Metabolomic Data Analysis with MetaboAnalyst 5.0

Name: guest9454368572409379571

November 24, 2021

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	5- Methylthioadenosine	5'-Methylthioadenosine	HMDB0001173	439176	C00170	CSC[C@@H]1[C@H]([C@H]([C@@H](O1)N2C=
2	Acetyl proline	NA	NA	NA	NA	NA
3	Adenine	Adenine	HMDB0000034	190	C00147	C1=NC2=C(N1)C(=NC=N2)N
4	Aspartate	L-Aspartic acid	HMDB0000191	5960	C00049	C([C@@H](C(=O)O)N)C(=O)O
5	CDP-Choline	Citicoline	HMDB0001413	13804	C00307	C[N+](C)(C)CCOP(=O)([O-])OP(=O)(O)OC[
6	Cystathionine	L-Cystathionine	HMDB0000099	439258	C02291	C(CSC[C@@H](C(=O)O)N)[C@@H](C(=O)O)
7	Feature 100	NA	NA	NA	NA	NA
8	Feature 101	NA	NA	NA	NA	NA
9	Feature 102	NA	NA	NA	NA	NA
10	Feature 103	NA	NA	NA	NA	NA
11	Feature 104	NA	NA	NA	NA	NA
12	Feature 105	NA	NA	NA	NA	NA
13	Feature 106	NA	NA	NA	NA	NA
14	Feature 107	NA	NA	NA	NA	NA
15	Feature 108	NA	NA	NA	NA	NA
16	Feature 109	NA	NA	NA	NA	NA
17	Feature 110	NA	NA	NA	NA	NA
18	Feature 112	NA	NA	NA	NA	NA
19	Feature 113	NA	NA	NA	NA	NA
20	Feature 114	NA	NA	NA	NA	NA
21	Feature 115	NA	NA	NA	NA	NA
22	Feature 116	NA	NA	NA	NA	NA
23	Feature 117	NA	NA	NA	NA	NA
24	Feature 118	NA	NA	NA	NA	NA
25	Feature 125	NA	NA	NA	NA	NA
26	Feature 126	NA	NA	NA	NA	NA
27	Feature 127	NA	NA	NA	NA	NA
28	Feature 128	NA	NA	NA	NA	NA
29	Feature 129	NA	NA	NA	NA	NA
30	Feature 130	NA	NA	NA	NA	NA
31	Feature 132	NA	NA	NA	NA	NA
32	Feature 133	NA	NA	NA	NA	NA
33	Feature 134	NA	NA	NA	NA	NA
34	Feature 137	NA	NA	NA	NA	NA
35	Feature 138	NA	NA	NA	NA	NA
36	Feature 139	NA	NA	NA	NA	NA
37	Feature 140	NA	NA	NA	NA	NA
38	Feature 141	NA	NA	NA	NA	NA
39	Feature 18	NA	NA	NA	NA	NA
40	Feature 19	NA	NA	NA	NA	NA
41	Feature 28	NA	NA	NA	NA	NA
42	Feature 29	NA	NA	NA	NA	NA
43	Feature 32	NA	NA	NA	NA	NA
44	Feature 34	NA	NA	NA	NA	NA
45	Feature 39	NA	NA	NA	NA	NA
46	Feature 5	NA	NA	NA	NA	NA
47	Feature 51	NA	NA	NA	NA	NA
48	Feature 52	NA	NA	NA	NA	NA
49	Feature 55	NA	NA	NA	NA	NA
50	Feature 6	NA	NA	NA	NA	NA
51	Feature 60	NA	NA	NA	NA	NA
52	Feature 68	NA	NA	NA	NA	NA
53	Feature 70	NA	NA	NA	NA	NA
54	Feature 71	NA	NA	NA	NA	NA
55	Feature 72	NA	NA	NA	NA	NA
56	Feature 73	NA	NA	NA	NA	NA
57	Feature 74	NA	NA	NA	NA	NA
58	Feature 76	NA	NA	NA	NA	NA
59	Feature 77	NA	NA	NA	NA	NA

