Metabolomic Data Analysis with MetaboAnalyst 5.0

Name: guest1100685848573538732

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 Map

	Query	Match	HMDB	PubChem	KEGG	SMILES
1	HMDB0000752	Methylglutaric acid	HMDB0000752	12284		CC(CC(=O)O)CC(=O)O
2	HMDB0000448	Adipic acid	HMDB0000448	196	C06104	C(CCC(=O)O)CC(=O)O
3	HMDB0000893	Suberic acid	HMDB0000893	10457	C08278	C(CCCC(=O)O)CCC(=O)
4	HMDB0001032	Dehydroepiandrosterone sulfate	HMDB0001032	12594	C04555	C[C@]12CC[C@H]3[C@H]([
5	HMDB0000262	Thymine	HMDB0000262	1135	C00178	CC1=CNC(=O)NC1=O
6	HMDB0000148	L-Glutamic acid	HMDB0000148	33032	C00025	C(CC(=O)O)[C@@H](C(=
7	HMDB0000391	7-Ketodeoxycholic acid	HMDB0000391	188292	C04643	C[C@H](CCC(=O)O)[C@H
8	HMDB0000784	Azelaic acid	HMDB0000784	2266	C08261	C(CCCC(=O)O)CCCC(=
9	HMDB0000244	Riboflavin	HMDB0000244	493570	C00255	CC1=CC2=C(C=C1C)N(C
10	HMDB0002172	N1,N12-Diacetylspermine	HMDB0002172	132680	C03413	CC(=O)NCCCNCCCCNCC
11	HMDB0000251	Taurine	HMDB0000251	1123	C00245	C(CS(=O)(=O)O)N
12	HMDB0000721	Glycylproline	HMDB0000721	79101		C1CC(N(C1)C(=O)CN)C(=
13	HMDB0000127	D-Glucuronic acid	HMDB0000127	94715	C00191	[C@@H]1([C@@H]([C@H](C
14	HMDB0000755	Hydroxyphenyllactic acid	HMDB0000755	9378	C03672	C1=CC(=CC=C1CC(C(=
15	HMDB0002643	3-(3-Hydroxyphenyl)-3-hydroxypropanoic acid	HMDB0002643	102959		C1=CC(=CC(=C1)O)C(CC
16	HMDB0000917	Ursocholic acid	HMDB0000917	122340	C17644	C[C@H](CCC(=O)O)[C@H
17	HMDB0000407	2-Hydroxy-3-methylbutyric acid	HMDB0000407	99823		CC(C)C(C(=O)O)O
18	HMDB0155722	NA	NA	NA	NA	NA
19	HMDB0000744	Malic acid	HMDB0000744	525	C03668	C(C(C(=O)O)O)C(=O)O
20	HMDB0033143	Pyrraline	HMDB0033143	14274616		C1=C(N(C(=C1)C=O)CCC
21	HMDB0000500	4-Hydroxybenzoic acid	HMDB0000500	135	C00156	C1=CC(=CC=C1C(=O)O)
22	HMDB0002123	1,3,7-Trimethyluric acid	HMDB0002123	79437	C16361	CN1C2=C(NC1=O)N(C(=
23	HMDB0000661	Glutaric acid	HMDB0000661	743	C00489	C(CC(=O)O)CC(=O)O
24	HMDB0000355	3-Hydroxymethylglutaric acid	HMDB0000355	1662	C03761	CC(CC(=O)O)(CC(=O)O)
