

Identification of Robust Biomarkers using the Biomarker Meta-Analysis Module

By: Jasmine Chong, Jeff Xia

Date: 14/02/2018

The aim of this tutorial is to demonstrate how the Biomarker Meta-Analysis module of MetaboAnalyst can be used to identify robust and novel biomarkers of disease/other conditions through the integration of individual metabolomics studies. The example data used in this tutorial comes from a subset of GCTOFMS data deposited in the Metabolomics Workbench from Fiehn et al. 2016 (doi: 10.21228/M85P57) which has been slightly modified to demonstrate the full capabilities of this module. The primary objective of their study is to identify biomarkers of adenocarcinoma (lung cancer) in blood.

Introduction to Biomarker Meta-Analysis

Biomarker identification is an important area of research in metabolomics, and their validation is challenging due to inconsistencies in identified biomarkers amongst similar experiments. With the wide application of metabolomics and the establishment of several metabolomics data repositories, there is an ever-increasing interest to perform secondary analysis of published metabolomics datasets collected under similar study conditions in similar populations, a practice often referred to as “meta-analysis”. When executed properly, meta-analysis can leverage the collective power of multiple studies to overcome study biases and small effect sizes to improve the precision in identifying true patterns and robust biomarkers within data. The primary goal of the Biomarker Meta-Analysis module is to provide a user-friendly tool for the integration of individual metabolomics studies to identify robust biomarkers. The main steps for Biomarker Meta-Analysis are as following and will be described in further detail:

- i. Upload individual datasets. Prior to uploading the data, clean the datasets to ensure consistency amongst feature names (compound IDs, spectral bins, or peaks) as well as consistency in the class labels across all included studies.
- ii. The module will perform differential enrichment analysis for each individual study to compute summary level-statistics for each metabolite feature (e.g. p-value).
- iii. The summary level-statistical results from all studies are combined, and meta-analysis is performed using one of several statistical options: combining of p-values, vote counting, or direct merging for very similar datasets.
- iv. The results can be visualized as a Venn diagram to view all possible combinations of shared features between the datasets.

Data Upload Preparation

Before uploading your data to the module, please make sure that the names of your features (compound names, spectral bins, peaks) are consistent between the individual studies. At least 25% of the features must match between the studies. Also make sure that the group labels are also

consistent between the studies, i.e. Cancer and Healthy. Finally, all uploaded sample identifiers must be unique. An example of what your dataset should look like is below:

	A	B	C	D	E	F	G	H	I
1	Samples	140225dlvsa44_1	140226dlvsa30_1	140226dlvsa36_1	140227dlvsa36_1	140227dlvsa47_1	140228dlvsa08_1	140228dlvsa17_1	140228dlvsa30_1
2	Class	Adenocarcinoma	Adenocarcinoma	Adenocarcinoma	Adenocarcinoma	Adenocarcinoma	Cancer	Cancer	Cancer
3	1_5-anhydroglucitol	6799	17473	38267	12027	19565	20174	8153	6878
4	1-monoolein	165	411	525	726	386	339	197	3225
5	1-monopalmitin	107	100	195	122	108	132	51	82
6	1-monostearin	67	125	209	200	108	113	71	162
7	2_3_5-trihydroxypyrazine NIST	34	54	45	107	55	46	41	61
8	2_3-dihydroxybutanoic acid NIST	74	146	183	152	202	211	101	143
9	2-deoxyerythritol	334	765	474	495	581	353	469	306
10	2-deoxytetronic acid NIST	762	1830	1356	1128	1803	772	758	1173
11	2-hydroxybutanoic acid	7786	15277	7794	9810	12725	18446	10663	13056
12	2-hydroxyglutaric acid	233	1274	1021	1533	2124	1191	197	926
13	2-hydroxyhippuric acid	84	82	107	109	103	78	322	74
14	2-hydroxyvaleric acid	1392	1047	876	1170	1658	453	498	1073
15	2-ketoisocaproic acid	2094	1229	1021	2840	1946	3155	2421	2535
16	3-aminoisobutyric acid	891	473	362	387	621	276	278	779
17	3-hydroxybutanoic acid	5015	2336	1509	4757	6757	13223	4428	2562
18	3-phosphoglycerate	131	61	52	980	67	46	60	93
19	4-hydroxyproline	1389	5112	4323	4082	3832	2080	1458	3146
20	5-hydroxyornithine NIST	197	142	264	178	427	116	64	200
21	5-methoxytryptamine	314	304	246	83	744	725	124	127
22	acetophenone NIST	299	971	801	1732	1246	1353	497	584
23	aconitic acid	132	99	149	223	139	142	131	104
24	adenosine-5-phosphate	155	283	163	193	156	152	220	201
25	adipic acid	192	695	1045	931	740	849	241	418
26	alanine	94588	183357	189323	158568	109405	125119	46592	188992
27	alpha ketoglutaric acid	244	92	67	60	130	96	129	151
28	aminomalonic acid	613	622	1598	1332	1430	1508	986	879
29	arabinose	159	308	500	294	132	248	126	321
30	arabitol	56	195	137	130	125	95	75	386
31	arachidic acid	2291	1067	748	769	1032	1468	3078	1206
32	arachidonic acid	230	7348	9097	18245	7869	7071	2006	550
33	asparagine	2168	3462	4609	4160	5327	5130	1414	4466
34	aspartic acid	561	1238	2026	2327	4619	2614	465	1060
35	azelaic acid	54	574	532	597	513	412	68	102
36	behenic acid	163	185	212	225	230	92	156	128

