

## Paired Analysis of Time-Series Studies Using Concentration Data

By Nick Psychogios ([psychogi@ualberta.ca](mailto:psychogi@ualberta.ca))

Last update: 4/15/2009

This tutorial shows how to perform paired analysis in time series data using univariate analysis (fold change analysis, t-test and volcano plots) and the high-dimensional feature selection methods of SAM and EBAM. The example used refers to compound concentration data obtained by targeted profiling of  $^1\text{H}$  NMR spectra of urine collected from 7 dairy cows. These cows were fed a diet containing high proportions (50%) of grain which are associated with high occurrence of multiple metabolic disorders. Morning urine samples were collected from each cow in **day 1** prior to and in **day 4** following the dietary intervention and the data set consists of 14  $^1\text{H}$ -NMR spectra. The goal here is to identify features (urinary metabolites) that are significantly affected due to the increased feeding levels of grain and its possible role in the development of metabolic disorders in dairy cows.

**Step 1.** Go to the **Data Formats** page (found by clicking on the Data Formats hyperlink on MetaboAnalyst's home page) right click the example link after the “Paired time-series concentration data” option, save the link to your local disk as “time\_series.csv”.

**Comma Separated Values (.csv) format ([show details](#)), including :**

- Compound concentration data ([download](#))
- Binned NMR/MS spectra data ([download](#))
- Paired time-series concentration data ([download](#))
- Processed peak intensity table ([download](#))

The csv file looks like the table below; empty cells refer to missing values:

1	Sample	c87_day1	143_day1	143_day1	163_day1	225_day1	239_day1	241_day1	87_day4	143_day4	143_day4	163_day4
2	Label	-1	-2	-3	-4	-5	-6	-7	1	2	3	4
3	2-Hydroxyphenylacetate	77.2				77.9		75.7		42.7		
4	3-Hydroxyisovalerate											
5	3-Methyl-2-oxovalerate	65.9	83.6	124.5	297.6	50.2	80.7	103.4		684.2	206.9	171.4
6	4-Hydroxyphenylacetate								171.4			
7	Acetaminophen	148	166.4	128.6	153.5	135.6	71			169.9	148.2	106.5
8	Acetate	510.2	730.4	1004.6	623.8	392.1	374	500.3	376.6	925.8	872.6	339.4
9	Acetoacetate	221.4	217.8	234.2	225.7	160.1	64.7	100		674.8	147	123.1
10	Acetylsalicylate		277	472.7	112.2	475.2	252.2				71.5	46.6
11	Adipate	115.8	136.6	91.1	46	23.8	47.4	32.3		466.1	51	79.8
12	Alanine	69.5	122.9	89.1			36	49.8		277.1	67.3	50.4
13	Allantoin	3478.6	4139.1	4126.9	3888.6	656.6	1766.1	4332.3		951	3948.1	2262.9
14	Betaine	36.2	385.6	165.4	54.7	63.3	54.3		245.7	766.2	108.8	33
15	Butyrate	170.9			64	57.8	62.1	44.2		744.5	106	60.7
16	Citrate								219.4			
17	Creatine	4647.1	1878.6	3052.2	1156.7	1495.6	571	2283.4	2523.3	7065.2	3693.7	437
18	Creatinine	10667.6	4219.8	5625.1	11150.2	4734.3	3991.5	6958.6	10188.2	13618.2	9355.7	5257
19	Dimethylamine	279	272.5	264.5	290.6	142.5	105.8	145.6	226.2	794.3	242.5	131
20	Ethanol	13721.1	9248.7	17324.7	24597.3	3488.8	1418.1	4887.6	19281.8	24398.6	14706.3	8451.5
21	Formate		232.2	191.6			43.3	154.1	173.2	773.3	179.9	79.1
22	Glycine	324.5	404.2	538.6	459.4	120.5	110.4	247.3	287.2	1273.2	594.9	135.5
23	Hippurate	8022.3	5206.5	5016.9	10152.9	3141.2	2153.9	4226	7454	14447.7	6204.4	3525.5

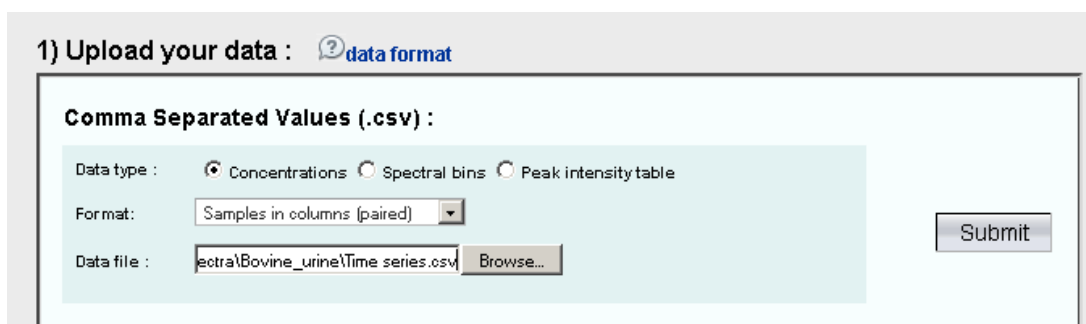
**Note:** The urine samples are reported in columns and the metabolite concentrations in rows and the labeling codes range from -7 to +7

**Hint:** Avoid the use of special characters ( $\mu$ ,  $\beta$ ,  $\ddot{e}$ ,  $\acute{e}$  etc) in your compound names/codes, only letters, numbers, underscore are allowed.

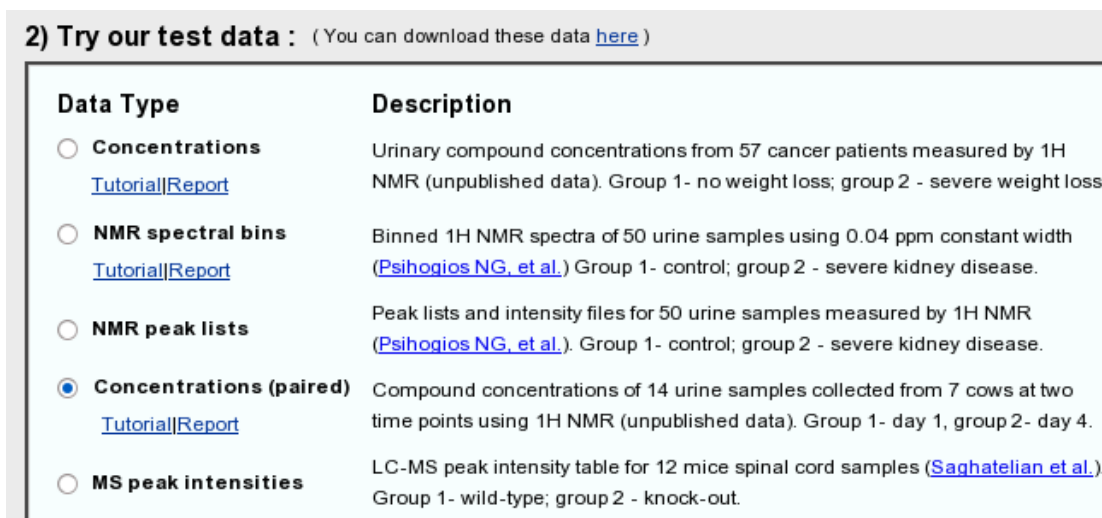
**Step 2.** Go the MetaboAnalyst **Home** page and click “click here to start” to enter the data upload page.



**Step 3.** In the **Upload** page, go to the “Upload your data” panel, select the options as indicated below [Samples in columns (paired)] and click “Submit”.

