Metabolomic Data Analysis with MetaboAnalyst 5.0

Name: guest10480732887989101541

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	HMDB0000244	Riboflavin	HMDB0000244	493570	C00255	CC1=CC2=C(C=C1C)N(C3=NC(=O)NC(
2	HMDB0000446	N-Alpha-acetyllysine	HMDB0000446	192590	C12989	CC(=O)NC(CCCCN)C(=O)O
3	HMDB0000251	Taurine	HMDB0000251	1123	C00245	C(CS(=O)(=O)O)N
4	HMDB0002172	N1,N12-Diacetylspermine	HMDB0002172	132680	C03413	CC(=O)NCCCNCCCCNCCCNC(=O)C
5	HMDB0000893	Suberic acid	HMDB0000893	10457	C08278	C(CCCC(=O)O)CCC(=O)O
6	HMDB0000626	Deoxycholic acid	HMDB0000626	222528	C04483	C[C@H](CCC(=O)O)[C@H]1CC[C@@H]2[C
7	HMDB0000157	Hypoxanthine	HMDB0000157	790	C00262	C1=NC2=C(N1)C(=O)N=CN2
8	HMDB0000752	Methylglutaric acid	HMDB0000752	12284		CC(CC(=O)O)CC(=O)O
9	HMDB0000448	Adipic acid	HMDB0000448	196	C06104	C(CCC(=O)O)CC(=O)O
10	HMDB0000929	L-Tryptophan	HMDB0000929	6305	C00078	C1=CC=C2C(=C1)C(=CN2)C[C@@H](C(
11	HMDB0000500	4-Hydroxybenzoic acid	HMDB0000500	135	C00156	C1=CC(=CC=C1C(=O)O)O
12	HMDB0000422	2-Methylglutaric acid	HMDB0000422	12046		CC(CCC(=O)O)C(=O)O
13	HMDB0003331	1-Methyladenosine	HMDB0003331	27476	C02494	CN1C=NC2=C(C1=N)N=CN2[C@H]3[C@
14	HMDB0000159	L-Phenylalanine	HMDB0000159	6140	C00079	C1=CC=C(C=C1)C[C@@H](C(=O)O)N
15	HMDB0000254	Succinic acid	HMDB0000254	1110	C00042	C(CC(=O)O)C(=O)O
16	HMDB0001138	N-Acetylglutamic acid	HMDB0001138	185	C00624	CC(=O)NC(CCC(=O)O)C(=O)O
17	HMDB0000661	Glutaric acid	HMDB0000661	743	C00489	C(CC(=O)O)CC(=O)O
18	HMDB0000631	Deoxycholic acid glycine conjugate	HMDB0000631	3035026	C05464	C[C@H](CCC(=O)NCC(=O)O)[C@H]1CCC
19	HMDB0000784	Azelaic acid	HMDB0000784	2266	C08261	C(CCCC(=O)O)CCCC(=O)O
20	HMDB0004620	N-a-Acetyl-L-arginine	HMDB0004620	67427		CC(=O)N[C@@H](CCCN=C(N)N)C(=O)O
21	HMDB0000678	Isovalerylglycine	HMDB0000678	546304		CC(C)CC(=O)NCC(=O)O
22	HMDB0001032	Dehydroepiandrosterone sulfate	HMDB0001032	12594	C04555	C[C@]12CC[C@H]3[C@H]([C@@H]1CCC2=
23	HMDB0011103	1,7-Dimethyluric acid	HMDB0011103	91611	C16356	CN1C2=C(NC1=O)NC(=O)N(C2=O)C
24	HMDB0000708	Glycoursodeoxycholic acid	HMDB0000708	12310288		C[C@H](CCC(=O)NCC(=O)O)[C@H]1CC[
25	HMDB0002123	1,3,7-Trimethyluric acid	HMDB0002123	79437	C16361	CN1C2=C(NC1=O)N(C(=O)N(C2=O)C)C
26	HMDB0003334	Symmetric dimethylarginine	HMDB0003334	169148		CNC(=NC)NCCC[C@@H](C(=O)O)N