25	HMDB0000292	Xanthine	HMDB0000292	1188	C00385	C1=NC2=C(N1)C(=O)NC
26	HMDB0001138	N-Acetylglutamic acid	HMDB0001138	185	C00624	CC(=O)NC(CCC(=O)O)C
27	HMDB0000446	N-Alpha-acetyllysine	HMDB0000446	192590	C12989	CC(=O)NC(CCCCN)C(=O
28	HMDB0000172	L-Isoleucine	HMDB0000172	6306	C00407	CC[C@H](C)[C@@H](C(=
29	HMDB0000512	N-Acetyl-L-phenylalanine	HMDB0000512	74839	C03519	CC(=O)N[C@@H](CC1=CC
30	HMDB0000687	L-Leucine	HMDB0000687	6106	C00123	CC(C)C[C@@H](C(=O)O)N
31	HMDB0000020	p-Hydroxyphenylacetic acid	HMDB0000020	127	C00642	C1=CC(=CC=C1CC(=O)O
32	HMDB0000641	L-Glutamine	HMDB0000641	5961	C00064	C(CC(=O)N)[C@@H](C(=
33	HMDB0012275	Phenylethylamine	HMDB0012275	1001	C05332	C1=CC=C(C=C1)CCN
34	HMDB0002035	4-Hydroxycinnamic acid	HMDB0002035	637542	C00811	C1=CC(=CC=C1/C=C/C(
35	HMDB0060015	Phenyl hydrogen sulfate	HMDB0060015	74426	C02180	OS(=O)(=O)OC1=CC=CC
36	HMDB0000956	Tartaric acid	HMDB0000956	444305	C00898	O[C@H]([C@@H](O)C(O)=
37	HMDB0003331	1-Methyladenosine	HMDB0003331	27476	C02494	CN1C=NC2=C(C1=N)N=C
38	HMDB0142137	NA	NA	NA	NA	NA
39	HMDB0000729	Alpha-Hydroxyisobutyric acid	HMDB0000729	11671		CC(C)(C(=O)O)O
40	HMDB0062640	3-hydroxy-2-isobutyrate	HMDB0062640	87	C01188	CC(CO)C(O)=O
41	HMDB0240751	NA	NA	NA	NA	NA
42	HMDB0013713	N-acetyltryptophan	HMDB0013713	700653		[H][C@@](CC1=CNC2=CC
43	HMDB0061384	NA	NA	NA	NA	NA
44	HMDB0255727	NA	NA	NA	NA	NA
45	HMDB0013676	2,6-Dihydroxybenzoic acid	HMDB0013676	9338	C21298	C1=CC(=C(C(=C1)O)C(=
46	HMDB0001991	7-Methylxanthine	HMDB0001991	68374	C16353	CN1C=NC2=C1C(=O)NC(
47	HMDB0000881	Xanthurenic acid	HMDB0000881	5699	C02470	C1=CC2=C(C(=C1)O)NC(
48	HMDB0061112	3-Carboxy-4-methyl-5-propyl-2-furanpropionic acid	HMDB0061112	123979		CCCC1=C(C)C(C(O)=O)=
49	HMDB0011635	p-Cresol sulfate	HMDB0011635	4615423		CC1=CC=C(C=C1)OS(=O
50	HMDB0000822	p-Hydroxymandelic acid	HMDB0000822	7721	C11527	C1=CC(=CC=C1C(C(=O)
51	HMDB0000678	Isovalerylglycine	HMDB0000678	546304		CC(C)CC(=O)NCC(=O)O
52	HMDB0005807	Gallic acid	HMDB0005807	370	C01424	C1=C(C=C(C(=C1O)O)O)
53	HMDB0013173	NA	NA	NA	NA	NA

