Metabolomic Data Analysis with MetaboAnalyst 5.0

Name: guest8235001142335868957

November 24, 2022

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 Compound Name Mapping

	Query	Match	HMDB	PubChem	KEGG	SMILES
1	HMDB0240751	NA	NA	NA	NA	NA
2	HMDB0003099	1-Methyluric acid	HMDB0003099	69726	C16359	CN1C(=O)C2=C(NC(=O)N2)NC1=O
3	HMDB0240545	N-Methylpyridinium	HMDB0240545	13597	C02724	C[N+]1=CC=CC=C1
4	HMDB0006029	N-Acetylglutamine	HMDB0006029	25561		CC(=O)NC(CCC(=O)N)C(=O)O
5	HMDB0001860	Paraxanthine	HMDB0001860	4687	C13747	CN1C=NC2=C1C(=O)N(C(=O)N2)C
6	HMDB0000017	4-Pyridoxic acid	HMDB0000017	6723	C00847	CC1=NC=C(C(=C1O)C(=O)O)CO
7	HMDB0011103	1,7-Dimethyluric acid	HMDB0011103	91611	C16356	CN1C2=C(NC1=O)NC(=O)N(C2=O)C
8	HMDB0013288	Nonanoylcarnitine	HMDB0013288	53481660		CCCCCCCCC(=O)OC(CC(=O)[O-])C[N+](C
9	HMDB0001138	N-Acetylglutamic acid	HMDB0001138	185	C00624	CC(=O)NC(CCC(=O)O)C(=O)O
10	HMDB0000244	Riboflavin	HMDB0000244	493570	C00255	CC1=CC2=C(C=C1C)N(C3=NC(=O)NC(=C
11	HMDB0004824	N2,N2-Dimethylguanosine	HMDB0004824	92919		CN(C)C1=NC(=O)C2=C(N1)N(C=N2)[C@H
12	HMDB0061384	NA	NA	NA	NA	NA
13	HMDB0000230	N-Acetylneuraminic acid	HMDB0000230	445063	C19910	CC(=O)N[C@@H]1[C@H](C[C@](O[C@H]1[C
14	HMDB0003072	Quinic acid	HMDB0003072	6508	C00296	OC1C[C@@](O)(C[C@@H](O)[C@H]1O)C(O)=
15	HMDB0155722	NA	NA	NA	NA	NA
16	HMDB0000158	L-Tyrosine	HMDB0000158	6057	C00082	C1=CC(=CC=C1C[C@@H](C(=O)O)N)O
17	HMDB0013676	2,6-Dihydroxybenzoic acid	HMDB0013676	9338	C21298	C1=CC(=C(C(=C1)O)C(=O)O)O
18	HMDB0000267	Pyroglutamic acid	HMDB0000267	7405	C01879	C1CC(=O)N[C@@H]1C(=O)O
19	HMDB0000812	N-Acetyl-L-aspartic acid	HMDB0000812	65065	C01042	CC(=O)N[C@@H](CC(=O)O)C(=O)O
20	HMDB0000439	2-Furoylglycine	HMDB0000439	21863		C1=COC(=C1)C(=O)NCC(=O)O
21	HMDB0000678	Isovalerylglycine	HMDB0000678	546304		CC(C)CC(=O)NCC(=O)O
22	HMDB0029992	Tetrahydropentoxyline	HMDB0029992	53481442		C1C(NC(C2=C1C3=CC=CC=C3N2)C(C(C(
23	HMDB0000289	Uric acid	HMDB0000289	1175	C00366	C12=C(NC(=O)N1)NC(=O)NC2=O
24	HMDB0000072	cis-Aconitic acid	HMDB0000072	643757	C00417	C(/C(=C/C(=O)O)/C(=O)O)C(=O)O
25	HMDB0000097	Choline	HMDB0000097	305	C00114	C[N+](C)(C)CCO
26	HMDB0000197	Indoleacetic acid	HMDB0000197	802	C00954	C1=CC=C2C(=C1)C(=CN2)CC(=O)O
27	HMDB0000193	Isocitric acid	HMDB0000193	1198	C00311	C(C(C(C(=O)O)O)C(=O)O)C(=O)O
28	HMDB0062555	hydroxyisovaleroyl carnitine	HMDB0062555	57357187		CC(C)CC(=O)OC(O)(CC([O-])=O)C[N+](C)
