Metabolomic Data Analysis with MetaboAnalyst 6.0

Name: guest9044362343704604219

January 7, 2025

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. [1]. [2]

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)*);

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

[2] 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)*)

You selected Over Representation Analysis (ORA) which requires a list of compound names as input.

4 Data Process

The first step is to standardize the compound labels. It is an essential step since the compound labels will be subsequently compared with compounds contained in the metabolite set library. MSEA has a built-in tool to convert between compound common names, synonyms, identifiers used in HMDB ID, PubChem, ChEBI, BiGG, METLIN, KEGG, or Reactome. **Table 1** shows the conversion results. Note: *1* indicates exact match, *2* indicates approximate match, and *0* indicates no match. A text file contain the result can be found the downloaded file *name_map.csv*

Table 1: Result from Compou

	Query	Match	HMDB	PubChem	KEGG	SMILES
1	C03017	Propionylcarnitine	HMDB0000824	188824	C03017	CCC(=O)O[C@H](CC(O)=O)C[N+](C)(C)C
2	C07005	Flunisolide	HMDB0014326	82153	C07005	[H][C@@]12C[C@@]3([H])[C@]4([H])C[C@H](
3	C00062	L-Arginine	HMDB0000517	6322	C00062	N[C@@H](CCCNC(N)=N)C(O)=O
4	C00003	NAD	HMDB0000902	5892	C00003	NC(=O)C1=C[N+](=CC=C1)[C@@H]1O[C@
5	C00307	Citicoline	HMDB0001413	13805	C00307	C[N+](C)(C)CCOP(O)(=O)OP(O)(=O)OC[
6	C00127	Oxidized glutathione	HMDB0003337	65359	C00127	N[C@@H](CCC(=O)N[C@@H](CSSC[C@H](
7	C00105	Uridine 5'-monophosphate	HMDB0000288	6030	C00105	O[C@H]1[C@@H](O)[C@@H](O[C@@H]1COP
8	C00015	Uridine 5'-diphosphate	HMDB0000295	6031	C00015	O[C@H]1[C@@H](O)[C@@H](O[C@@H]1COP
9	C00020	Adenosine monophosphate	HMDB0000045	6083	C00020	NC1=C2N=CN([C@@H]3O[C@H](COP(O)(O
10	C02494	1-Methyladenosine	HMDB0003331	27476	C02494	CN1C=NC2=C(N=CN2[C@@H]2O[C@H](CO
11	C20387	Biotin sulfone	HMDB0004818	21252323	C20387	[H][C@]12CS(=O)(=O)[C@@H](CCCCC(O)=
12	C01586	Hippuric acid	HMDB0000714	464	C01586	OC(=O)CNC(=O)C1=CC=CC=C1
13	C02301	O-Acylcarnitine		5355	C02301	
14	C11430	9-Hydroxyphenanthrene	HMDB0059801	10229	C11430	OC1=CC2=CC=CC=C2C2=CC=CC=C12
15	C16207	Anthraquinone	HMDB0248468	6780	C16207	O=C1C2=CC=CC=C2C(=O)C2=CC=CC=
16	C22599	NA	NA	NA	NA	NA
17	C00319	Sphingosine	HMDB0000252	5280335	C00319	CCCCCCCCCCCCC\C=C\[C@@H](O)[C@@
18	C00836	Sphinganine	HMDB0000269	91486	C00836	CCCCCCCCCCCCCCC[C@@H](O)[C@@H](
19	C02990	Palmitoylcarnitine	HMDB0000222	11953816	C02990	CCCCCCCCCCCCCCCC(=O)O[C@H](CC(
