Metabolomic Data Analysis with MetaboAnalyst 5.0

Name: guest4862966280389371363

November 24, 2022

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 M

	Query	Match	HMDB	PubChem	KEGG	SMILES
1	HMDB0000912	Succinyladenosine	HMDB0000912	20849086		C1=NC2=C(C(=N1)N[C@@H]
2	HMDB0029992	Tetrahydropentoxyline	HMDB0029992	53481442		C1C(NC(C2=C1C3=CC=CC=
3	HMDB0004824	N2,N2-Dimethylguanosine	HMDB0004824	92919		CN(C)C1=NC(=O)C2=C(N1)N
4	HMDB0001860	Paraxanthine	HMDB0001860	4687	C13747	CN1C=NC2=C1C(=O)N(C(=
5	HMDB0000193	Isocitric acid	HMDB0000193	1198	C00311	C(C(C(C(=O)O)O)C(=O)O)C
6	HMDB0000072	cis-Aconitic acid	HMDB0000072	643757	C00417	C(/C(=C/C(=O)O)/C(=O)O)
7	HMDB0000058	Cyclic AMP	HMDB0000058	6076	C00575	C1[C@@H]2[C@H]([C@H]([C@
8	HMDB0000299	Xanthosine	HMDB0000299	64959	C01762	C1=NC2=C(N1[C@H]3[C@@H
9	HMDB0000230	N-Acetylneuraminic acid	HMDB0000230	445063	C19910	CC(=O)N[C@@H]1[C@H](C[C
10	HMDB0000893	Suberic acid	HMDB0000893	10457	C08278	C(CCCC(=O)O)CCC(=O)O
11	HMDB0000355	3-Hydroxymethylglutaric acid	HMDB0000355	1662	C03761	CC(CC(=O)O)(CC(=O)O)O
12	HMDB0155722	NA	NA	NA	NA	NA
13	HMDB0000440	3-Hydroxyphenylacetic acid	HMDB0000440	12122	C05593	C1=CC(=CC(=C1)O)CC(=O)
14	HMDB0061384	NA	NA	NA	NA	NA
15	HMDB0000729	Alpha-Hydroxyisobutyric acid	HMDB0000729	11671		CC(C)(C(=O)O)O
16	HMDB0062640	3-hydroxy-2-isobutyrate	HMDB0062640	87	C01188	CC(CO)C(O)=O
17	HMDB0003157	Guanidinosuccinic acid	HMDB0003157	439918	C03139	C([C@@H](C(=O)O)N=C(N)N
18	HMDB0002730	Nicotinamide N-oxide	HMDB0002730	72661		C1=CC(=C[N+](=C1)[O-])C(=
19	HMDB0000625	Gluconic acid	HMDB0000625	10690	C00257	C([C@H]([C@H]([C@@H]([C@H
20	HMDB0011103	1,7-Dimethyluric acid	HMDB0011103	91611	C16356	CN1C2=C(NC1=O)NC(=O)N(
21	HMDB0001107	7-Methylguanosine	HMDB0001107	445404		CN1C=[N+](C2=C1C(=O)N=
22	HMDB0000292	Xanthine	HMDB0000292	1188	C00385	C1=NC2=C(N1)C(=O)NC(=C
23	HMDB0001406	Niacinamide	HMDB0001406	936	C00153	C1=CC(=CN=C1)C(=O)N
24	HMDB0001987	2-Hydroxy-2-methylbutyric acid	HMDB0001987	95433		CCC(C)(C(=O)O)O
25	HMDB0003072	Quinic acid	HMDB0003072	6508	C00296	OC1C[C@@](O)(C[C@@H](O)[
26	HMDB0000812	N-Acetyl-L-aspartic acid	HMDB0000812	65065	C01042	CC(=O)N[C@@H](CC(=O)O)C
27	HMDB0000191	L-Aspartic acid	HMDB0000191	5960	C00049	C([C@@H](C(=O)O)N)C(=O)C