29	HMDB0003157	Guanidinosuccinic acid	HMDB0003157	439918	C03139	C([C@@H](C(=O)O)N=C(N)N)C(=O)O
30	HMDB0000893	Suberic acid	HMDB0000893	10457	C08278	C(CCCC(=O)O)CCC(=O)O
31	HMDB0000132	Guanine	HMDB0000132	764	C00242	C1=NC2=C(N1)C(=O)N=C(N2)N
32	HMDB0012296	Trimethylaminoacetone	HMDB0012296	151806		CC(=O)C[N+](C)(C)C
33	HMDB0000058	Cyclic AMP	HMDB0000058	6076	C00575	C1[C@@H]2[C@H]([C@H]([C@@H](O2)N3C=N
34	HMDB0000912	Succinyladenosine	HMDB0000912	20849086		C1=NC2=C(C(=N1)N[C@@H](CC(=O)O)C(=
35	HMDB0000875	Trigonelline	HMDB0000875	5570	C01004	C[N+]1=CC=CC(=C1)C(=O)[O-]
36	HMDB0000929	L-Tryptophan	HMDB0000929	6305	C00078	C1=CC=C2C(=C1)C(=CN2)C[C@@H](C(=O
37	HMDB0244966	NA	NA	NA	NA	NA
38	HMDB0000440	3-Hydroxyphenylacetic acid	HMDB0000440	12122	C05593	C1=CC(=CC(=C1)O)CC(=O)O
39	HMDB0000235	Thiamine	HMDB0000235	1130	C00378	CC1=C(SC=[N+]1CC2=CN=C(N=C2N)C)C
40	HMDB0002432	Sumiki's acid	HMDB0002432	80642	C20448	C1=C(OC(=C1)C(=O)O)CO
41	HMDB0010319	Inodxyl glucuronide	HMDB0010319	2733785	C03033	C1=CC=C2C(=C1)C(=CN2)O[C@H]3[C@@H
42	HMDB0001987	2-Hydroxy-2-methylbutyric acid	HMDB0001987	95433		CCC(C)(C(=O)O)O
43	HMDB0002024	Imidazoleacetic acid	HMDB0002024	96215	C02835	C1=C(NC=N1)CC(=O)O
44	HMDB0001406	Niacinamide	HMDB0001406	936	C00153	C1=CC(=CN=C1)C(=O)N
45	HMDB0000201	L-Acetylcarnitine	HMDB0000201	7045767	C02571	CC(=O)OC(CC(=O)[O-])C[N+](C)(C)C
46	HMDB0001713	m-Coumaric acid	HMDB0001713	637541	C12621	C1=CC(=CC(=C1)O)/C=C/C(=O)O
47	HMDB60001	NA	NA	NA	NA	NA
48	HMDB0000842	Quinaldic acid	HMDB0000842	7124	C06325	C1=CC=C2C(=C1)C=CC(=N2)C(=O)O
49	HMDB0000355	3-Hydroxymethylglutaric acid	HMDB0000355	1662	C03761	CC(CC(=O)O)(CC(=O)O)O
50	HMDB0000254	Succinic acid	HMDB0000254	1110	C00042	C(CC(=O)O)C(=O)O
51	HMDB0000138	Glycocholic acid	HMDB0000138	23617285	C01921	C[C@H](CCC(=O)NCC(=O)O)[C@H]1CC[C@
52	HMDB0000418	18-Hydroxycortisol	HMDB0000418	44263343		C[C@]12CCC(=O)C=C1CC[C@@H]3[C@@H]2
53	HMDB0000262	Thymine	HMDB0000262	1135	C00178	CC1=CNC(=O)NC1=O
54	HMDB0000669	Ortho-Hydroxyphenylacetic acid	HMDB0000669	11970	C05852	C1=CC=C(C(=C1)CC(=O)O)O
55	HMDB0000133	Guanosine	HMDB0000133	6802	C00387	C1=NC2=C(N1[C@H]3[C@@H]([C@@H]([C@@
56	HMDB0000630	Cytosine	HMDB0000630	597	C00380	C1=C(NC(=O)N=C1)N
57	HMDB0004148	Dopamine 4-sulfate	HMDB0004148	123932	C13691	C1=CC(=C(C=C1CCN)O)OS(=O)(=O)O
58	HMDB0000625	Gluconic acid	HMDB0000625	10690	C00257	C([C@H]([C@H]([C@@H]([C@H](C(=O)O)O)O

