Metabolomic Data Analysis with MetaboAnalyst 6.0

Name: guest10196768329384931757

May 20, 2024

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 Com

	Query	Match	HMDB	PubChem	KEGG	SMILES
1	O-Butanoylcarnitine	Butyrylcarnitine	HMDB0002013	213144	C02862	CCCC(=O)O[C@H](CC([O-])
2	Glutaryl-carnitine	Glutarylcarnitine	HMDB0013130	53481699		C[N+](C)(C)C[C@@H](CC(O
3	Arginine-Glutamine	Arginylglutamine	HMDB0028707	7019985		N[C@@H](CCCNC(N)=N)C(
4	Deoxycarnitine	4-Trimethylammoniobutanoic acid	HMDB0001161	725	C01181	C[N+](C)(C)CCCC([O-])=O
5	N-carbomoyl-L-aspartate	Ureidosuccinic acid	HMDB0000828	93072	C00438	NC(=O)N[C@@H](CC(O)=O
6	Deoxyadenosine	Deoxyadenosine	HMDB0000101	13730	C00559	NC1=C2N=CN([C@H]3C[C@
7	5-Methylcytosine	5-Methylcytosine	HMDB0002894	65040	C02376	CC1=C(N)NC(=O)N=C1
8	Raffinose	Raffinose	HMDB0003213	439242	C00492	OC[C@H]1O[C@@](CO)(O[C
9	Glucose	D-Glucose	HMDB0000122	5793	C00031	OC[C@H]1O[C@@H](O)[C@H
10	L-Octanoylcarnitine	Octanoylcarnitine	HMDB0000791	11953814	C02838	CCCCCCCC(=O)O[C@H](C
11	Malate	Malic acid	HMDB0000156	222656	C00149	O[C@@H](CC(O)=O)C(O)=O
12	Creatinine	Creatinine	HMDB0000562	588	C00791	CN1CC(=O)NC1=N
13	Lysine-Glutamine	Lysylglutamine	HMDB0028949	196305		NCCCC[C@H](N)C(=O)N[C@
14	Serine	Serine	HMDB0000187	5951	C00065	N[C@@H](CO)C(O)=O
15	Arginine-Valine	Arginylvaline	HMDB0028722	6992654		CC(C)[C@H](NC(=O)[C@@H
16	Guanidinoacetate	Guanidoacetic acid	HMDB0000128	763	C00581	NC(=N)NCC(O)=O
17	Uric acid	Uric acid	HMDB0000289	1175	C00366	O=C1NC2=C(N1)C(=O)NC(
18	AMP	Adenosine monophosphate	HMDB0000045	6083	C00020	NC1=C2N=CN([C@@H]3O[C
19	Hypotaurine	Hypotaurine	HMDB0000965	107812	C00519	NCCS(O)=O
20	Guanine	Guanine	HMDB0000132	764	C00242	NC1=NC(=O)C2=C(N1)N=C
21	Asparagine	L-Asparagine	HMDB0000168	6267	C00152	N[C@@H](CC(N)=O)C(O)=O
22	Phenethylamine	Phenylethylamine	HMDB0012275	1001	C05332	NCCC1=CC=CC=C1
23	Thiamine	Thiamine	HMDB0000235	1130	C00378	CC1=C(CCO)SC=[N+]1CC1
24	Histidine	Histidine	HMDB0000177	6274	C00135	N[C@@H](CC1=CN=CN1)C(
25	Cyclic AMP	Cyclic AMP	HMDB0000058	6076	C00575	[H][C@@]12COP(O)(=O)O[C
26	Hypoxanthine	Hypoxanthine	HMDB0000157	790	C00262	OC1=NC=NC2=C1NC=N2
27	2-Hydroxyglutarate	2-Hydroxyglutarate	HMDB0059655	43	C02630	OC(CCC(O)=O)C(O)=O
