

# Metabolomic Data Analysis with MetaboAnalyst 6.0

Name: guest14586620120739914885

July 1, 2024

## 1 Background

MSEA or Metabolite Set Enrichment Analysis is a way to identify biologically meaningful patterns that are significantly enriched in quantitative metabolomic data. In conventional approaches, metabolites are evaluated individually for their significance under conditions of study. Those compounds that have passed certain significance level are then combined to see if any meaningful patterns can be discerned. In contrast, MSEA directly investigates if a set of functionally related metabolites without the need to preselect compounds based on some arbitrary cut-off threshold. It has the potential to identify subtle but consistent changes among a group of related compounds, which may go undetected with the conventional approaches.

Essentially, MSEA is a metabolomic version of the popular GSEA (Gene Set Enrichment Analysis) software with its own collection of metabolite set libraries as well as an implementation of user-friendly web-interfaces. GSEA is widely used in genomics data analysis and has proven to be a powerful alternative to conventional approaches. For more information, please refer to the original paper by Subramanian A, and a nice review paper by Nam D, Kim SY.<sup>1, 2</sup>

## 2 MSEA Overview

Metabolite set enrichment analysis consists of four steps - data input, data processing, data analysis, and results download. Different analysis procedures are performed based on different input types. In addition, users can also browse and search the metabolite set libraries as well as upload their self-defined metabolite sets for enrichment analysis. Users can also perform metabolite name mapping between a variety of compound names, synonyms, and major database identifiers.

## 3 Data Input

There are three enrichment analysis algorithms offered by MSEA. Accordingly, three different types of data inputs are required by these three approaches:

- A list of important compound names - entered as a one column data (*Over Representation Analysis (ORA)*);
- A single measured biofluid (urine, blood, CSF) sample- entered as tab separated two-column data with the first column for compound name, and the second for concentration values (*Single Sample Profiling (SSP)*);

---

<sup>1</sup>Subramanian A. *Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.*, Proc Natl Acad Sci USA. 2005 102(43): 15545-50

<sup>2</sup>Nam D, Kim SY. *Gene-set approach for expression pattern analysis*, Briefings in Bioinformatics. 2008 9(3): 189-197.

- A compound concentration table - entered as a comma separated (.csv) file with the each sample per row and each metabolite concentration per column. The first column is sample names and the second column for sample phenotype labels (*Quantitative Enrichment Analysis (QEA)*)

