Metabolomic Data Analysis with MetaboAnalyst 5.0

Name: guest66266503738778779 87

April 19, 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 Quantitative Enrichment Analysis (QEA) which requires a concentration table. This is the most common data format generated from quantitative metabolomics studies. The phenotype label can be can be discrete (binary or multi-class) or continuous.

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	1,6-Anhydro-beta-D-glucose	Levoglucosan	HMDB0000640	2724705		C1[C@@H]2[C@H]([C@@H]([C@H](O
2	1-Methylnicotinamide	1-Methylnicotinamide	HMDB0000699	457	C02918	C[N+]1=CC=CC(=C1)C(=O)N
3	2-Aminobutyrate	L-Alpha-aminobutyric acid	HMDB0000452	80283	C02356	CC[C@@H](C(=O)O)N
4	2-Hydroxyisobutyrate	Alpha-Hydroxyisobutyric acid	HMDB0000729	11671		CC(C)(C(=O)O)O
5	2-Oxoglutarate	Oxoglutaric acid	HMDB0000208	51	C00026	C(CC(=O)O)C(=O)C(=O)O
6	3-Aminoisobutyrate	3-Aminoisobutanoic acid	HMDB0003911	64956	C05145	CC(CN)C(=O)O
7	3-Hydroxybutyrate	3-Hydroxybutyric acid	HMDB0000357	441	C01089	CC(CC(=O)O)O
8	3-Hydroxyisovalerate	3-Hydroxyisovaleric acid	HMDB0000754	69362	C20827	CC(C)(CC(=O)O)O
9	3-Indoxylsulfate	Indoxyl sulfate	HMDB0000682	10258		C1=CC=C2C(=C1)C(=CN2)OS(=
10	4-Hydroxyphenylacetate	p-Hydroxyphenylacetic acid	HMDB0000020	127	C00642	C1=CC(=CC=C1CC(=O)O)O
11	Acetate	Acetic acid	HMDB0000042	176	C00033	CC(=O)O
12	Acetone	Acetone	HMDB0001659	180	C00207	CC(=O)C
13	Adipate	Adipic acid	HMDB0000448	196	C06104	C(CCC(=O)O)CC(=O)O
14	Alanine	L-Alanine	HMDB0000161	5950	C00041	C[C@@H](C(=O)O)N
15	Asparagine	L-Asparagine	HMDB0000168	6267	C00152	C([C@@H](C(=O)O)N)C(=O)N
16	Betaine	Betaine	HMDB0000043	247	C00719	C[N+](C)(C)CC(=O)[O-]
17	Carnitine	L-Carnitine	HMDB0000062	2724480	C00318	C[N+](C)(C)C[C@H](CC(=O)[O-])
18	Citrate	Citric acid	HMDB0000094	311	C00158	C(C(=O)O)C(CC(=O)O)(C(=O)O
19	Creatine	Creatine	HMDB0000064	586	C00300	CN(CC(=O)O)C(=N)N
20	Creatinine	Creatinine	HMDB0000562	588	C00791	CN1CC(=O)N=C1N
21	Dimethylamine	Dimethylamine	HMDB0000087	674	C00543	CNC
22	Ethanolamine	Ethanolamine	HMDB0000149	700	C00189	C(CO)N
23	Formate	Formic acid	HMDB0000142	284	C00058	C(=O)O
24	Fucose	L-Fucose	HMDB0000174	25310	C01019	C[C@H]1[C@H]([C@H]([C@@H](C(