The screenshot shows the "1) Upload your data" panel. It includes a "data format" help icon. Under "Comma Separated Values (.csv)", there are three radio buttons for "Data type": "Concentrations" (selected), "Spectral bins", and "Peak intensity table". The "Format" dropdown is set to "Samples in columns (paired)". The "Data file" field contains "ectra\Bovine\_urine\Time series.csv" with a "Browse..." button. A "Submit" button is located to the right.

**Note:** Alternatively, you can directly select the #4 option in the “Try our test data” without downloading the example

The screenshot shows the "2) Try our test data" panel. It contains a table with two columns: "Data Type" and "Description". The "Concentrations (paired)" option is selected with a blue radio button. Each row includes a "Tutorial|Report" link.

Data Type	Description
<input type="radio"/> Concentrations <a href="#">Tutorial Report</a>	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 <a href="#">Tutorial Report</a>	Binned 1H NMR spectra of 50 urine samples using 0.04 ppm constant width ( <a href="#">Psihogios NG, et al.</a> ) 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 ( <a href="#">Psihogios NG, et al.</a> ). Group 1- control; group 2 - severe kidney disease.
<input checked="" type="radio"/> Concentrations (paired) <a href="#">Tutorial Report</a>	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 ( <a href="#">Saghatelian et al.</a> ). Group 1- wild-type; group 2 - knock-out.

**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 data groups), determine whether the data is paired or unpaired and detect any missing or zero values. In this case, 200 missing values were detected in the data. By default, MetaboAnalyst will replace these values with half of the minimum positive number detected in the data, assuming the concentrations were below the detection limit. In this case, we use this default method, click “Skip” to go to normalization step.

**Data processing information**

Checking data content ...passed

Two groups were detected based on the sample labels.

Samples are paired.

The labels of paired samples passed sanity check.

A total of 7 pairs were detected.

All data values are numeric.

All data values are non-negative.

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

A total of 200 , ( 34 %) 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

Alternatively, other methods available at MetaboAnalyst could have been used, such as replacement by mean/median, Probabilistic PCA (PPCA), Bayesian PCA (BPCA) or Singular Value Decomposition (SVD) to impute the missing values. In addition, variables with any percentage of missing values can

be automatically removed or the user can manually select which variables to remove. Click on “Missing value imputation” to use any of the above methods on Step 1 or Step 2, as shown below.

**Step 1. Remove features with too many missing values :**

Automatically  variables with >  (%) of missing values.

Manually specify which variables to remove ( [Click here](#) )

**Step 2: Calculate the remaining missing values :**

Exclude variables with missing values

Replace by a small value (half of the minimum positive value in the original data)

Replace by the  of each column.

Impute missing values by

**Step 5.** Now we arrive at the data normalization step. The internal data structure is transformed now to a table with each column representing a urine sample (from a dairy cow) and each row 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 statistic analysis. Row-wise normalization aims to normalize each sample (row) so that it is comparable to the other. In the present example, even though the samples are tabled in columns we still refer to this normalization procedure as row-wise normalization. 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) or 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 the dilution effect (select option 4 - Normalization by a reference feature and choose “creatinine” from the drop-down menu). After deciding to normalize by reference feature 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.

**Row-wise normalization**

None

Normalization by sum

Normalization by a reference sample

Normalization by a reference feature

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

**Column-wise normalization**

None

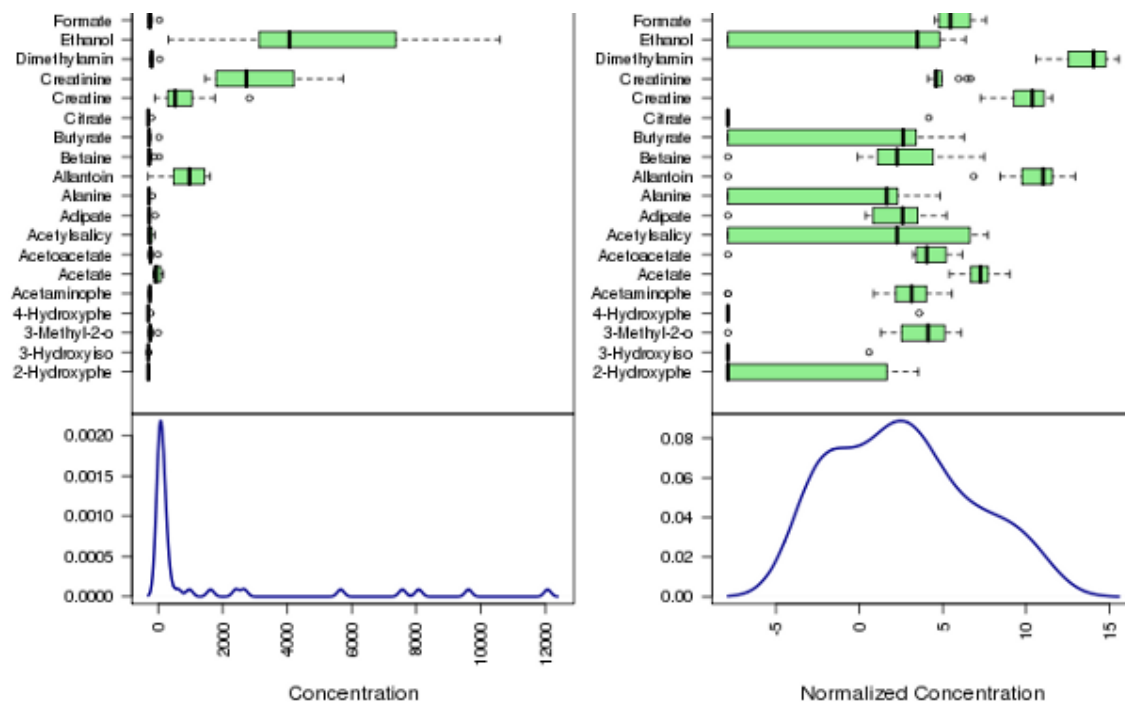
**Log** (log<sub>2</sub> transformation)

**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)

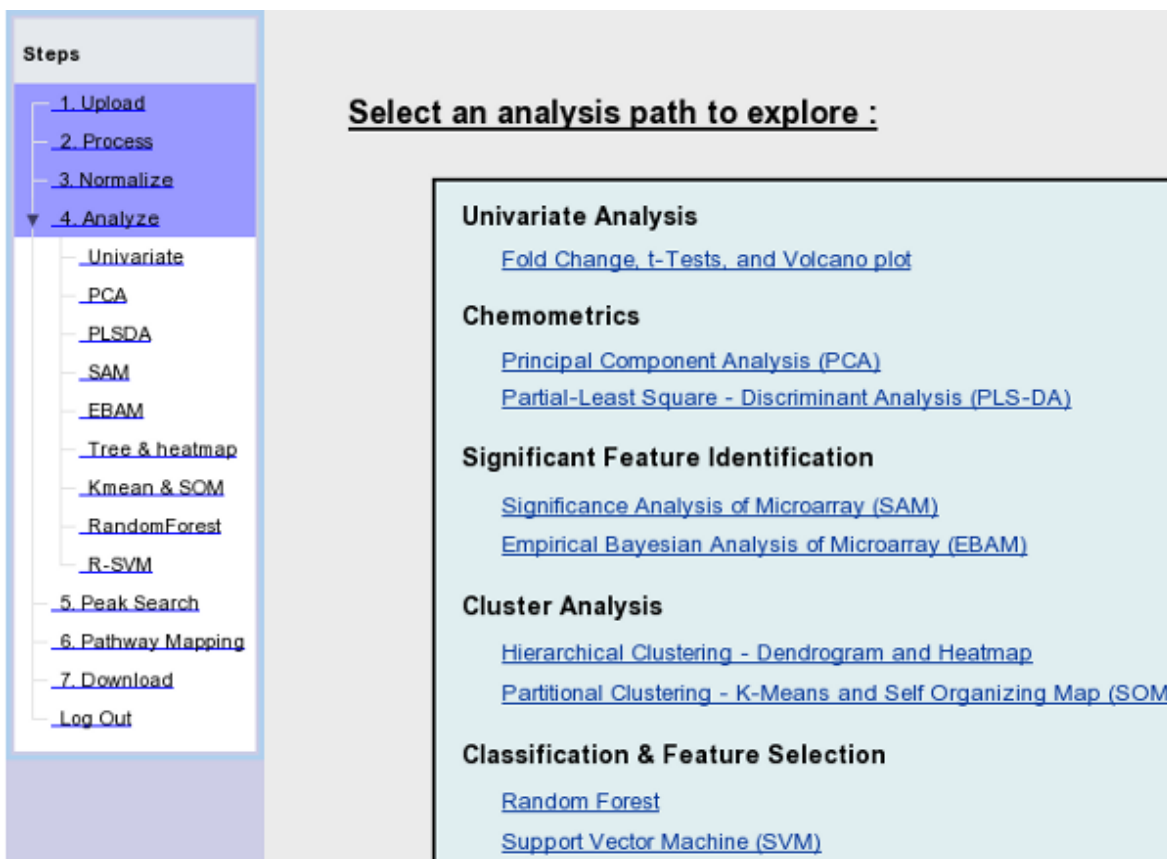
**Range Scaling** (mean-centered and divided by the range of each variable)