20	C00199	D-Ribulose 5-phosphate	HMDB0000618	439184	C00199	OCC(=O)[C@H](O)[C@H](O)COP(O)(O)=O
21	C05382	D-Sedoheptulose 7-phosphate	HMDB0001068	92042786	C05382	OC[C@]1(O)O[C@H](COP(O)(O)=O)[C@@H
22	C00362	2'-Deoxyguanosine 5'-monophosphate	HMDB0001044	65059	C00362	NC1=NC2=C(N=CN2[C@H]2C[C@H](O)[C@
23	C03546	D-myo-Inositol 4-phosphate	HMDB0001313	440043	C03546	O[C@@H]1[C@H](O)[C@H](O)[C@@H](OP(O
24	C00052	Uridine diphosphategalactose	HMDB0000302	18068	C00052	OC[C@H]1O[C@H](OP(O)(=O)OP(O)(=O)O
25	C00167	Uridine diphosphate glucuronic acid	HMDB0000935	17473	C00167	O[C@@H]1[C@@H](COP(O)(=O)OP(O)(=O
26	C00262	Hypoxanthine	HMDB0000157	790	C00262	OC1=NC=NC2=C1NC=N2
27	C00364	5-Thymidylic acid	HMDB0001227	9700	C00364	CC1=CN([C@H]2C[C@H](O)[C@@H](COP(O
28	C00158	Citric acid	HMDB0000094	311	C00158	OC(=O)CC(O)(CC(O)=O)C(O)=O
29	C03794	Adenylsuccinic acid	HMDB0000536	447145	C03794	O[C@@H]1[C@@H](COP(O)(O)=O)O[C@H]
30	C00360	Deoxyadenosine monophosphate	HMDB0000905	12599	C00360	NC1=NC=NC2=C1N=CN2[C@H]1C[C@H](
31	C10438	Cinnamic acid	HMDB0000567	5372954	C10438	OC(=O)\C=C/C1=CC=CC=C1
32	C00157	Phosphatidylcholine			C00157	
33	C01495	3-Hydroxyflavone	HMDB0031816	11349	C01495	OC1=C(OC2=CC=CC=C2C1=O)C1=CC=
34	C00242	Guanine	HMDB0000132	764	C00242	NC1=NC(=O)C2=C(N1)N=CN2
35	C02862	Butyrylcarnitine	HMDB0002013	213144	C02862	CCCC(=O)O[C@H](CC([O-])=O)C[N+](C)(
36	C00864	Pantothenic acid	HMDB0000210	6613	C00864	CC(C)(CO)[C@@H](O)C(=O)NCCC(O)=O
37	C02571	L-Acetylcarnitine	HMDB0000201	7045767	C02571	CC(=O)O[C@H](CC(O)=O)C[N+](C)(C)C
38	C00029	Uridine diphosphate glucose	HMDB0000286	8629	C00029	OC[C@H]1O[C@H](OP(O)(=O)OP(O)(=O)O
39	C00147	Adenine	HMDB0000034	190	C00147	NC1=C2NC=NC2=NC=N1
40	C00284	Edetic Acid	HMDB0015109	6049	C00284	OC(=O)CN(CCN(CC(O)=O)CC(O)=O)CC(
41	C02305	Phosphocreatine	HMDB0001511	587	C02305	CN(CC(O)=O)C(=N)NP(O)(O)=O
42	C00570	CDP-ethanolamine	HMDB0001564	123727	C00570	NCCOP(O)(=O)OP(O)(=O)OC[C@H]1O[C@
43	C05526	S-Glutathionyl-L-cysteine	METPA0607		C05526	
44	C00946	Adenosine 2'-phosphate	HMDB0011617	53481006	C00946	NC1=NC=NC2=C1N=CN2C1O[C@H](CO)[
45	C05282	gamma-Glutamylglutamic acid	HMDB0011737	92865	C05282	N[C@@H](CCC(=O)N[C@@H](CCC(O)=O)(
46	C00055	Cytidine monophosphate	HMDB0000095	6131	C00055	NC1=NC(=O)N(C=C1)[C@@H]1O[C@H](CO
47	C00043	Uridine diphosphate-N-acetylglucosamine	HMDB0000290	445675	C00043	CC(=O)N[C@@H]1[C@@H](O)[C@H](O)[C@
48	C00103	Glucose 1-phosphate	HMDB0001586	65533	C00103	OC[C@H]1O[C@H](OP(O)(O)=O)[C@H](O)

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.