25	Fumarate	Fumaric acid	HMDB0000134	444972	C00122	C(=C/C(=O)O)\C(=O)O
26	Glucose	D-Glucose	HMDB0000122	5793	C00221	C([C@@H]1[C@H]([C@@H]([C@H](
27	Glutamine	L-Glutamine	HMDB0000641	5961	C00064	C(CC(=O)N)[C@@H](C(=O)O)N
28	Glycine	Glycine	HMDB0000123	750	C00037	C(C(=O)O)N
29	Glycolate	Glycolate		3460	C00160	
30	Guanidoacetate	Guanidoacetic acid	HMDB0000128	763	C00581	C(C(=O)O)N=C(N)N
31	Hippurate	Hippuric acid	HMDB0000714	464	C01586	C1=CC=C(C=C1)C(=O)NCC(=O
32	Histidine	L-Histidine	HMDB0000177	6274	C00135	C1=C(NC=N1)C[C@@H](C(=O)O
33	Hypoxanthine	Hypoxanthine	HMDB0000157	790	C00262	C1=NC2=C(N1)C(=O)N=CN2
34	Isoleucine	L-Isoleucine	HMDB0000172	6306	C00407	CC[C@H](C)[C@@H](C(=O)O)N
35	Lactate	L-Lactic acid	HMDB0000190	61503	C00186	C[C@@H](C(=O)O)O
36	Leucine	L-Leucine	HMDB0000687	6106	C00123	CC(C)C[C@@H](C(=O)O)N
37	Lysine	L-Lysine	HMDB0000182	5962	C00047	C(CCN)C[C@@H](C(=O)O)N
38	Methylamine	Methylamine	HMDB0000164	6329	C00218	CN
39	Methylguanidine	Methylguanidine	HMDB0001522	10111	C02294	CN=C(N)N
40	N,N-Dimethylglycine	Dimethylglycine	HMDB0000092	673	C01026	CN(C)CC(=O)O
41	O-Acetylcarnitine	L-Acetylcarnitine	HMDB0000201	7045767	C02571	CC(=O)OC(CC(=O)[O-])C[N+](C
42	Pantothenate	Pantothenic acid	HMDB0000210	6613	C00864	CC(C)(CO)C(C(=O)NCCC(=O)O
43	Pyroglutamate	Pyroglutamic acid	HMDB0000267	7405	C01879	C1CC(=O)N[C@@H]1C(=O)O
44	Pyruvate	Pyruvic acid	HMDB0000243	1060	C00022	CC(=O)C(=O)O
45	Quinolinate	Quinolinic acid	HMDB0000232	1066	C03722	C1=CC(=C(N=C1)C(=O)O)C(=C
46	Serine	L-Serine	HMDB0000187	5951	C00065	C([C@@H](C(=O)O)N)O
47	Succinate	Succinic acid	HMDB0000254	1110	C00042	C(CC(=O)O)C(=O)O
48	Sucrose	Sucrose	HMDB0000258	5988	C00089	C([C@@H]1[C@H]([C@@H]([C@H](
49	Tartrate	D-Glyceraldehyde 3-phosphate		3418	C00118	
50	Taurine	Taurine	HMDB0000251	1123	C00245	C(CS(=O)(=O)O)N
51	Threonine	L-Threonine	HMDB0000167	6288	C00188	C[C@H]([C@@H](C(=O)O)N)O
52	Trigonelline	Trigonelline	HMDB0000875	5570	C01004	C[N+]1=CC=CC(=C1)C(=O)[O-]
53	Trimethylamine N-oxide	Trimethylamine N-oxide	HMDB0000925	1145	C01104	C[N+](C)(C)[O-]
54	Tryptophan	L-Tryptophan	HMDB0000929	6305	C00078	C1=CC=C2C(=C1)C(=CN2)C[C@
55	Tyrosine	L-Tyrosine	HMDB0000158	6057	C00082	C1=CC(=CC=C1C[C@@H](C(=O
56	Uracil	Uracil	HMDB0000300	1174	C00106	C1=CNC(=O)NC1=O