28	HMDB0002802	Cortisone	HMDB0002802	225609	C00762	C[C@]12CCC(=O)C=C1CC[C@
29	HMDB0001713	m-Coumaric acid	HMDB0001713	637541	C12621	C1=CC(=CC(=C1)O)/C=C/C
30	HMDB0000822	p-Hydroxymandelic acid	HMDB0000822	7721	C11527	C1=CC(=CC=C1C(C(=O)O)C
31	HMDB0000842	Quinaldic acid	HMDB0000842	7124	C06325	C1=CC=C2C(=C1)C=CC(=N
32	HMDB0000730	Isobutyrylglycine	HMDB0000730	10855600		CC(C)C(=O)NCC(=O)O
33	HMDB0033143	Pyrraline	HMDB0033143	14274616		C1=C(N(C(=C1)C=O)CCCCC
34	HMDB0000491	3-Methyl-2-oxovaleric acid	HMDB0000491	47	C00671	CCC(C)C(=O)C(=O)O
35	HMDB0000262	Thymine	HMDB0000262	1135	C00178	CC1=CNC(=O)NC1=O
36	HMDB0000201	L-Acetylcarnitine	HMDB0000201	7045767	C02571	CC(=O)OC(CC(=O)[O-])C[N+
37	HMDB0000407	2-Hydroxy-3-methylbutyric acid	HMDB0000407	99823		CC(C)C(C(=O)O)O
38	HMDB0003099	1-Methyluric acid	HMDB0003099	69726	C16359	CN1C(=O)C2=C(NC(=O)N2)N
39	HMDB0013713	N-acetyltryptophan	HMDB0013713	700653		[H][C@@](CC1=CNC2=CC=C
40	HMDB0002035	4-Hydroxycinnamic acid	HMDB0002035	637542	C00811	C1=CC(=CC=C1/C=C/C(=O
41	HMDB0000138	Glycocholic acid	HMDB0000138	23617285	C01921	C[C@H](CCC(=O)NCC(=O)O
42	HMDB0000669	Ortho-Hydroxyphenylacetic acid	HMDB0000669	11970	C05852	C1=CC=C(C(=C1)CC(=O)O)
43	HMDB0000306	Tyramine	HMDB0000306	5610	C00483	C1=CC(=CC=C1CCN)O
44	HMDB0000133	Guanosine	HMDB0000133	6802	C00387	C1=NC2=C(N1[C@H]3[C@@H
45	No result	NA	NA	NA	NA	NA
46	HMDB0000162	L-Proline	HMDB0000162	145742	C00148	C1C[C@H](NC1)C(=O)O
47	HMDB0000172	L-Isoleucine	HMDB0000172	6306	C00407	CC[C@H](C)[C@@H](C(=O)O)
48	HMDB0002432	Sumiki's acid	HMDB0002432	80642	C20448	C1=C(OC(=C1)C(=O)O)CO
49	HMDB0002024	Imidazoleacetic acid	HMDB0002024	96215	C02835	C1=C(NC=N1)CC(=O)O
50	HMDB0028933	Leucyl-Leucine	HMDB0028933	76807	C11332	CC(C)CC(N)C(=O)NC(CC(C)
51	HMDB0000661	Glutaric acid	HMDB0000661	743	C00489	C(CC(=O)O)CC(=O)O
52	HMDB60001	NA	NA	NA	NA	NA
53	HMDB0011180	L-prolyl-L-proline	HMDB0011180	263469		C1CC(NC1)C(=O)N2CCCC2C
54	HMDB0062179	NA	NA	NA	NA	NA
55	HMDB0000752	Methylglutaric acid	HMDB0000752	12284		CC(CC(=O)O)CC(=O)O
56	HMDB0000732	Hydroxykynurenine	HMDB0000732	89	C02794	C1=CC(=C(C(=C1)O)N)C(=C
57	HMDB0000020	p-Hydroxyphenylacetic acid	HMDB0000020	127	C00642	C1=CC(=CC=C1CC(=O)O)O
58	HMDB0244966	NA	NA	NA	NA	NA

