Metabolomic Data Analysis with MetaboAnalyst 6.0

Name: guest7267300299023638096

January 8, 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 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 categorical (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 C

	Query	Match	HMDB	PubChem	KEGG	SMILES
1	C00003	NAD	HMDB0000902	5892	C00003	NC(=O)C1=C[N+](=CC=C1)[C@@H]1O[C
2	C00005	NADPH	HMDB0000221	5884	C00005	NC(=O)C1=CN(C=CC1)[C@@H]1O[C@H](
3	C00008	ADP	HMDB0001341	6022	C00008	NC1=NC=NC2=C1N=CN2[C@@H]1O[C@H
4	C00015	Uridine 5'-diphosphate	HMDB0000295	6031	C00015	O[C@H]1[C@@H](O)[C@@H](O[C@@H]1CO
5	C00016	FAD	HMDB0001248	643975	C00016	CC1=CC2=C(C=C1C)N(C[C@H](O)[C@H](
6	C00019	S-Adenosylmethionine	HMDB0001185	34756	C00019	C[S+](CC[C@H](N)C(O)=O)C[C@H]1O[C@
7	C00020	Adenosine monophosphate	HMDB0000045	6083	C00020	NC1=C2N=CN([C@@H]3O[C@H](COP(O)(
8	C00029	Uridine diphosphate glucose	HMDB0000286	8629	C00029	OC[C@H]1O[C@H](OP(O)(=O)OP(O)(=O)
9	C00043	Uridine diphosphate-N-acetylglucosamine	HMDB0000290	445675	C00043	CC(=O)N[C@@H]1[C@@H](O)[C@H](O)[C@
10	C00051	Glutathione	HMDB0000125	124886	C00051	N[C@@H](CCC(=O)N[C@@H](CS)C(=O)N
11	C00052	Uridine diphosphategalactose	HMDB0000302	18068	C00052	OC[C@H]1O[C@H](OP(O)(=O)OP(O)(=O)
12	C00055	Cytidine monophosphate	HMDB0000095	6131	C00055	NC1=NC(=O)N(C=C1)[C@@H]1O[C@H](C
13	C00061	Flavin mononucleotide	HMDB0001520	643976	C00061	CC1=CC2=C(C=C1C)N(C[C@H](O)[C@H](
14	C00062	L-Arginine	HMDB0000517	6322	C00062	N[C@@H](CCCNC(N)=N)C(O)=O
15	C00082	L-Tyrosine	HMDB0000158	6057	C00082	N[C@@H](CC1=CC=C(O)C=C1)C(O)=O
16	C00103	Glucose 1-phosphate	HMDB0001586	65533	C00103	OC[C@H]1O[C@H](OP(O)(O)=O)[C@H](O)
17	C00105	Uridine 5'-monophosphate	HMDB0000288	6030	C00105	O[C@H]1[C@@H](O)[C@@H](O[C@@H]1CO
18	C00118	Glyceraldehyde 3-phosphate	HMDB0001112	439168	C00118	O[C@H](COP(O)(O)=O)C=O
19	C00120	Biotin	HMDB0000030	171548	C00120	[H][C@]12CS[C@@H](CCCCC(O)=O)[C@@]
20	C00127	Oxidized glutathione	HMDB0003337	65359	C00127	N[C@@H](CCC(=O)N[C@@H](CSSC[C@H](
21	C00130	Inosinic acid	HMDB0000175	8582	C00130	O[C@@H]1[C@@H](COP(O)(O)=O)O[C@H]
22	C00144	Guanosine monophosphate	HMDB0001397	6804	C00144	NC1=NC2=C(N=CN2[C@@H]2O[C@H](CO
23	C00147	Adenine	HMDB0000034	190	C00147	NC1=C2NC=NC2=NC=N1
24	C00157	Phosphatidylcholine			C00157	
25	C00158	Citric acid	HMDB0000094	311	C00158	OC(=O)CC(O)(CC(O)=O)C(O)=O