57	Valine	L-Valine	HMDB0000883	6287	C00183	CC(C)[C@@H](C(=O)O)N
58	Xylose	D-Xylose	HMDB0000098	135191	C00181	C1[C@H]([C@@H]([C@H](C(O1)O)
59	cis-Aconitate	cis-Aconitic acid	HMDB0000072	643757	C00417	C(/C(=C/C(=O)O)/C(=O)O)C(=
60	myo-Inositol	myo-Inositol	HMDB0000211		C00137	O[C@H]1[C@H](O)[C@@H](O)[C@
61	trans-Aconitate	trans-Aconitic acid	HMDB0000958	444212	C02341	C(/C(=C\C(=O)O)/C(=O)O)C(=
62	pi-Methylhistidine	1-Methylhistidine	HMDB0000001	92105	C01152	CN1C=C(N=C1)C[C@@H](C(=O)
63	tau-Methylhistidine	3-Methylhistidine	HMDB0000479	64969	C01152	CN1C=NC=C1C[C@@H](C(=O)O

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

Quantitative enrichment analysis (QEA) will be performed when the user uploads a concentration table. The enrichment analysis is performed using package **globaltest** [3]. It uses a generalized linear model to estimate a *Q-statistic* for each metabolite set, which describes the correlation between compound concentration profiles, X, and clinical outcomes, Y. The *Q statistic* for a metabolite set is the average of the Q statistics for each metabolite in the set. **Figure 2** below summarizes the result.

[3] Jelle J. Goeman, Sara A. van de Geer, Floor de Kort and Hans C. van Houwelingen. *A global test for groups of genes: testing association with a clinical outcome*, Bioinformatics Vol. 20 no. 1 2004, pages 93-99

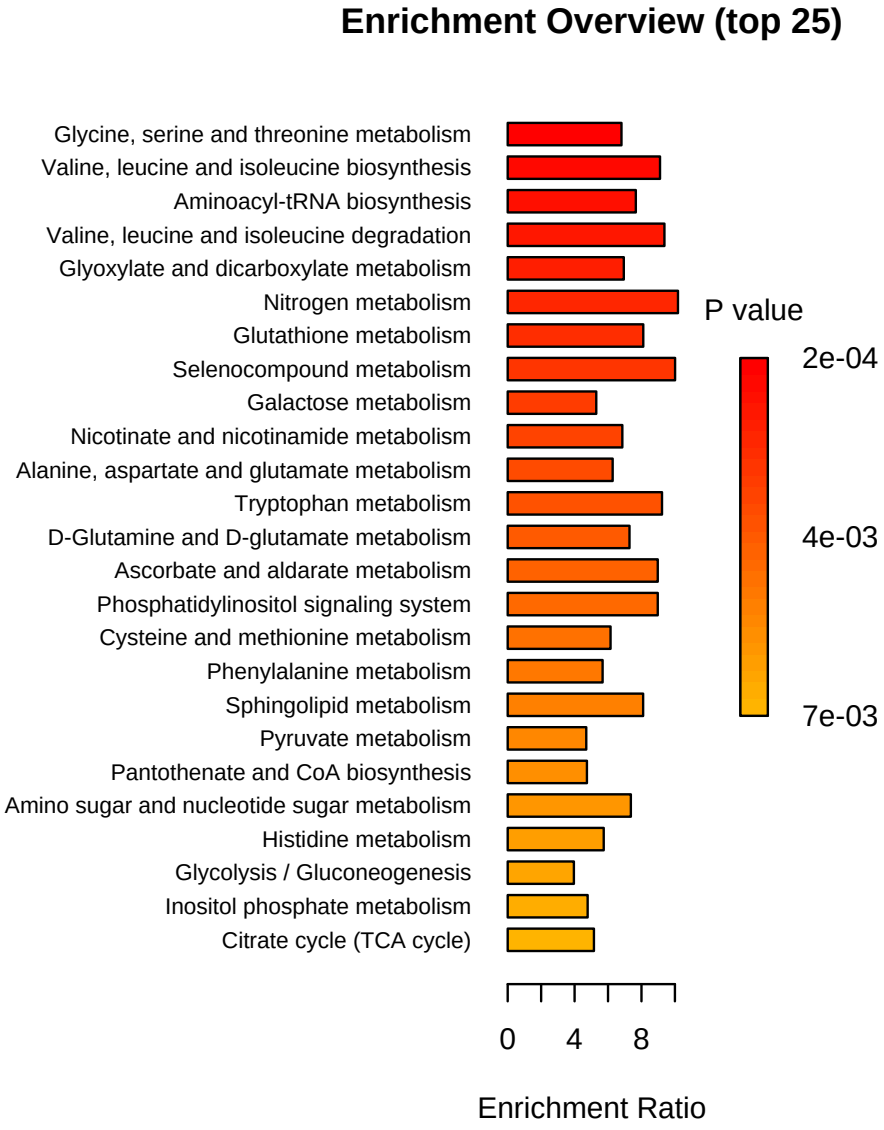

Figure 1: Summary plot for Quantitative Enrichment Analysis (QEA).