Click “Process” to perform the normalization procedures. The result is summarized 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) with the log-transformed curve on the right, the log normalization step along with the reference feature normalization makes the concentration data reasonably “normal”. You can also try other normalization approaches and compare their results. Click “Next” to proceed to the analysis.

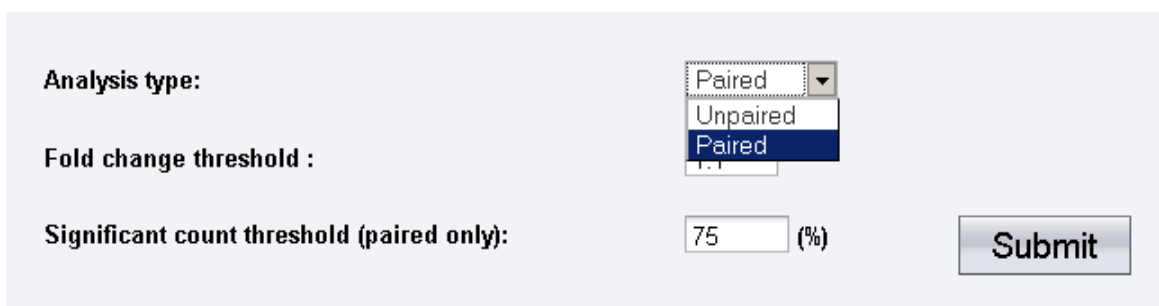




**Step 6.** Now we have finished data processing and normalization and the data are suitable for different statistical analyses. There are many feature selection methods available in MetaboAnalyst. Since our data set contains paired samples, we will use two methods that support paired analysis - Univariate analysis (fold change analysis, t-Tests and Volcano Plot) and Significant Feature Identification (SAM and EBAM). The screen shot below shows the analysis view. Please note the navigation panel on the left, where 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.

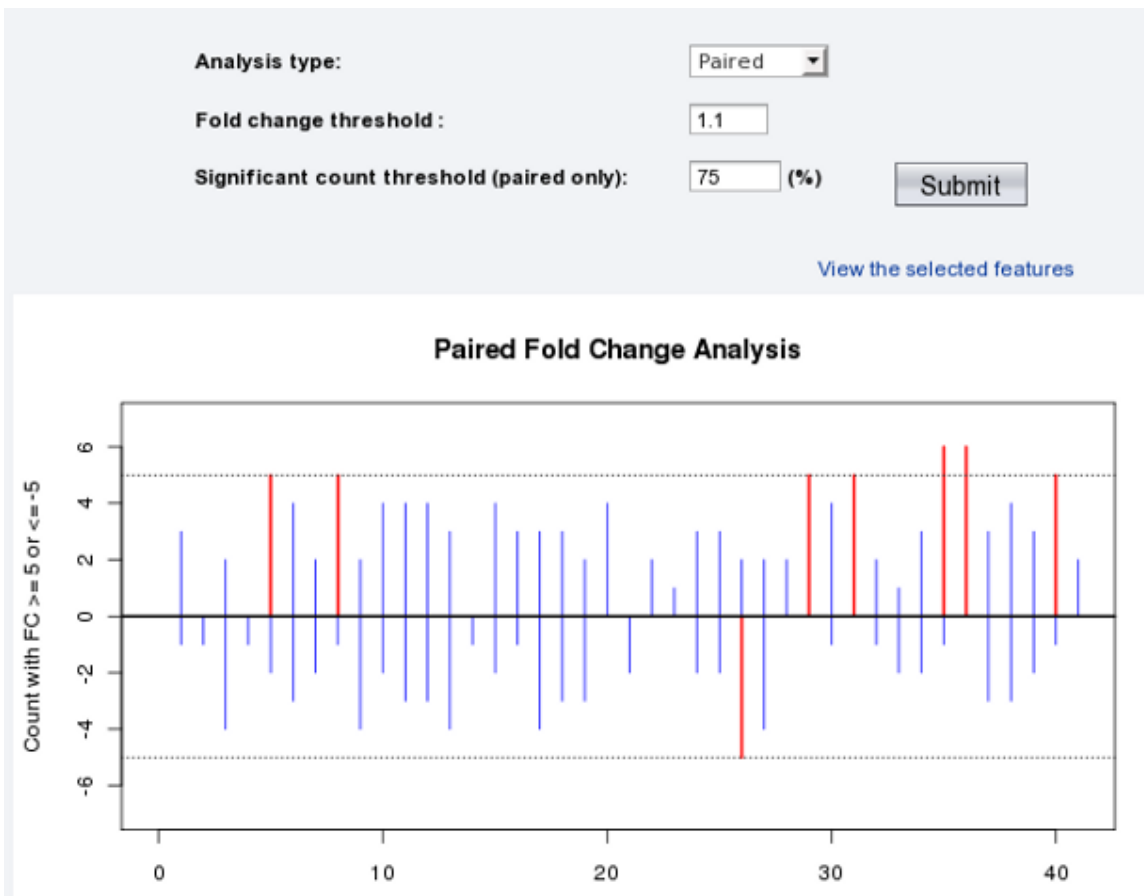


**Step 7.** Generally the simplest kind of analysis that can be performed on this type of metabolomic data is univariate data analysis. Univariate analyses are often first used to obtain an overview or rough ranking of potentially important features before applying more sophisticated data analysis. Univariate analysis examines each variable separately and does not consider 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. From here we will perform a fold change analysis. The purpose of fold change analysis is to compare absolute value change between two group averages. In paired fold change analysis, researchers aim to find some features that change consistently between two groups. The consistency is measured as percentage - (# of pairs that are above/below (not both) above a given threshold) / (total # of paired samples). To start a fold change analysis, click the “Fold Change” tab. Since the current dataset refers to paired samples, change the type of analysis from “Unpaired” to “Paired” at the drop-down menu and then click “Submit”.



The screenshot shows a web form with three input fields and a submit button. The first field is a dropdown menu labeled 'Analysis type:' with 'Paired' selected. The second field is a text input labeled 'Fold change threshold :'. The third field is a text input labeled 'Significant count threshold (paired only):' with the value '75' and a percentage sign '(%)' next to it. A 'Submit' button is located to the right of the third field.