59	HMDB0002894	5-Methylcytosine	HMDB0002894	65040	C02376	CC1=C(NC(=O)N=C1)N
60	HMDB0000162	L-Proline	HMDB0000162	145742	C00148	C1C[C@H](NC1)C(=O)O
61	HMDB0028933	Leucyl-Leucine	HMDB0028933	76807	C11332	CC(C)CC(N)C(=O)NC(CC(C)C)C(O)=O
62	HMDB0060015	Phenyl hydrogen sulfate	HMDB0060015	74426	C02180	OS(=O)(=O)OC1=CC=CC=C1
63	HMDB0000661	Glutaric acid	HMDB0000661	743	C00489	C(CC(=O)O)CC(=O)O
64	HMDB0013713	N-acetyltryptophan	HMDB0013713	700653		[H][C@@](CC1=CNC2=CC=CC=C12)(N=C(
65	HMDB0002730	Nicotinamide N-oxide	HMDB0002730	72661		C1=CC(=C[N+](=C1)[O-])C(=O)N
66	HMDB0000714	Hippuric acid	HMDB0000714	464	C01586	C1=CC=C(C=C1)C(=O)NCC(=O)O
67	HMDB0001297	Norcotinine	HMDB0001297	413		C1CC(=O)NC1C2=CN=CC=C2
68	HMDB0000172	L-Isoleucine	HMDB0000172	6306	C00407	CC[C@H](C)[C@@H](C(=O)O)N
69	HMDB00191	L-Aspartic acid	HMDB0000191	5960	C00049	C([C@@H](C(=O)O)N)C(=O)O
70	HMDB0000020	p-Hydroxyphenylacetic acid	HMDB0000020	127	C00642	C1=CC(=CC=C1CC(=O)O)O
71	HMDB0002035	4-Hydroxycinnamic acid	HMDB0002035	637542	C00811	C1=CC(=CC=C1/C=C/C(=O)O)O
72	HMDB0001886	3-Methylxanthine	HMDB0001886	70639	C16357	CN1C2=C(C(=O)NC1=O)NC=N2

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.