## 4 Data Process

Table 1: Results

59	CMP	Cytidine monophosphate	HMDB0000095	6131	C00055	<chem>NC1=NC(=O)N(C=C1)OP(=O)(O)O</chem>
60	O-Phosphorylethanolamine	O-Phosphoethanolamine	HMDB0000224	1015	C00346	<chem>NCCOP(=O)(O)O</chem>
61	GDP-Mannose	Guanosine diphosphate mannose	HMDB0001163	18396	C00096	<chem>NC1=NC2=C(N=CN2)[C@@H](O)[C@H](O)[C@@H](O)[C@@H](O)[C@@H]1O</chem>
62	Malonic acid	Malonic acid	HMDB0000691	867	C00383	<chem>OC(=O)CC(=O)O</chem>
63	4-Guanidinobutanoic acid	4-Guanidinobutanoic acid	HMDB0003464	500	C01035	<chem>NC(=N)NCCCC(=O)O</chem>
64	GDP	Guanosine diphosphate	HMDB0001201	8977	C00035	<chem>NC1=NC2=C(N=CN2)[C@@H](O)[C@H](O)[C@@H](O)[C@@H](O)[C@@H]1O</chem>
65	Fructose	D-Fructose	HMDB0000660	439709	C00095	<chem>OC[C@H]1O[C@](O)(CO)[C@@H](O)[C@@H](O)[C@@H]1O</chem>
66	Glutamate	Glutamic acid	HMDB0000148	33032	C00025	<chem>N[C@@H](CCC(=O)O)C(=O)O</chem>
67	L-Octanoylcarnitine	Octanoylcarnitine	HMDB0000791	11953814	C02838	<chem>CCCCCCCC(=O)O[C@@H](O)[C@H](O)[C@@H](O)[C@@H](O)[C@@H](O)C(=O)N</chem>
68	Nicotinamide riboside	Nicotinamide riboside	HMDB0000855	439924	C03150	<chem>NC(=O)C1=C[N+](=CC1)N</chem>
69	1-Methylhistidine	1-Methylhistidine	HMDB0000001	92105	C01152	<chem>CN1C=NC(C[C@H](N)C1)=O</chem>
70	Proline	Proline	HMDB0000162	145742	C00148	<chem>OC(=O)[C@@H]1CCCN1</chem>
71	Glutaric acid	Glutaric acid	HMDB0000661	743	C00489	<chem>OC(=O)CCCC(=O)O</chem>
72	5-Hydroxytryptophan	5-Hydroxy-L-tryptophan	HMDB0000472	439280	C00643	<chem>N[C@@H](CC1=CN(C2=CC=CC=C2)C(=O)N1)C(=O)O</chem>
73	UDP-N-acetyl-glucosamine	Uridine diphosphate-N-acetylglucosamine	HMDB0000290	445675	C00043	<chem>CC(=O)N[C@@H]1[C@@H](O)[C@H](O)[C@@H](O)[C@@H]1O</chem>
74	UDP-D-Glucose	Uridine diphosphate glucose	HMDB0000286	8629	C00029	<chem>OC[C@H]1O[C@H](OP(=O)(O)O)[C@H](O)[C@@H](O)[C@@H]1O</chem>
75	Pantothenate	Pantothenic acid	HMDB0000210	6613	C00864	<chem>CC(C)(CO)[C@@H](O)C(=O)O</chem>
76	NADP	NADP	HMDB0000217	5885	C00006	<chem>NC(=O)C1=C[N+](=CC1)N</chem>
77	Homocitrulline	Homocitrulline	HMDB0000679	65072	C02427	<chem>N[C@@H](CCCCNC(N)=O)C(=O)O</chem>
78	Acetyl CoA	Acetoacetyl-CoA	HMDB0001484	92153	C00332	<chem>CC(=O)CC(=O)SCCNC(=O)O</chem>
79	Glutathione	Glutathione	HMDB0000125	124886	C00051	<chem>N[C@@H](CCC(=O)N)C(=O)O</chem>
80	S-Adenosyl-homocysteine	S-Adenosylhomocysteine	HMDB0000939	439155	C00021	<chem>N[C@@H](CCSC[C@H]1O[C@H](COP(=O)(O)O)[C@@H](O)[C@@H]1O)C(=O)O</chem>
81	Pirbuterol	Pirbuterol	HMDB0015407	4845	C07807	<chem>CC(C)(C)NCC(=O)C1=NC(=O)C=C1</chem>
82	Orotate	Orotic acid	HMDB0000226	967	C00295	<chem>OC(=O)C1=CC(=O)NC(=O)N1</chem>

The second step is to check concentration values. For SSP analysis, the concentration must be measured in *umol* for blood and CSF samples. The urinary concentrations must be first converted to *umol/mmol\_creatinine* in order to compare with reported concentrations in literature. No missing or negative values are allowed in SSP analysis. The concentration data for QEA analysis is more flexible. Users can upload either the original concentration data or normalized data. Missing or negative values are allowed (coded as *NA*) for QEA.