26	C00167	Uridine diphosphate glucuronic acid	HMDB0000935	17473	C00167	O[C@@H]1[C@@H](COP(O)(=O)OP(O)(=O
27	C00170	5'-Methylthioadenosine	HMDB0001173	439176	C00170	CSC[C@H]1O[C@H]([C@H](O)[C@@H]1O)N
28	C00199	D-Ribulose 5-phosphate	HMDB0000618	439184	C00199	OCC(=O)[C@H](O)[C@H](O)COP(O)(O)=
29	C00242	Guanine	HMDB0000132	764	C00242	NC1=NC(=O)C2=C(N1)N=CN2
30	C00257	Gluconic acid	HMDB0000625	10690	C00257	OC[C@@H](O)[C@@H](O)[C@H](O)[C@@H]
31	C00262	Hypoxanthine	HMDB0000157	790	C00262	OC1=NC=NC2=C1NC=N2
32	C00294	Inosine	HMDB0000195	6021	C00294	OC[C@H]1O[C@H]([C@H](O)[C@@H]1O)N1
33	C00299	Uridine	HMDB0000296	6029	C00299	OC[C@H]1O[C@H]([C@H](O)[C@@H]1O)N1
34	C00307	Citicoline	HMDB0001413	13805	C00307	C[N+](C)(C)CCOP(O)(=O)OP(O)(=O)OC[
35	C00319	Sphingosine	HMDB0000252	5280335	C00319	CCCCCCCCCCCCC\C=C\[C@@H](O)[C@
36	C00325	GDP-L-fucose	HMDB0001095	439211	C00325	C[C@@H]1OC(OP(O)(=O)OP(O)(=O)OC[C
37	C00350	PE(14:0/20:1(11Z))	HMDB0008834	52924120	C00350	[H][C@@](COC(=O)CCCCCCCCCCCCC)(C
38	C00360	Deoxyadenosine monophosphate	HMDB0000905	12599	C00360	NC1=NC=NC2=C1N=CN2[C@H]1C[C@H](
39	C00362	2'-Deoxyguanosine 5'-monophosphate	HMDB0001044	65059	C00362	NC1=NC2=C(N=CN2[C@H]2C[C@H](O)[C
40	C00364	5-Thymidylic acid	HMDB0001227	9700	C00364	CC1=CN([C@H]2C[C@H](O)[C@@H](COP(
41	C00378	Thiamine	HMDB0000235	1130	C00378	CC1=C(CCO)SC=[N+]1CC1=CN=C(C)N=
42	C00387	Guanosine	HMDB0000133	6802	C00387	NC1=NC2=C(N=CN2[C@@H]2O[C@H](CO
43	C00463	Indole	HMDB0000738	798	C00463	N1C=CC2=C1C=CC=C2
44	C00487	L-Carnitine	HMDB0000062	10917	C00487	C[N+](C)(C)C[C@H](O)CC(O)=O
45	C00491	L-Cystine	HMDB0000192	67678	C00491	N[C@@H](CSSC[C@H](N)C(O)=O)C(O)=O
46	C00570	CDP-ethanolamine	HMDB0001564	123727	C00570	NCCOP(O)(=O)OP(O)(=O)OC[C@H]1O[C
47	C00588	Phosphorylcholine	HMDB0001565	1014	C00588	C[N+](C)(C)CCOP(O)(O)=O
48	C00612	N1-Acetylspermidine	HMDB0001276	496	C00612	CC(=O)NCCCNCCCCN
49	C00670	Glycerophosphocholine	HMDB0000086	657272	C00670	C[N+](C)(C)CCOP([O-])(=O)OC[C@H](O)C
50	C00836	Sphinganine	HMDB0000269	91486	C00836	CCCCCCCCCCCCCCC[C@@H](O)[C@@H](
51	C00864	Pantothenic acid	HMDB0000210	6613	C00864	CC(C)(CO)[C@@H](O)C(=O)NCCC(O)=O
52	C00946	Adenosine 2'-phosphate	HMDB0011617	53481006	C00946	NC1=NC=NC2=C1N=CN2C1O[C@H](CO)[
53	C01586	Hippuric acid	HMDB0000714	464	C01586	OC(=O)CNC(=O)C1=CC=CC=C1
54	C01657	N-Acetyl-L-tyrosine	HMDB0000866	68310	C01657	CC(=O)N[C@@H](CC1=CC=C(O)C=C1)C(
55	C02301	O-Acylcarnitine		5355	C02301	