28	Nicotinamide riboside	Nicotinamide riboside	HMDB0000855	439924	C03150	NC(=O)C1=C[N+](=CC=C1
29	Glutamine	Glutamine	HMDB0000641	5961	C00064	N[C@@H](CCC(N)=O)C(O)=
30	N-acetyl-L-ornithine	N2-Acetylornithine	HMDB0003357	439232	C00437	CC(=O)N[C@@H](CCCN)C(
31	NADH	NADH	HMDB0001487	439153	C00004	NC(=O)C1=CN(C=CC1)[C@
32	Cellobiose	Cellobiose	HMDB0000055	10712	C00185	OC[C@H]1O[C@@H](O[C@H]
33	Norvaline	Norvaline	HMDB0013716	439575	C01799	CCC[C@@H](N)C(O)=O
34	Indole	Indole	HMDB0000738	798	C00463	N1C=CC2=C1C=CC=C2
35	Orotate	Orotic acid	HMDB0000226	967	C00295	OC(=O)C1=CC(=O)NC(=O
36	Fructose	D-Fructose	HMDB0000660	439709	C00095	OC[C@H]1O[C@](O)(CO)[C@
37	Nicotinate	Nicotinic acid	HMDB0001488	938	C00253	OC(=O)C1=CN=CC=C1
38	5- Methylthioadenosine	5'-Methylthioadenosine	HMDB0001173	439176	C00170	CSC[C@H]1O[C@H]([C@H](C
39	Ornithine	Ornithine	HMDB0000214	6262	C00077	NCCC[C@H](N)C(O)=O
40	Uridine	Uridine	HMDB0000296	6029	C00299	OC[C@H]1O[C@H]([C@H](O)
41	Phosphocholine	Phosphorylcholine	HMDB0001565	1014	C00588	C[N+](C)(C)CCOP(O)(O)=O
42	Cytosine	Cytosine	HMDB0000630	597	C00380	NC1=CC=NC(=O)N1
43	L-Palmitoylcarnitine	L-Palmitoylcarnitine	HMDB0240774	16902		[H][C@](CC([O-])=O)(C[N+](
44	UMP	Uridine 5'-monophosphate	HMDB0000288	6030	C00105	O[C@H]1[C@@H](O)[C@@H](
45	Glucuronic acid	D-Glucuronic acid	HMDB0000127	444791	C00191	O[C@H]1O[C@@H]([C@@H](O
46	Carnitine	L-Carnitine	HMDB0000062	10917	C00487	C[N+](C)(C)C[C@H](O)CC(O
47	ADP-D-Glucose	ADP-glucose	HMDB0006557	16500	C00498	NC1=C2N=CN([C@@H]3O[C
48	Montiporic Acid C	NA	NA	NA	NA	NA
49	Montiporic Acid D	NA	NA	NA	NA	NA
50	Dihydroxyacetone phosphate	Dihydroxyacetone phosphate	HMDB0001473	668	C00111	OCC(=O)COP(O)(O)=O
51	Montiporic Acid A	NA	NA	NA	NA	NA
52	Nicotinamide ribotide	Nicotinamide ribotide	HMDB0000229	14181	C00455	NC(=O)C1=C[N+](=CC=C1
53	Montiporic Acid B	NA	NA	NA	NA	NA
54	Porphobilinogen	Porphobilinogen	HMDB0000245	1021	C00931	NCC1=C(CC(O)=O)C(CCC(
55	Tryptophan	L-Tryptophan	HMDB0000929	6305	C00078	N[C@@H](CC1=CNC2=C1C=
56	a-ketoglutarate	Oxoglutaric acid	HMDB0000208	51	C00026	OC(=O)CCC(=O)C(O)=O
57	Citrate	Citric acid	HMDB0000094	311	C00158	OC(=O)CC(O)(CC(O)=O)C
58	ADP	ADP	HMDB0001341	6022	C00008	NC1=NC=NC2=C1N=CN2[C