No significant features were detected with the default thresholds. By reducing the “Fold change threshold” from 2 to 1.1 (“Significant pairs threshold” = 75%) and resubmitting the data, the following plot appears, where 8 features are marked with red vertical lines with only 2 crossing the positive 75% threshold. This means that these two features (each contains 7 paired concentrations) have over 75% or >5 pairs with a consistent fold change above the 1.1 threshold. The other 6 features are equal to the threshold and could be characterized as “marginally” significant.



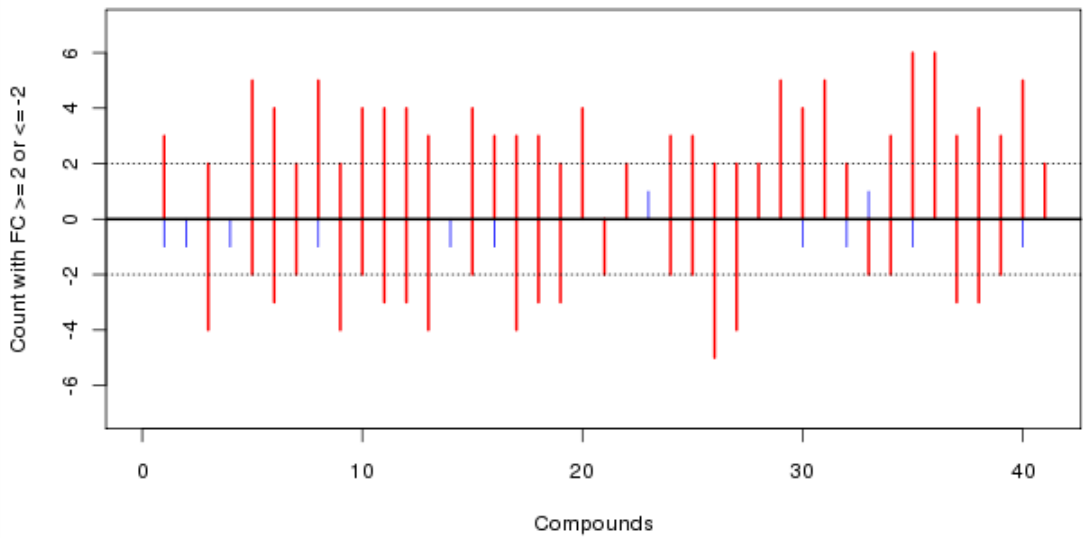
Click the “View the selected features” link to view the names/identities of the 8 features. The table below shows that 2 are significant (tyramine and trimethylamine N-oxide) and 6 are marginally significant (acetylsalicylate, N-nitrosodimethylamine, pyridoxine, p-cresol, acetaminophen and methanol).

Compounds	Count (up)	Count (down)
Tyramine	6.0	0.0
N-Nitrosodimethylamine	5.0	0.0
Pyridoxine	5.0	0.0
Trimethylamine N-oxide	6.0	1.0
Acetylsalicylate	5.0	1.0
p-Cresol	5.0	1.0
Acetaminophen	5.0	2.0
Methanol	2.0	5.0

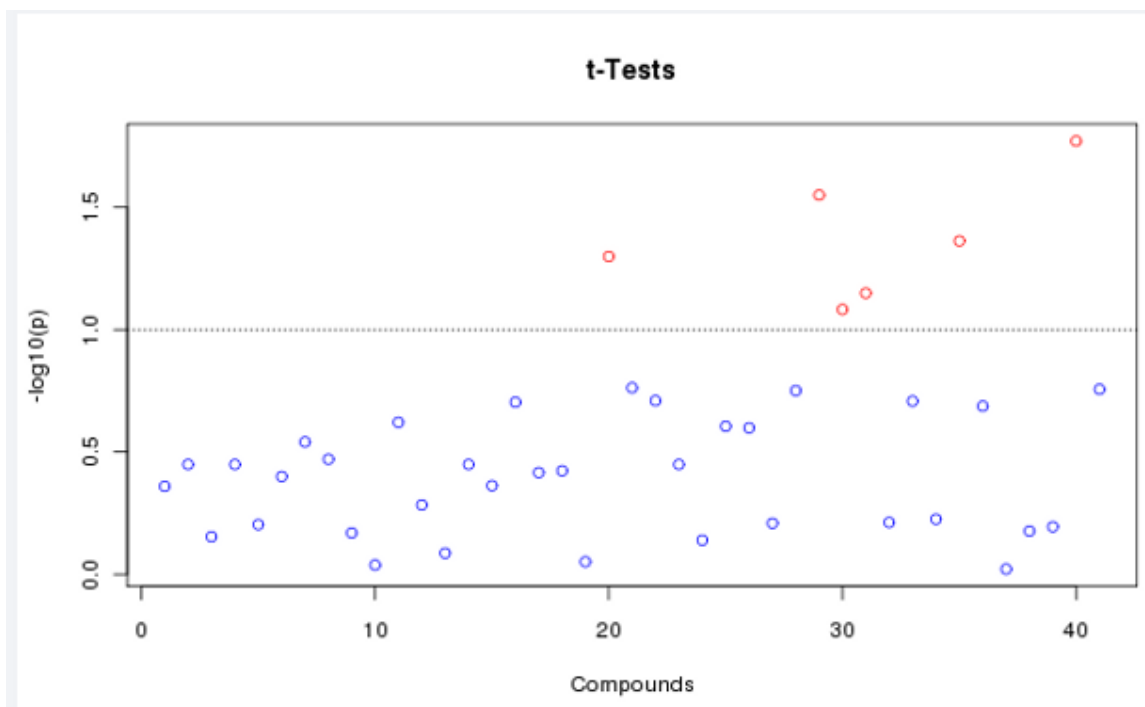
**Note:** Be cautious when you reduce the “Significant pairs threshold” too much below 50%. Because you can potentially get results that some features are BOTH up-regulated and down-regulated, like the one shown below. We only want to see features that are consistently up- or consistently down-regulated

**Analysis type:**    
**Fold change threshold :**    
**Significant count threshold (paired only):**  (%)    
[View the selected features](#)

**Paired Fold Change Analysis**



**Step 8.** We finished the fold change analysis and we want to perform t-tests analysis. In a t-tests analysis one attempts to determine whether the means of two groups are distinct. Once a t-value is determined, a p-value can be calculated that can be used to determine whether this distinction is statistically significant. Both paired (same individuals measured before and after an intervention) and unpaired (individuals randomly assigned to two groups) analyses are supported. Since we are performing univariate analysis, click on the “t-Tests” tab. In the “t-Tests” page change the type of analysis from “unpaired” to “paired” and click on “Submit”. Since we are performing analysis of paired samples, selecting “unequal” or “equal” in the drop-down menu of “Group variance” will not affect the results. The plot shows that 6 features are significant (marked with red circles) for a  $p$  level  $< 0.1$ .



