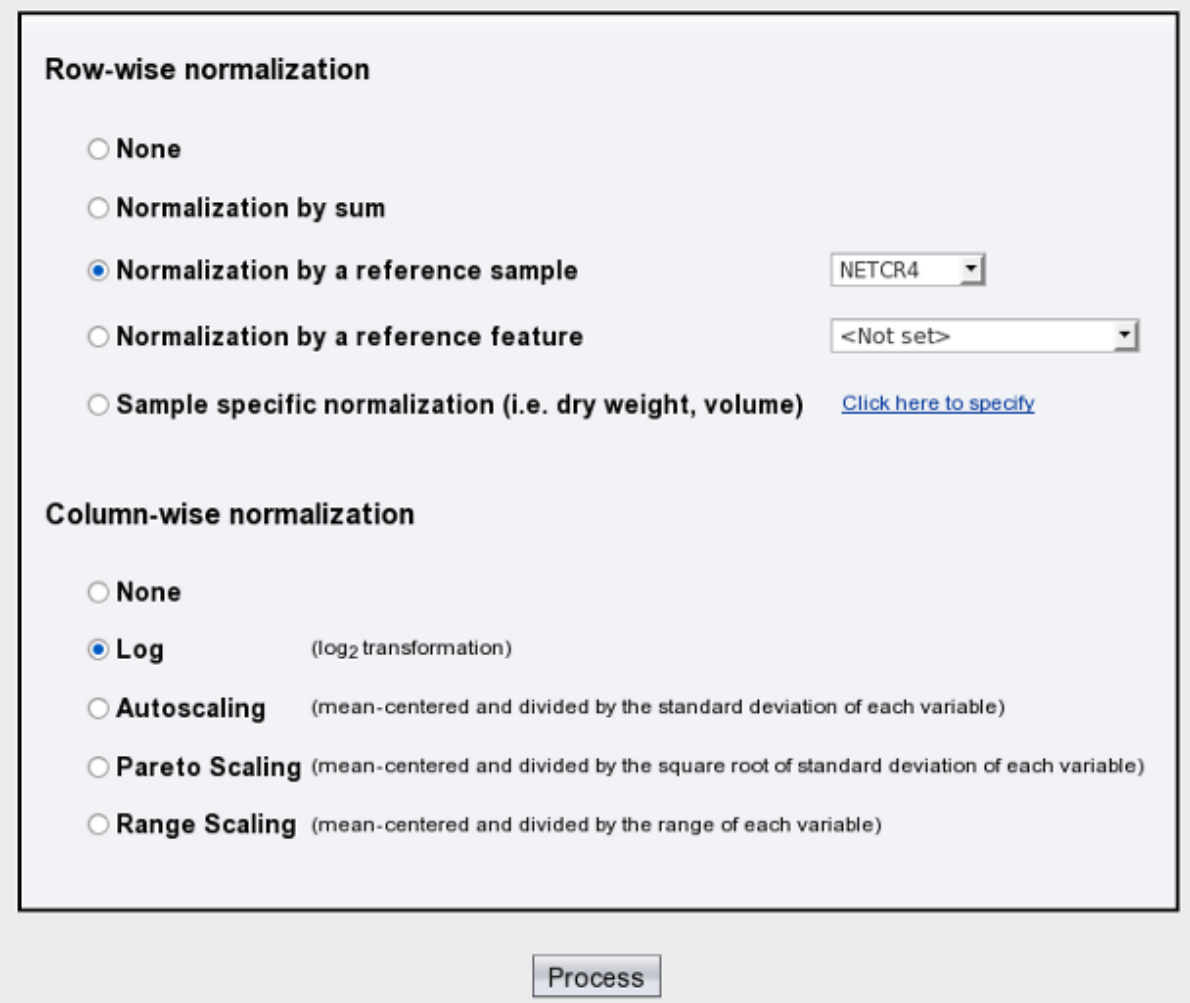
Step 4. The data integrity check will run automatically and the result is shown below. For lists of concentrations the data integrity check will assess the content (look for consistent formatting and the presence of two groups), determine whether the data is paired or determine if negative numbers exists. In this case, 869 zero values and no missing values were identified in the data. Since zero values may cause some algorithms not to work properly, MetaboAnalyst will replace these zero values with a small positive value (the half of the minimum positive number detected in the data). Click "Skip" to go to normalization step. If missing values had been detected, then the most appropriate from a variety of methods provided by MetboAnalyst could have been used to deal with this issue (for such an example, see MetaboAnalyst Tutorial 4).