Biomarker Meta-Analysis Step-By-Step

Step 1: On the MetaboAnalyst home page, press “**click here to start**” to enter the module overview.



MetaboAnalyst-- a comprehensive tool for metabolomics analysis and interpretation

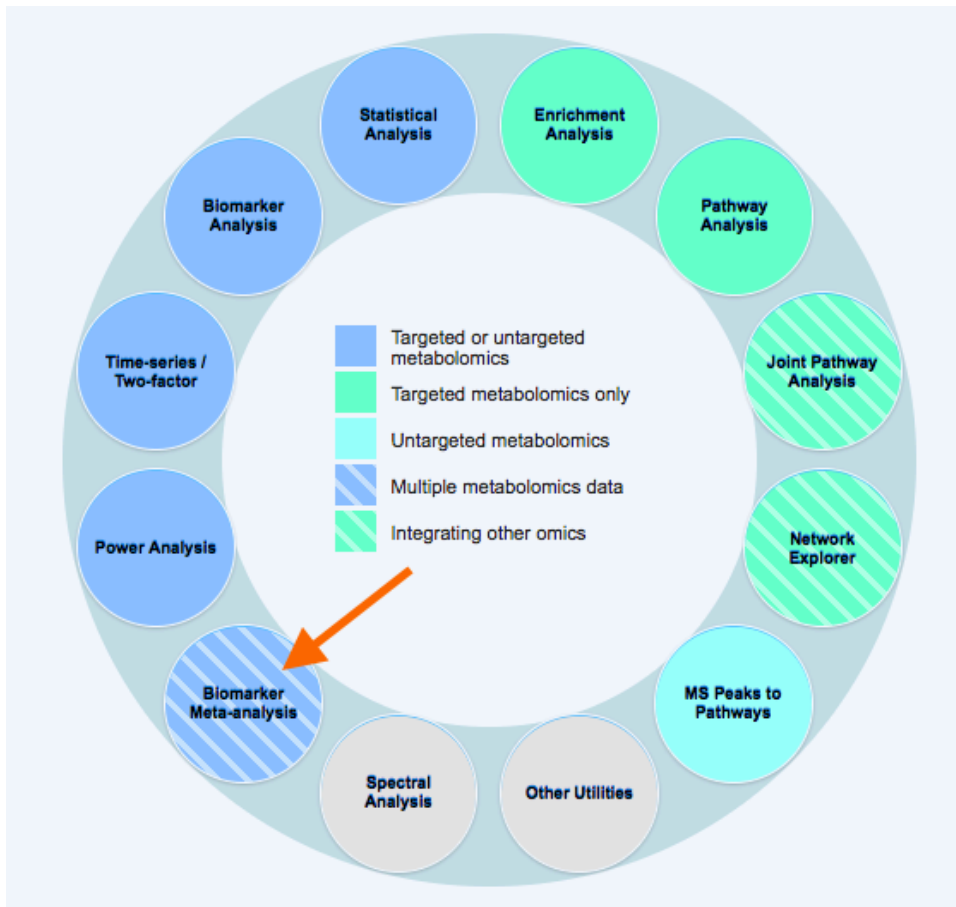
Welcome >> [click here to start](#) <<

Home
Overview
Data Formats

News & Updates

- Added a new tutorial for the three new modules (02/09/2018); **NEW**
- Minor bug fixes and feature enhancements based on user feedback (02/06/2018); **NEW**
- Release of **MetaboAnalyst 4.0** together with a companion R package **MetaboAnalystR**. You can still access [version 3.0 here](#) (01/29/2018); **NEW**
- Updated the interface for module selection (01/22/2018); **NEW**

Step 2: On the Module View page, click the “**Biomarker Meta-Analysis**” circle to begin.



