Metabolomic Data Analysis with MetaboAnalyst 5.0

Name: guest10606626162765222506

May 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 Comp

	Query	Match	HMDB	PubChem	KEGG
1	HMDB0029726	2-Vinylthiophene	HMDB0029726	519642	
2	HMDB0033746	Benzenethiol	HMDB0033746	7969	
3	HMDB0029730	2-Ethyl-4-methylthiazole	HMDB0029730	27440	
4	HMDB0040062	NA	NA	NA	NA
5	HMDB0040096	NA	NA	NA	NA
6	HMDB0033155	Trimethylthiazole	HMDB0033155	61653	
7	HMDB0031337	Thiiranebutanenitrile	HMDB0031337	148822	
8	HMDB0029731	4-Ethyl-2-methylthiazole	HMDB0029731	520568	
9	HMDB0029732	5-Ethyl-4-methylthiazole	HMDB0029732	40380	
10	HMDB0034300	5-Isothiocyanato-1-pentene	HMDB0034300	87436	
11	HMDB0172134	NA	NA	NA	NA
12	HMDB0172135	NA	NA	NA	NA
13	HMDB0172136	NA	NA	NA	NA
14	HMDB0165771	NA	NA	NA	NA
15	HMDB0165773	NA	NA	NA	NA
16	HMDB0165770	NA	NA	NA	NA
17	HMDB0165772	NA	NA	NA	NA
18	HMDB0240308	NA	NA	NA	NA
19	HMDB0029737	1H-Indole-3-carboxaldehyde	HMDB0029737	10256	C0849
20	HMDB0181148	NA	NA	NA	NA
21	HMDB0060537	NA	NA	NA	NA
22	HMDB0060826	NA	NA	NA	NA
23	HMDB0004185	5-Hydroxyindoleacetylglycine	HMDB0004185	440806	C0583
24	HMDB0003066	Chalcone	HMDB0003066	637760	C1558
25	HMDB0135598	NA	NA	NA	NA
26	HMDB0060323	NA	NA	NA	NA
27	HMDB0032357	N-Lactoyl ethanolamine phosphate	HMDB0032357	11550267	
28	HMDB0155289	NA	NA	NA	NA
29	HMDB0060757	NA	NA	NA	NA
30	HMDB0000684	L-Kynurenine	HMDB0000684	161166	C0032
31	HMDB0012948	Formyl-5-hydroxykynurenamine	HMDB0012948	440743	C0564
32	HMDB0014350	Pyrimethamine	HMDB0014350	4993	C0739
33	HMDB0060321	NA	NA	NA	NA
34	HMDB0062613	NA	NA	NA	NA
35	HMDB0012458	7alpha-Hydroxy-3-oxo-4-cholestenoate	HMDB0012458	3081085	C1733
36	HMDB0060128	NA	NA	NA	NA
37	HMDB0060127	NA	NA	NA	NA
38	HMDB0030066	Australigenin	HMDB0030066	4483037	
39	HMDB0036249	NA	NA	NA	NA
40	HMDB0034403	Barogenin	HMDB0034403	101688	
41	HMDB0011590	MG(24:6(6Z,9Z,12Z,15Z,18Z,21Z)/0:0/0:0)	HMDB0011590	53480998	
42	HMDB0166149	NA	NA	NA	NA
43	HMDB0174217	NA	NA	NA	NA
44	HMDB0157634	NA	NA	NA	NA
45	HMDB0157631	NA	NA	NA	NA
46	HMDB0157632	NA	NA	NA	NA
47	HMDB0157633	NA	NA	NA	NA
48	HMDB0157635	NA	NA	NA	NA
49	HMDB0157638	NA	NA	NA	NA
50	HMDB0164013	NA	NA	NA	NA
51	HMDB0160505	NA	NA	NA	NA
52	HMDB0160504	NA	NA	NA	NA
53	HMDB0157644	NA	NA	NA	NA
54	HMDB0157646	NA	NA	NA	NA
55	HMDB0157643	NA	NA	NA	NA
56	HMDB0157645	NA	NA	NA	NA
57	HMDB0157640	NA	NA	NA	NA
58	HMDB0157641	NA	NA	NA	NA
59	HMDB0157639	NA	NA	NA	NA