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

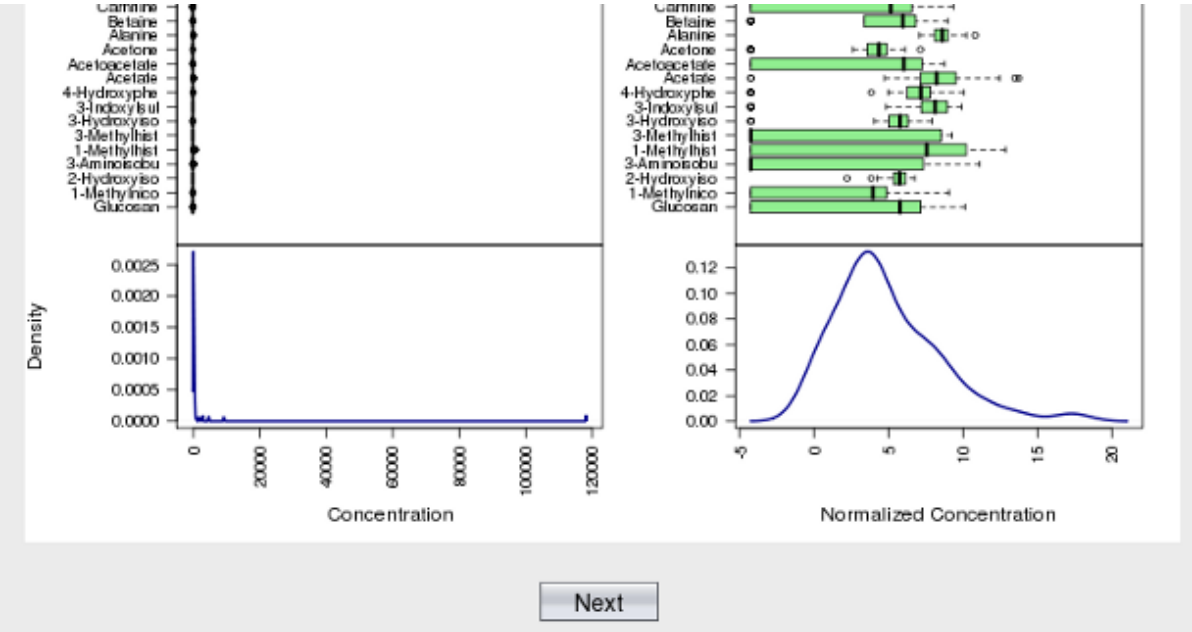


Step 5. Now we arrive at the data normalization step. The internal data structure is transformed now to a table with each row representing a urine sample (from a patient) and each column representing a feature (a compound with a concentration). With the data structured in this format, two types of data 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 that they are comparable to each other. For row-wise normalization MetaboAnalyst supports 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 each other. Four widely-used methods are offered in MetaboAnalyst - log transformation, auto-scaling, Pareto scaling, and range scaling. Urine concentrations are usually normalized by creatinine concentration to adjust for dilution effects (select option 4 - Normalization by a reference feature and choose “creatinine”). However, in this case, creatinine is the product of protein breakdown and is related to the skeletal muscle loss. Since normalizing to a metabolite that might be important for understanding this disorder might introduce biological bias, we need to choose another kind of normalization process. As a result we choose to normalize by a reference sample “NETCR4” (a general rule is to choose a sample in the control group with the fewest missing values). After deciding to normalize by reference sample for our row-wise (sample) normalization we then choose “Log normalization” for our column normalization to make the metabolite concentration values more comparable among different compounds.

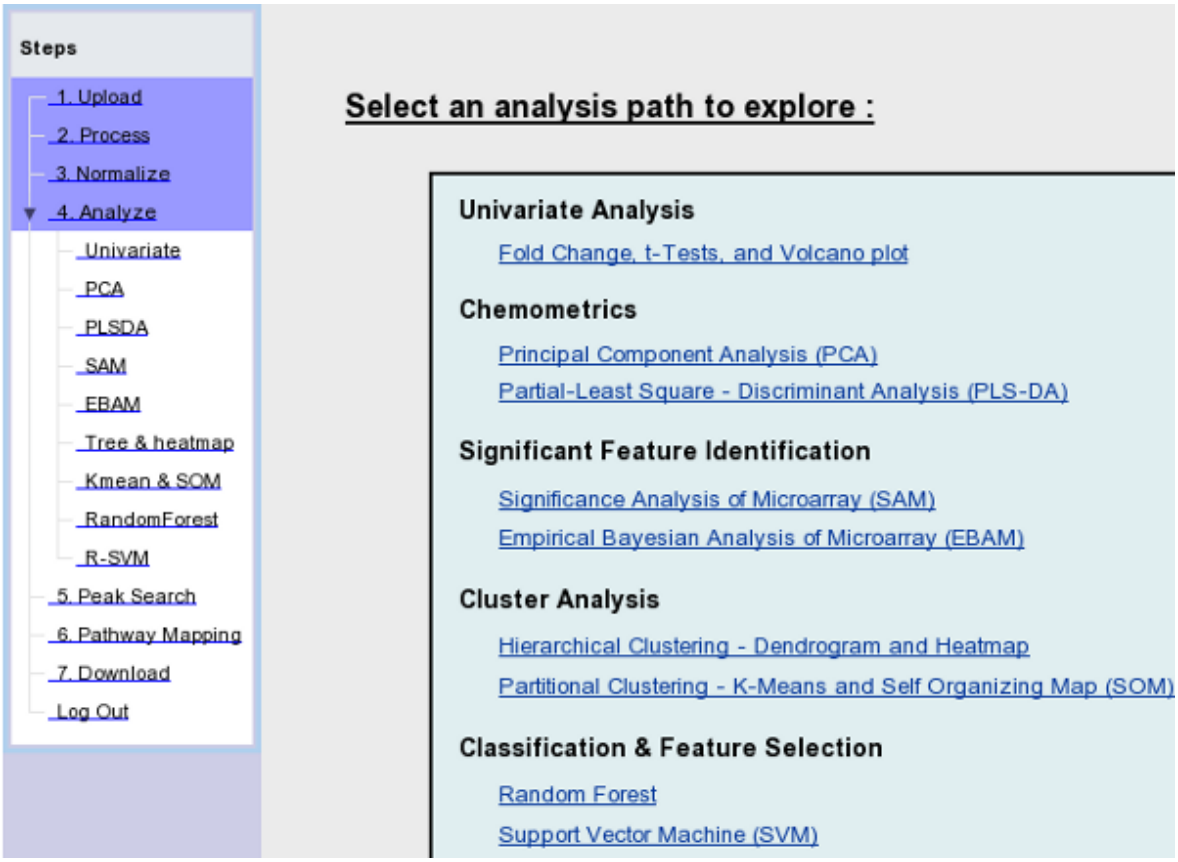
Hint: Remember to click on Normalization by a reference sample and not only select NETCR4



