Metabolomic Data Analysis with MetaboAnalyst 5.0

Name: guest14826309385831893896

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

	Query	Match	HMDB	PubChem	KEGG	SMILES
1	2-Aminoethylphosphonate	Ciliatine	HMDB0011747	339	C03557	C(CP(=O)(O)O)N
2	4-Guanidinobutanoic acid	4-Guanidinobutanoic acid	HMDB0003464	500	C01035	C(CC(=O)O)CN=C(N)N
3	Arginine	L-Arginine	HMDB0000517	6322	C00062	C(C[C@@H](C(=O)O)N)
4	D-2-Aminobutyric acid	D-Alpha-aminobutyric acid	HMDB0000650	439691	C02261	CC[C@H](C(=O)O)N
5	Feature 10	NA	NA	NA	NA	NA
6	Feature 11	NA	NA	NA	NA	NA
7	Feature 12	NA	NA	NA	NA	NA
8	Feature 13	NA	NA	NA	NA	NA
9	Feature 14	NA	NA	NA	NA	NA
10	Feature 15	NA	NA	NA	NA	NA
11	Feature 16	NA	NA	NA	NA	NA
12	Feature 17	NA	NA	NA	NA	NA
13	Feature 22	NA	NA	NA	NA	NA
14	Feature 23	NA	NA	NA	NA	NA
15	Feature 25	NA	NA	NA	NA	NA
16	Feature 26	NA	NA	NA	NA	NA
17	Feature 27	NA	NA	NA	NA	NA
18	Feature 35	NA	NA	NA	NA	NA
19	Feature 36	NA	NA	NA	NA	NA
20	Feature 37	NA	NA	NA	NA	NA
21	Feature 40	NA	NA	NA	NA	NA
22	Feature 41	NA	NA	NA	NA	NA
23	Feature 42	NA	NA	NA	NA	NA
24	Feature 43	NA	NA	NA	NA	NA
25	Feature 45	NA	NA	NA	NA	NA
26	Feature 47	NA	NA	NA	NA	NA
27	Feature 50	NA	NA	NA	NA	NA
28	Feature 53	NA	NA	NA	NA	NA
29	Feature 54	NA	NA	NA	NA	NA
30	Feature 56	NA	NA	NA	NA	NA
31	Feature 57	NA	NA	NA	NA	NA
32	Feature 58	NA	NA	NA	NA	NA
33	Feature 59	NA	NA	NA	NA	NA
34	Feature 61	NA	NA	NA	NA	NA
35	Feature 62	NA	NA	NA	NA	NA
36	Feature 63	NA	NA	NA	NA	NA
37	Feature 64	NA	NA	NA	NA	NA
38	Feature 65	NA	NA	NA	NA	NA
39	Feature 66	NA	NA	NA	NA	NA
40	Feature 67	NA	NA	NA	NA	NA
41	Feature 69	NA	NA	NA	NA	NA
42	Feature 7	NA	NA	NA	NA	NA
43	Feature 75	NA	NA	NA	NA	NA
44	Feature 8	NA	NA	NA	NA	NA
45	Feature 84	NA	NA	NA	NA	NA
46	Feature 89	NA	NA	NA	NA	NA
47	Feature 9	NA	NA	NA	NA	NA
48	Glucose-6-phosphate	Glucose 6-phosphate	HMDB0001401	5958	C00092	C([C@@H]1[C@H]([C@@
49	Glycine	Glycine	HMDB0000123	750	C00037	C(C(=O)O)N
50	Histidine	L-Histidine	HMDB0000177	6274	C00135	C1=C(NC=N1)C[C@@H
51	Homoarginine	Homo-L-arginine	HMDB0000670	9085	C01924	C(CCN=C(N)N)C[C@@H
52	Lysine	L-Lysine	HMDB0000182	5962	C00047	C(CCN)C[C@@H](C(=O
53	N-acetyl-L-ornithine	N-Acetylornithine	HMDB0003357	439232	C00437	CC(=O)N[C@@H](CCCN
54	NADPH	NADPH	HMDB0000221	22833512	C00005	C1C=CN(C=C1C(=O)N
55	NG-dimethyl-L-arginine	Asymmetric dimethylarginine	HMDB0001539	123831	C03626	CN(C)C(=NCCC[C@@H
56	O-Phosphorylethanolamine	O-Phosphoethanolamine	HMDB0000224	1015	C00346	C(COP(=O)(O)O)N
57	Ornithine	Ornithine	HMDB0000214	6262	C00077	C(C[C@@H](C(=O)O)N)
58	Phenylpropanolamine	Phenylpropanolamine	HMDB0001942	26934	C02343	C[C@H]([C@H](C1=CC=
59	Phosphocholine	Phosphorylcholine	HMDB0001565	8691	C00588	C[N+](C)(C)CCOP(=O)

60	UDP-N-acetyl-glucosamine	Uridine diphosphate-N-acetylglucosamine	HMDB0000290	9547196	C00043	CC(=O)N[C@@H]1[C@H
61	UMP	Uridine 5'-monophosphate	HMDB0000288	6030	C00105	C1=CN(C(=O)NC1=O)[
62	Uric acid	Uric acid	HMDB0000289	1175	C00366	C12=C(NC(=O)N1)NC(=
63	Xanthosine-5-phosphate	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.