60	HMDB0157642	NA	NA	NA	NA
61	HMDB0157637	NA	NA	NA	NA
62	HMDB0157636	NA	NA	NA	NA
63	HMDB0034293	Asperagenin	HMDB0034293	71435521	
64	HMDB0032783	Porrigenin A	HMDB0032783	12312669	
65	HMDB0157653	NA	NA	NA	NA
66	HMDB0157650	NA	NA	NA	NA
67	HMDB0157649	NA	NA	NA	NA
68	HMDB0157652	NA	NA	NA	NA
69	HMDB0157654	NA	NA	NA	NA
70	HMDB0157651	NA	NA	NA	NA
71	HMDB0157648	NA	NA	NA	NA
72	HMDB0160506	NA	NA	NA	NA
73	HMDB0157647	NA	NA	NA	NA
74	HMDB0161140	NA	NA	NA	NA
75	HMDB0000708	Glycoursodeoxycholic acid	HMDB0000708	12310288	
76	HMDB0161141	NA	NA	NA	NA
77	HMDB0006898	Chenodeoxyglycocholic acid	HMDB0006898	53477907	C0546
78	HMDB0000631	Deoxycholic acid glycine conjugate	HMDB0000631	3035026	C0546
79	HMDB0000637	Chenodeoxycholic acid glycine conjugate	HMDB0000637	22833540	C0546
80	HMDB0184641	NA	NA	NA	NA
81	HMDB0161139	NA	NA	NA	NA
82	HMDB0173226	NA	NA	NA	NA
83	HMDB0173227	NA	NA	NA	NA
84	HMDB0173228	NA	NA	NA	NA
85	HMDB0173225	NA	NA	NA	NA
86	HMDB0173223	NA	NA	NA	NA
87	HMDB0173224	NA	NA	NA	NA
88	HMDB0012516	11'-Carboxy-alpha-tocotrienol	HMDB0012516	53481452	
89	HMDB0030140	Adlupulone	HMDB0030140		
90	HMDB0030041	Lupulone	HMDB0030041	51397980	C1070
91	HMDB0155347	NA	NA	NA	NA
92	HMDB0155348	NA	NA	NA	NA
93	HMDB0034064	2-Angeloyl-9-(3-methyl-2E-pentenoyl)-2b,9a-dihydroxy-4Z,10(14)-oplopadien-3-one	HMDB0034064	131751519	
94	HMDB0012936	Dynorphin B (10-13)	HMDB0012936	53481556	

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.