Step 3: In the Biomarker Meta-Analysis upload page, press the “**Upload**” button to begin uploading your individual metabolomic datasets. Please specify the format of your data (samples in columns or in rows) and select your data. Click “**Submit**” to upload the data. A message will pop up in the top corner of your screen informing you if the data upload was successful.

Data Upload	Sanity Check	Visualization	Normalization	DE Analysis	Data Summary	Include
Upload	Process	View	Normalize	Analyze	Detail	<input checked="" type="checkbox"/>
						Add New

Adjust study batch effect

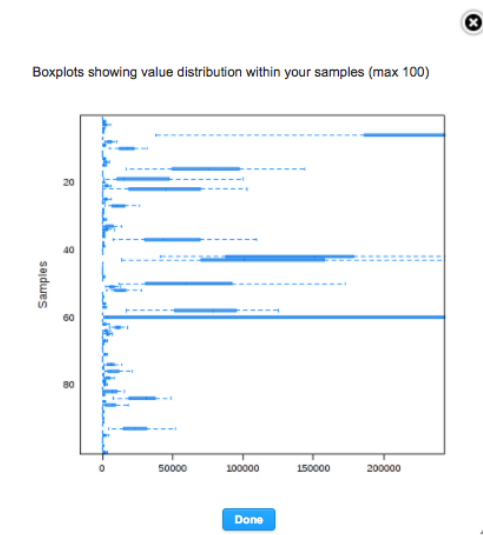
Following uploading of your data, the remaining buttons are now clickable. Press “[Process](#)” to perform a sanity check on your uploaded data, such as verifying sample names are unique, identifying group labels, and checking the number of samples and features uploaded (see example in the screenshot below). Click “Done” to close the Sanity Check.

Sanity Check

Data processing information:
Checking data content ...passed
Samples are in columns and features in rows.
No empty rows were found in your data.
No empty labels were found in your data.
Two groups found: Adenocarcinoma and Adenocarcinoma
All sample names are unique.
No empty feature names found
All feature names are unique
All sample names are OK
All feature names are OK
A total of 86 samples were found.
A total of 152 features were found.

[Done](#)

Next, click “[View](#)” to see boxplots of your data showing the distribution of the first 100 samples (screenshot below). As you can see from this example, the distributions are very uneven amongst the samples. Click “Done” to close the box.



Click “[Normalize](#)” to choose to perform Log₂ transformation and auto-scaling (see screenshot below). As the data was unevenly distributed, both will be performed. Click “[Submit](#)” to perform the normalization.

Data Normalization

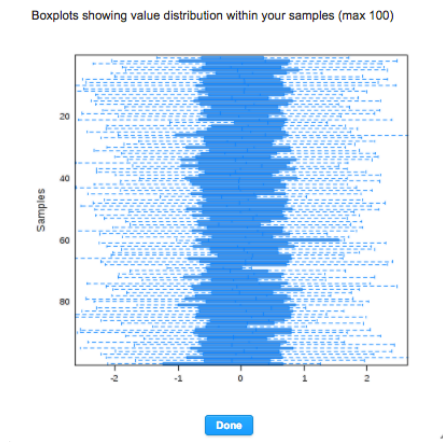
We suggest all datasets to be compared at log scales with similar distribution across different datasets. You can use **boxplots** (available in the previous step - data visualization plots) to help you determine if the data is already on the log scale.

Normalization procedure: Log2 transformation

Perform auto-scaling

Submit

Now, if you click the “[View](#)” button once again, we can see a visualization of the results of the data normalization. From the screenshot below, the data is much more evenly distributed.



Remember that the module will perform differential analysis (DE) on each individual dataset uploaded. Click “[Analyze](#)” to perform DE using linear models (limma) on your dataset. Please input the p-value (FDR) cut-off and the fold-change cut-off and then click “Submit”.

DE Analysis

You can perform differential analysis using linear models (limma). Note, this is for current exploratory analysis. It will be re-computed using the p-value cutoff as specified in the meta-analysis (next page) and for the Venn Diagram comparison.