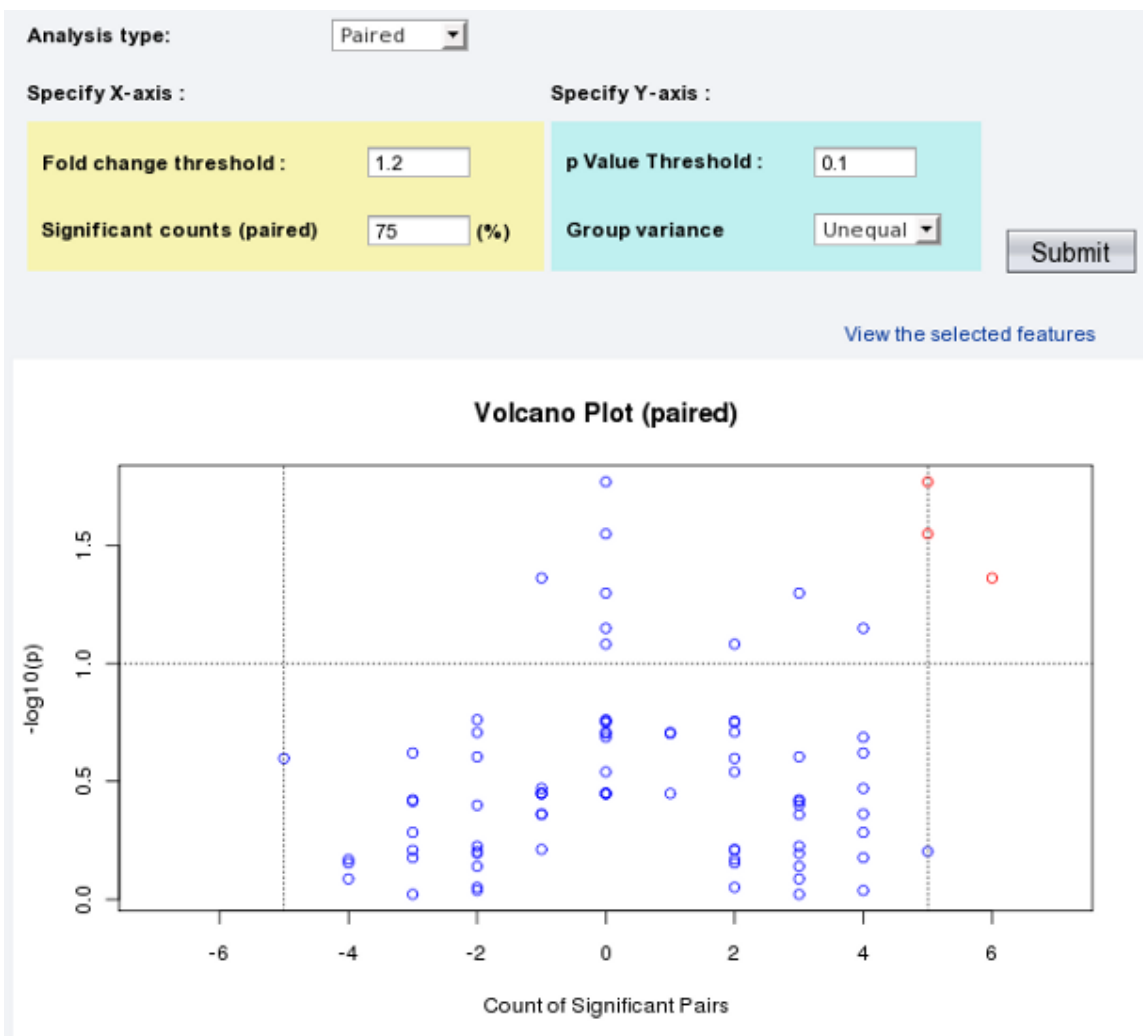
If we click on “View the selected features” you will see the table shown below that lists the 6 significant features with the corresponding  $p$  values in a ranking order starting from the most significant ones (lower  $p$  value). Go back to the t-tests webpage.

Compounds	p.value	$-\log_{10}(p)$
p-Cresol	0.01695	1.77079
N-Nitrosodimethylamine	0.02819	1.54989
Trimethylamine N-oxide	0.04339	1.36257
Hippurate	0.05022	1.29913
Pyridoxine	0.07085	1.14968
Phenylacetylglutamine	0.08269	1.08254

If you change to a lower  $p$  value (e.g.  $p < 0.05$ ), which is often used in t-tests analysis, and click on “Submit” then only 3 features are highlighted as significant. Click on “View the selected features (p-Cresol, N-nitrosodimethylamine, trimethylamine-N-oxide), which not surprisingly are the 3 top features in the  $p < 0.1$  t-tests analysis previously mentioned.

Compounds	p.value	$-\log_{10}(p)$
p-Cresol	0.01695	1.77079
N-Nitrosodimethylamine	0.02819	1.54989
Trimethylamine N-oxide	0.04339	1.36257

**Step 9.** We finished the t-Tests analysis and we want to perform Volcano plots 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, change the analysis type to “paired”, adjust the fold change threshold from 2 to 1.2, with 0.1 as p value threshold and click on “Submit”. As seen by the figure below, three features are detected. These significant features are colored in red.





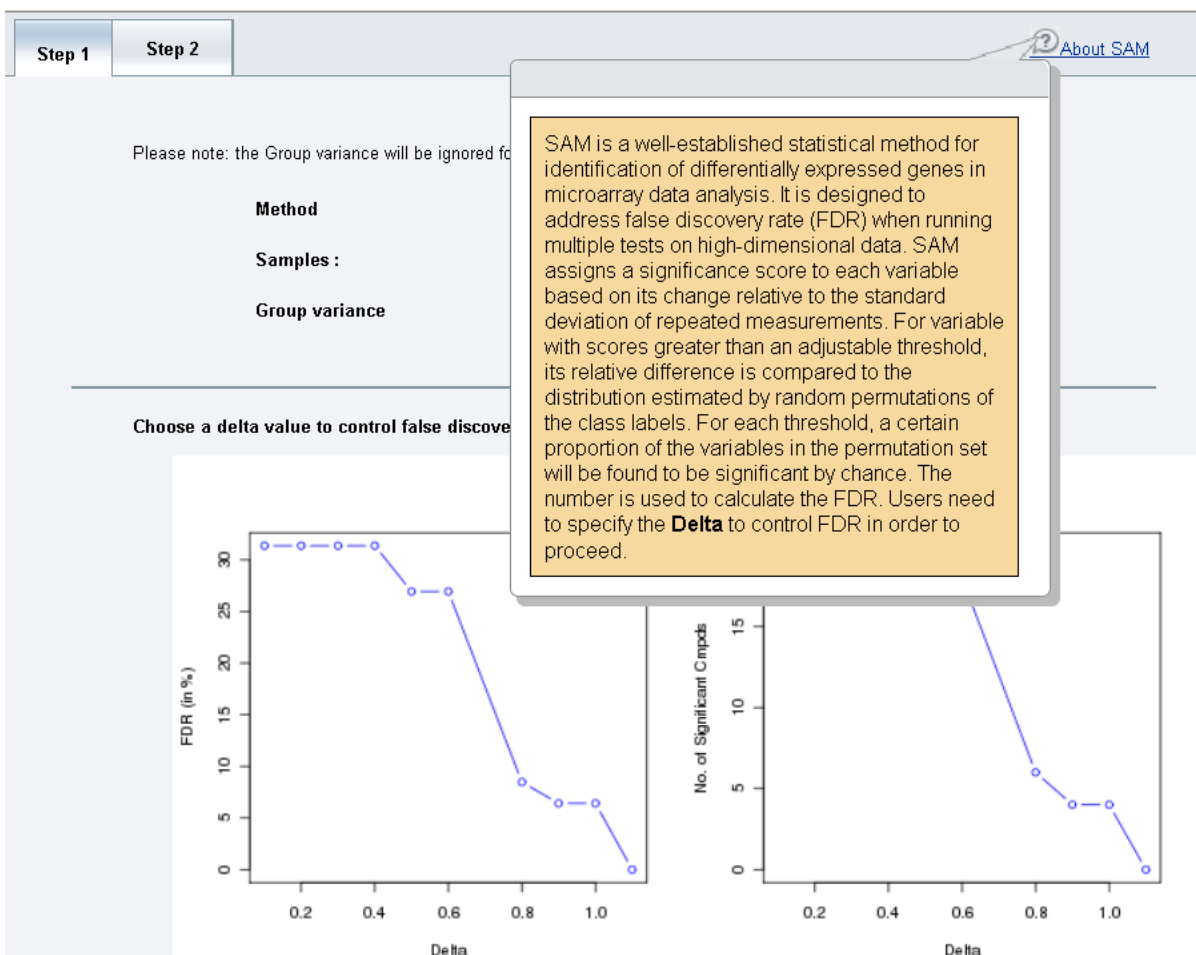
Click the “View the selected features” to view the names/identities of these features. The table below shows these three features (p-cresol, trimethylamine N-oxide, N-nitrosodimethylamine).