59	Thymidine	Thymidine	HMDB0000273	5789	C00214	CC1=CN([C@H]2C[C@H](O)
60	N-octanoylglycine	Capryloylglycine	HMDB0000832	84290		CCCCCCCC(=O)NCC(O)=O
61	ADP-ribose	Adenosine diphosphate ribose	HMDB0001178	192	C00301	NC1=C2N=CN(C3OC(COP(
62	1-Methylimidazole acetic acid	Methylimidazoleacetic acid	HMDB0002820	75810	C05828	CN1C=NC(CC(O)=O)=C1
63	GMP	Guanosine monophosphate	HMDB0001397	6804	C00144	NC1=NC2=C(N=CN2[C@@H
64	Cresol	p-Cresol	HMDB0001858	2879	C01468	CC1=CC=C(O)C=C1
65	Itaconic acid	Itaconic acid	HMDB0002092	811	C00490	OC(=O)CC(=C)C(O)=O
66	Carbamoyl phosphate	Carbamoyl phosphate	HMDB0001096	278	C00169	NC(=O)OP(O)(O)=O
67	Phenylalanine	Phenylalanine	HMDB0000159	6140	C00079	N[C@@H](CC1=CC=CC=C1
68	5-methylthioadenosine	5'-Methylthioadenosine	HMDB0001173	439176	C00170	CSC[C@H]1O[C@H]([C@H](C
69	Isoleucine	Isoleucine	HMDB0000172	6306	C00407	CC[C@H](C)[C@H](N)C(O)=
70	4-Aminobutyrate	gamma-Aminobutyric acid	HMDB0000112	119	C00334	NCCCC(O)=O
71	GTP	Guanosine triphosphate	HMDB0001273	6830	C00044	NC1=NC2=C(N=CN2[C@@H
72	Glutamate	Glutamic acid	HMDB0000148	33032	C00025	N[C@@H](CCC(O)=O)C(O)=
73	ATP	Adenosine triphosphate	HMDB0000538	5957	C00002	NC1=NC=NC2=C1N=CN2[C
74	Hippuric acid	Hippuric acid	HMDB0000714	464	C01586	OC(=O)CNC(=O)C1=CC=C
75	Methionine	Methionine	HMDB0000696	6137	C00073	CSCC[C@H](N)C(O)=O
76	Pterin	Pterin	HMDB0000802	73000	C00715	NC1=NC2=NC=CN=C2C(=
77	Glucose-6-phosphate	Glucose 6-phosphate	HMDB0001401	5958	C00092	OC1O[C@H](COP(O)(O)=O)
78	Lumichrome	NA	NA	NA	NA	NA
79	Succinate	Succinic acid	HMDB0000254	1110	C00042	OC(=O)CCC(O)=O
80	Coenzyme A	Coenzyme A	HMDB0001423	87642	C00010	CC(C)(COP(O)(=O)OP(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

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.