59	HMDB0000630	Cytosine	HMDB0000630	597	C00380	C1=C(NC(=O)N=C1)N
60	HMDB0000118	Homovanillic acid	HMDB0000118	1738	C05582	COC1=C(C=CC(=C1)CC(=O
61	HMDB0002721	1-Methylinosine	HMDB0002721	65095		CN1C=NC2=C(C1=O)N=CN2
62	HMDB0000881	Xanthurenic acid	HMDB0000881	5699	C02470	C1=CC2=C(C(=C1)O)NC(=C
63	HMDB0001297	Norcotinine	HMDB0001297	413		C1CC(=O)NC1C2=CN=CC=C
64	HMDB0000512	N-Acetyl-L-phenylalanine	HMDB0000512	74839	C03519	CC(=O)N[C@@H](CC1=CC=C
65	HMDB0000714	Hippuric acid	HMDB0000714	464	C01586	C1=CC=C(C=C1)C(=O)NCC
66	HMDB0000152	Gentisic acid	HMDB0000152	3469	C00628	C1=CC(=C(C=C1O)C(=O)O)
67	HMDB0000132	Guanine	HMDB0000132	764	C00242	C1=NC2=C(N1)C(=O)N=C(N
68	HMDB0000500	4-Hydroxybenzoic acid	HMDB0000500	135	C00156	C1=CC(=CC=C1C(=O)O)O
69	HMDB0240756	NA	NA	NA	NA	NA
70	HMDB0000254	Succinic acid	HMDB0000254	1110	C00042	C(CC(=O)O)C(=O)O
71	HMDB0000259	Serotonin	HMDB0000259	5202	C00780	C1=CC2=C(C=C1O)C(=CN2)
72	HMDB0000784	Azelaic acid	HMDB0000784	2266	C08261	C(CCCC(=O)O)CCCC(=O)O
73	HMDB0001476	3-Hydroxyanthranilic acid	HMDB0001476	86	C00632	C1=CC(=C(C(=C1)O)N)C(=C
74	HMDB0000754	3-Hydroxyisovaleric acid	HMDB0000754	69362	C20827	CC(C)(CC(=O)O)O
75	HMDB0000296	Uridine	HMDB0000296	6029	C00299	C1=CN(C(=O)NC1=O)[C@H]2
76	HMDB0142137	NA	NA	NA	NA	NA
77	HMDB0000687	L-Leucine	HMDB0000687	6106	C00123	CC(C)C[C@@H](C(=O)O)N
78	HMDB0002172	N1,N12-Diacetylspermine	HMDB0002172	132680	C03413	CC(=O)NCCCNCCCCNCCCN
79	HMDB0094713	NA	NA	NA	NA	NA
80	HMDB0001434	3-Methoxytyrosine	HMDB0001434	1670		COC1=C(C=CC(=C1)CC(C(=
81	HMDB0002894	5-Methylcytosine	HMDB0002894	65040	C02376	CC1=C(NC(=O)N=C1)N
82	HMDB0000875	Trigonelline	HMDB0000875	5570	C01004	C[N+]1=CC=CC(=C1)C(=O)[
83	HMDB0000755	Hydroxyphenyllactic acid	HMDB0000755	9378	C03672	C1=CC(=CC=C1CC(C(=O)O
84	HMDB0000821	Phenylacetylglycine	HMDB0000821	68144	C05598	C1=CC=C(C=C1)CC(=O)NC
85	HMDB0061684	N-Acetylisoleucine	HMDB0061684	7036275		CC[C@H](C)[C@H](NC(C)=O)
86	HMDB0240545	N-Methylpyridinium	HMDB0240545	13597	C02724	C[N+]1=CC=CC=C1
87	HMDB0000418	18-Hydroxycortisol	HMDB0000418	44263343		C[C@]12CCC(=O)C=C1CC[C@
88	HMDB0006116	3-Hydroxyhippuric acid	HMDB0006116	450268		C1=CC(=CC(=C1)O)C(=O)N
89	HMDB0002643	3-(3-Hydroxyphenyl)-3-hydroxypropanoic acid	HMDB0002643	102959		C1=CC(=CC(=C1)O)C(CC(=
90	HMDB0000206	N6-Acetyl-L-lysine	HMDB0000206	92832	C02727	CC(=O)NCCCC[C@@H](C(=C
91	HMDB0000251	Taurine	HMDB0000251	1123	C00245	C(CS(=O)(=O)O)N
92	HMDB0003334	Symmetric dimethylarginine	HMDB0003334	169148		CNC(=NC)NCCC[C@@H](C(=
93	HMDB0000097	Choline	HMDB0000097	305	C00114	C[N+](C)(C)CCO
94	HMDB0000684	L-Kynurenine	HMDB0000684	161166	C00328	C1=CC=C(C(=C1)C(=O)C[C
95	HMDB0002825	Theobromine	HMDB0002825	5429	C07480	CN1C=NC2=C1C(=O)NC(=O
96	HMDB0001276	N1-Acetylspermidine	HMDB0001276	496	C00612	CC(=O)NCCCNCCCCN
97	HMDB0000303	Tryptamine	HMDB0000303	1150	C00398	C1=CC=C2C(=C1)C(=CN2)C
98	HMDB0240751	NA	NA	NA	NA	NA
99	HMDB0240295	NA	NA	NA	NA	NA
100	HMDB0000050	Adenosine	HMDB0000050	60961	C00212	C1=NC2=C(C(=N1)N)N=CN2