56	C02305	Phosphocreatine	HMDB0001511	587	C02305	CN(CC(O)=O)C(=N)NP(O)(O)=O
57	C02494	1-Methyladenosine	HMDB0003331	27476	C02494	CN1C=NC2=C(N=CN2[C@@H]2O[C@H](C
58	C02567	N1-Acetylspermine	HMDB0001186	916	C02567	CC(=O)NCCCNCCCCNCCCN
59	C02571	L-Acetylcarnitine	HMDB0000201	7045767	C02571	CC(=O)O[C@H](CC(O)=O)C[N+](C)(C)C
60	C02862	Butyrylcarnitine	HMDB0002013	213144	C02862	CCCC(=O)O[C@H](CC([O-])=O)C[N+](C)(
61	C02990	Palmitoylcarnitine	HMDB0000222	11953816	C02990	CCCCCCCCCCCCCCCC(=O)O[C@H](CC(
62	C03017	Propionylcarnitine	HMDB0000824	188824	C03017	CCC(=O)O[C@H](CC(O)=O)C[N+](C)(C)C
63	C03546	D-myo-Inositol 4-phosphate	HMDB0001313	440043	C03546	O[C@@H]1[C@H](O)[C@H](O)[C@@H](OP(
64	C03742	(S)-4-Hydroxymandelonitrile	HMDB0060318	440104	C03742	O[C@H](C#N)C1=CC=C(O)C=C1
65	C03794	Adenylsuccinic acid	HMDB0000536	447145	C03794	O[C@@H]1[C@@H](COP(O)(O)=O)O[C@H]
66	C03889	NA	NA	NA	NA	NA
67	C04100	NA	NA	NA	NA	NA
68	C04230	CE(22:5(7Z,10Z,13Z,16Z,19Z))	HMDB0010375	24779458	C04230	CC\C=C/C\C=C/C\C=C/C\C=C/C\C=C
69	C05282	gamma-Glutamylglutamic acid	HMDB0011737	92865	C05282	N[C@@H](CCC(=O)N[C@@H](CCC(O)=O)
70	C05382	D-Sedoheptulose 7-phosphate	HMDB0001068	92042786	C05382	OC[C@]1(O)O[C@H](COP(O)(O)=O)[C@@H
71	C05526	S-Glutathionyl-L-cysteine	METPA0607		C05526	
72	C05551	Penicillin G	HMDB0015186	5904	C05551	[H][C@]12SC(C)(C)[C@@H](N1C(=O)[C@H]
73	C05635	5-Hydroxyindoleacetic acid	HMDB0000763	1826	C05635	OC(=O)CC1=CNC2=C1C=C(O)C=C2
74	C06525	Gentianine	HMDB0303030	354616	C06525	C=CC1=C2CCOC(=O)C2=CN=C1
75	C07005	Flunisolide	HMDB0014326	82153	C07005	[H][C@@]12C[C@@]3([H])[C@]4([H])C[C@H](
76	C07471	Methacholine	HMDB0015654	1993	C07471	CC(C[N+](C)(C)C)OC(C)=O
77	C07968	Diethylcarbamazine	HMDB0014849	3052	C07968	CCN(CC)C(=O)N1CCN(C)CC1
78	C10598	NA	NA	NA	NA	NA
79	C13916	Cer(d18:1/14:0)	HMDB0011773	5282310	C13916	CCCCCCCCCCCCC\C=C\[C@@H](O)[C@
80	C14550	4-Nonylphenol	HMDB0038982	1752	C14550	CCCCCCCCCC1=CC=C(O)C=C1
81	C14785	1,4-Dihydroxynaphthalene	HMDB0255445	11305	C14785	OC1=CC=C(O)C2=CC=CC=C12
82	C17925	Methylnoradrenaline	HMDB0002832	3917	C17925	CC(N)C(O)C1=CC(O)=C(O)C=C1
83	C19434	NA	NA	NA	NA	NA
84	C19463	1,5-Naphthalenediamine	HMDB0244231	16720	C19463	NC1=CC=CC2=C1C=CC=C2N
85	C19670	Oleamide	HMDB0002117	5283387	C19670	CCCCCCCC\C=C/CCCCCCCC(N)=O
86	C20387	Biotin sulfone	HMDB0004818	21252323	C20387	[H][C@]12CS(=O)(=O)[C@@H](CCCCC(O)=
87	C21484	LysoPE(18:0/0:0)	HMDB0011130	9547068	C21484	[H][C@@](O)(COC(=O)CCCCCCCCCCCCC