Set p value (FDR) cutoff:

Set fold change (FC) cutoff:

Submit

The number of significant and non-significant features will appear in a pie chart below the “Submit” button (see screenshot below). Click “Done” to close the pop-up box.

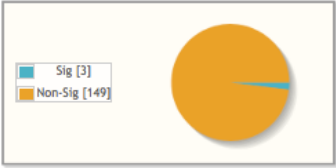
DE Analysis

You can perform differential analysis using linear models (limma). Note, this is for current exploratory analysis. It will be re-computed using the p-value cutoff as specified in the meta-analysis (next page) and for the Venn Diagram comparison.

Set p value (FDR) cutoff

Set fold change (FC) cutoff

Submit



Done

To view an over-view of the data processing performed on your uploaded data, click “[Detail](#)”. As you can see in the example below, the number of features and samples are shown, as are the group labels, which normalization procedures were applied, and how many features were significantly different between the two groups. Click “[Done](#)” to close the pop-up box.

Data name: data1

Number of features: 152

Number of samples: 86

Group labels: Adenocarcinoma vs. Control

Normalization procedures used: Log2 transform followed by autoscale

Number of significant hits: 3

Done

Finally, click “[Add new](#)” to add a new data set and perform the same steps as above for each new dataset uploaded.

Data Upload	Sanity Check	Visualization	Normalization	DE Analysis	Data Summary	Include
<input checked="" type="checkbox"/> data4	<input checked="" type="checkbox"/> Process	<input type="checkbox"/> View	<input checked="" type="checkbox"/> Normalize	<input checked="" type="checkbox"/> Analyze	<input type="checkbox"/> Detail	<input checked="" type="checkbox"/>
<input type="checkbox"/> Upload	<input type="checkbox"/> Process	<input type="checkbox"/> View	<input type="checkbox"/> Normalize	<input type="checkbox"/> Analyze	<input type="checkbox"/> Detail	<input checked="" type="checkbox"/>

Add New

Adjust study batch effect

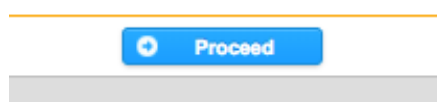
Please note that a maximum of 1000 samples are allowed. Also, after uploading all your datasets, you can check the boxes in the “Include” column of the data-upload page to include a study or not in the meta-analysis. As well, check the “Adjust study batch effect” if you would like batch-effect to be accounted for in the meta-analysis.

Use Case: Click the “Try Examples” button on the bottom of the page to use the example dataset detailed above (Adenocarcinoma vs. Control). A dialogue will appear, click “Yes”, and then the four datasets will be automatically uploaded and processed.

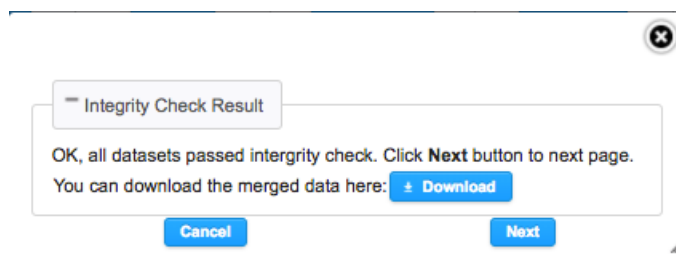
Example Datasets

Datasets	Data Type	Description	Phenotype
data1 data2 data3 data4	Untargeted metabolomics data (GC-TOF-MS).	Four test datasets were created from a subset of data from a study investigating biomarkers of adenocarcinoma in blood (For more details: Fiehn et al.)	Groups: Adenocarcinoma and Control

Following the uploading of your data/example data, click “Proceed” on the bottom of the page to move onto the next step.



This will initiate a final data integrity check on all of your uploaded data that is included for meta-analysis. It will check if group labels are consistent, if at least 25% of features are common, the number of uploaded samples, etc. Click “Next” to continue.



Step 4: Following data uploading and processing, we can choose from 1 of 3 methods to perform meta-analysis. In this tutorial, we will go through the results of the “Combining p-values” method. The screenshot below highlights the 3 methods available, paired with a detailed description. Please note that on the right-hand side is the R Command history, which reveals the step-by-step of your analysis in real-time. This R Command History can be used to reproduce your analysis locally in R

with the MetaboAnalystR package. Also, on the left-hand side is the navigation panel, which can directly navigate you through different steps of the Biomarker Meta-Analysis module.