## 5 Selection of Metabolite Set Library

Before proceeding to enrichment analysis, a metabolite set library has to be chosen. There are seven built-in libraries offered by MSEA:

- Metabolic pathway associated metabolite sets (*currently contains 99 entries*);
- Disease associated metabolite sets (reported in blood) (*currently contains 344 entries*);
- Disease associated metabolite sets (reported in urine) (*currently contains 384 entries*);
- Disease associated metabolite sets (reported in CSF) (*currently contains 166 entries*);
- Metabolite sets associated with SNPs (*currently contains 4598 entries*);
- Predicted metabolite sets based on computational enzyme knockout model (*currently contains 912 entries*);
- Metabolite sets based on locations (*currently contains 73 entries*);
- Drug pathway associated metabolite sets (*currently contains 461 entries*);

In addition, MSEA also allows user-defined metabolite sets to be uploaded to perform enrichment analysis on arbitrary groups of compounds which researchers want to test. The metabolite set library is simply a two-column comma separated text file with the first column for metabolite set names and the second column for its compound names (**must use HMDB compound name**) separated by "; ". Please note, the built-in libraries are mainly from human studies. The functional grouping of metabolites may not be valid. Therefore, for data from subjects other than human being, users are suggested to upload their self-defined metabolite set libraries for enrichment analysis.

## 6 Enrichment Analysis

Over Representation Analysis (ORA) is performed when a list of compound names is provided. The list of compound list can be obtained through conventional feature selection methods, or from a clustering algorithm, or from the compounds with abnormal concentrations detected in SSP, to investigate if some biologically meaningful patterns can be identified.

ORA was implemented using the *hypergeometric test* to evaluate whether a particular metabolite set is represented more than expected by chance within the given compound list. One-tailed p values are provided after adjusting for multiple testing. **Figure 2** below summarizes the result.

