Metabolomic Data Analysis with MetaboAnalyst 6.0

Name: guest2880469127969304155

January 7, 2025

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	C03017	Propionylcarnitine	HMDB0000824	188824	C03017	CCC(=O)O[C@H](CC(O)=O)C[N+](C)(C)C
2	C00307	Citicoline	HMDB0001413	13805	C00307	C[N+](C)(C)CCOP(O)(=O)OP(O)(=O)OC[C@H
3	C00127	Oxidized glutathione	HMDB0003337	65359	C00127	N[C@@H](CCC(=O)N[C@@H](CSSC[C@H](NC(
4	C02990	Palmitoylcarnitine	HMDB0000222	11953816	C02990	CCCCCCCCCCCCCCCC(=O)O[C@H](CC(O)=
5	C02301	O-Acylcarnitine		5355	C02301	
6	C05284	11b-Hydroxyandrost-4-ene-3,17-dione	HMDB0006773	4762	C05284	[H][C@@]12CCC(=O)[C@@]1(C)C[C@H](O)[C@
7	C07074	Lovastatin	HMDB0014372	53232	C07074	[H][C@]12[C@H](C[C@@H](C)C=C1C=C[C@H](
8	C21484	LysoPE(18:0/0:0)	HMDB0011130	9547068	C21484	[H][C@@](O)(COC(=O)CCCCCCCCCCCCCCCC
9	C14635	NA	NA	NA	NA	NA
10	C01657	N-Acetyl-L-tyrosine	HMDB0000866	68310	C01657	CC(=O)N[C@@H](CC1=CC=C(O)C=C1)C(O)=
11	C00051	Glutathione	HMDB0000125	124886	C00051	N[C@@H](CCC(=O)N[C@@H](CS)C(=O)NCC(O
12	C03672	Hydroxyphenyllactic acid	HMDB0000755	9378	C03672	OC(CC1=CC=C(O)C=C1)C(O)=O
13	C00468	Estrone	HMDB0000145	5870	C00468	[H][C@@]12CCC(=O)[C@@]1(C)CC[C@]1([H])C3
14	C09633	1,4-Dimethyl-7-ethylazulene	HMDB0036470	10719	C09633	CCC1=CC2=C(C)C=CC2=C(C)C=C1
15	C10598	NA	NA	NA	NA	NA
16	C00147	Adenine	HMDB0000034	190	C00147	NC1=C2NC=NC2=NC=N1
17	C00170	5'-Methylthioadenosine	HMDB0001173	439176	C00170	CSC[C@H]1O[C@H]([C@H](O)[C@@H]1O)N1C=
18	C19670	Oleamide	HMDB0002117	5283387	C19670	CCCCCCCC\C=C/CCCCCCCC(N)=O
19	C02571	L-Acetylcarnitine	HMDB0000201	7045767	C02571	CC(=O)O[C@H](CC(O)=O)C[N+](C)(C)C
20	C00570	CDP-ethanolamine	HMDB0001564	123727	C00570	NCCOP(O)(=O)OP(O)(=O)OC[C@H]1O[C@H](

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.