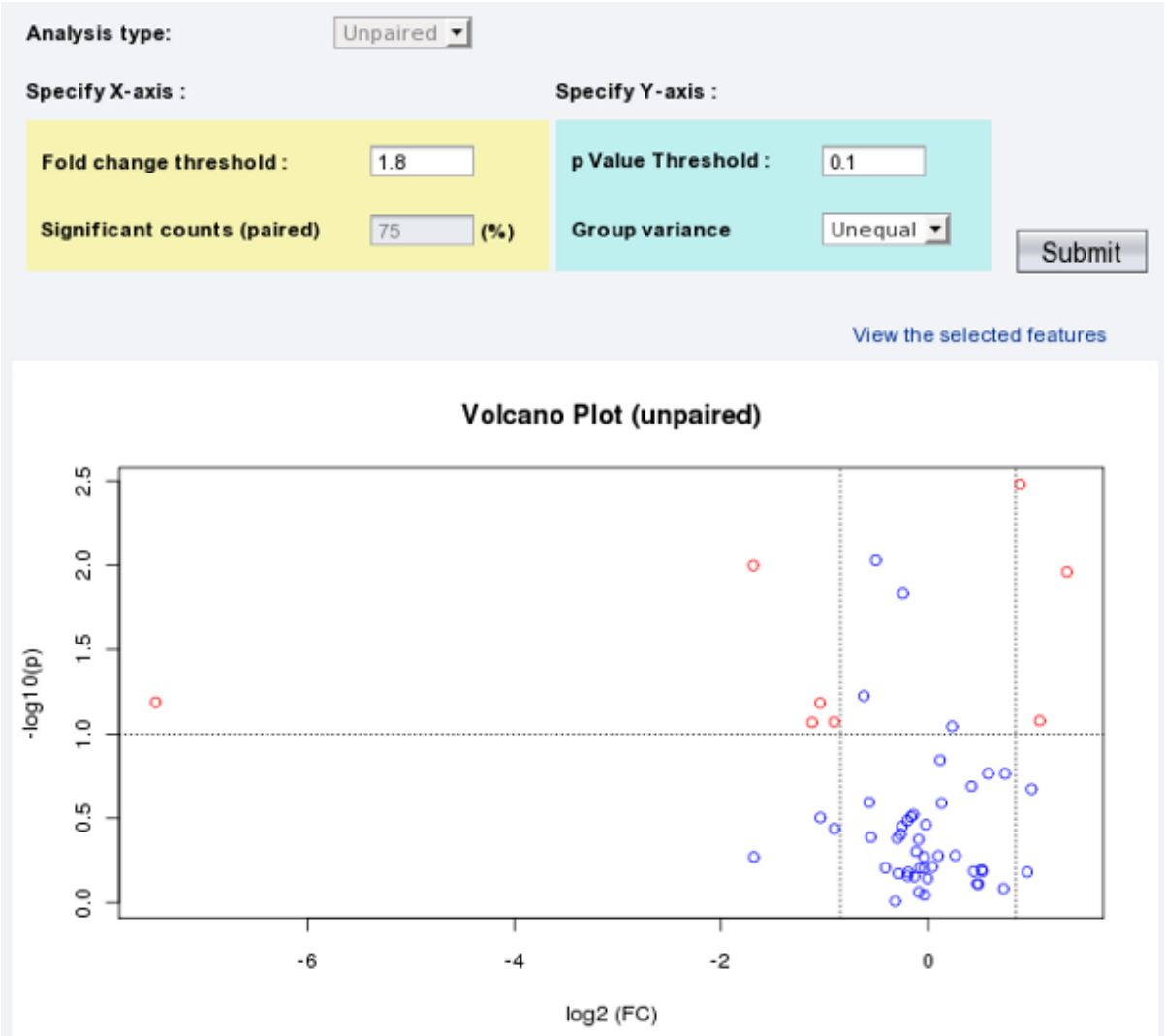
The normalization result is shown below. On the left is a plot (box-whisker plot on top, linear distribution plot on the bottom) of the data prior to normalization. On the right is a plot (box-whisker plot on top, linear distribution plot on the bottom) of the data after normalization. As can be seen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



Step 6. Now we have finished data processing and normalization. The data are now suitable for different statistical analyses. There are many feature selection methods available in MetaboAnalyst. Here we will only show results from Volcano Plot, PLS-DA and SAM methods. The screen shot below shows MetaboAnalyst's analysis view. Please note the navigation panel on the left. A color change indicates the corresponding step has been successfully performed. All the data analysis methods can be directly accessed by clicking the corresponding hyperlink.