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

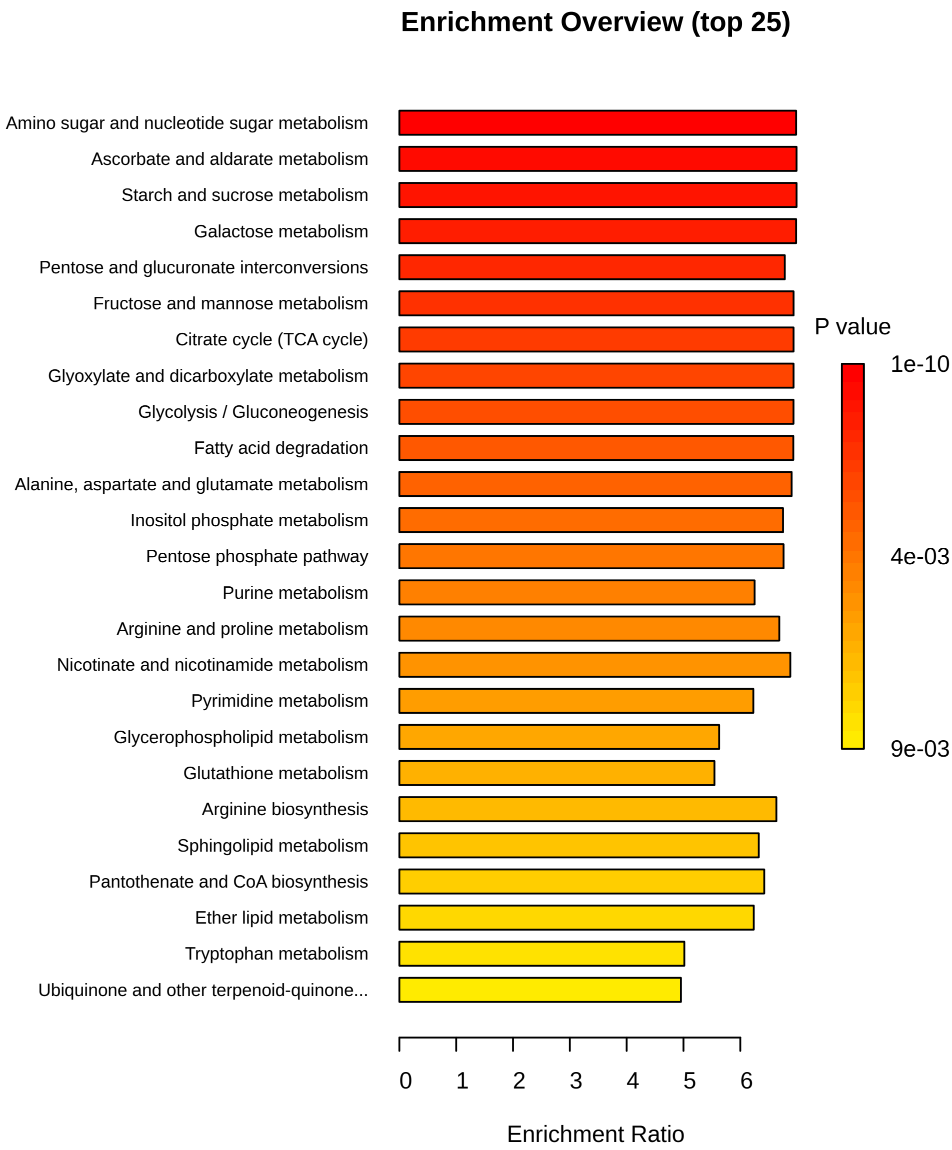

Figure 1: Summary plot for Quantitative Enrichment Analysis (QEA).