Compounds	Counts (up)	Counts (down)	p.value	-log <sub>10</sub> (p)
p-Cresol	5.0	0.0	0.017	1.771
N-Nitrosodimethylamine	5.0	0.0	0.028	1.55
Trimethylamine N-oxide	6.0	1.0	0.043	1.363

**Step 10.** With the univariate analysis complete, we can try another approach to select interesting or significant features that distinguish between control and corn-fed dairy cows. Here we will 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. The number is used to calculate the FDR.

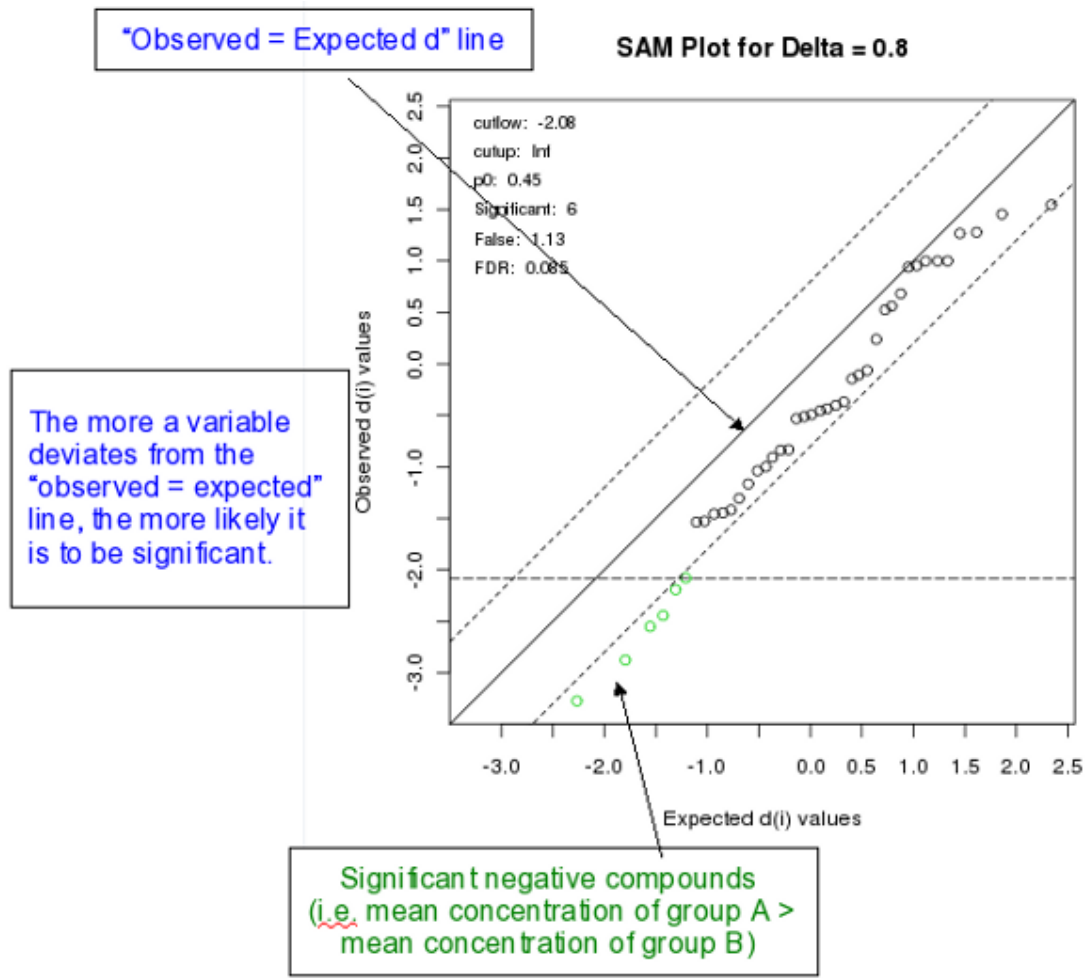
To use SAM analysis, go back and click the “SAM” link on the navigational panel and you will see the following set of “Delta” plots. The Delta plots are a visualization of the table generated by SAM that contains the estimated FDR and the number of identified metabolites for a set of Delta values. Note the pop-up help balloon when you place the mouse over “About SAM”. Change the “Samples” type from

“Unpaired” to “Paired” and click “Submit” in the upper panel. As seen in the picture below, you see 2 graphs: “FDR (%)” vs. “Delta” (left) and “Number of significant compounds” vs. Delta (right). The default Delta value of 0.8 has an FDR of ~ 8 % and identifies ~6 compounds above this threshold. Click 'Submit' in the bottom panel to go to Step 2 to view the result.



The Step 2 tab shows a typical SAM plot with Delta = 0.8. Click the “View details of the selected features” button to see the SAM results table. 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 the grain-fed dairy cows (none in this example). In the negative metabolite set, lower levels of

these metabolites correlate with the grain-fed dairy cows. Six compounds were identified above the chosen threshold. The term “rawp” refers to the raw p-values from regular t-tests.



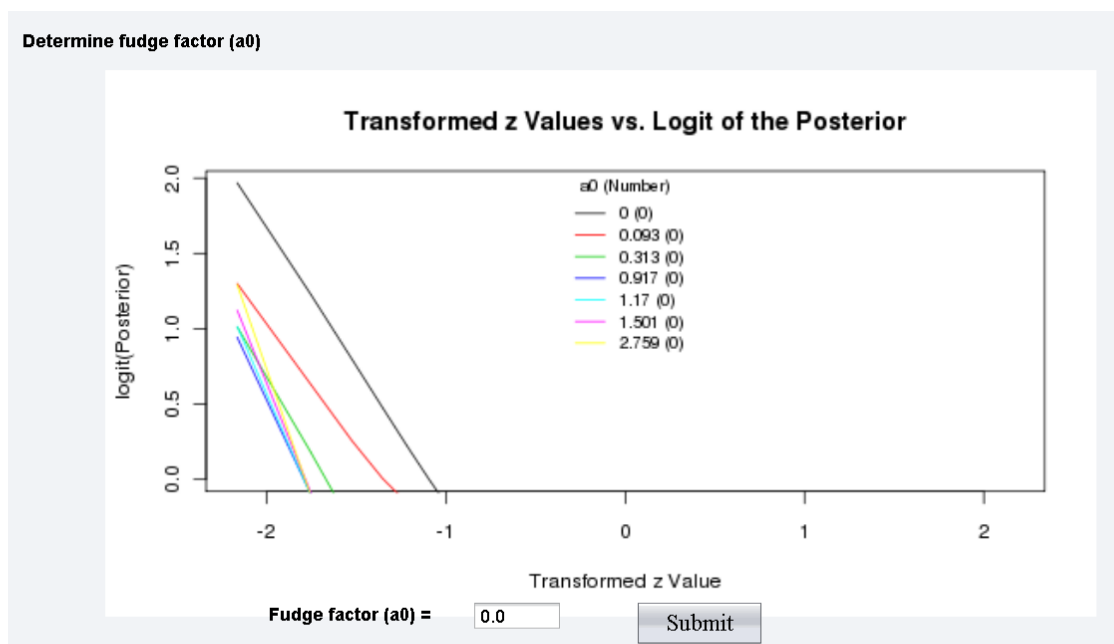
Click on the “View details of the selected features” link to get the table that lists these six features.