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.

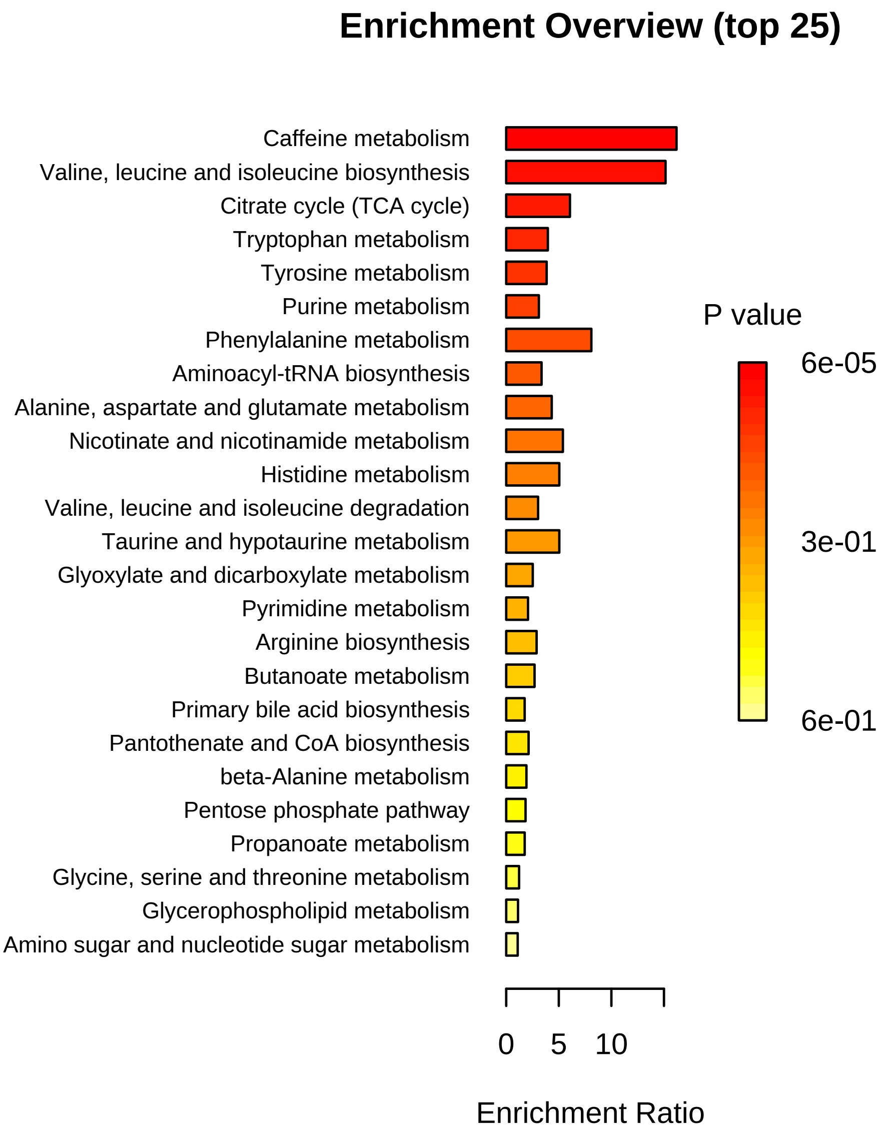

Figure 1: Summary Plot for Over Representation Analysis (ORA)