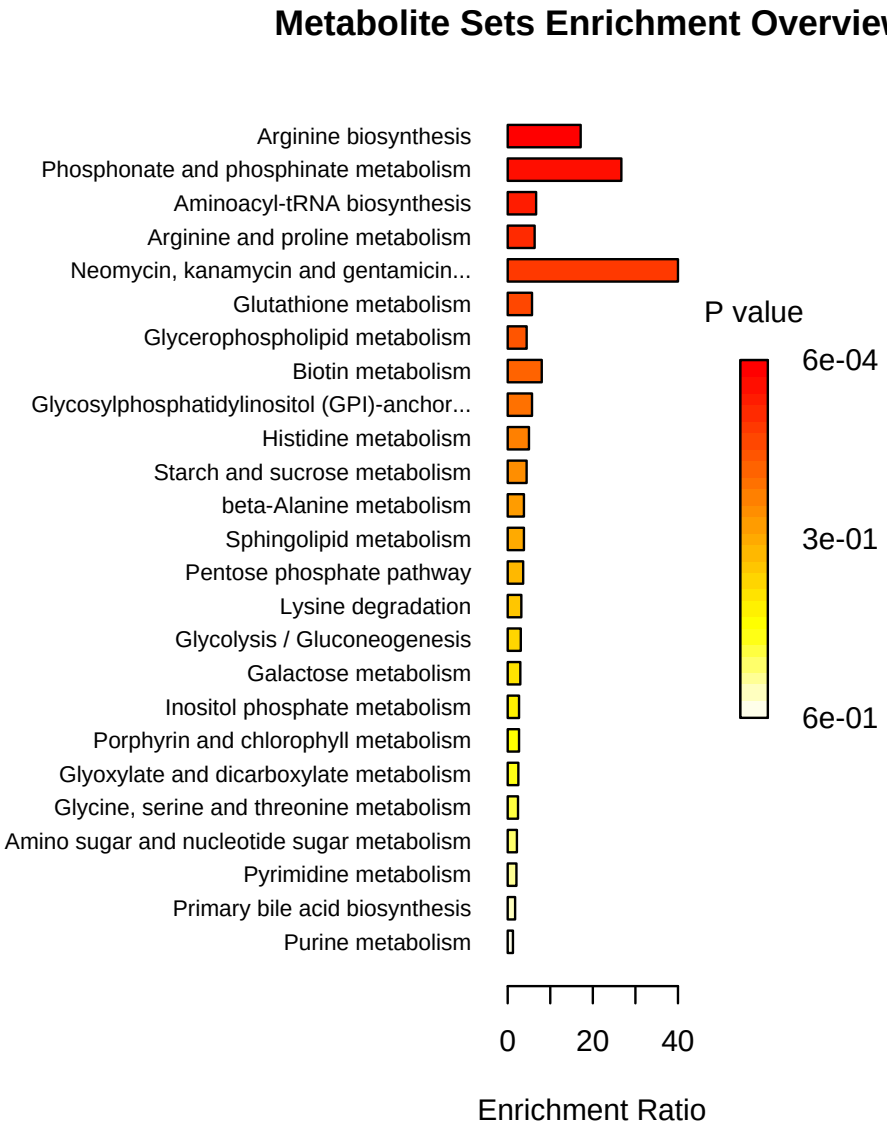

Figure 1: Summary Plot for Over Representation Analysis (ORA)