Table 2: Result from Quantitative Enrichment Analysis

	Total Cmpd	Hits	Statistic Q	Expected Q	Raw p	Holm p	FDR
Glycine, serine and threonine metabolism	33	8	8.95	1.32	2.46E-04	1.13E-02	6.74E-03
Valine, leucine and isoleucine biosynthesis	8	4	11.99	1.32	4.15E-04	1.87E-02	6.74E-03
Aminoacyl-tRNA biosynthesis	48	13	10.08	1.32	5.94E-04	2.61E-02	6.74E-03
Valine, leucine and isoleucine degradation	40	3	12.33	1.32	6.53E-04	2.81E-02	6.74E-03
Glyoxylate and dicarboxylate metabolism	32	8	9.14	1.32	8.17E-04	3.43E-02	6.74E-03
Nitrogen metabolism	6	1	13.40	1.32	1.06E-03	4.35E-02	6.74E-03
Glutathione metabolism	28	2	10.67	1.32	1.16E-03	4.63E-02	6.74E-03
Selenocompound metabolism	20	1	13.17	1.32	1.18E-03	4.63E-02	6.74E-03
Galactose metabolism	27	3	6.97	1.32	1.49E-03	5.65E-02	6.74E-03
Nicotinate and nicotinamide metabolism	15	2	9.02	1.32	1.49E-03	5.65E-02	6.74E-03
Alanine, aspartate and glutamate metabolism	28	8	8.26	1.32	1.71E-03	6.14E-02	6.74E-03
Tryptophan metabolism	41	1	12.14	1.32	1.90E-03	6.64E-02	6.74E-03
D-Glutamine and D-glutamate metabolism	6	3	9.58	1.32	1.91E-03	6.64E-02	6.74E-03
Ascorbate and aldarate metabolism	8	1	11.81	1.32	2.21E-03	7.31E-02	6.79E-03
Phosphatidylinositol signaling system	28	1	11.81	1.32	2.21E-03	7.31E-02	6.79E-03
Cysteine and methionine metabolism	33	3	8.09	1.32	2.38E-03	7.39E-02	6.85E-03
Phenylalanine metabolism	10	2	7.47	1.32	3.68E-03	1.10E-01	9.60E-03
Sphingolipid metabolism	21	1	10.66	1.32	3.76E-03	1.10E-01	9.60E-03
Pyruvate metabolism	22	4	6.18	1.32	4.14E-03	1.16E-01	1.00E-02
Pantothenate and CoA biosynthesis	19	3	6.24	1.32	5.16E-03	1.39E-01	1.19E-02
Amino sugar and nucleotide sugar metabolism	37	1	9.69	1.32	5.85E-03	1.52E-01	1.26E-02
Histidine metabolism	16	2	7.56	1.32	6.30E-03	1.58E-01	1.26E-02
Glycolysis / Gluconeogenesis	26	4	5.21	1.32	6.32E-03	1.58E-01	1.26E-02
Inositol phosphate metabolism	30	2	6.29	1.32	7.04E-03	1.62E-01	1.33E-02
Citrate cycle (TCA cycle)	20	6	6.79	1.32	7.43E-03	1.64E-01	1.33E-02
Purine metabolism	65	2	7.56	1.32	7.67E-03	1.64E-01	1.33E-02
Pyrimidine metabolism	39	2	6.95	1.32	7.79E-03	1.64E-01	1.33E-02
Arginine biosynthesis	14	3	6.68	1.32	1.12E-02	2.12E-01	1.64E-02
Arginine and proline metabolism	38	3	5.04	1.32	1.12E-02	2.12E-01	1.64E-02
Ubiquinone and other terpenoid-quinone biosynthesis	9	1	8.26	1.32	1.13E-02	2.12E-01	1.64E-02
Phenylalanine, tyrosine and tryptophan biosynthesis	4	1	8.26	1.32	1.13E-02	2.12E-01	1.64E-02
Propanoate metabolism	23	1	8.24	1.32	1.14E-02	2.12E-01	1.64E-02
Primary bile acid biosynthesis	46	2	6.11	1.32	1.20E-02	2.12E-01	1.68E-02
Tyrosine metabolism	42	4	5.22	1.32	1.31E-02	2.12E-01	1.77E-02
Butanoate metabolism	15	2	5.10	1.32	2.15E-02	2.58E-01	2.82E-02
Glycerophospholipid metabolism	36	1	6.39	1.32	2.66E-02	2.92E-01	3.32E-02
Fructose and mannose metabolism	20	2	5.23	1.32	2.67E-02	2.92E-01	3.32E-02
Porphyrin and chlorophyll metabolism	30	1	6.26	1.32	2.81E-02	2.92E-01	3.41E-02
Starch and sucrose metabolism	18	2	4.55	1.32	3.01E-02	2.92E-01	3.55E-02
Taurine and hypotaurine metabolism	8	1	5.96	1.32	3.23E-02	2.92E-01	3.72E-02
Neomycin, kanamycin and gentamicin biosynthesis	2	1	5.90	1.32	3.33E-02	2.92E-01	3.74E-02
beta-Alanine metabolism	21	2	4.41	1.32	4.29E-02	2.92E-01	4.70E-02
Lysine degradation	25	2	3.05	1.32	9.93E-02	3.97E-01	1.06E-01
Pentose and glucuronate interconversions	18	1	2.04	1.32	2.15E-01	6.46E-01	2.25E-01
Biotin metabolism	10	1	1.54	1.32	2.82E-01	6.46E-01	2.88E-01
Pentose phosphate pathway	22	1	0.77	1.32	4.46E-01	6.46E-01	4.46E-01