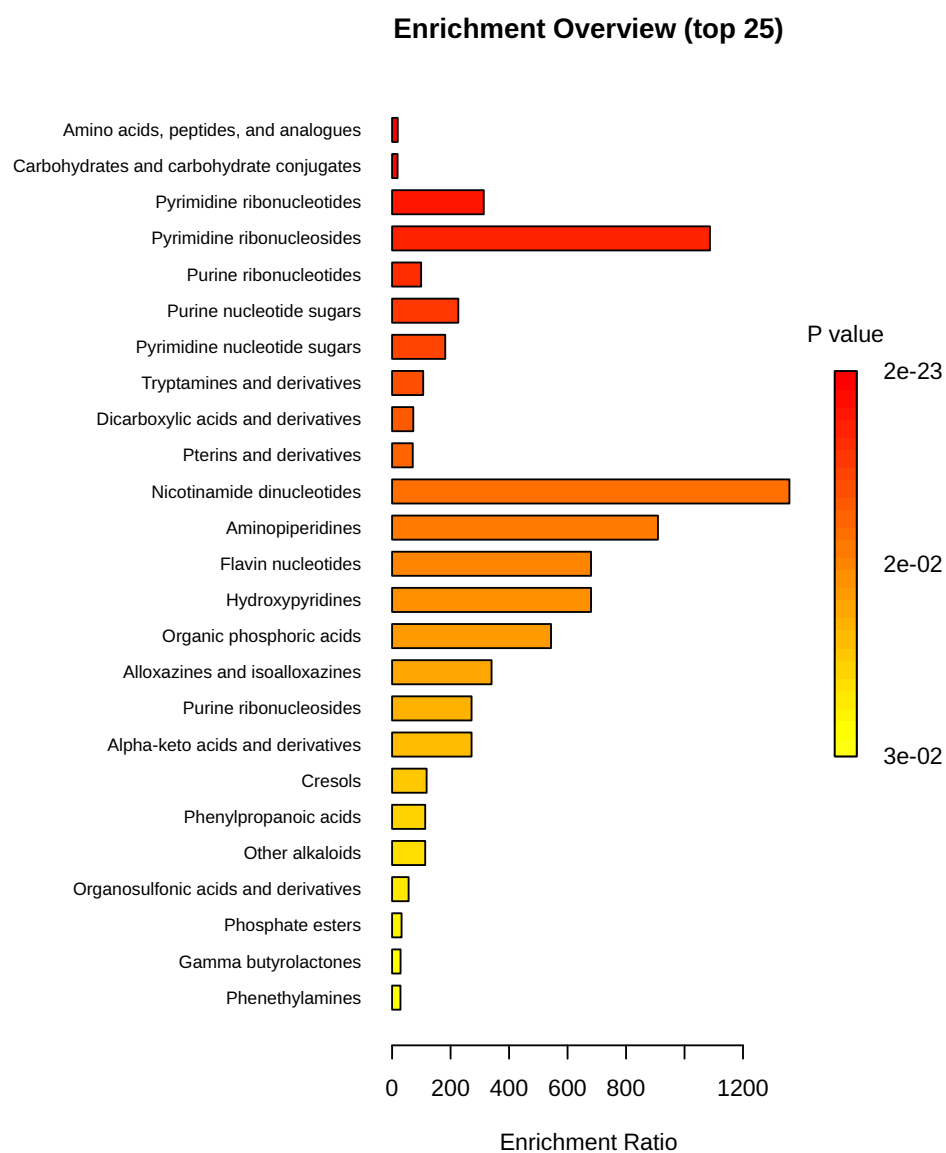


Figure 1: Summary Plot for Over Representation Analysis (ORA)

Table 2: Result from Over Representation Analysis

	total	expected	hits	Raw p	Holm p	FDR
Amino acids, peptides, and analogues	3220	1.19	23	1.89E-23	1.68E-20	1.68E-20
Carbohydrates and carbohydrate conjugates	1600	0.59	11	2.03E-11	1.80E-08	9.02E-09
Pyrimidine ribonucleotides	26	0.01	3	1.24E-07	1.10E-04	3.67E-05
Pyrimidine ribonucleosides	5	0.00	2	1.34E-06	1.18E-03	2.97E-04
Purine ribonucleotides	82	0.03	3	4.15E-06	3.68E-03	7.39E-04
Purine nucleotide sugars	24	0.01	2	3.67E-05	3.25E-02	5.44E-03
Pyrimidine nucleotide sugars	30	0.01	2	5.78E-05	5.11E-02	7.34E-03
Tryptamines and derivatives	51	0.02	2	1.68E-04	1.49E-01	1.87E-02
Dicarboxylic acids and derivatives	75	0.03	2	3.65E-04	3.22E-01	3.42E-02
Pterins and derivatives	77	0.03	2	3.84E-04	3.38E-01	3.42E-02
Nicotinamide dinucleotides	2	0.00	1	7.36E-04	6.48E-01	5.96E-02
Aminopiperidines	3	0.00	1	1.10E-03	9.70E-01	8.19E-02
Flavin nucleotides	4	0.00	1	1.47E-03	1.00E+00	9.35E-02
Hydroxypyridines	4	0.00	1	1.47E-03	1.00E+00	9.35E-02