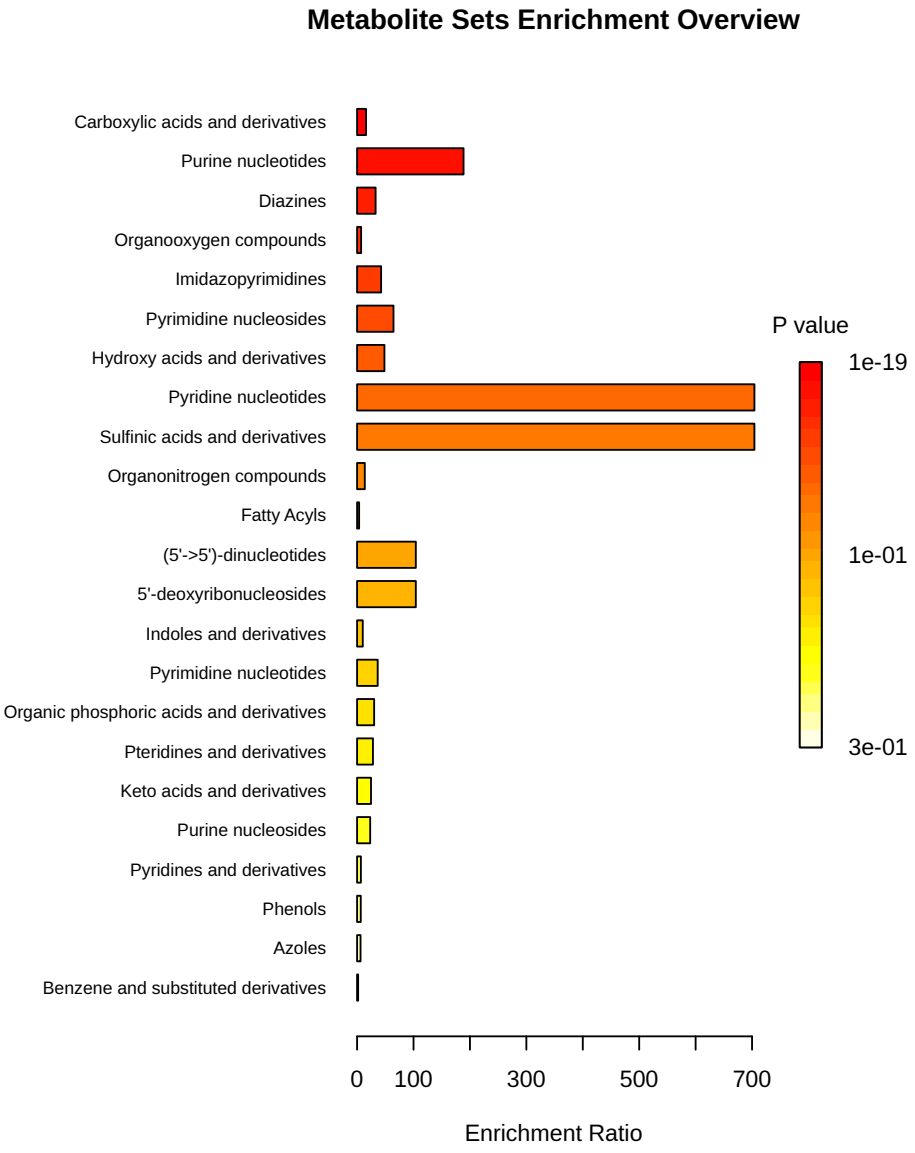

Figure 1: Summary Plot for Over Representation Analysis (ORA)

Table 2: Result from Over Representation Analysis

	total	expected	hits	Raw p	Holm p	FDR
Carboxylic acids and derivatives	3740	1.33	21	1.34E-19	6.37E-17	6.37E-17
Purine nucleotides	134	0.05	9	1.56E-18	7.39E-16	3.70E-16
Diazines	342	0.12	4	7.56E-06	3.58E-03	1.20E-03
Organooxygen compounds	3160	1.12	8	1.73E-05	8.16E-03	2.05E-03
Imidazopyrimidines	198	0.07	3	5.25E-05	2.48E-02	4.99E-03
Pyrimidine nucleosides	87	0.03	2	4.58E-04	2.16E-01	3.64E-02
Hydroxy acids and derivatives	116	0.04	2	8.12E-04	3.82E-01	5.52E-02
Pyridine nucleotides	4	0.00	1	1.42E-03	6.67E-01	6.89E-02
Sulfinic acids and derivatives	4	0.00	1	1.42E-03	6.67E-01	6.89E-02
Organonitrogen compounds	618	0.22	3	1.45E-03	6.76E-01	6.89E-02
Fatty Acyls	4680	1.67	6	6.56E-03	1.00E+00	2.84E-01
(5'->5')-dinucleotides	27	0.01	1	9.57E-03	1.00E+00	3.50E-01
5'-deoxyribonucleosides	27	0.01	1	9.57E-03	1.00E+00	3.50E-01
Indoles and derivatives	559	0.20	2	1.72E-02	1.00E+00	5.83E-01
Pyrimidine nucleotides	77	0.03	1	2.70E-02	1.00E+00	8.58E-01
Organic phosphoric acids and derivatives	93	0.03	1	3.26E-02	1.00E+00	9.69E-01
Pteridines and derivatives	100	0.04	1	3.50E-02	1.00E+00	9.79E-01
Keto acids and derivatives	114	0.04	1	3.98E-02	1.00E+00	1.00E+00
Purine nucleosides	121	0.04	1	4.22E-02	1.00E+00	1.00E+00
Pyridines and derivatives	418	0.15	1	1.38E-01	1.00E+00	1.00E+00
Phenols	434	0.15	1	1.43E-01	1.00E+00	1.00E+00
Azoles	462	0.16	1	1.52E-01	1.00E+00	1.00E+00
Benzene and substituted derivatives	3050	1.09	2	2.96E-01	1.00E+00	1.00E+00