Compounds	d.value	stdev	rawp	q.value	R.fold
p-Cresol	-3.274	0.232	0.0080	0.133	NaN
N-Nitrosodimethylamine	-2.876	1.459	0.014	0.133	NaN
Trimethylamine N-oxide	-2.552	0.197	0.027	0.137	NaN
Hippurate	-2.444	0.113	0.03	0.137	NaN
Pyridoxine	-2.192	1.81	0.045	0.166	NaN
Phenylacetylglycine	-2.08	0.093	0.06	0.183	NaN

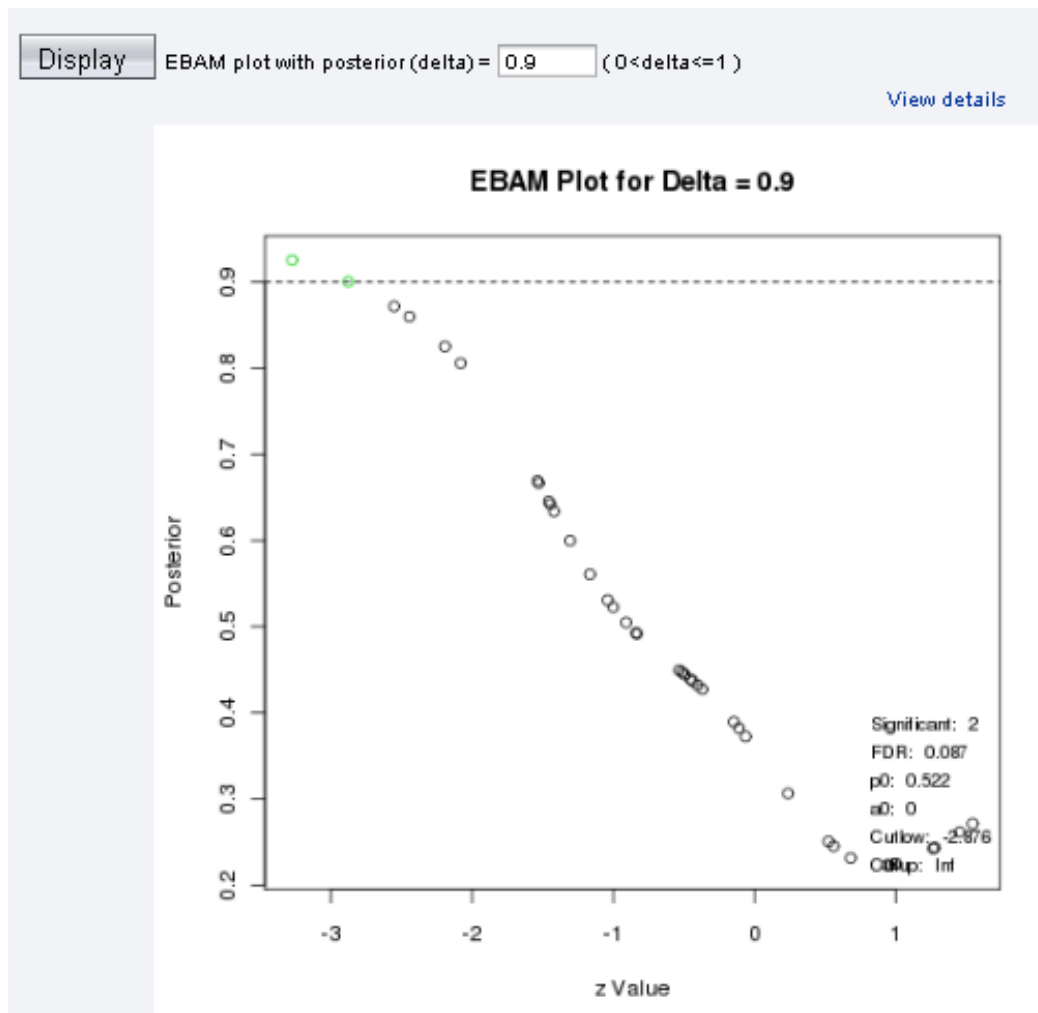
**Step 11.** For feature selection we can also perform Empirical Bayesian Analysis of Microarray (EBAM). EBAM uses a two-group mixture model for both null and differentially expressed metabolites. The prior probability and density parameters are estimated from the data. The EBAM algorithm is essentially a variation of the SAM method. The only difference is that EBAM uses a modified t-statistic in calculating the score. Click the “EBAM” link on the navigation panel to go to the “Step 1” page. Change the type of “Samples” from the default “Unpaired” to “Paired” and click on “Submit” in the upper panel.

**Note:** For the current example, the option for “Group variance” will be ignored since paired samples are being analyzed.

The following plot of the “Transformed z Values versus the Logit of the Posterior” for the default fudge factor ( $a_0 = 0$ ) appears, which shows the distribution of 7 fudge factors ranging from 0 to 2.759. The fudge factor helps to find the largest number of significant features.



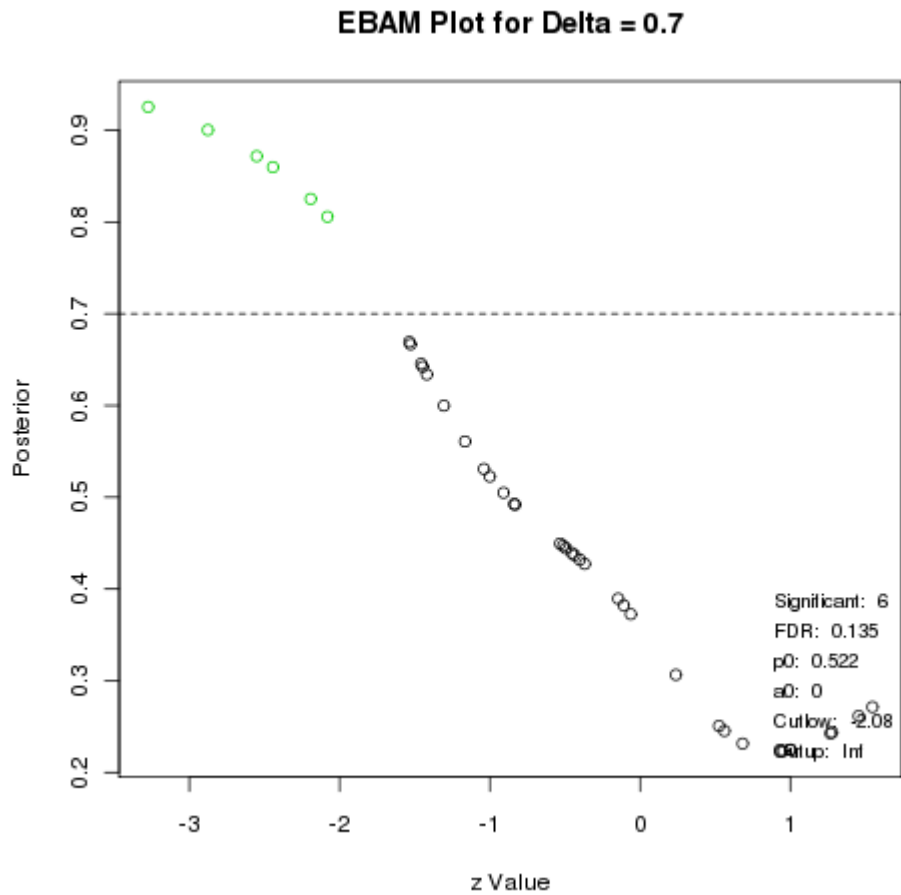
Keep the default value 0 for the fudge factor ( $a_0$ ) and click on “Submit”. This will lead to “Step 2” where the following “EBAM Plot” for the default value of  $\Delta = 0.9$  is shown.