Table 2: Result from Over Representation Analysis

	total	expected	hits	Raw p	Holm p	FDR
Arginine biosynthesis	14	0.17	3	5.51E-04	4.63E-02	4.63E-02
Phosphonate and phosphinate metabolism	6	0.07	2	2.15E-03	1.79E-01	6.68E-02
Aminoacyl-tRNA biosynthesis	48	0.60	4	2.38E-03	1.96E-01	6.68E-02
Arginine and proline metabolism	38	0.47	3	1.06E-02	8.55E-01	2.22E-01
Neomycin, kanamycin and gentamicin biosynthesis	2	0.03	1	2.48E-02	1.00E+00	4.17E-01
Glutathione metabolism	28	0.35	2	4.60E-02	1.00E+00	6.44E-01
Glycerophospholipid metabolism	36	0.45	2	7.23E-02	1.00E+00	8.68E-01
Biotin metabolism	10	0.12	1	1.18E-01	1.00E+00	1.00E+00
Glycosylphosphatidylinositol (GPI)-anchor biosynthesis	14	0.17	1	1.62E-01	1.00E+00	1.00E+00
Histidine metabolism	16	0.20	1	1.83E-01	1.00E+00	1.00E+00
Starch and sucrose metabolism	18	0.23	1	2.03E-01	1.00E+00	1.00E+00
beta-Alanine metabolism	21	0.26	1	2.33E-01	1.00E+00	1.00E+00
Sphingolipid metabolism	21	0.26	1	2.33E-01	1.00E+00	1.00E+00
Pentose phosphate pathway	22	0.28	1	2.43E-01	1.00E+00	1.00E+00
Lysine degradation	25	0.31	1	2.71E-01	1.00E+00	1.00E+00
Glycolysis / Gluconeogenesis	26	0.33	1	2.81E-01	1.00E+00	1.00E+00
Galactose metabolism	27	0.34	1	2.90E-01	1.00E+00	1.00E+00
Inositol phosphate metabolism	30	0.38	1	3.16E-01	1.00E+00	1.00E+00
Porphyrin and chlorophyll metabolism	30	0.38	1	3.16E-01	1.00E+00	1.00E+00
Glyoxylate and dicarboxylate metabolism	32	0.40	1	3.34E-01	1.00E+00	1.00E+00
Glycine, serine and threonine metabolism	33	0.41	1	3.42E-01	1.00E+00	1.00E+00
Amino sugar and nucleotide sugar metabolism	37	0.46	1	3.75E-01	1.00E+00	1.00E+00
Pyrimidine metabolism	39	0.49	1	3.91E-01	1.00E+00	1.00E+00
Primary bile acid biosynthesis	46	0.57	1	4.44E-01	1.00E+00	1.00E+00
Purine metabolism	65	0.81	1	5.66E-01	1.00E+00	1.00E+00

7 Appendix: R Command History

```
 [1] "mSet<-InitDataObjects(\"conc\", \"msetora\", FALSE)"
 [2] "cmpd.vec<-c(\"2-Aminoethylphosphonate\",\"4-Guanidinobutanoic acid\",\"Arginine\",\"D-2-Aminobu
 [3] "mSet<-Setup.MapData(mSet, cmpd.vec);"
 [4] "mSet<-CrossReferencing(mSet, \"name\");"
 [5] "mSet<-CreateMappingResultTable(mSet)"
 [6] "mSet<-PerformDetailMatch(mSet, \"N-acetyl-L-ornithine\");"
 [7] "mSet<-GetCandidateList(mSet);"
 [8] "mSet<-SetCandidate(mSet, \"N-acetyl-L-ornithine\", \"N-Acetylornithine\");"
 [9] "mSet<-PerformDetailMatch(mSet, \"NG-dimethyl-L-arginine\");"
[10] "mSet<-GetCandidateList(mSet);"
[11] "mSet<-SetCandidate(mSet, \"NG-dimethyl-L-arginine\", \"Asymmetric dimethylarginine\");"
[12] "mSet<-PerformDetailMatch(mSet, \"Xanthosine-5-phosphate\");"
[13] "mSet<-GetCandidateList(mSet);"
[14] "mSet<-SetCandidate(mSet, \"Xanthosine-5-phosphate\", \"Xanthylic acid\");"
[15] "mSet<-SetMetabolomeFilter(mSet, F);"
[16] "mSet<-SetCurrentMsetLib(mSet, \"kegg_pathway\", 2);"
[17] "mSet<-CalculateHyperScore(mSet)"
[18] "mSet<-PlotORA(mSet, \"ora_0_\", \"net\", \"png\", 72, width=NA)"
[19] "mSet<-PlotEnrichDotPlot(mSet, \"ora\", \"ora_dot_0_\", \"png\", 72, width=NA)"
[20] "mSet<-CalculateHyperScore(mSet)"
[21] "mSet<-PlotORA(mSet, \"ora_1_\", \"net\", \"png\", 72, width=NA)"
[22] "mSet<-PlotEnrichDotPlot(mSet, \"ora\", \"ora_dot_1_\", \"png\", 72, width=NA)"
[23] "mSet<-SaveTransformedData(mSet)"
[24] "mSet<-PreparePDFReport(mSet, \"guest14826309385831893896\")\n"
```

The report was generated on Wed Nov 24 14:51:48 2021 with R version 4.1.1 (2021-08-10).