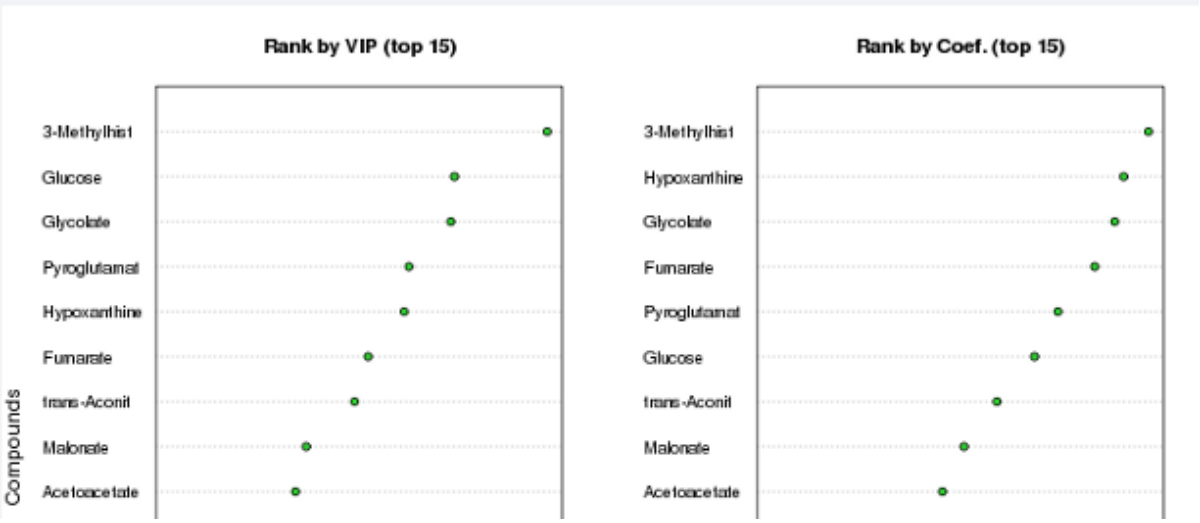


Step 7. Generally the simplest kind of analysis that can be performed on this type of metabolomic data is univariate data analysis. Univariate analyses are often first used to obtain an overview of the data or a rough ranking of potentially important features before applying more sophisticated data analysis tools. Univariate analysis examines each variable separately without taking into account the effect of multiple comparisons. MetaboAnalyst's univariate analysis path supports three commonly used methods - fold-change analysis, t-tests, and volcano plots. To begin the univariate analysis, click the "Univariate" link on the navigation panel to the left. From here we will perform a volcano plot analysis. Volcano plots are used to compare the size of the fold change to the statistical significance level. The X axis plots the fold change between the two groups (on a log scale), while the Y axis represents the p-value for a t-test of differences between samples (on a negative log scale). To start a volcano plot click the "Volcano" tab. Adjust the fold change (FC) threshold to 1.8 and click "Submit". As can be seen by the figure below, eight features are detected as significant and colored in red. Click the "View the selected features" link to view the names/identities of these features. The table at the bottom shows these eight features. These include glycolate, fumarate, 3-methylhistidine, glucose, *etc.*



Compounds	FC	log2(FC)	p.value	-log10(p)
Glycolate	1.856	0.893	0.0030	2.478
Fumarate	0.31	-1.688	0.01	1.999
3-Methylhistidine	2.543	1.347	0.011	1.961
Glucose	0.0060	-7.471	0.065	1.189
Pyroglutamate	0.485	-1.043	0.065	1.184
Acetoacetate	2.116	1.081	0.084	1.078
Choline	0.533	-0.908	0.085	1.071
Malonate	0.46	-1.119	0.086	1.068

Step 8. The Volcano plot has provided some intriguing results. We may now want to examine whether these metabolites are also detected as being significant using a slightly more sophisticated analysis tool. In particular we will use Partial-Least Squares Discriminant Analysis (PLS-DA). As a supervised method, PLS-DA can perform both classification and feature selection. The algorithm uses cross-validation to select an optimal number of components for classification. Two feature importance measures are commonly used in PLS-DA. Variable Importance in Projection or VIP score is a weighted sum of squares of the PLS loadings. The weights are based on the amount of explained Y-variance in each dimension. The other importance measure is based on the weighted sum of PLS-regression coefficients. The weights are a function of the reduction of the sums of squares across the number of PLS components. More details about these two methods can be obtained by placing your mouse over the "About PLS" link. Go back to the Analysis window and click the "PLSDA" link on the navigation panel and then click the "Var.Importance" tab. You will see the result as shown below. The graphs rank the different metabolites (the top 15) according to the VIP score on the left and according to the coefficient score on the right.