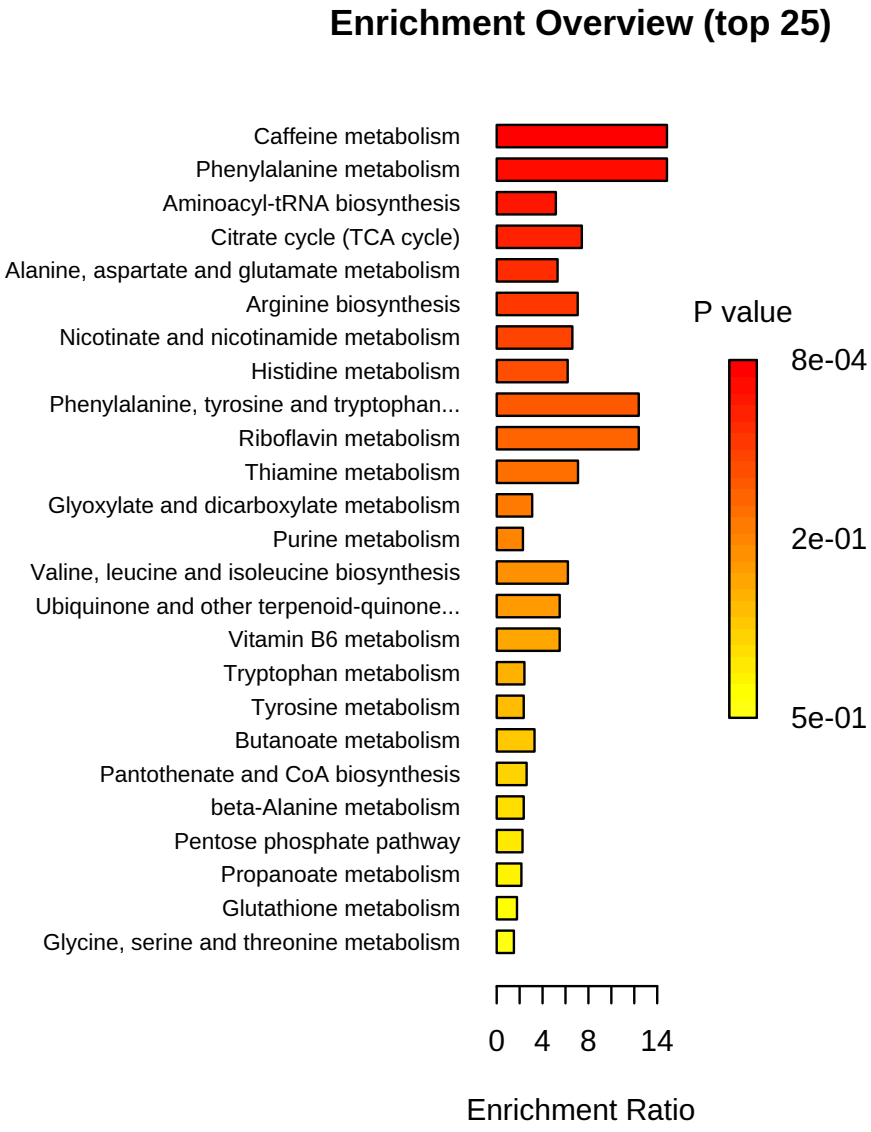

Figure 1: Summary Plot for Over Representation Analysis (ORA)

Table 2: Result from Over Representation Analysis

	total	expected	hits	Raw p	Holm p	FDR
Caffeine metabolism	10	0.20	3	8.13E-04	6.83E-02	3.41E-02
Phenylalanine metabolism	10	0.20	3	8.13E-04	6.83E-02	3.41E-02
Aminoacyl-tRNA biosynthesis	48	0.97	5	2.22E-03	1.82E-01	6.23E-02
Citrate cycle (TCA cycle)	20	0.40	3	6.73E-03	5.45E-01	1.41E-01
Alanine, aspartate and glutamate metabolism	28	0.56	3	1.73E-02	1.00E+00	2.91E-01
Arginine biosynthesis	14	0.28	2	3.09E-02	1.00E+00	4.17E-01
Nicotinate and nicotinamide metabolism	15	0.30	2	3.52E-02	1.00E+00	4.17E-01
Histidine metabolism	16	0.32	2	3.97E-02	1.00E+00	4.17E-01
Phenylalanine, tyrosine and tryptophan biosynthesis	4	0.08	1	7.84E-02	1.00E+00	6.58E-01
Riboflavin metabolism	4	0.08	1	7.84E-02	1.00E+00	6.58E-01
Thiamine metabolism	7	0.14	1	1.33E-01	1.00E+00	8.82E-01
Glyoxylate and dicarboxylate metabolism	32	0.65	2	1.35E-01	1.00E+00	8.82E-01
Purine metabolism	65	1.31	3	1.40E-01	1.00E+00	8.82E-01
Valine, leucine and isoleucine biosynthesis	8	0.16	1	1.51E-01	1.00E+00	8.82E-01
Ubiquinone and other terpenoid-quinone biosynthesis	9	0.18	1	1.68E-01	1.00E+00	8.82E-01
Vitamin B6 metabolism	9	0.18	1	1.68E-01	1.00E+00	8.82E-01
Tryptophan metabolism	41	0.83	2	1.99E-01	1.00E+00	9.66E-01
Tyrosine metabolism	42	0.85	2	2.07E-01	1.00E+00	9.66E-01
Butanoate metabolism	15	0.30	1	2.65E-01	1.00E+00	1.00E+00
Pantothenate and CoA biosynthesis	19	0.38	1	3.23E-01	1.00E+00	1.00E+00
beta-Alanine metabolism	21	0.42	1	3.50E-01	1.00E+00	1.00E+00
Pentose phosphate pathway	22	0.44	1	3.63E-01	1.00E+00	1.00E+00
Propanoate metabolism	23	0.46	1	3.76E-01	1.00E+00	1.00E+00
Glutathione metabolism	28	0.56	1	4.38E-01	1.00E+00	1.00E+00
Glycine, serine and threonine metabolism	33	0.67	1	4.93E-01	1.00E+00	1.00E+00
Glycerophospholipid metabolism	36	0.73	1	5.24E-01	1.00E+00	1.00E+00
Amino sugar and nucleotide sugar metabolism	37	0.75	1	5.34E-01	1.00E+00	1.00E+00
Arginine and proline metabolism	38	0.77	1	5.44E-01	1.00E+00	1.00E+00
Pyrimidine metabolism	39	0.79	1	5.53E-01	1.00E+00	1.00E+00
Valine, leucine and isoleucine degradation	40	0.81	1	5.62E-01	1.00E+00	1.00E+00
Primary bile acid biosynthesis	46	0.93	1	6.14E-01	1.00E+00	1.00E+00

