Metabolomic Data Analysis with MetaboAnalyst 6.0

Name: guest3628632976710346520

April 27, 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 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*

	Query	Match	HMDB	PubChem	KEGG	SMILES
1	C22350	NA	NA	NA	NA	NA
2	C00379	D-Xylitol	HMDB0002917	6912	C00379	OC[C@H](O)[C@@H](O)[C@H](O)CO
3	C00474	Ribitol	HMDB0000508		C00474	OC[C@H](O)[C@H](O)[C@H](O)CO
4	C01487	Allose	HMDB0001151	12285879	C01487	[H][C@@]1(CO)OC(O)[C@@]([H])(O)[C@@]([H])(O)[C@@]1([H])O
5	C08374	Capsicoside E	HMDB0031443	8182	C08374	[H][C@]12C[C@@]3([H])[C@]4([H])CC[C@@]5([H])C[C@@H](O[C@@
6	C01697	Galactitol	HMDB0000107	11850	C01697	OC[C@H](O)[C@@H](O)[C@@H](O)[C@H](O)CO
7	C00089	Sucrose	HMDB0000258	5988	C00089	OC[C@H]1O[C@@](CO)(O[C@H]2O[C@H](CO)[C@@H](O)[C@H](
8	C01380	Polyethylene glycol	HMDB0037790	174	C01380	[H]OCCO[H]
9	C00031	D-Glucose	HMDB0000122	5793	C00031	OC[C@H]1O[C@@H](O)[C@H](O)[C@@H](O)[C@@H]1O
10	C00137	myo-Inositol	HMDB0000211		C00137	O[C@H]1[C@H](O)[C@@H](O)[C@H](O)[C@H](O)[C@@H]1O
11	C02457	Propane-1,3-diol	METPA0292		C02457	
12	C00095	D-Fructose	HMDB0000660	439709	C00095	OC[C@H]1O[C@](O)(CO)[C@@H](O)[C@@H]1O
13	C14214	Dibutyl phthalate	HMDB0033244	3026	C14214	CCCCOC(=O)C1=CC=CC=C1C(=O)OCCCC
14	C00116	Glycerol	HMDB0000131	753	C00116	OCC(O)CO
15	C01432	NA	NA	NA	NA	NA
16	C00243	Alpha-Lactose	HMDB0000186	84571	C00243	OC[C@H]1O[C@@H](O[C@H]2[C@H](O)[C@@H](O)[C@@H](O)O[
17	C00249	Palmitic acid	HMDB0000220	985	C00249	CCCCCCCCCCCCCCCC(O)=O
18	C06467	NA	NA	NA	NA	NA