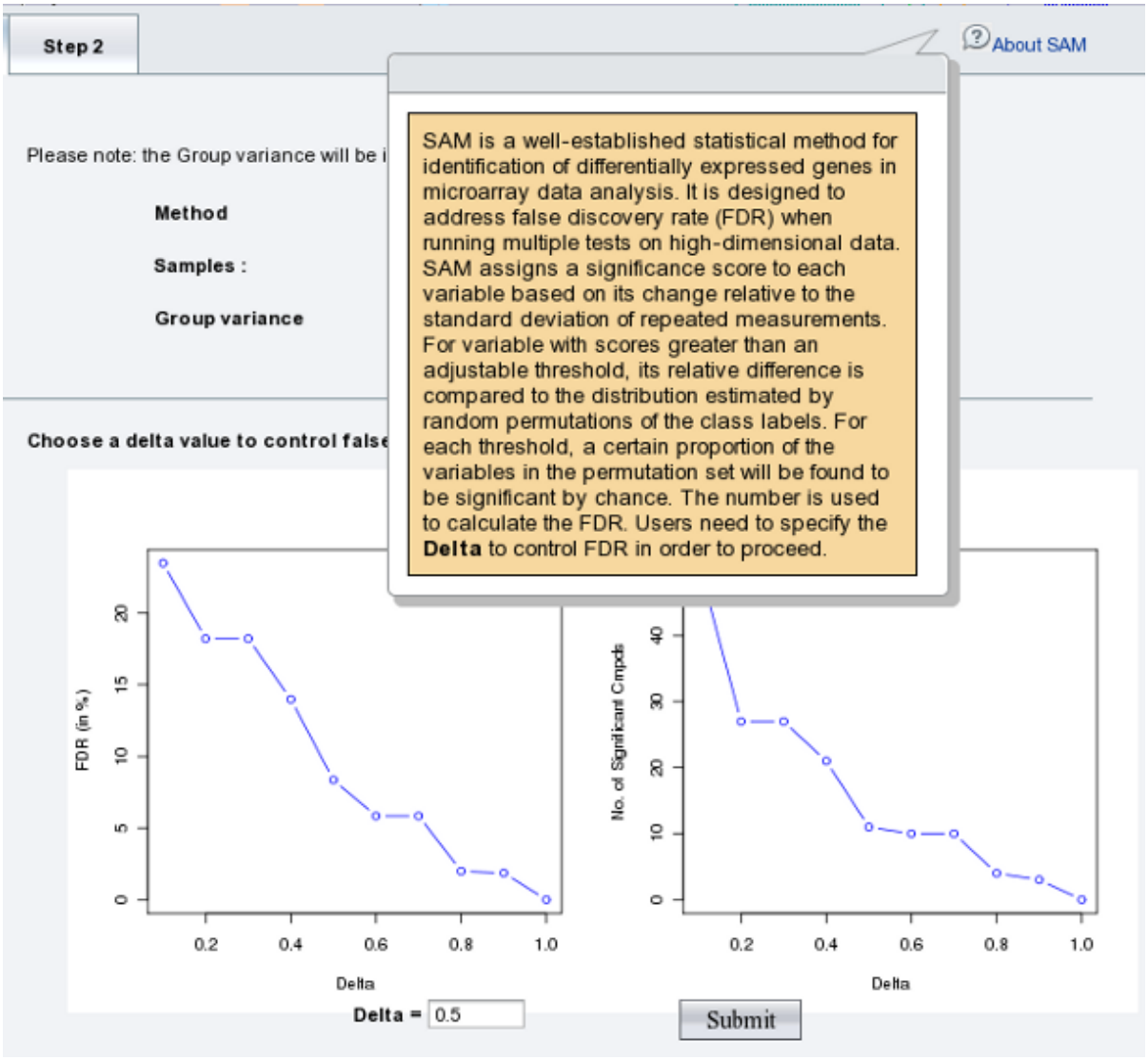


Click the “View details” link to see the data table that was used to produce the graph. Note that the VIP score is not normalized. The VIP scores tend to be 200X larger than the coefficient scores, but the relative ranking of “significant” metabolites is largely the same. VIP is a weighted sum of squares of the PLS weight, which indicates the importance of the variable to the whole model. In many studies VIP values >2.0 are selected and used for further data analysis, but this cut-off depends on the number of variables used. Since the number of variables in this study is less than 100, we can use a more relaxed VIP cutoff of around 1.0.

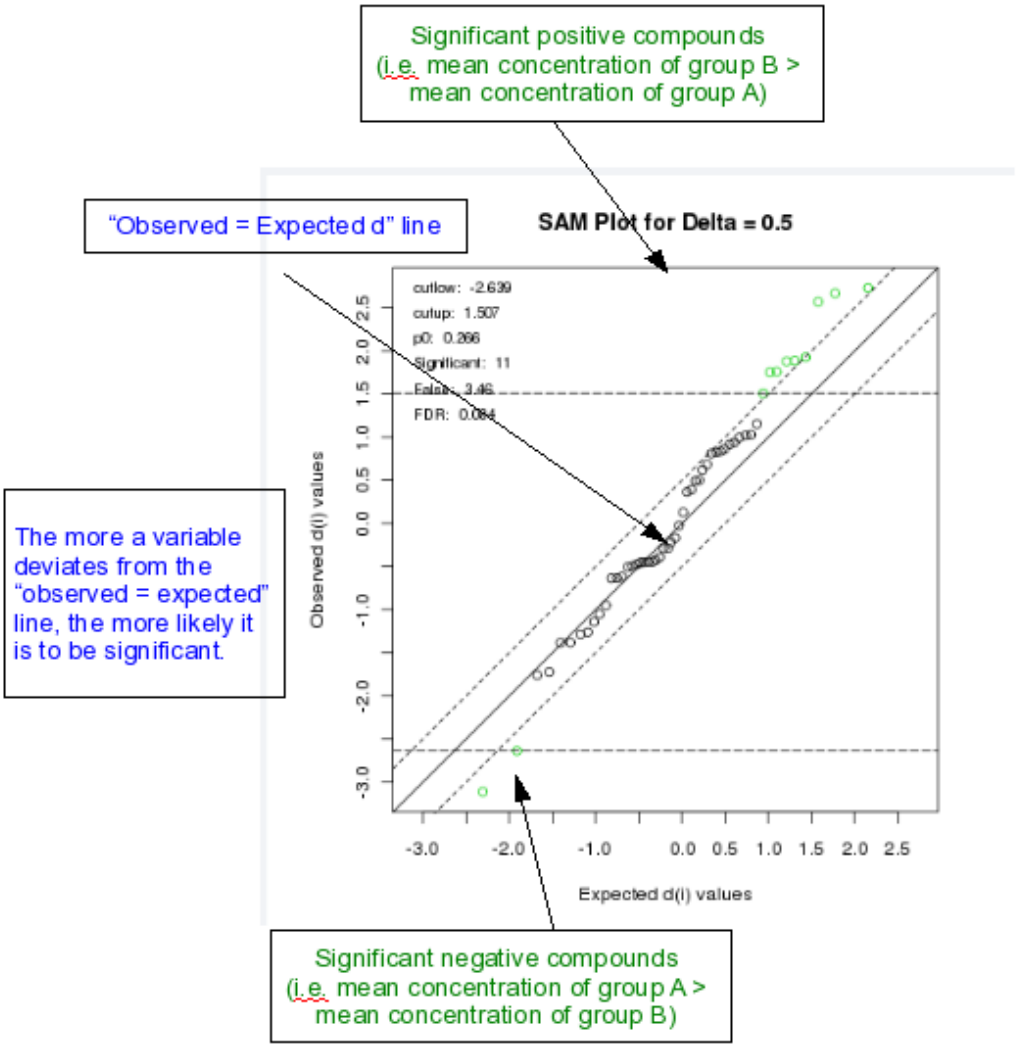
Compounds	VIP score
3-Methylhistidine	2.4964
Glucose	2.1209
Glycolate	2.1064
Pyroglutamate	1.9364
Hypoxanthine	1.9172
Fumarate	1.7711
trans-Aconitate	1.7167
Malonate	1.5208
Acetoacetate	1.4774
O-Acetylcarnitine	1.2359
Choline	1.2335
1-Methylhistidine	1.0894