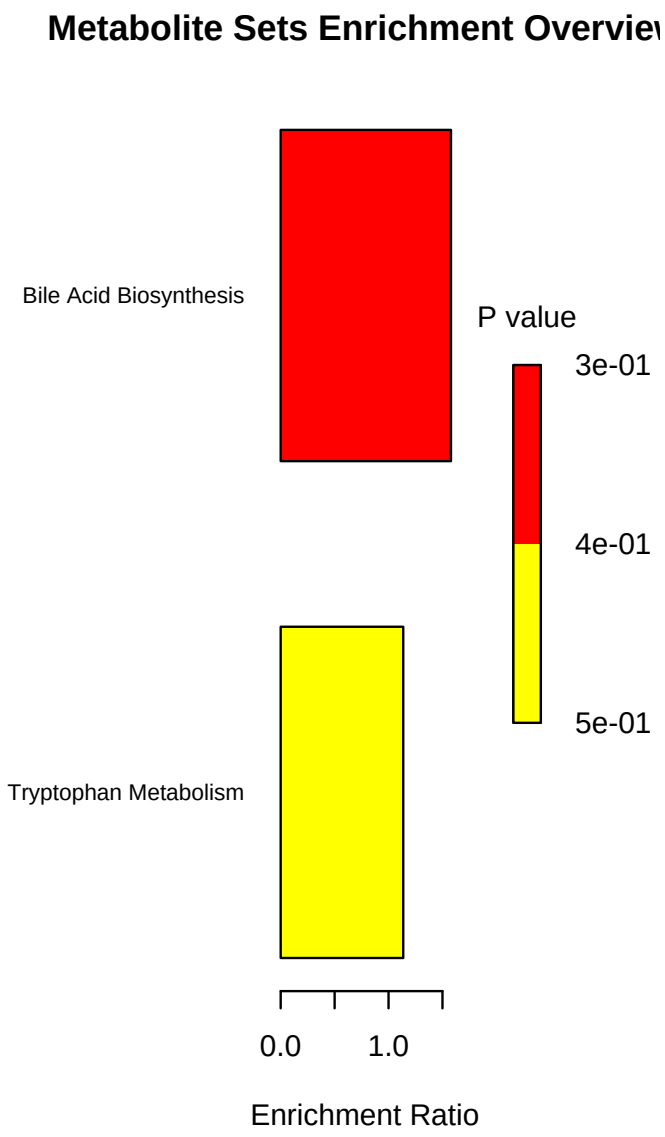

Figure 1: Summary Plot for Over Representation Analysis (ORA)

Table 2: Result from Over Representation Analysis

	total	expected	hits	Raw p	Holm p	FDR
Bile Acid Biosynthesis	65	1.90	3	2.96E-01	1.00E+00	1.00E+00
Tryptophan Metabolism	60	1.76	2	5.35E-01	1.00E+00	1.00E+00

7 Appendix: R Command History

```
 [1] "mSet<-InitDataObjects(\"conc\", \"pathora\", FALSE)"
 [2] "cmpd.vec<-c(\"HMDB0029726\",\"HMDB0033746\",\"HMDB0029730\",\"HMDB0040062\",\"HMDB0040096\",\"
 [3] "mSet<-Setup.MapData(mSet, cmpd.vec);"
 [4] "mSet<-CrossReferencing(mSet, \"hmdb\");"
 [5] "mSet<-CreateMappingResultTable(mSet)"
 [6] "mSet<-SetKEGG.PathLib(mSet, \"hsa\", \"current\")"
 [7] "mSet<-SetMetabolomeFilter(mSet, F);"
 [8] "mSet<-CalculateOraScore(mSet, \"rbc\", \"hyperg\")"
 [9] "mSet<-PlotPathSummary(mSet, F, \"path_view_0_\", \"png\", 72, width=NA)"
[10] "mSet<-SaveTransformedData(mSet)"
[11] "UpdateDataObjects(\"conc\", \"msetora\", FALSE)"
[12] "mSet<-SetMetabolomeFilter(mSet, F);"
[13] "mSet<-SetCurrentMsetLib(mSet, \"smpdb_pathway\", 2);"
[14] "mSet<-CalculateHyperScore(mSet)"
[15] "mSet<-PlotORA(mSet, \"ora_0_\", \"net\", \"png\", 72, width=NA)"
[16] "mSet<-PlotEnrichDotPlot(mSet, \"ora\", \"ora_dot_0_\", \"png\", 72, width=NA)"
[17] "mSet<-CalculateHyperScore(mSet)"
[18] "mSet<-PlotORA(mSet, \"ora_1_\", \"net\", \"png\", 72, width=NA)"
[19] "mSet<-PlotEnrichDotPlot(mSet, \"ora\", \"ora_dot_1_\", \"png\", 72, width=NA)"
[20] "mSet<-SaveTransformedData(mSet)"
[21] "mSet<-PreparePDFReport(mSet, \"guest10606626162765222506\")\n"
```

The report was generated on Mon May 24 12:45:30 2021 with R version 4.0.2 (2020-06-22).