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

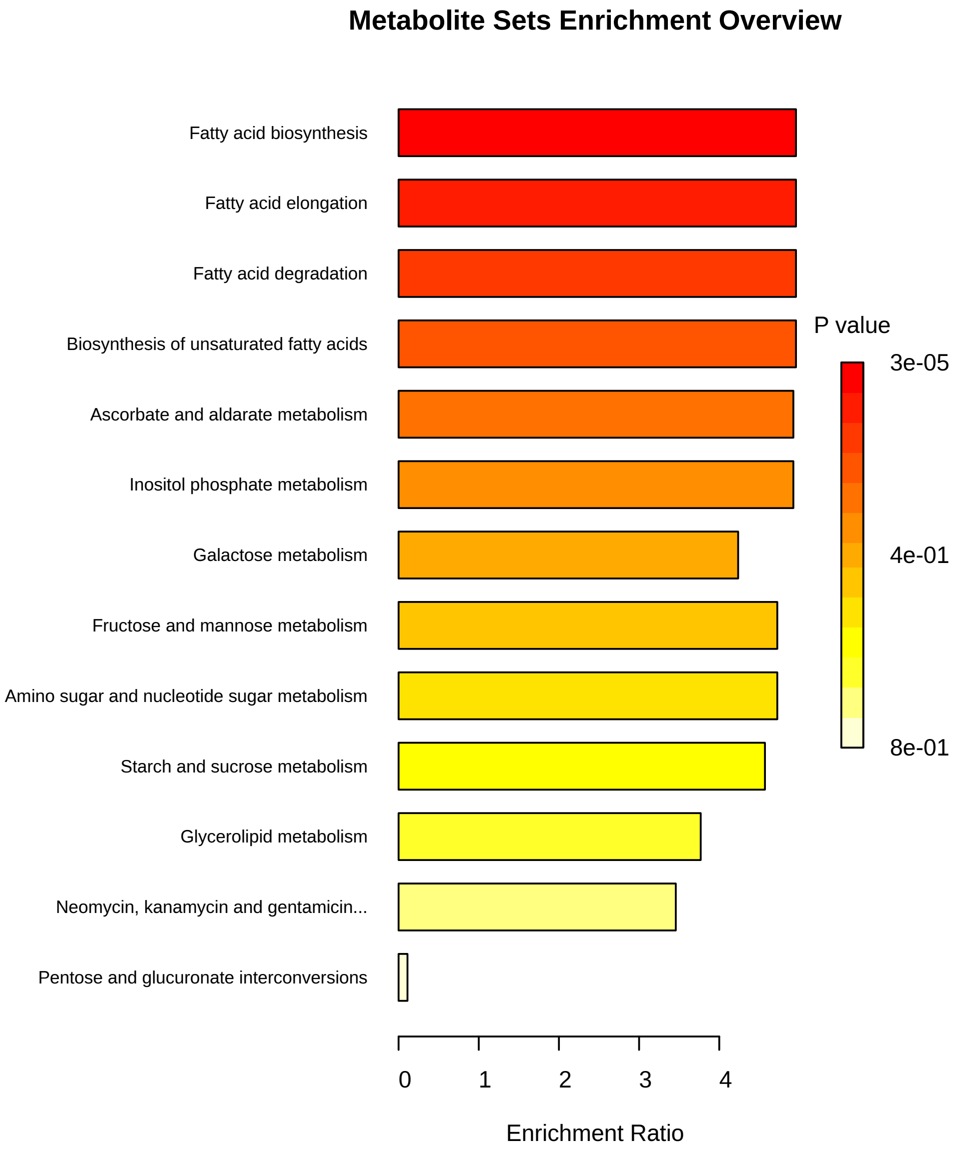

Figure 1: Summary plot for Quantitative Enrichment Analysis (QEA).

Table 2: Result from Quantitative Enrichment Analysis

	Total Cmpd	Hits	Statistic Q	Expected Q	Raw p	Holm p	FDR
Fatty acid biosynthesis	47	1	99.14	20.00	2.76E-05	3.59E-04	8.97E-05
Fatty acid elongation	38	1	99.14	20.00	2.76E-05	3.59E-04	8.97E-05
Fatty acid degradation	39	1	99.14	20.00	2.76E-05	3.59E-04	8.97E-05
Biosynthesis of unsaturated fatty acids	36	1	99.14	20.00	2.76E-05	3.59E-04	8.97E-05
Ascorbate and aldarate metabolism	9	1	98.49	20.00	8.57E-05	7.72E-04	1.86E-04
Inositol phosphate metabolism	30	1	98.49	20.00	8.57E-05	7.72E-04	1.86E-04
Galactose metabolism	27	7	84.71	20.00	7.04E-04	4.93E-03	1.31E-03
Fructose and mannose metabolism	20	1	94.48	20.00	1.16E-03	6.98E-03	1.68E-03
Amino sugar and nucleotide sugar metabolism	42	1	94.48	20.00	1.16E-03	6.98E-03	1.68E-03
Starch and sucrose metabolism	18	3	91.40	20.00	2.07E-03	8.30E-03	2.70E-03
Glycerolipid metabolism	16	1	75.38	20.00	2.49E-02	7.47E-02	2.94E-02
Neomycin, kanamycin and gentamicin biosynthesis	2	1	69.16	20.00	4.01E-02	8.03E-02	4.35E-02
Pentose and glucuronate interconversions	19	1	2.22	20.00	7.78E-01	7.78E-01	7.78E-01