7 Appendix: R Command History

```
 [1] "mSet<-InitDataObjects(\"conc\", \"msetora\", FALSE)"
 [2] "cmpd.vec<-c(\"O-Butanoylcarnitine\",\"Glutaryl-carnitine\",\"Arginine-Glutamine\",\"Deoxycarni
 [3] "mSet<-Setup.MapData(mSet, cmpd.vec);"
 [4] "mSet<-CrossReferencing(mSet, \"name\");"
 [5] "mSet<-CreateMappingResultTable(mSet)"
 [6] "mSet<-PerformDetailMatch(mSet, \"Glutaryl-carnitine\");"
 [7] "mSet<-GetCandidateList(mSet);"
 [8] "mSet<-SetCandidate(mSet, \"Glutaryl-carnitine\", \"Glutarylcarnitine\");"
 [9] "mSet<-PerformDetailMatch(mSet, \"Arginine-Glutamine\");"
[10] "mSet<-GetCandidateList(mSet);"
[11] "mSet<-SetCandidate(mSet, \"Arginine-Glutamine\", \"Arginylglutamine\");"
[12] "mSet<-PerformDetailMatch(mSet, \"N-carbomoyl-L-aspartate\");"
[13] "mSet<-GetCandidateList(mSet);"
[14] "mSet<-SetCandidate(mSet, \"N-carbomoyl-L-aspartate\", \"Ureidosuccinic acid\");"
[15] "mSet<-PerformDetailMatch(mSet, \"Lysine-Glutamine\");"
[16] "mSet<-GetCandidateList(mSet);"
[17] "mSet<-SetCandidate(mSet, \"Lysine-Glutamine\", \"Lysylglutamine\");"
[18] "mSet<-PerformDetailMatch(mSet, \"Arginine-Valine\");"
[19] "mSet<-GetCandidateList(mSet);"
[20] "mSet<-SetCandidate(mSet, \"Arginine-Valine\", \"Arginylvaline\");"
[21] "mSet<-PerformDetailMatch(mSet, \"5- Methylthioadenosine\");"
[22] "mSet<-GetCandidateList(mSet);"
[23] "mSet<-SetCandidate(mSet, \"5- Methylthioadenosine\", \"5'-Methylthioadenosine\");"
[24] "mSet<-PerformDetailMatch(mSet, \"1-Methylimidazole acetic acid\");"
[25] "mSet<-GetCandidateList(mSet);"
[26] "mSet<-SetCandidate(mSet, \"1-Methylimidazole acetic acid\", \"Methylimidazoleacetic acid\");"
[27] "mSet<-PerformDetailMatch(mSet, \"Cresol\");"
[28] "mSet<-GetCandidateList(mSet);"
[29] "mSet<-SetCandidate(mSet, \"Cresol\", \"p-Cresol\");"
[30] "mSet<-PerformDetailMatch(mSet, \"Lumichrome\");"
[31] "mSet<-GetCandidateList(mSet);"
[32] "mSet<-SetMetabolomeFilter(mSet, F);"
[33] "mSet<-SetCurrentMsetLib(mSet, \"main_class\", 2);"
[34] "mSet<-CalculateHyperScore(mSet)"
[35] "mSet<-PlotORA(mSet, \"ora_0_\", \"net\", \"png\", 72, width=NA)"