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.

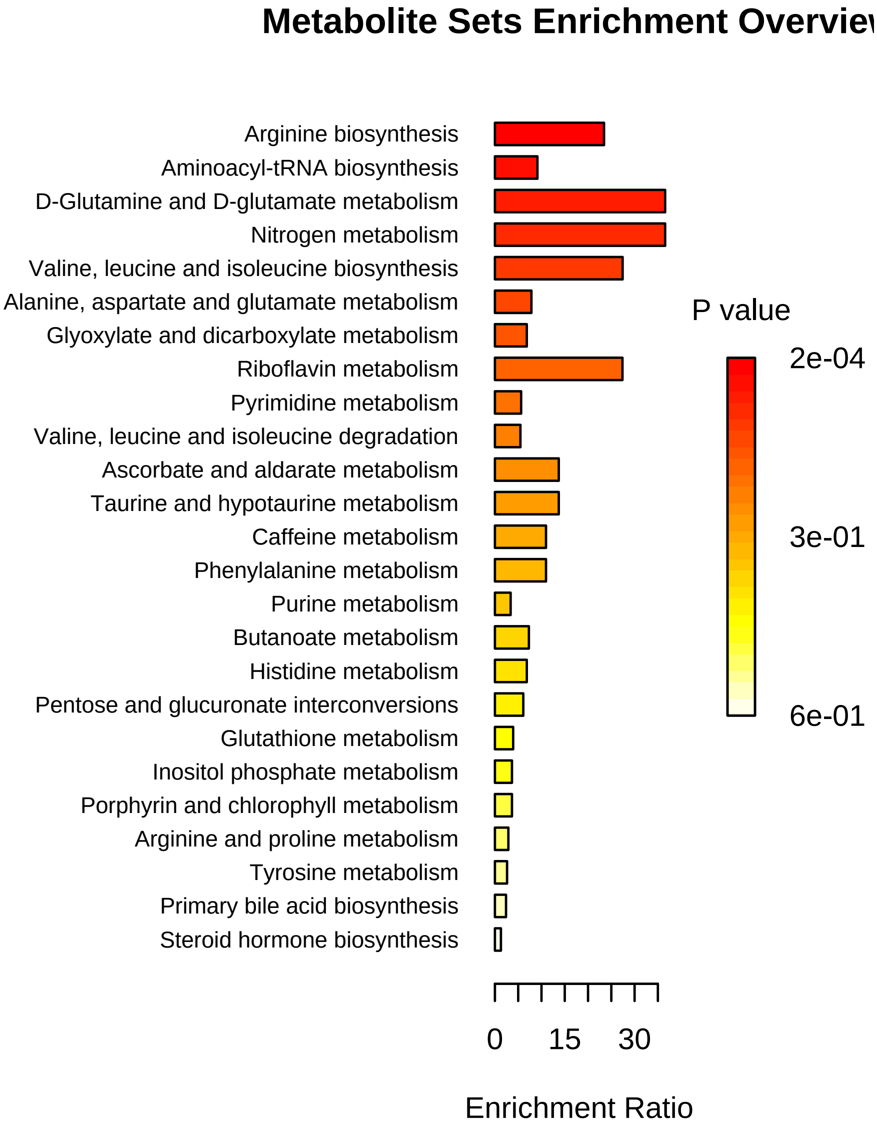

Figure 1: Summary Plot for Over Representation Analysis (ORA)

Table 2: Result from Over Representation Analysis

	total	expected	hits	Raw p	Holm p	FDR
Arginine biosynthesis	14	0.13	3	2.07E-04	1.74E-02	1.74E-02
Aminoacyl-tRNA biosynthesis	48	0.44	4	6.69E-04	5.55E-02	2.38E-02
D-Glutamine and D-glutamate metabolism	6	0.05	2	1.13E-03	9.30E-02	2.38E-02
Nitrogen metabolism	6	0.05	2	1.13E-03	9.30E-02	2.38E-02
Valine, leucine and isoleucine biosynthesis	8	0.07	2	2.09E-03	1.68E-01	3.52E-02
Alanine, aspartate and glutamate metabolism	28	0.26	2	2.55E-02	1.00E+00	3.57E-01
Glyoxylate and dicarboxylate metabolism	32	0.29	2	3.27E-02	1.00E+00	3.78E-01
Riboflavin metabolism	4	0.04	1	3.60E-02	1.00E+00	3.78E-01
Pyrimidine metabolism	39	0.35	2	4.72E-02	1.00E+00	4.15E-01
Valine, leucine and isoleucine degradation	40	0.36	2	4.94E-02	1.00E+00	4.15E-01
Ascorbate and aldarate metabolism	8	0.07	1	7.08E-02	1.00E+00	4.96E-01
Taurine and hypotaurine metabolism	8	0.07	1	7.08E-02	1.00E+00	4.96E-01
Caffeine metabolism	10	0.09	1	8.77E-02	1.00E+00	5.26E-01
Phenylalanine metabolism	10	0.09	1	8.77E-02	1.00E+00	5.26E-01
Purine metabolism	65	0.59	2	1.16E-01	1.00E+00	6.48E-01
Butanoate metabolism	15	0.14	1	1.29E-01	1.00E+00	6.76E-01
Histidine metabolism	16	0.15	1	1.37E-01	1.00E+00	6.76E-01
Pentose and glucuronate interconversions	18	0.16	1	1.53E-01	1.00E+00	7.13E-01
Glutathione metabolism	28	0.26	1	2.28E-01	1.00E+00	9.69E-01
Inositol phosphate metabolism	30	0.27	1	2.42E-01	1.00E+00	9.69E-01
Porphyrin and chlorophyll metabolism	30	0.27	1	2.42E-01	1.00E+00	9.69E-01
Arginine and proline metabolism	38	0.35	1	2.97E-01	1.00E+00	1.00E+00
Tyrosine metabolism	42	0.38	1	3.23E-01	1.00E+00	1.00E+00
Primary bile acid biosynthesis	46	0.42	1	3.48E-01	1.00E+00	1.00E+00
Steroid hormone biosynthesis	85	0.78	1	5.51E-01	1.00E+00	1.00E+00

7 Appendix: R Command History

```
 [1] "mSet<-InitDataObjects(\"conc\", \"msetora\", FALSE)"
 [2] "cmpd.vec<-c(\"HMDB0000752\",\"HMDB0000448\",\"HMDB0000893\",\"HMDB0001032\",\"HMDB0000262\",\"
 [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, \"guest11006858485735387 32\")\n"
```

The report was generated on Thu Nov 24 05:32:41 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 .