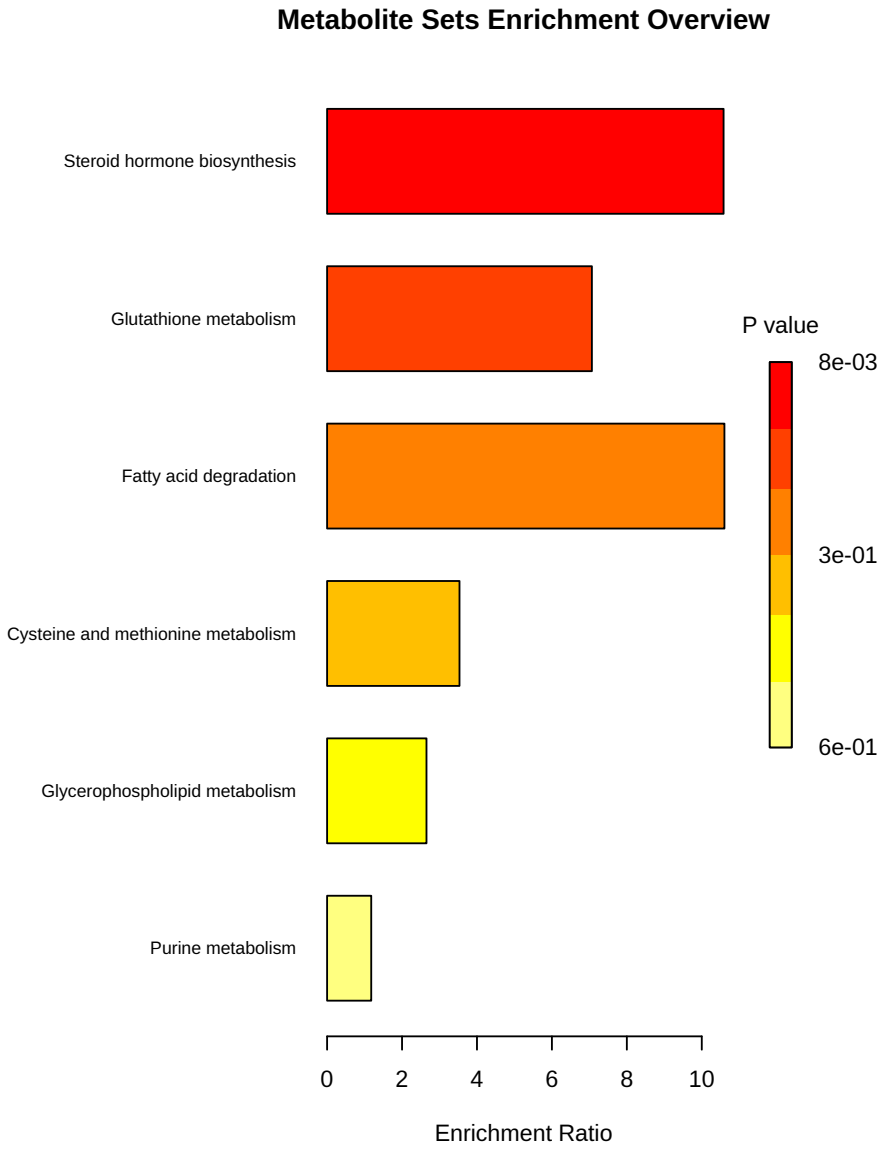

Figure 1: Summary Plot for Over Representation Analysis (ORA)

Table 2: Result from Over Representation Analysis

	total	expected	hits	Raw p	Holm p	FDR
Steroid hormone biosynthesis	2	0.19	2	8.36E-03	6.69E-01	6.69E-01
Glutathione metabolism	3	0.28	2	2.37E-02	1.00E+00	9.48E-01
Fatty acid degradation	1	0.09	1	9.43E-02	1.00E+00	1.00E+00
Cysteine and methionine metabolism	3	0.28	1	2.59E-01	1.00E+00	1.00E+00
Glycerophospholipid metabolism	4	0.38	1	3.30E-01	1.00E+00	1.00E+00
Purine metabolism	9	0.85	1	6.00E-01	1.00E+00	1.00E+00

7 Appendix: R Command History

```
 [1] "mSet<-InitDataObjects(\"conc\", \"msetora\", FALSE)"
 [2] "cmpd.vec<-c(\"C03017\",\"C00307\",\"C00127\",\"C02990\",\"C02301\",\"C05284\",\"C07074\",\"C21
 [3] "mSet<-Setup.MapData(mSet, cmpd.vec);"
 [4] "mSet<-CrossReferencing(mSet, \"kegg\");"
 [5] "mSet<-CreateMappingResultTable(mSet)"
 [6] "mSet<-Setup.HMDBReferenceMetabolome(mSet, \"Comp.txt\");"
 [7] "mSet<-SetMetabolomeFilter(mSet, T);"
 [8] "mSet<-SetCurrentMsetLib(mSet, \"kegg_pathway\", 2);"
 [9] "mSet<-CalculateHyperScore(mSet)"
[10] "mSet<-PlotORA(mSet, \"ora_0_\", \"net\", \"png\", 72, width=NA)"
[11] "mSet<-PlotEnrichDotPlot(mSet, \"ora\", \"ora_dot_0_\", \"png\", 72, width=NA)"
[12] "mSet<-CalculateHyperScore(mSet)"
[13] "mSet<-PlotORA(mSet, \"ora_1_\", \"net\", \"png\", 72, width=NA)"
[14] "mSet<-PlotEnrichDotPlot(mSet, \"ora\", \"ora_dot_1_\", \"png\", 72, width=NA)"
[15] "mSet<-CalculateHyperScore(mSet)"
[16] "mSet<-PlotORA(mSet, \"ora_2_\", \"net\", \"png\", 72, width=NA)"
[17] "mSet<-PlotEnrichDotPlot(mSet, \"ora\", \"ora_dot_2_\", \"png\", 72, width=NA)"
[18] "mSet<-CalculateHyperScore(mSet)"
[19] "mSet<-PlotORA(mSet, \"ora_3_\", \"net\", \"png\", 72, width=NA)"
[20] "mSet<-PlotEnrichDotPlot(mSet, \"ora\", \"ora_dot_3_\", \"png\", 72, width=NA)"
[21] "mSet<-SaveTransformedData(mSet)"
[22] "mSet<-PreparePDFReport(mSet, \"guest2880469127969304155\")\n"
```

The report was generated on Tue Jan 7 09:23:51 2025 with R version 4.3.2 (2023-10-31), OS system: Linux.