The current EBAM plot was calculated for  $a_0 = 0$  and  $\delta = 0.9$ . The FDR is 0.087 and 2 significant features have been detected. You can view these 2 features by clicking on the “View details” button.

Compounds	z.value	posterior	local.fdr
p-Cresol	-3.274	0.925	0.075
N-Nitrosodimethylamine	-2.876	0.9	0.1

Decrease the posterior (delta) level to 0.7 (above the graph) and click on “Display”. In the new plot, 6 significant features have been detected, which can be viewed by clicking the “View details” button.



Compounds	z.value	posterior	local.fdr
p-Cresol	-3.274	0.925	0.075
N-Nitrosodimethylamine	-2.876	0.9	0.1
Trimethylamine N-oxide	-2.552	0.872	0.128
Hippurate	-2.444	0.86	0.14
Pyridoxine	-2.192	0.825	0.175
Phenylacetylglycine	-2.08	0.806	0.194

**Step 12.** Based on the result from Fold Change Analysis, t-tests, Volcano Plot and SAM, several compounds appear to be significant. Let us use p-Cresol, trimethylamine-*N*-oxide and tyramine as an example and further check which pathways these are involved into. To do so, we can use MetaboAnalyst's data annotation tools. Click the "Pathway mapping" link on the left navigation panel and enter the three compound names separated by ";", then click the "Search" button. The result shows only the pathways for Tyramine. No entry found for p-Cresol and triemthylamine-*N*-oxide 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 tyramine (Metabocard)..

**Pathway mapping :**

Please enter compound names (separated by semi-colon):

p-cresol; trimethylamine-N-oxide; tyramine

Search

[View all library hits](#)

Pathway Name	Members
<a href="#">Tyrosine Metabolism a.html</a>	Tyramine
<a href="#">Alkaloid Biosynthesis a.html</a>	Tyramine



**Step 13.** Now, we want to find out the biological function of trimethylamine-*N*-oxide. Go to HMDB ([www.hmdb.ca](http://www.hmdb.ca)), enter “trimethylamine-*N*-oxide” and click “Search” button. The result is shown below.

## Human Metabolome Database

Version 2.0 | [Version 1.0](#)Search:   [\[Advanced\]](#)

### Search Results

Search for "trimethylamine-N-oxide" returned 2 results

Showing 1-2 out of 2 hits

HMDB ID	Name	Formula	Weight
HMDB00925 <a href="#">MetaboCard</a>	<b>Trimethylamine oxide</b>	C <sub>3</sub> H <sub>9</sub> NO	75.109703
	... N-oxide; TMAO; TMA-oxide; <b>Trimethylamine N-oxide</b> ; <b>Trimethylamine-n-oxide</b> ; Trimethylamine Oxide; Triox Osmolyte Aliphatic Amines Mammalian Metabolite Organic Endogenous N-oxide Tertiary ...		

Click the “MetaboCard” on the left panel, the result is shown below:

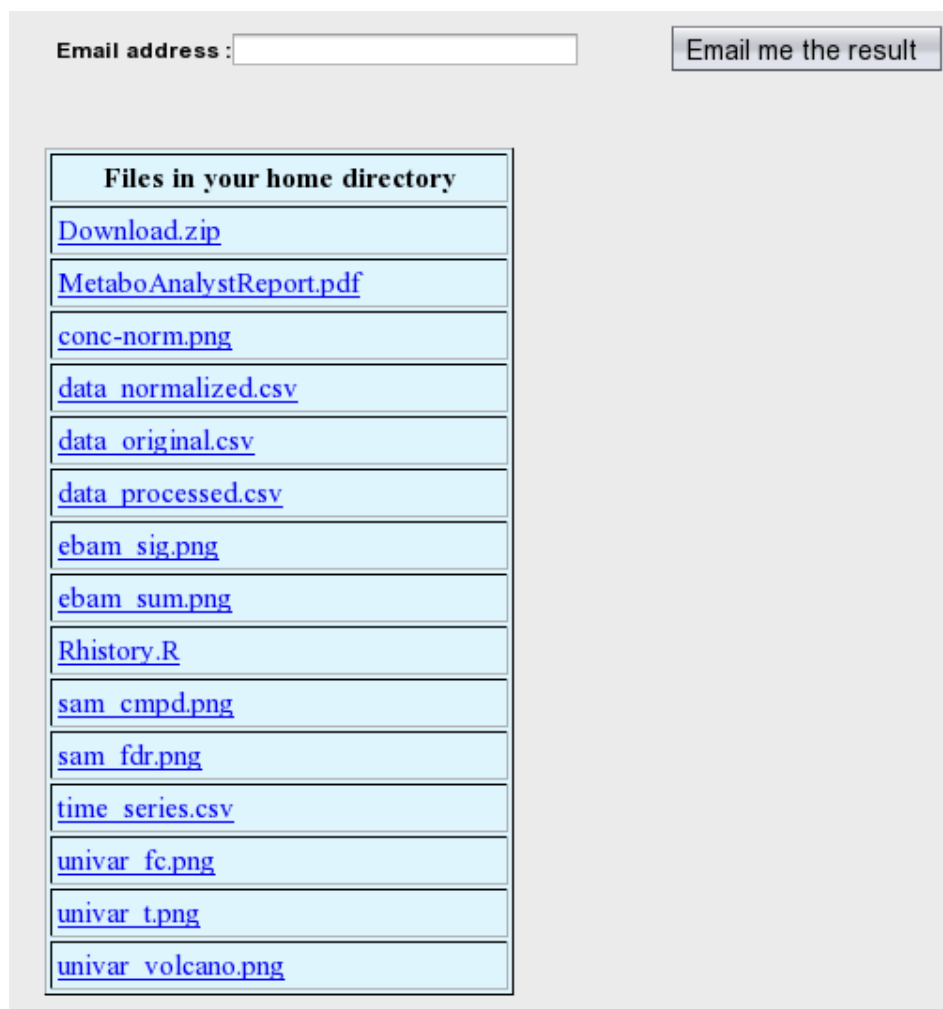
### Showing metabocard for Trimethylamine oxide (HMDB00925)

Legend:  metabolite field  enzyme field

Version	2.0
Creation Date	2005-11-16 15:48:42
Update Date	2005-11-16 15:48:42
Accession Number	HMDB00925
Common Name	<b>Trimethylamine oxide</b>
Description	Trimethylamine oxide (TMAO) is an oxidation product of trimethylamine and a common metabolite in animals and humans. TMAO decomposes to trimethylamine (TMA), which is the main odorant that is characteristic of degrading seafood. TMAO is an osmolyte that the body will use to counter-act the effects of increased concentrations of urea (due to kidney failure) and can be used as a biomarker for kidney problems. Fish odor syndrome or trimethylaminuria is a defect in the production of the enzyme flavin containing monooxygenase 3 (FMO3) causing incomplete breakdown of trimethylamine from choline-containing food into trimethylamine oxide. Trimethylamine then builds up and is released in the person's sweat, urine, and breath, giving off a strong fishy odor.

As indicated, this compound can be used as a biomarker for kidney problems, which might be relevant to the metabolic diseases that cows develop when are fed diets with high proportions of grain.

**Step 14.** 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. The data will remain on the server for 72 hours before being automatically deleted.



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