[36] "mSet<-PlotEnrichDotPlot(mSet, \"ora\", \"ora_dot_0_\", \"png\", 72, width=NA)"
[37] "mSet<-PlotEnrichPieChart(mSet, \"ora\", \"ora_pie_0_\", \"png\", 72)"
[38] "mSet<-CalculateHyperScore(mSet)"
[39] "mSet<-PlotORA(mSet, \"ora_1_\", \"net\", \"png\", 72, width=NA)"
[40] "mSet<-PlotEnrichDotPlot(mSet, \"ora\", \"ora_dot_1_\", \"png\", 72, width=NA)"
[41] "mSet<-PlotEnrichPieChart(mSet, \"ora\", \"ora_pie_1_\", \"png\", 72)"
[42] "mSet<-CalculateHyperScore(mSet)"
[43] "mSet<-PlotORA(mSet, \"ora_2_\", \"net\", \"png\", 72, width=NA)"
[44] "mSet<-PlotEnrichDotPlot(mSet, \"ora\", \"ora_dot_2_\", \"png\", 72, width=NA)"
[45] "mSet<-PlotEnrichPieChart(mSet, \"ora\", \"ora_pie_2_\", \"png\", 72)"
[46] "mSet<-CalculateHyperScore(mSet)"
[47] "mSet<-PlotORA(mSet, \"ora_3_\", \"net\", \"png\", 72, width=NA)"
[48] "mSet<-PlotEnrichDotPlot(mSet, \"ora\", \"ora_dot_3_\", \"png\", 72, width=NA)"
[49] "mSet<-PlotEnrichPieChart(mSet, \"ora\", \"ora_pie_3_\", \"png\", 72)"
[50] "mSet<-CalculateHyperScore(mSet)"
[51] "mSet<-PlotORA(mSet, \"ora_4_\", \"net\", \"png\", 72, width=NA)"
[52] "mSet<-PlotEnrichDotPlot(mSet, \"ora\", \"ora_dot_4_\", \"png\", 72, width=NA)"
[53] "mSet<-PlotEnrichPieChart(mSet, \"ora\", \"ora_pie_4_\", \"png\", 72)"
[54] "mSet<-CalculateHyperScore(mSet)"
[55] "mSet<-PlotORA(mSet, \"ora_5_\", \"net\", \"png\", 72, width=NA)"
[56] "mSet<-PlotEnrichDotPlot(mSet, \"ora\", \"ora_dot_5_\", \"png\", 72, width=NA)"
```

```
[57] "mSet<-PlotEnrichPieChart(mSet, \"ora\", \"ora_pie_5_\", \"png\", 72)"
[58] "mSet<-CalculateHyperScore(mSet)"
[59] "mSet<-PlotORA(mSet, \"ora_6_\", \"net\", \"png\", 72, width=NA)"
[60] "mSet<-PlotEnrichDotPlot(mSet, \"ora\", \"ora_dot_6_\", \"png\", 72, width=NA)"
[61] "mSet<-PlotEnrichPieChart(mSet, \"ora\", \"ora_pie_6_\", \"png\", 72)"
[62] "mSet<-SaveTransformedData(mSet)"
[63] "mSet<-PreparePDFReport(mSet, \"guest10196768329384931757\")\n"
```

The report was generated on Mon May 20 16:41:56 2024 with R version 4.3.2 (2023-10-31), OS system: Linux, version: -Ubuntu SMP Tue Mar 5 20:16:58 UTC 2024 .