Compounds	Coef. score
3-Methylhistidine	0.0127
Hypoxanthine	0.0122
Glycolate	0.012
Fumarate	0.0116
Pyroglutamate	0.0108
Glucose	0.0104
trans-Aconitate	0.0096
Malonate	0.0090
Acetoacetate	0.0085
Choline	0.0070
O-Acetylcarnitine	0.0065
Ethanol	0.0058

Step 9. With the completion of the PLS-DA analysis, we can try another approach to select interesting or significant features that distinguish between cachexic and non-cachexic patients. Here we'll attempt to use Significance Analysis of Microarray (SAM). SAM is designed to address False Discovery Rate (FDR) problems when running multiple tests on high-dimensional data. It first assigns a significance score to each variable based on its change relative to the standard deviation of repeated measurements. Then it chooses variables with scores greater than an adjustable threshold and compares their relative difference to the distribution estimated by random permutations of the class labels. For each threshold, a certain proportion of the variables in the permutation set will be found to be significant by chance. This number is used to calculate the FDR. To use SAM analysis, go back and click the "SAM" link on the navigational panel; you will see the following set of "Delta" plots. The Delta plots are a visualization of the table generated by SAM that contains the estimated FDR and the number of identified metabolites for a set of Delta values. Note the pop-up help balloon when you place the mouse over "About SAM". The default Delta value (0.5) has an FDR of 12% and identifies ~10 significant compounds above this threshold, as seen on the left plot. You can increase the Delta to reduce the FDR. The figure below shows that when Delta is above 0.7, FDR approaches to 0 (left panel). However, no significant compound will be identified (right panel). The default 0.5 is a compromise between the FDR and the number of compounds detected. Click 'Submit' in the bottom panel to go to Step 2 to view the result.


Step 2
About SAM
Please note: the Group variance will be i
Method
Samples :
Group variance
SAM is a well-established statistical method for identification of differentially expressed genes in microarray data analysis. It is designed to address false discovery rate (FDR) when running multiple tests on high-dimensional data. SAM assigns a significance score to each variable based on its change relative to the standard deviation of repeated measurements. For variable with scores greater than an adjustable threshold, its relative difference is compared to the distribution estimated by random permutations of the class labels. For each threshold, a certain proportion of the variables in the permutation set will be found to be significant by chance. The number is used to calculate the FDR. Users need to specify the Delta to control FDR in order to proceed.
Choose a delta value to control false
FDR (in %)
0
5
10
15
20
0.2
0.4
0.6
0.8
1.0
Delta
No. of Significant Cmpds
0
10
20
30
40
0.2
0.4
0.6
0.8
1.0
Delta
Delta = 0.5
Submit

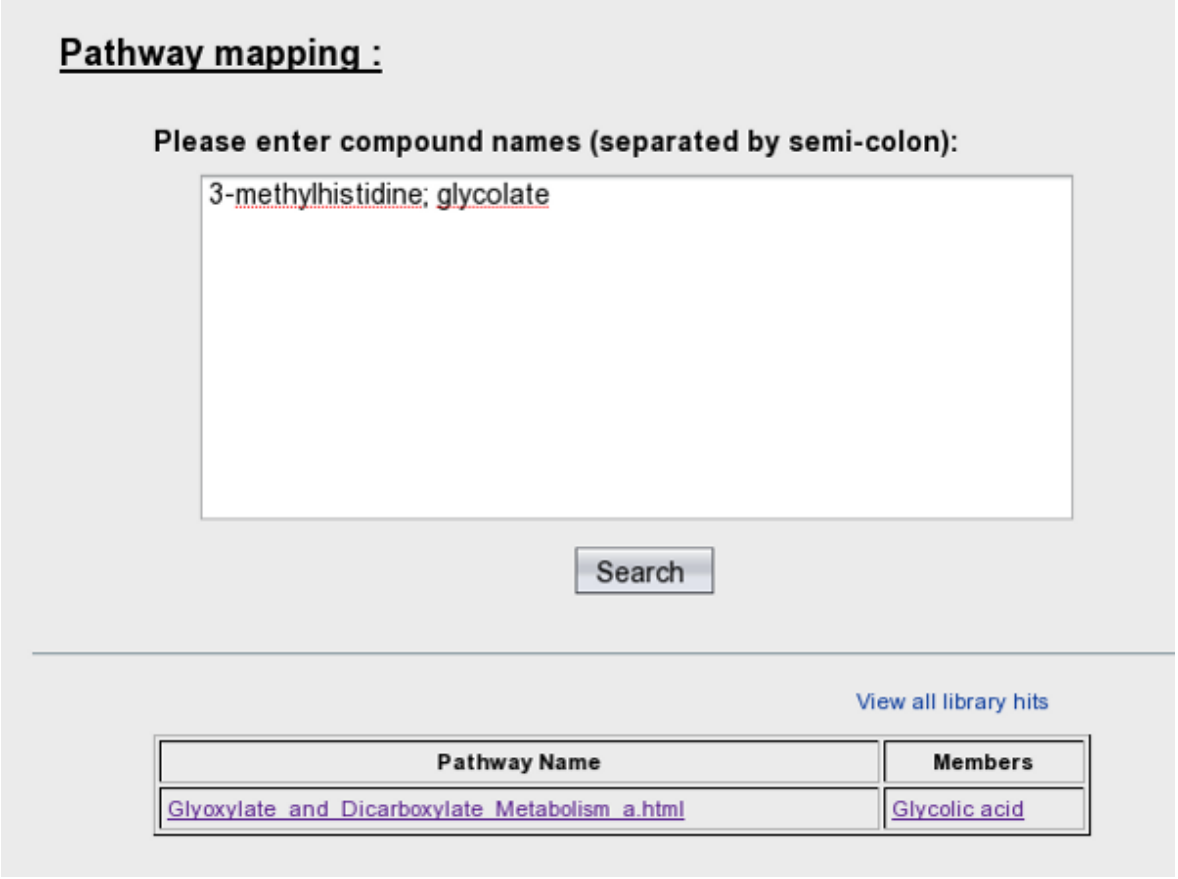
The Step 2 tab shows a typical SAM plot with Delta = 0.5. Click the “View details of the ...” button to see the SAM results table (next page). A SAM plot displays a “positive” metabolite set and a “negative” metabolite set. In the positive metabolite set, higher levels of these metabolites correlate with higher values for the cachexia phenotype. In the negative metabolite set, lower levels of these metabolites correlate with higher values for the cachexia phenotype. A total of 11 compounds were identified above the chosen threshold.