60	Feature 78	NA	NA	NA	NA	NA
61	Feature 79	NA	NA	NA	NA	NA
62	Feature 81	NA	NA	NA	NA	NA
63	Feature 82	NA	NA	NA	NA	NA
64	Feature 83	NA	NA	NA	NA	NA
65	Feature 85	NA	NA	NA	NA	NA
66	Feature 86	NA	NA	NA	NA	NA
67	Feature 87	NA	NA	NA	NA	NA
68	Feature 88	NA	NA	NA	NA	NA
69	Feature 90	NA	NA	NA	NA	NA
70	Feature 91	NA	NA	NA	NA	NA
71	Feature 92	NA	NA	NA	NA	NA
72	Feature 93	NA	NA	NA	NA	NA
73	Feature 94	NA	NA	NA	NA	NA
74	Feature 95	NA	NA	NA	NA	NA
75	Feature 96	NA	NA	NA	NA	NA
76	Glycerophosphocholine	Glycerophosphocholine	HMDB0000086	71920	C00670	C[N+](C)(C)CCOP(=O)([O-])OC[C@H](CO)O
77	Hypoxanthine	Hypoxanthine	HMDB0000157	790	C00262	C1=NC2=C(N1)C(=O)N=CN2
78	L-Palmitoylcarnitine	L-Palmitoylcarnitine	HMDB0000222	11953816	C02990	CCCCCCCCCCCCCCCC(=O)O[C@H](CC(=O
79	N-acetyl-glutamine	N-Acetylglutamine	HMDB0006029	25561		CC(=O)NC(CCC(=O)N)C(=O)O
80	Nicotinamide riboside	Nicotinamide riboside	HMDB0000855	439924	C03150	C1=CC(=C[N+](=C1)[C@H]2[C@@H]([C@@H
81	Nicotinate	Nicotinic acid	HMDB0001488	938	C00253	C1=CC(=CN=C1)C(=O)O
82	o-acetyl-L-serine	O-Acetylserine	HMDB0003011	99478	C00979	CC(=O)OC[C@@H](C(=O)O)N
83	O-Decanoyl-L-carnitine	O-decanoyl-L-carnitine	HMDB0062631	11953821	C03299	[H][C@@](CC([O-])=O)(C[N+](C)(C)C)OC(=
84	Taurine	Taurine	HMDB0000251	1123	C00245	C(CS(=O)(=O)O)N

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.

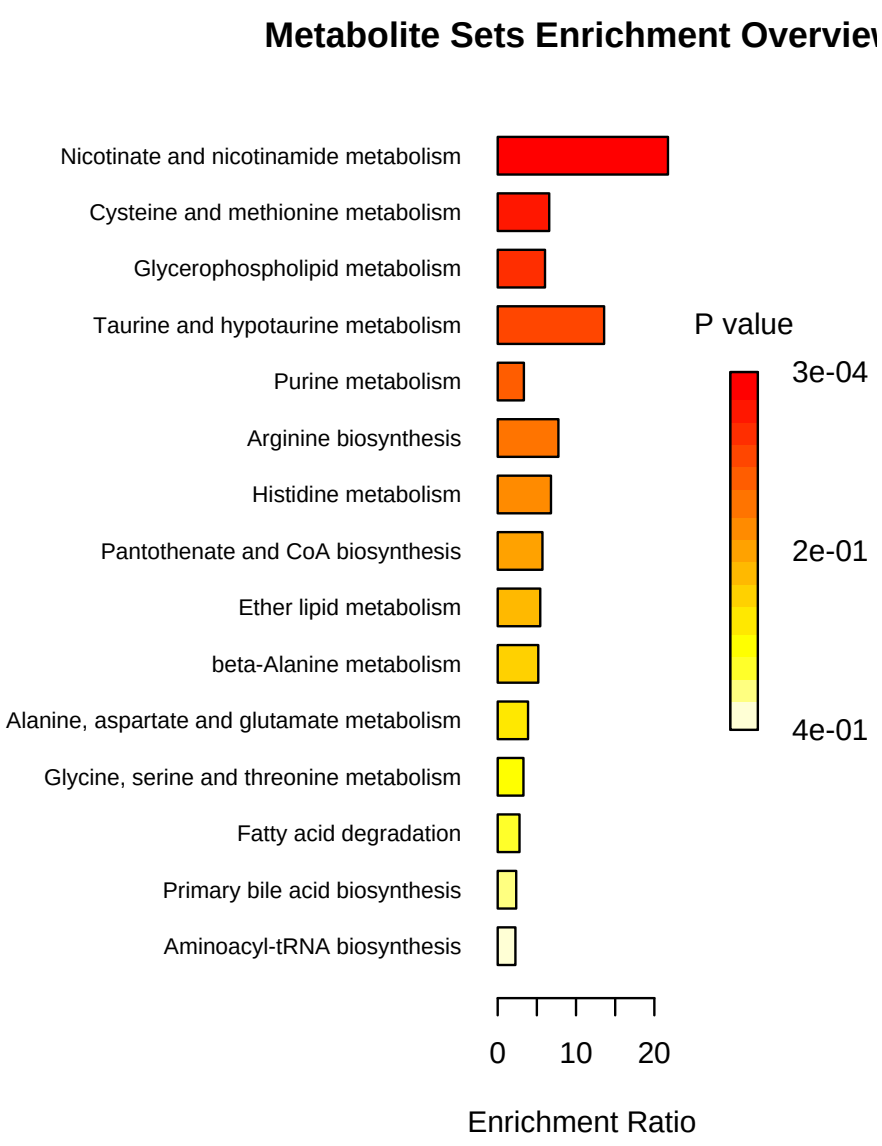

Figure 1: Summary Plot for Over Representation Analysis (ORA)