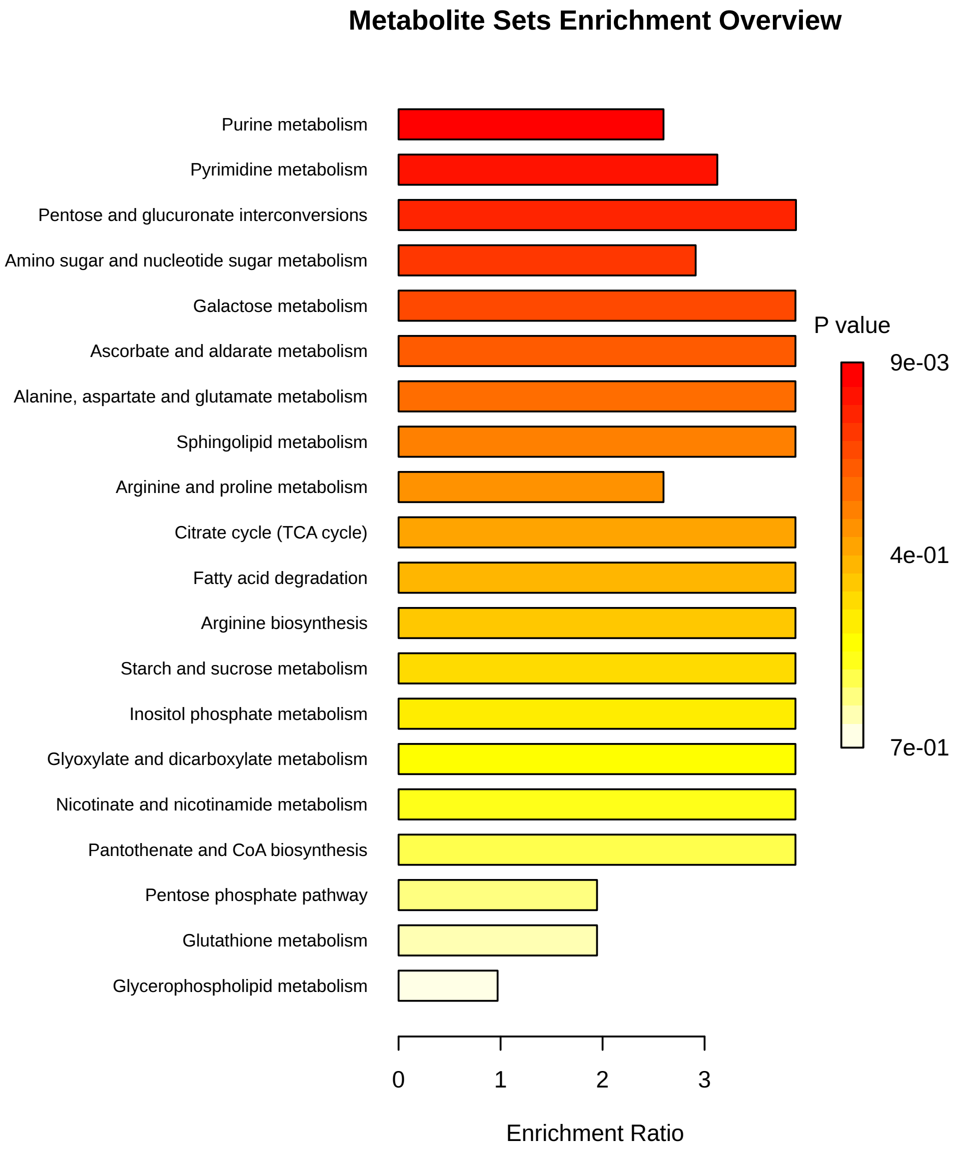

Figure 1: Summary Plot for Over Representation Analysis (ORA)