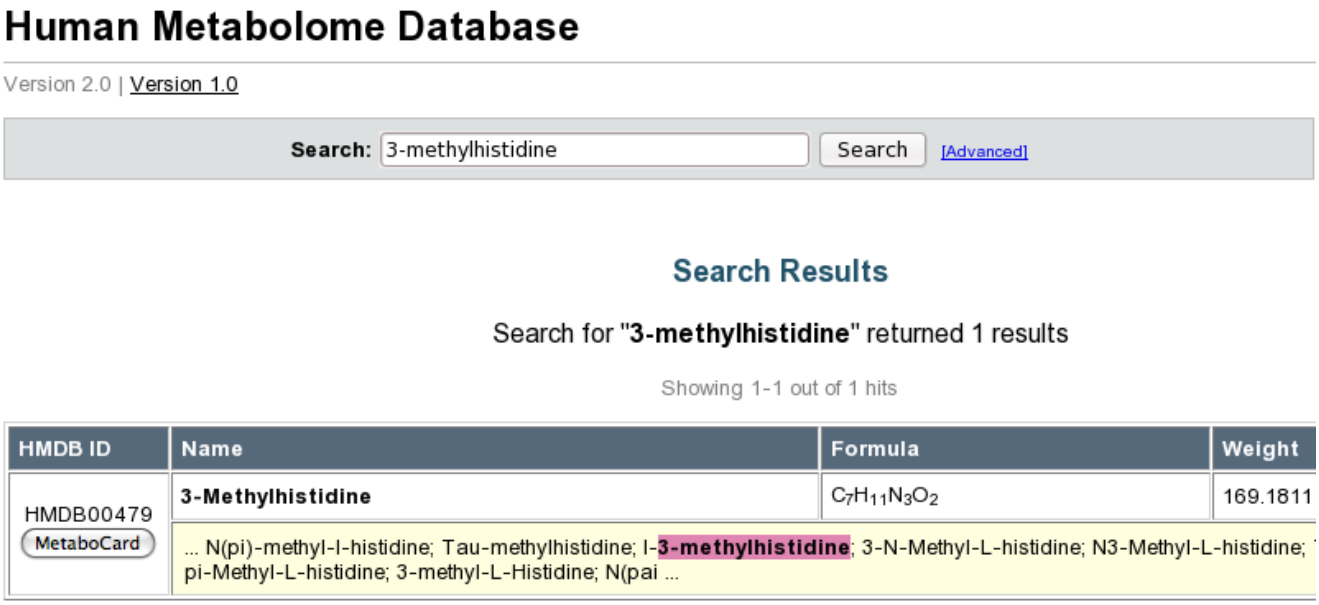
Significant compounds identified by SAM with delta = 0.5. Note that the term “rawp” refers to the raw p-values from regular t-tests.

Compounds	d.value	stdev	rawp	q.value	R.fold
Glycolate	-3.114	0.986	0.0020	0.027	0.119
trans-Aconitate	2.729	0.919	0.0050	0.027	5.69
Fumarate	2.667	0.803	0.0070	0.027	4.414
3-Methylhistidine	-2.639	1.35	0.0080	0.027	0.085
Hypoxanthine	2.569	0.93	0.01	0.028	5.235
Trigonelline	1.925	0.361	0.056	0.102	1.618
Glucose	1.887	1.541	0.061	0.102	7.505
Pyroglutamate	1.879	1.418	0.063	0.102	6.343
Choline	1.756	1.026	0.081	0.102	3.487
Malonate	1.751	1.246	0.082	0.102	4.536
3-Hydroxyisovalerate	1.507	0.618	0.13	0.144	1.908