7 Appendix: R Command History

```
 [1] "mSet<-InitDataObjects(\"conc\", \"msetora\", FALSE)"
 [2] "cmpd.vec<-c(\"HMDB0240751\",\"HMDB0003099\",\"HMDB0240545\",\"HMDB0006029\",\"HMDB0001860\",\"
 [3] "mSet<-Setup.MapData(mSet, cmpd.vec);"
 [4] "mSet<-CrossReferencing(mSet, \"hmdb\");"
 [5] "mSet<-CreateMappingResultTable(mSet)"
 [6] "mSet<-SetMetabolomeFilter(mSet, F);"
 [7] "mSet<-SetCurrentMsetLib(mSet, \"kegg_pathway\", 2);"
 [8] "mSet<-CalculateHyperScore(mSet)"
 [9] "mSet<-PlotORA(mSet, \"ora_0_\", \"net\", \"png\", 72, width=NA)"
[10] "mSet<-PlotEnrichDotPlot(mSet, \"ora\", \"ora_dot_0_\", \"png\", 72, width=NA)"
[11] "mSet<-CalculateHyperScore(mSet)"
[12] "mSet<-PlotORA(mSet, \"ora_1_\", \"net\", \"png\", 72, width=NA)"
[13] "mSet<-PlotEnrichDotPlot(mSet, \"ora\", \"ora_dot_1_\", \"png\", 72, width=NA)"
[14] "mSet<-SetMetabolomeFilter(mSet, F);"
[15] "mSet<-SetCurrentMsetLib(mSet, \"kegg_pathway\", 2);"
[16] "mSet<-CalculateHyperScore(mSet)"
[17] "mSet<-PlotORA(mSet, \"ora_2_\", \"net\", \"png\", 72, width=NA)"
[18] "mSet<-PlotEnrichDotPlot(mSet, \"ora\", \"ora_dot_2_\", \"png\", 72, width=NA)"
[19] "mSet<-CalculateHyperScore(mSet)"
[20] "mSet<-PlotORA(mSet, \"ora_3_\", \"net\", \"png\", 72, width=NA)"
[21] "mSet<-PlotEnrichDotPlot(mSet, \"ora\", \"ora_dot_3_\", \"png\", 72, width=NA)"
[22] "mSet<-SaveTransformedData(mSet)"
[23] "mSet<-PreparePDFReport(mSet, \"guest8235001142335868957\")\n"
```

The report was generated on Thu Nov 24 07:31:52 2022 with R version 4.2.2 (2022-10-31), OS system: Linux, version: -Ubuntu SMP Thu Oct 13 08:03:55 UTC 2022 .