Table 2: Result from Quantitative Enrichment Analysis

	Total Cmpd	Hits	Statistic Q	Expected Q	Raw p	Holm p	FDR
Amino sugar and nucleotide sugar metabolism	42	4	99.75	14.29	1.49E-10	5.36E-09	3.82E-09
Ascorbate and aldarate metabolism	9	2	99.86	14.29	2.12E-10	7.42E-09	3.82E-09
Starch and sucrose metabolism	18	1	99.86	14.29	8.03E-10	2.73E-08	9.32E-09
Galactose metabolism	27	2	99.75	14.29	1.04E-09	3.42E-08	9.32E-09
Pentose and glucuronate interconversions	19	3	96.90	14.29	2.99E-08	9.56E-07	2.15E-07
Fructose and mannose metabolism	20	2	99.11	14.29	1.53E-07	4.74E-06	9.18E-07
Citrate cycle (TCA cycle)	20	1	99.11	14.29	2.23E-07	6.70E-06	9.40E-07
Glyoxylate and dicarboxylate metabolism	31	1	99.11	14.29	2.23E-07	6.70E-06	9.40E-07
Glycolysis / Gluconeogenesis	26	1	99.09	14.29	2.35E-07	6.70E-06	9.40E-07
Fatty acid degradation	39	1	99.05	14.29	2.72E-07	7.33E-06	9.78E-07
Alanine, aspartate and glutamate metabolism	28	2	98.62	14.29	4.04E-07	1.05E-05	1.26E-06
Inositol phosphate metabolism	30	2	96.46	14.29	4.19E-07	1.05E-05	1.26E-06
Pentose phosphate pathway	23	3	96.60	14.29	4.89E-07	1.17E-05	1.35E-06
Purine metabolism	70	12	89.32	14.29	7.22E-07	1.66E-05	1.86E-06
Arginine and proline metabolism	36	3	95.52	14.29	8.06E-07	1.77E-05	1.94E-06
Nicotinate and nicotinamide metabolism	15	1	98.30	14.29	1.56E-06	3.28E-05	3.51E-06
Pyrimidine metabolism	39	5	89.00	14.29	1.40E-05	2.80E-04	2.96E-05
Glycerophospholipid metabolism	36	5	80.40	14.29	3.51E-05	6.67E-04	7.03E-05
Glutathione metabolism	28	3	79.22	14.29	4.29E-05	7.72E-04	8.13E-05
Arginine biosynthesis	14	1	94.79	14.29	4.52E-05	7.72E-04	8.13E-05
Sphingolipid metabolism	32	2	90.36	14.29	1.33E-04	2.13E-03	2.28E-04
Pantothenate and CoA biosynthesis	20	1	91.74	14.29	1.82E-04	2.73E-03	2.98E-04
Ether lipid metabolism	20	1	89.10	14.29	4.22E-04	5.91E-03	6.61E-04
Tryptophan metabolism	41	1	71.65	14.29	8.04E-03	1.05E-01	1.13E-02
Ubiquinone and other terpenoid-quinone biosynthesis	18	1	70.79	14.29	8.83E-03	1.06E-01	1.13E-02
Tyrosine metabolism	42	1	70.79	14.29	8.83E-03	1.06E-01	1.13E-02
Phenylalanine metabolism	8	1	70.79	14.29	8.83E-03	1.06E-01	1.13E-02
Phenylalanine, tyrosine and tryptophan biosynthesis	4	1	70.79	14.29	8.83E-03	1.06E-01	1.13E-02
Thiamine metabolism	7	1	66.95	14.29	1.30E-02	1.06E-01	1.62E-02
Riboflavin metabolism	4	2	44.16	14.29	6.23E-02	4.36E-01	7.46E-02
Lysine degradation	30	1	46.07	14.29	6.42E-02	4.36E-01	7.46E-02
Biotin metabolism	10	1	40.53	14.29	8.96E-02	4.48E-01	1.01E-01
Arachidonic acid metabolism	44	1	25.70	14.29	2.00E-01	7.99E-01	2.05E-01
Linoleic acid metabolism	5	1	25.70	14.29	2.00E-01	7.99E-01	2.05E-01
alpha-Linolenic acid metabolism	13	1	25.70	14.29	2.00E-01	7.99E-01	2.05E-01
Cysteine and methionine metabolism	33	3	19.28	14.29	2.58E-01	7.99E-01	2.58E-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<-ReplaceMin(mSet);"
 [5] "mSet<-CrossReferencing(mSet, \"kegg\");"
 [6] "mSet<-CreateMappingResultTable(mSet)"
 [7] "mSet<-PreparePrenormData(mSet)"
 [8] "mSet<-SanityCheckData(mSet)"
 [9] "mSet<-FilterVariable(mSet, \"F\", 25, \"iqr\", 0, \"mean\", 0)"
[10] "mSet<-PreparePrenormData(mSet)"
[11] "mSet<-Normalization(mSet, \"NULL\", \"NULL\", \"NULL\", ratio=FALSE, ratioNum=20)"
[12] "mSet<-PlotNormSummary(mSet, \"norm_0_\", \"png\", 72, width=NA)"
[13] "mSet<-PlotSampleNormSummary(mSet, \"snorm_0_\", \"png\", 72, width=NA)"
[14] "mSet<-SetMetabolomeFilter(mSet, F);"
[15] "mSet<-SetCurrentMsetLib(mSet, \"kegg_pathway\", 2);"
[16] "mSet<-CalculateGlobalTestScore(mSet)"
[17] "mSet<-PlotQEA.Overview(mSet, \"qea_0_\", \"net\", \"png\", 72, width=NA)"
[18] "mSet<-PlotEnrichDotPlot(mSet, \"qea\", \"qea_dot_0_\", \"png\", 72, width=NA)"
[19] "mSet<-PlotQEA.MetSet(mSet, \"Amino sugar and nucleotide sugar metabolism\", \"png\", 72, width
[20] "mSet<-PlotQEA.MetSet(mSet, \"Amino sugar and nucleotide sugar metabolism\", \"png\", 72, width
[21] "mSet<-PlotQEA.MetSet(mSet, \"Pentose and glucuronate interconversions\", \"png\", 72, width=NA
[22] "mSet<-SaveTransformedData(mSet)"
[23] "mSet<-PreparePDFReport(mSet, \"guest7267300299023638096\")\n"
```

The report was generated on Wed Jan 8 09:34:45 2025 with R version 4.3.2 (2023-10-31), OS system: Linux.