The screenshot displays the MetaboAnalyst 4.0 web interface. The main title is "MetaboAnalyst-- a comprehensive tool for metabolomics analysis and interpretation". On the left, a navigation panel includes options: Upload, Meta analysis, Result table, Venn diagram, Download, and Exit. The central area is divided into three sections: "Combining P Values", "Vote Counting", and "Direct Merging". The "Combining P Values" section is highlighted with a red box and contains the following text: "There are two widely used methods to combine p values from multiple studies for information integration - the Fisher's method (-2*ΣLog(p)) and the Stouffer's method (based on inverse normal transformation). Stouffer's method incorporates weight (i.e. based on sample sizes) into the calculation; while Fisher's method is known as a 'weight-free' method. They usually have very similar performance. However, in metabolomic meta-analysis, larger sample sizes do not warrant larger weights as the quality of each study can vary. Users should choose to apply Stouffer's method only when all studies are of similar qualities (i.e. same analytical platforms with similar levels of missing values)." Below this text, there is a dropdown menu for "Select a method" set to "Fisher's method" and a text input for "Set a significance level" set to "0.05", with a "Submit" button. The "Vote Counting" section has a "Set a significance level" of "0.05" and "Set the minimal number of votes" of "2", with a "Submit" button. The "Direct Merging" section has a "Set a significance level" of "0.05" and a "Submit" button. On the right, an "R Command History" panel shows a list of 18 R commands. At the bottom, there are "Previous" and "Proceed" buttons.

Combining P Values: In the Combining P Values box, click the “Select a method” box to choose either the Fisher’s method or the Stouffer’s method for combining p-values. As the same GCTOFMS instrumentation was used, and all studies have no missing values, we will select the Stouffer’s method. By default the significance level in “Set a significance level” is 0.05, which we will keep for the meta-analysis. Click “Submit” to perform p-value combination. A dialogue box will appear in the top-right corner of your screen, informing you of the total number of significant features identified in the meta-analysis (screenshot below). Click “Proceed” on the bottom right corner of your screen to continue to the detailed results of the meta-analysis.

The screenshot shows a blue dialog box with an information icon (i) and the text: "OK A total of 34 significant features found. Click Proceed button to view the result." The dialog box is positioned in the top-right corner of the screen.

Step 5: Remember that the goal of Biomarker Meta-Analysis is to identify robust metabolic biomarkers of disease/conditions. Using the example dataset, we aim to identify consistent markers of adenocarcinoma in blood across all four datasets. The results of the “Combining P Values” meta-analysis are listed in a table as in the screenshot below. In the table, the features are ranked by their combined p-value from lowest to highest. Here, we see that adenosine-5-phosphate, pyrophosphate, and pyruvic acid are the top 3 biomarkers identified, all with a combined p-value of 0.

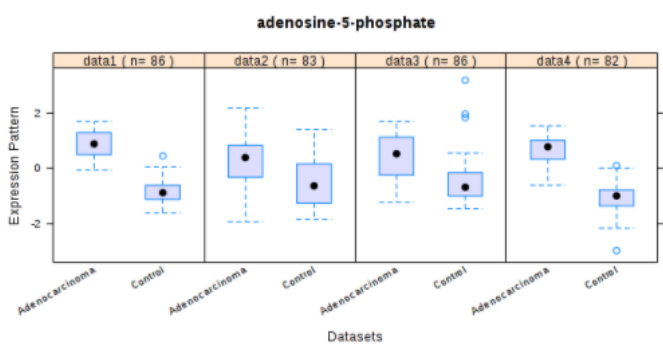
The statistics from individual data analysis are given in columns with the corresponding dataset names. You can either adjust its content or sort the table.

Data summary: Log fold change | Sort by: CombinedPval | Order: (Ascending) | Update | Search | Download