Table 2: Result from Over Representation Analysis

	total	expected	hits	Raw p	Holm p	FDR
Caffeine metabolism	10	0.25	4	6.03E-05	5.06E-03	5.06E-03
Valine, leucine and isoleucine biosynthesis	8	0.20	3	7.19E-04	5.97E-02	3.02E-02
Citrate cycle (TCA cycle)	20	0.49	3	1.19E-02	9.77E-01	2.77E-01
Tryptophan metabolism	41	1.01	4	1.67E-02	1.00E+00	2.77E-01
Tyrosine metabolism	42	1.04	4	1.82E-02	1.00E+00	2.77E-01
Purine metabolism	65	1.61	5	1.98E-02	1.00E+00	2.77E-01
Phenylalanine metabolism	10	0.25	2	2.37E-02	1.00E+00	2.79E-01
Aminoacyl-tRNA biosynthesis	48	1.19	4	2.84E-02	1.00E+00	2.79E-01
Alanine, aspartate and glutamate metabolism	28	0.69	3	2.99E-02	1.00E+00	2.79E-01
Nicotinate and nicotinamide metabolism	15	0.37	2	5.11E-02	1.00E+00	4.29E-01
Histidine metabolism	16	0.40	2	5.75E-02	1.00E+00	4.39E-01
Valine, leucine and isoleucine degradation	40	0.99	3	7.36E-02	1.00E+00	5.15E-01
Taurine and hypotaurine metabolism	8	0.20	1	1.82E-01	1.00E+00	1.00E+00
Glyoxylate and dicarboxylate metabolism	32	0.79	2	1.86E-01	1.00E+00	1.00E+00
Pyrimidine metabolism	39	0.96	2	2.51E-01	1.00E+00	1.00E+00
Arginine biosynthesis	14	0.35	1	2.97E-01	1.00E+00	1.00E+00
Butanoate metabolism	15	0.37	1	3.14E-01	1.00E+00	1.00E+00
Primary bile acid biosynthesis	46	1.14	2	3.16E-01	1.00E+00	1.00E+00
Pantothenate and CoA biosynthesis	19	0.47	1	3.80E-01	1.00E+00	1.00E+00
beta-Alanine metabolism	21	0.52	1	4.11E-01	1.00E+00	1.00E+00
Pentose phosphate pathway	22	0.54	1	4.26E-01	1.00E+00	1.00E+00
Propanoate metabolism	23	0.57	1	4.40E-01	1.00E+00	1.00E+00
Glycine, serine and threonine metabolism	33	0.82	1	5.66E-01	1.00E+00	1.00E+00
Glycerophospholipid metabolism	36	0.89	1	5.98E-01	1.00E+00	1.00E+00
Amino sugar and nucleotide sugar metabolism	37	0.92	1	6.09E-01	1.00E+00	1.00E+00
Arginine and proline metabolism	38	0.94	1	6.19E-01	1.00E+00	1.00E+00
Steroid hormone biosynthesis	85	2.10	1	8.88E-01	1.00E+00	1.00E+00

7 Appendix: R Command History

```
 [1] "mSet<-InitDataObjects(\"conc\", \"msetora\", FALSE)"
 [2] "cmpd.vec<-c(\"HMDB0000912\",\"HMDB0029992\",\"HMDB0004824\",\"HMDB0001860\",\"HMDB0000193\",\"
 [3] "mSet<-Setup.MapData(mSet, cmpd.vec);"
 [4] "mSet<-CrossReferencing(mSet, \"hmdb\");"
 [5] "mSet<-CreateMappingResultTable(mSet)"
 [6] "mSet<-SetMetabolomeFilter(mSet, F);"
 [7] "mSet<-SetCurrentMsetLib(mSet, \"kegg_pathway\", 2);"
 [8] "mSet<-CalculateHyperScore(mSet)"
 [9] "mSet<-PlotORA(mSet, \"ora_0_\", \"net\", \"png\", 72, width=NA)"
[10] "mSet<-PlotEnrichDotPlot(mSet, \"ora\", \"ora_dot_0_\", \"png\", 72, width=NA)"
[11] "mSet<-CalculateHyperScore(mSet)"
[12] "mSet<-PlotORA(mSet, \"ora_1_\", \"net\", \"png\", 72, width=NA)"
[13] "mSet<-PlotEnrichDotPlot(mSet, \"ora\", \"ora_dot_1_\", \"png\", 72, width=NA)"
[14] "mSet<-SaveTransformedData(mSet)"
[15] "mSet<-PreparePDFReport(mSet, \"guest4862966280389371363\")\n"
```

The report was generated on Thu Nov 24 07:21:14 2022 with R version 4.2.2 (2022-10-31), OS system: Linux, version: -Ubuntu SMP Thu Oct 13 08:03:55 UTC 2022 .