Organic phosphoric acids	5	0.00	1	1.84E-03	1.00E+00	1.09E-01
Alloxazines and isoalloxazines	8	0.00	1	2.94E-03	1.00E+00	1.64E-01
Purine ribonucleosides	10	0.00	1	3.67E-03	1.00E+00	1.82E-01
Alpha-keto acids and derivatives	10	0.00	1	3.67E-03	1.00E+00	1.82E-01
Cresols	23	0.01	1	8.43E-03	1.00E+00	3.73E-01
Phenylpropanoic acids	24	0.01	1	8.80E-03	1.00E+00	3.73E-01
Other alkaloids	24	0.01	1	8.80E-03	1.00E+00	3.73E-01
Organosulfonic acids and derivatives	48	0.02	1	1.75E-02	1.00E+00	7.09E-01
Phosphate esters	84	0.03	1	3.05E-02	1.00E+00	1.00E+00
Gamma butyrolactones	94	0.03	1	3.40E-02	1.00E+00	1.00E+00
Phenethylamines	95	0.04	1	3.44E-02	1.00E+00	1.00E+00
Benzenediols	105	0.04	1	3.79E-02	1.00E+00	1.00E+00
Indoles	133	0.05	1	4.78E-02	1.00E+00	1.00E+00
Pyrimidines and pyrimidine derivatives	210	0.08	1	7.44E-02	1.00E+00	1.00E+00
Alcohols and polyols	309	0.11	1	1.08E-01	1.00E+00	1.00E+00
Amines	350	0.13	1	1.21E-01	1.00E+00	1.00E+00
Fatty acid esters	1650	0.61	2	1.24E-01	1.00E+00	1.00E+00
Fatty acids and conjugates	941	0.35	1	2.93E-01	1.00E+00	1.00E+00