7 Appendix: R Command History

```
 [1] "mSet<-InitDataObjects(\"conc\", \"msetqea\", FALSE)"
 [2] "mSet<-Read.TextData(mSet, \"Replacing_with_your_file_path\", \"rowu\", \"disc\");"
 [3] "mSet<-SanityCheckData(mSet)"
 [4] "mSet<-ContainMissing(mSet)"
 [5] "mSet<-ReplaceMin(mSet);"
 [6] "mSet<-CrossReferencing(mSet, \"name\");"
 [7] "mSet<-CreateMappingResultTable(mSet)"
 [8] "mSet<-PreparePrenormData(mSet)"
 [9] "mSet<-Normalization(mSet, \"NULL\", \"NULL\", \"AutoNorm\", ratio=FALSE, ratioNum=20)"
[10] "mSet<-PlotNormSummary(mSet, \"norm_0_\", \"png\", 72, width=NA)"
[11] "mSet<-PlotSampleNormSummary(mSet, \"snorm_0_\", \"png\", 72, width=NA)"
[12] "mSet<-SetMetabolomeFilter(mSet, F);"
[13] "mSet<-SetCurrentMsetLib(mSet, \"kegg_pathway\", 2);"
[14] "mSet<-CalculateGlobalTestScore(mSet)"
[15] "mSet<-PlotQEA.Overview(mSet, \"qea_0_\", \"net\", \"png\", 72, width=NA)"
[16] "mSet<-PlotEnrichDotPlot(mSet, \"qea\", \"qea_dot_0_\", \"png\", 72, width=NA)"
[17] "mSet<-PreparePDFReport(mSet, \"guest66266503738778779 87\")\n"
```

The report was generated on Mon Apr 19 23:24:40 2021 with R version 4.0.2 (2020-06-22).