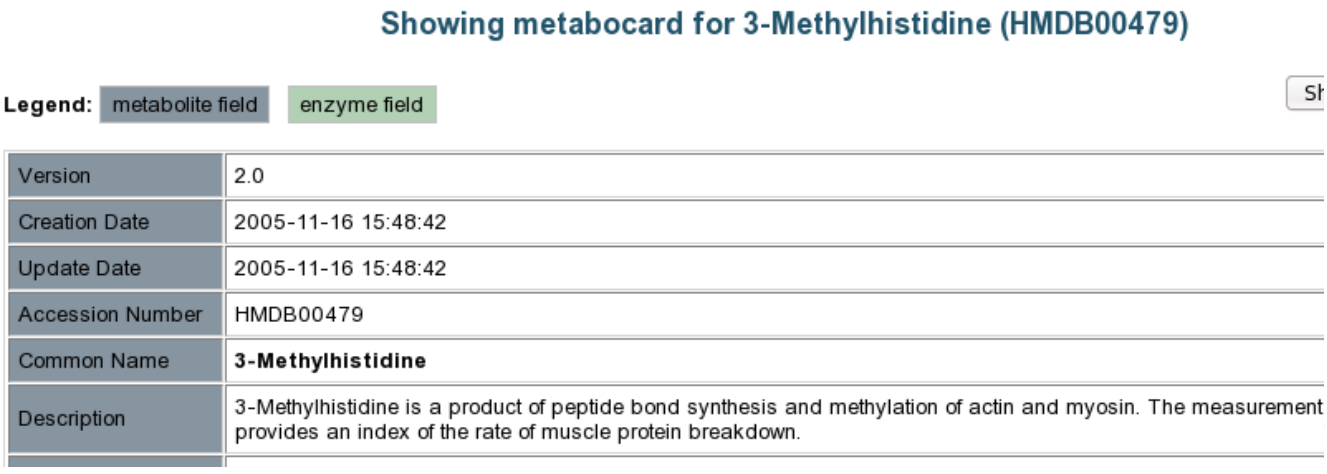
Step 10. Based on the result from the Volcano Plot, PLS-DA and SAM, several compounds are consistently identified as being significant by different approaches. Using 3-methylhistine and glycolate as examples, let us further check which pathways they are involved in. To do so, we can use MetaboAnalyst's data annotation tools. Click the "Pathway mapping" link on the left navigation panel, enter the two compound names separated by a semicolon (i.e. ;) and then click the "Search" button. The result shows only the pathway for Glycolate. No entry was found for 3-Methylhistidine in the pathway library of the Human Metabolome Database (HMDB). By clicking on the relative links in the resulting table, you will access the corresponding pathway as well as detailed information about the metabolite (Metabocard).



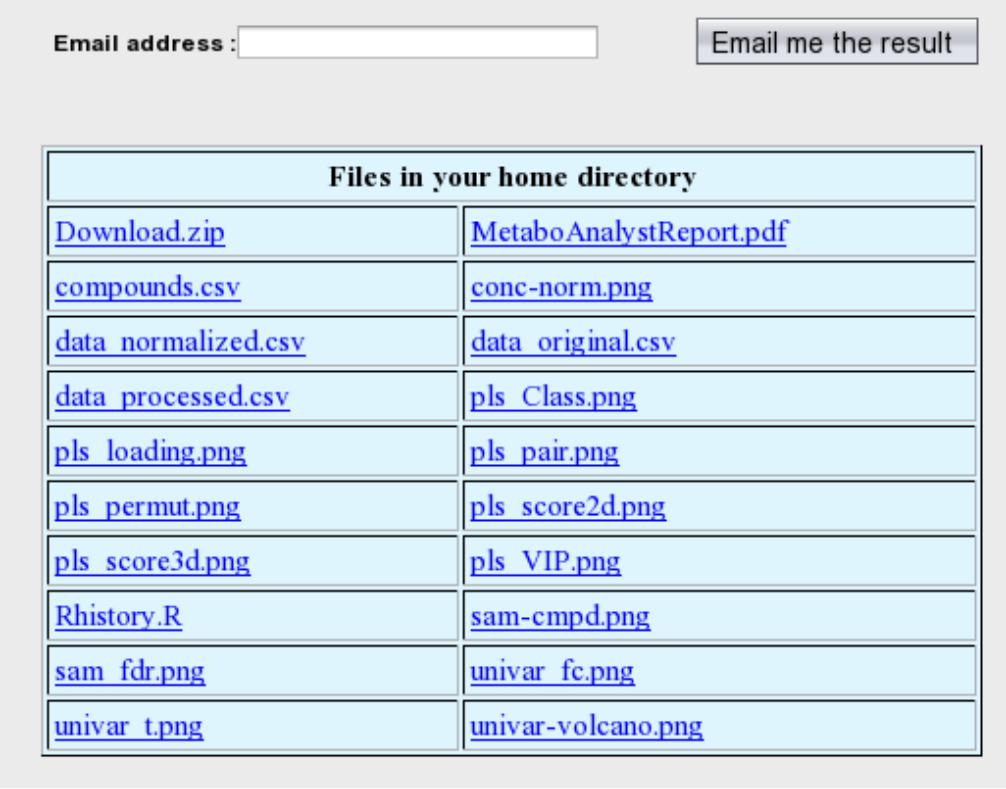
Step 11. Now, we want to find out about the biological function of 3-melthylhistidine. Go to the HMDB (www.hmdb.ca), and enter "3-methylhistdine" and click "Search" button. The result is shown below.



Click the "MetaboCard" on the left panel; the result is shown below. As indicated, this compound can be used as an index of muscle protein breakdown which is relevant to the cachexia patients with significant skeletal muscle loss.



Step 12. Now, assume we have finished the analysis. Click the “Download” link on the navigation panel. A detailed analysis report will be generated (MetaboAnalystReport.pdf) containing introductions and results from every steps you have performed. Now, you can directly click “Download.zip” file to download 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. The data will remain on the server for 72 hours before being automatically deleted.



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