## 7 Appendix: R Command History

```
[1] "mSet<-InitDataObjects(\"conc\", \"msetora\", FALSE)"
[2] "cmpd.vec<-c(\"Montiporic Acid C\", \"Montiporic Acid B\", \"Montiporic Acid D\", \"Montiporic Acid E\")"
[3] "mSet<-Setup.MapData(mSet, cmpd.vec);"
[4] "mSet<-CrossReferencing(mSet, \"name\");"
[5] "mSet<-CreateMappingResultTable(mSet)"
[6] "mSet<-PerformDetailMatch(mSet, \"Cresol\");"
[7] "mSet<-GetCandidateList(mSet);"
[8] "mSet<-SetCandidate(mSet, \"Cresol\", \"p-Cresol\");"
[9] "mSet<-PerformDetailMatch(mSet, \"O-Decanoyl-L-carnitine\");"
[10] "mSet<-GetCandidateList(mSet);"
[11] "mSet<-SetCandidate(mSet, \"O-Decanoyl-L-carnitine\", \"Decanoylcarnitine\");"
[12] "mSet<-PerformDetailMatch(mSet, \"Lysine-Glutamine\");"
[13] "mSet<-GetCandidateList(mSet);"
[14] "mSet<-SetCandidate(mSet, \"Lysine-Glutamine\", \"Lysylglutamine\");"
[15] "mSet<-PerformDetailMatch(mSet, \"NG-dimethyl-L-arginine\");"
[16] "mSet<-GetCandidateList(mSet);"
[17] "mSet<-SetCandidate(mSet, \"NG-dimethyl-L-arginine\", \"Asymmetric dimethylarginine\");"
[18] "mSet<-PerformDetailMatch(mSet, \"Arginine-Alanine\");"
[19] "mSet<-GetCandidateList(mSet);"
[20] "mSet<-SetCandidate(mSet, \"Arginine-Alanine\", \"Arginylalanine\");"
[21] "mSet<-PerformDetailMatch(mSet, \"Arginine-Glutamine\");"
[22] "mSet<-GetCandidateList(mSet);"
[23] "mSet<-SetCandidate(mSet, \"Arginine-Glutamine\", \"Arginylglutamine\");"
[24] "mSet<-PerformDetailMatch(mSet, \"N N N-Trimethyllysine\");"
[25] "mSet<-GetCandidateList(mSet);"
[26] "mSet<-PerformDetailMatch(mSet, \"N-Acetyl-Glucosamine\");"
[27] "mSet<-GetCandidateList(mSet);"
[28] "mSet<-SetCandidate(mSet, \"N-Acetyl-Glucosamine\", \"N-Acetyl-D-Glucosamine 6-Phosphate\");"
[29] "mSet<-PerformDetailMatch(mSet, \"Acetyl glycine\");"
[30] "mSet<-GetCandidateList(mSet);"
[31] "mSet<-SetCandidate(mSet, \"Acetyl glycine\", \"Phenylacetylglutamine\");"
[32] "mSet<-PerformDetailMatch(mSet, \"Acetyl CoA\");"
[33] "mSet<-GetCandidateList(mSet);"
[34] "mSet<-SetCandidate(mSet, \"Acetyl CoA\", \"Acetoacetyl-CoA\");"
[35] "mSet<-SetMetabolomeFilter(mSet, F);"
[36] "mSet<-SetCurrentMsetLib(mSet, \"main_class\", 2);"
[37] "mSet<-CalculateHyperScore(mSet)"
[38] "mSet<-PlotORA(mSet, \"ora_0\", \"net\", \"png\", 72, width=NA)"
[39] "mSet<-PlotEnrichDotPlot(mSet, \"ora\", \"ora_dot_0\", \"png\", 72, width=NA)"
[40] "mSet<-PlotEnrichPieChart(mSet, \"ora\", \"ora_pie_0\", \"png\", 72)"
[41] "mSet<-CalculateHyperScore(mSet)"
[42] "mSet<-PlotORA(mSet, \"ora_1\", \"net\", \"png\", 72, width=NA)"
[43] "mSet<-PlotEnrichDotPlot(mSet, \"ora\", \"ora_dot_1\", \"png\", 72, width=NA)"
[44] "mSet<-PlotEnrichPieChart(mSet, \"ora\", \"ora_pie_1\", \"png\", 72)"
[45] "mSet<-CalculateHyperScore(mSet)"
[46] "mSet<-PlotORA(mSet, \"ora_2\", \"net\", \"png\", 72, width=NA)"
[47] "mSet<-PlotEnrichDotPlot(mSet, \"ora\", \"ora_dot_2\", \"png\", 72, width=NA)"
[48] "mSet<-PlotEnrichPieChart(mSet, \"ora\", \"ora_pie_2\", \"png\", 72)"
[49] "mSet<-CalculateHyperScore(mSet)"
[50] "mSet<-PlotORA(mSet, \"ora_3\", \"net\", \"png\", 72, width=NA)"
[51] "mSet<-PlotEnrichDotPlot(mSet, \"ora\", \"ora_dot_3\", \"png\", 72, width=NA)"
[52] "mSet<-PlotEnrichPieChart(mSet, \"ora\", \"ora_pie_3\", \"png\", 72)"
[53] "mSet<-SaveTransformedData(mSet)"
[54] "mSet<-PreparePDFReport(mSet, \"guest14586620120739914885\")\n"
[55] "mSet<-SetMetabolomeFilter(mSet, F);"
[56] "mSet<-SetCurrentMsetLib(mSet, \"sub_class\", 2);"
```

```

[57] "mSet<-CalculateHyperScore(mSet)"
[58] "mSet<-PlotORA(mSet, \"ora_4_\", \"net\", \"png\", 72, width=NA)"
[59] "mSet<-PlotEnrichDotPlot(mSet, \"ora\", \"ora_dot_4_\", \"png\", 72, width=NA)"
[60] "mSet<-PlotEnrichPieChart(mSet, \"ora\", \"ora_pie_4_\", \"png\", 72)"
[61] "mSet<-CalculateHyperScore(mSet)"
[62] "mSet<-PlotORA(mSet, \"ora_5_\", \"net\", \"png\", 72, width=NA)"
[63] "mSet<-PlotEnrichDotPlot(mSet, \"ora\", \"ora_dot_5_\", \"png\", 72, width=NA)"
[64] "mSet<-PlotEnrichPieChart(mSet, \"ora\", \"ora_pie_5_\", \"png\", 72)"
[65] "mSet<-SaveTransformedData(mSet)"
[66] "mSet<-PreparePDFReport(mSet, \"guest14586620120739914885\")\n"

```

---

The report was generated on Mon Jul 1 16:04:59 2024 with R version 4.3.2 (2023-10-31), OS system: Linux, version: -Ubuntu SMP Tue Mar 5 20:16:58 UTC 2024 .