Table 2: Result from Over Representation Analysis

	total	expected	hits	Raw p	Holm p	FDR
Purine metabolism	9	2.31	6	9.24E-03	7.39E-01	4.25E-01
Pyrimidine metabolism	5	1.28	4	1.56E-02	1.00E+00	4.25E-01
Pentose and glucuronate interconversions	3	0.77	3	1.59E-02	1.00E+00	4.25E-01
Amino sugar and nucleotide sugar metabolism	4	1.03	3	5.22E-02	1.00E+00	6.46E-01
Galactose metabolism	2	0.51	2	6.46E-02	1.00E+00	6.46E-01
Ascorbate and aldarate metabolism	2	0.51	2	6.46E-02	1.00E+00	6.46E-01
Alanine, aspartate and glutamate metabolism	2	0.51	2	6.46E-02	1.00E+00	6.46E-01
Sphingolipid metabolism	2	0.51	2	6.46E-02	1.00E+00	6.46E-01
Arginine and proline metabolism	3	0.77	2	1.62E-01	1.00E+00	1.00E+00
Citrate cycle (TCA cycle)	1	0.26	1	2.57E-01	1.00E+00	1.00E+00
Fatty acid degradation	1	0.26	1	2.57E-01	1.00E+00	1.00E+00
Arginine biosynthesis	1	0.26	1	2.57E-01	1.00E+00	1.00E+00
Starch and sucrose metabolism	1	0.26	1	2.57E-01	1.00E+00	1.00E+00
Inositol phosphate metabolism	1	0.26	1	2.57E-01	1.00E+00	1.00E+00
Glyoxylate and dicarboxylate metabolism	1	0.26	1	2.57E-01	1.00E+00	1.00E+00
Nicotinate and nicotinamide metabolism	1	0.26	1	2.57E-01	1.00E+00	1.00E+00
Pantothenate and CoA biosynthesis	1	0.26	1	2.57E-01	1.00E+00	1.00E+00
Pentose phosphate pathway	2	0.51	1	4.49E-01	1.00E+00	1.00E+00
Glutathione metabolism	2	0.51	1	4.49E-01	1.00E+00	1.00E+00
Glycerophospholipid metabolism	4	1.03	1	6.99E-01	1.00E+00	1.00E+00

7 Appendix: R Command History

```
 [1] "mSet<-InitDataObjects(\"conc\", \"msetora\", FALSE)"
 [2] "cmpd.vec<-c(\"C03017\",\"C07005\",\"C00062\",\"C00003\",\"C00307\",\"C00127\",\"C00105\",\"C00
 [3] "mSet<-Setup.MapData(mSet, cmpd.vec);"
 [4] "mSet<-CrossReferencing(mSet, \"kegg\");"
 [5] "mSet<-CreateMappingResultTable(mSet)"
 [6] "mSet<-Setup.HMDBReferenceMetabolome(mSet, \"Phena.txt\");"
 [7] "mSet<-SetMetabolomeFilter(mSet, T);"
 [8] "mSet<-SetCurrentMsetLib(mSet, \"kegg_pathway\", 2);"
 [9] "mSet<-CalculateHyperScore(mSet)"
[10] "mSet<-PlotORA(mSet, \"ora_0_\", \"net\", \"png\", 72, width=NA)"
[11] "mSet<-PlotEnrichDotPlot(mSet, \"ora\", \"ora_dot_0_\", \"png\", 72, width=NA)"
[12] "mSet<-CalculateHyperScore(mSet)"
[13] "mSet<-PlotORA(mSet, \"ora_1_\", \"net\", \"png\", 72, width=NA)"
[14] "mSet<-PlotEnrichDotPlot(mSet, \"ora\", \"ora_dot_1_\", \"png\", 72, width=NA)"
[15] "mSet<-CalculateHyperScore(mSet)"
[16] "mSet<-PlotORA(mSet, \"ora_2_\", \"net\", \"png\", 72, width=NA)"
[17] "mSet<-PlotEnrichDotPlot(mSet, \"ora\", \"ora_dot_2_\", \"png\", 72, width=NA)"
[18] "mSet<-CalculateHyperScore(mSet)"
[19] "mSet<-PlotORA(mSet, \"ora_3_\", \"net\", \"png\", 72, width=NA)"
[20] "mSet<-PlotEnrichDotPlot(mSet, \"ora\", \"ora_dot_3_\", \"png\", 72, width=NA)"
[21] "mSet<-CalculateHyperScore(mSet)"
[22] "mSet<-PlotORA(mSet, \"ora_4_\", \"net\", \"png\", 72, width=NA)"
[23] "mSet<-PlotEnrichDotPlot(mSet, \"ora\", \"ora_dot_4_\", \"png\", 72, width=NA)"
[24] "mSet<-SaveTransformedData(mSet)"
[25] "mSet<-PreparePDFReport(mSet, \"guest9044362343704604219\")\n"
```

The report was generated on Tue Jan 7 09:57:03 2025 with R version 4.3.2 (2023-10-31), OS system: Linux.