27	HMDB0013677	3,5-Dihydroxybenzoic acid	HMDB0013677	7424	C00180	C1=C(C=C(C=C1O)O)C(=O)O
28	HMDB0028942	Leucyl-Valine	HMDB0028942	6993116		CC(C)CC(N)C(=O)NC(C(C)C)C(O)=O
29	HMDB0001844	Methylsuccinic acid	HMDB0001844	10349	C08645	CC(CC(=O)O)C(=O)O
30	HMDB0244966	NA	NA	NA	NA	NA
31	HMDB0255727	NA	NA	NA	NA	NA
32	HMDB0000730	Isobutyrylglycine	HMDB0000730	10855600		CC(C)C(=O)NCC(=O)O
33	HMDB0005807	Gallic acid	HMDB0005807	370	C01424	C1=C(C=C(C(=C1O)O)O)C(=O)O
34	HMDB0000687	L-Leucine	HMDB0000687	6106	C00123	CC(C)C[C@@H](C(=O)O)N
35	HMDB0000881	Xanthurenic acid	HMDB0000881	5699	C02470	C1=CC2=C(C(=C1)O)NC(=CC2=O)C(=
36	HMDB0000956	Tartaric acid	HMDB0000956	444305	C00898	O[C@H]([C@@H](O)C(O)=O)C(O)=O
37	HMDB0000729	Alpha-Hydroxyisobutyric acid	HMDB0000729	11671		CC(C)(C(=O)O)O
38	HMDB0062640	3-hydroxy-2-isobutyrate	HMDB0062640	87	C01188	CC(CO)C(O)=O
39	HMDB0001991	7-Methylxanthine	HMDB0001991	68374	C16353	CN1C=NC2=C1C(=O)NC(=O)N2
40	HMDB0061384	NA	NA	NA	NA	NA
41	HMDB0013713	N-acetyltryptophan	HMDB0013713	700653		[H][C@@](CC1=CNC2=CC=CC=C12)(N=
42	HMDB0000152	Gentisic acid	HMDB0000152	3469	C00628	C1=CC(=C(C=C1O)C(=O)O)O
43	HMDB0000301	Urocanic acid	HMDB0000301	736715	C00785	C1=C(NC=N1)/C=C/C(=O)O
44	HMDB0001847	Caffeine	HMDB0001847	2519	C07481	CN1C=NC2=C1C(=O)N(C(=O)N2C)C
45	HMDB0000822	p-Hydroxymandelic acid	HMDB0000822	7721	C11527	C1=CC(=CC=C1C(C(=O)O)O)O
46	HMDB0001406	Niacinamide	HMDB0001406	936	C00153	C1=CC(=CN=C1)C(=O)N
47	HMDB0012275	Phenylethylamine	HMDB0012275	1001	C05332	C1=CC=C(C=C1)CCN
48	HMDB0000226	Orotic acid	HMDB0000226	967	C00295	C1=C(NC(=O)NC1=O)C(=O)O
49	HMDB0006029	N-Acetylglutamine	HMDB0006029	25561		CC(=O)NC(CCC(=O)N)C(=O)O
50	HMDB0001325	N6,N6,N6-Trimethyl-L-lysine	HMDB0001325	440120	C03793	C[N+](C)(C)CCCC[C@@H](C(=O)[O-])N
51	HMDB0000235	Thiamine	HMDB0000235	1130	C00378	CC1=C(SC=[N+]1CC2=CN=C(N=C2N)C)
52	HMDB0013676	2,6-Dihydroxybenzoic acid	HMDB0013676	9338	C21298	C1=CC(=C(C(=C1)O)C(=O)O)O
53	HMDB0000721	Glycylproline	HMDB0000721	79101		C1CC(N(C1)C(=O)CN)C(=O)O
54	HMDB0000641	L-Glutamine	HMDB0000641	5961	C00064	C(CC(=O)N)[C@@H](C(=O)O)N
55	HMDB0000158	L-Tyrosine	HMDB0000158	6057	C00082	C1=CC(=CC=C1C[C@@H](C(=O)O)N)O
56	HMDB0000355	3-Hydroxymethylglutaric acid	HMDB0000355	1662	C03761	CC(CC(=O)O)(CC(=O)O)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.