ID	data1	data2	data3	data4	CombinedTstat	CombinedPval	View
adenosine-5-phosphate	-1.7535	-0.81954	-0.89017	-1.6839	-187.97	0.0	View
pyrophosphate	-1.6539	-0.68801	-1.0106	-1.6181	-174.04	0.0	View
pyruvic acid	-1.7275	-0.02462	-1.154	-0.18216	-123.22	0.0	View
maltotriose	-0.57042	-0.78002	-0.62125	-0.34062	-45.974	8.2354E-6	View
glutamine	0.25055	0.58333	0.9243	0.23142	42.535	2.9447E-5	View
lactamide	-0.16134	-0.33998	-0.99086	-0.14049	-37.845	1.8363E-4	View
citruiline	0.17019	0.70856	0.64683	0.23439	34.702	5.1871E-4	View
lactic acid	-0.048584	-0.13516	-1.0411	0.010808	-34.79	5.1871E-4	View
alpha ketoglutaric acid	-0.52005	-0.28152	-0.58456	-0.40263	-32.543	0.0011327	View
cystine	0.21575	0.80808	0.38177	0.20646	31.319	0.0015355	View
taurine	0.0050866	-0.21137	-0.88704	-0.27037	-31.389	0.0015355	View
maltose	-0.35858	-0.69713	-0.38259	-0.22187	-30.371	0.0020746	View
fructose	0.54515	0.28723	0.60315	0.1174	29.697	0.0025194	View
asparagine	0.26279	0.37151	0.66667	0.28026	29.053	0.0030375	View
oxalic acid	-0.24872	-0.55199	-0.44125	-0.32291	-27.222	0.0059119	View

(1 of 3) | 1 | 2 | 3 | 15

Click “View” to view a box-plot of the expression pattern of the compounds per each uploaded dataset. In the screenshot below of adenosine-5-phosphate (A5P), we can clearly see that A5P is consistently higher in patients with adenocarcinoma than in healthy controls.



Step 6: We can then visualize the results of the meta-analysis in a Venn diagram (click “Venn Diagram”). A dialogue box will appear, listing all of the datasets and the results of the meta-analysis (as meta_dat), the number of differentially expressed features per each dataset, and a box for users to click on under the “Include” header to include or not to include datasets in the venn diagram. Note that a maximum of four datasets can be visualized using a Venn diagram. In this case, we will not include dataset 4 (screenshot below). Click “Submit” to continue.

The max number of datasets that can be compared is **four**.
 Datasets without significant hits will be excluded.

Name	DE #	Include
data1	3	<input checked="" type="checkbox"/>
data2	12	<input checked="" type="checkbox"/>
data3	21	<input checked="" type="checkbox"/>
data4	2	<input checked="" type="checkbox"/>
meta_dat	34	<input checked="" type="checkbox"/>

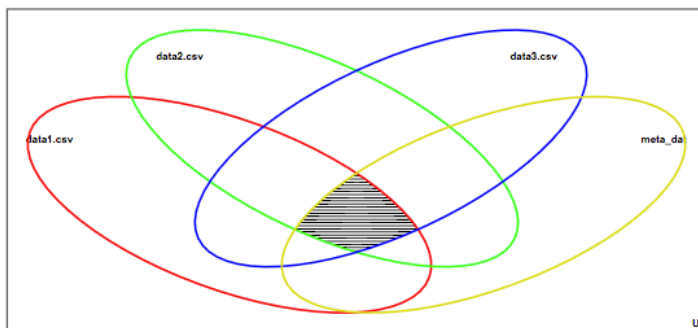
Cancel Submit

Step 7: On this page, we can view the resulting Venn diagram. Each area of the Venn diagram is clickable. If we click directly in the center of the overlapping datasets, we can see that a total of 2 features (A5P and pyrophosphate) overlap between datasets 1, 2, 3, and the meta-analysis. Click on all possible areas to view all possible combinations of features.

Different regions in the Venn diagram represent all possible comparisons among the data sets.

- At most **four** datasets can be compared at the same time. Datasets without hits will **NOT** be shown here;
- The area of the region does **NOT** relate to the number of features;
- **Click** an area to show the corresponding features on the left panel.

- Total:2
- adenosine-5-phosphate
- pyrophosphate



Step 8: On the left-hand side navigation panel, click “Download” to view all images and tables generated throughout the Biomarker Meta-Analysis, as well as to download the Analysis Report, which contains all the details of each step of your analysis as well as all of the results. You can also directly download all the results in a zip file by clicking “Download.zip”.

Result Download

Please download the PDF Analysis Report and results (tables and images) below. The “Download.zip” contains all the files in your home directory. The PDF report may not be generated sometimes. You can try to re-generate PDF using an alternative approach using the button below.

[Regenerate](#) [Analysis Report](#)

Download.zip	SigFeatures_data4.csv
Rhistory.R	data4.csv
venn_diagram.png	data2.csv
MetaboAnalyst_merged_data.csv	adenosine-5-phosphate.png
data1.csv	SigFeatures_data2.csv
SigFeatures_data3.csv	data3.csv
meta_sig_features_metap.csv	SigFeatures_data1.csv

[Logout](#)

--End of tutorial --