Table 2: Result from Over Representation Analysis

	total	expected	hits	Raw p	Holm p	FDR
Nicotinate and nicotinamide metabolism	15	0.14	3	2.65E-04	2.22E-02	2.22E-02
Cysteine and methionine metabolism	33	0.30	2	3.53E-02	1.00E+00	1.00E+00
Glycerophospholipid metabolism	36	0.33	2	4.14E-02	1.00E+00	1.00E+00
Taurine and hypotaurine metabolism	8	0.07	1	7.14E-02	1.00E+00	1.00E+00
Purine metabolism	65	0.60	2	1.18E-01	1.00E+00	1.00E+00
Arginine biosynthesis	14	0.13	1	1.22E-01	1.00E+00	1.00E+00
Histidine metabolism	16	0.15	1	1.38E-01	1.00E+00	1.00E+00
Pantothenate and CoA biosynthesis	19	0.17	1	1.62E-01	1.00E+00	1.00E+00
Ether lipid metabolism	20	0.18	1	1.70E-01	1.00E+00	1.00E+00
beta-Alanine metabolism	21	0.19	1	1.77E-01	1.00E+00	1.00E+00
Alanine, aspartate and glutamate metabolism	28	0.26	1	2.30E-01	1.00E+00	1.00E+00
Glycine, serine and threonine metabolism	33	0.30	1	2.65E-01	1.00E+00	1.00E+00
Fatty acid degradation	39	0.36	1	3.06E-01	1.00E+00	1.00E+00
Primary bile acid biosynthesis	46	0.42	1	3.50E-01	1.00E+00	1.00E+00
Aminoacyl-tRNA biosynthesis	48	0.44	1	3.63E-01	1.00E+00	1.00E+00

7 Appendix: R Command History

```
 [1] "mSet<-InitDataObjects(\"conc\", \"msetora\", FALSE)"
 [2] "cmpd.vec<-c(\"5- Methylthioadenosine\",\"Acetyl proline\",\"Adenine\",\"Aspartate\",\"CDP-Choli
 [3] "mSet<-Setup.MapData(mSet, cmpd.vec);"
 [4] "mSet<-CrossReferencing(mSet, \"name\");"
 [5] "mSet<-CreateMappingResultTable(mSet)"
 [6] "mSet<-PerformDetailMatch(mSet, \"5- Methylthioadenosine\");"
 [7] "mSet<-GetCandidateList(mSet);"
 [8] "mSet<-SetCandidate(mSet, \"5- Methylthioadenosine\", \"5'-Methylthioadenosine\");"
 [9] "mSet<-PerformDetailMatch(mSet, \"Acetyl proline\");"
[10] "mSet<-GetCandidateList(mSet);"
[11] "mSet<-PerformDetailMatch(mSet, \"N-acetyl-glutamine\");"
[12] "mSet<-GetCandidateList(mSet);"
[13] "mSet<-SetCandidate(mSet, \"N-acetyl-glutamine\", \"N-Acetylglutamine\");"
[14] "mSet<-SetMetabolomeFilter(mSet, F);"
[15] "mSet<-SetCurrentMsetLib(mSet, \"kegg_pathway\", 2);"
[16] "mSet<-CalculateHyperScore(mSet)"
[17] "mSet<-PlotORA(mSet, \"ora_0_\", \"net\", \"png\", 72, width=NA)"
[18] "mSet<-PlotEnrichDotPlot(mSet, \"ora\", \"ora_dot_0_\", \"png\", 72, width=NA)"
[19] "mSet<-CalculateHyperScore(mSet)"
[20] "mSet<-PlotORA(mSet, \"ora_1_\", \"net\", \"png\", 72, width=NA)"
[21] "mSet<-PlotEnrichDotPlot(mSet, \"ora\", \"ora_dot_1_\", \"png\", 72, width=NA)"
[22] "mSet<-SaveTransformedData(mSet)"
[23] "mSet<-PreparePDFReport(mSet, \"guest9454368572409379571\")\n"
```

The report was generated on Wed Nov 24 14:59:31 2021 with R version 4.0.2 (2020-06-22).