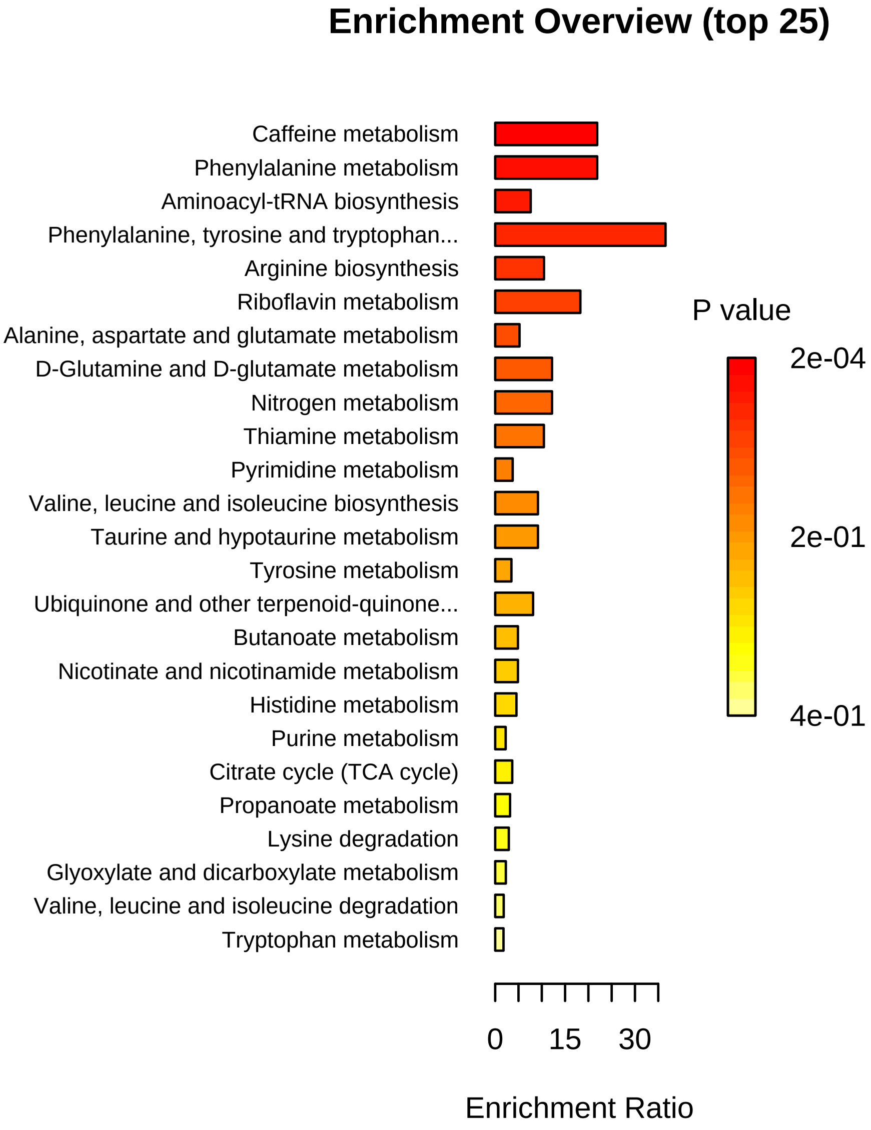

Figure 1: Summary Plot for Over Representation Analysis (ORA)

Table 2: Result from Over Representation Analysis

	total	expected	hits	Raw p	Holm p	FDR
Caffeine metabolism	10	0.14	3	2.49E-04	2.09E-02	9.44E-03
Phenylalanine metabolism	10	0.14	3	2.49E-04	2.09E-02	9.44E-03
Aminoacyl-tRNA biosynthesis	48	0.66	5	3.37E-04	2.76E-02	9.44E-03
Phenylalanine, tyrosine and tryptophan biosynthesis	4	0.05	2	1.05E-03	8.52E-02	2.21E-02
Arginine biosynthesis	14	0.19	2	1.47E-02	1.00E+00	2.47E-01
Riboflavin metabolism	4	0.05	1	5.36E-02	1.00E+00	6.52E-01
Alanine, aspartate and glutamate metabolism	28	0.38	2	5.44E-02	1.00E+00	6.52E-01
D-Glutamine and D-glutamate metabolism	6	0.08	1	7.94E-02	1.00E+00	6.54E-01
Nitrogen metabolism	6	0.08	1	7.94E-02	1.00E+00	6.54E-01
Thiamine metabolism	7	0.10	1	9.20E-02	1.00E+00	6.54E-01
Pyrimidine metabolism	39	0.53	2	9.74E-02	1.00E+00	6.54E-01
Valine, leucine and isoleucine biosynthesis	8	0.11	1	1.05E-01	1.00E+00	6.54E-01
Taurine and hypotaurine metabolism	8	0.11	1	1.05E-01	1.00E+00	6.54E-01
Tyrosine metabolism	42	0.57	2	1.11E-01	1.00E+00	6.54E-01
Ubiquinone and other terpenoid-quinone biosynthesis	9	0.12	1	1.17E-01	1.00E+00	6.54E-01
Butanoate metabolism	15	0.20	1	1.87E-01	1.00E+00	9.26E-01
Nicotinate and nicotinamide metabolism	15	0.20	1	1.87E-01	1.00E+00	9.26E-01
Histidine metabolism	16	0.22	1	1.99E-01	1.00E+00	9.27E-01
Purine metabolism	65	0.89	2	2.22E-01	1.00E+00	9.82E-01
Citrate cycle (TCA cycle)	20	0.27	1	2.42E-01	1.00E+00	1.00E+00
Propanoate metabolism	23	0.31	1	2.73E-01	1.00E+00	1.00E+00
Lysine degradation	25	0.34	1	2.93E-01	1.00E+00	1.00E+00
Glyoxylate and dicarboxylate metabolism	32	0.44	1	3.59E-01	1.00E+00	1.00E+00
Valine, leucine and isoleucine degradation	40	0.55	1	4.28E-01	1.00E+00	1.00E+00
Tryptophan metabolism	41	0.56	1	4.36E-01	1.00E+00	1.00E+00
Primary bile acid biosynthesis	46	0.63	1	4.74E-01	1.00E+00	1.00E+00
Steroid hormone biosynthesis	85	1.16	1	7.00E-01	1.00E+00	1.00E+00

7 Appendix: R Command History

```
 [1] "mSet<-InitDataObjects(\"conc\", \"msetora\", FALSE)"
 [2] "cmpd.vec<-c(\"HMDB0000244\",\"HMDB0000446\",\"HMDB0000251\",\"HMDB0002172\",\"HMDB0000893\",\"
 [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<-SaveTransformedData(mSet)"
[15] "mSet<-PreparePDFReport(mSet, \"guest10480732887989101541\")\n"
```

The report was generated on Thu Nov 24 06:25:26 2022 with R version 4.1.3 (2022-03-10), OS system: Linux, version: 20.04.2-Ubuntu SMP Wed Aug 17 02:46:40 UTC 2022 .