7 Appendix: R Command History

```
 [1] "mSet<-InitDataObjects(\"conc\", \"pathqea\", FALSE)"
 [2] "mSet<-Read.TextData(mSet, \"Replacing_with_your_file_path\", \"colu\", \"cont\");"
 [3] "mSet<-SanityCheckData(mSet)"
 [4] "mSet<-ReplaceMin(mSet);"
 [5] "mSet<-CrossReferencing(mSet, \"kegg\");"
 [6] "mSet<-CreateMappingResultTable(mSet)"
 [7] "mSet<-PreparePrenormData(mSet)"
 [8] "mSet<-Normalization(mSet, \"NULL\", \"NULL\", \"NULL\", ratio=FALSE, ratioNum=20)"
 [9] "mSet<-PlotNormSummary(mSet, \"norm_0_\", \"png\", 72, width=NA)"
[10] "mSet<-PlotSampleNormSummary(mSet, \"snorm_0_\", \"png\", 72, width=NA)"
[11] "mSet<-SetKEGG.PathLib(mSet, \"ath\", \"current\")"
[12] "mSet<-SetMetabolomeFilter(mSet, F);"
[13] "mSet<-CalculateQeaScore(mSet, \"rbc\", \"gt\")"
[14] "mSet<-PlotPathSummary(mSet, F, \"path_view_0_\", \"png\", 72, width=NA, NA, NA )"
[15] "mSet<-PlotKEGGPath(mSet, \"Starch and sucrose metabolism\",576, 480, \"png\", NULL)"
[16] "mSet<-RerenderMetPAGraph(mSet, \"zoom1714223022417.png\",576.0, 480.0, 100.0)"
[17] "mSet<-RerenderMetPAGraph(mSet, \"zoom1714223142349.png\",576.0, 480.0, 100.0)"
[18] "mSet<-RerenderMetPAGraph(mSet, \"zoom1714223146041.png\",576.0, 480.0, 100.0)"
[19] "mSet<-RerenderMetPAGraph(mSet, \"zoom1714223147546.png\",576.0, 480.0, 100.0)"
[20] "mSet<-RerenderMetPAGraph(mSet, \"zoom1714223149873.png\",576.0, 480.0, 100.0)"
[21] "mSet<-RerenderMetPAGraph(mSet, \"zoom1714223150861.png\",576.0, 480.0, 100.0)"
[22] "mSet<-RerenderMetPAGraph(mSet, \"zoom1714223152260.png\",576.0, 480.0, 100.0)"
[23] "mSet<-RerenderMetPAGraph(mSet, \"zoom1714223152897.png\",576.0, 480.0, 100.0)"
[24] "mSet<-RerenderMetPAGraph(mSet, \"zoom1714223153546.png\",576.0, 480.0, 100.0)"
[25] "mSet<-RerenderMetPAGraph(mSet, \"zoom1714223154381.png\",576.0, 480.0, 100.0)"
[26] "mSet<-RerenderMetPAGraph(mSet, \"zoom1714223155095.png\",576.0, 480.0, 100.0)"
[27] "mSet<-RerenderMetPAGraph(mSet, \"zoom1714223155906.png\",576.0, 480.0, 100.0)"
[28] "mSet<-PlotKEGGPath(mSet, \"Starch and sucrose metabolism\",576, 480, \"png\", NULL)"
[29] "mSet<-PlotKEGGPath(mSet, \"Galactose metabolism\",576, 480, \"png\", NULL)"
[30] "mSet<-PlotKEGGPath(mSet, \"Glycerolipid metabolism\",576, 480, \"png\", NULL)"
[31] "mSet<-PlotKEGGPath(mSet, \"Inositol phosphate metabolism\",576, 480, \"png\", NULL)"
[32] "mSet<-PlotKEGGPath(mSet, \"Fructose and mannose metabolism\",576, 480, \"png\", NULL)"
[33] "mSet<-PlotKEGGPath(mSet, \"Amino sugar and nucleotide sugar metabolism\",576, 480, \"png\", NU
[34] "mSet<-PlotKEGGPath(mSet, \"Fatty acid biosynthesis\",576, 480, \"png\", NULL)"
[35] "mSet<-PlotKEGGPath(mSet, \"Starch and sucrose metabolism\",576, 480, \"png\", NULL)"
[36] "mSet<-PlotKEGGPath(mSet, \"Galactose metabolism\",576, 480, \"png\", NULL)"
[37] "mSet<-PlotKEGGPath(mSet, \"Glycerolipid metabolism\",576, 480, \"png\", NULL)"
[38] "mSet<-PlotKEGGPath(mSet, \"Galactose metabolism\",576, 480, \"png\", NULL)"
[39] "mSet<-PlotKEGGPath(mSet, \"Starch and sucrose metabolism\",576, 480, \"png\", NULL)"
[40] "mSet<-SaveTransformedData(mSet)"
[41] "mSet<-PreparePDFReport(mSet, \"guest3628632976710346520\")\n"
[42] "UpdateDataObjects(\"conc\", \"msetqea\", FALSE)"
[43] "mSet<-SanityCheckData(mSet)"
[44] "mSet<-ReplaceMin(mSet);"
[45] "mSet<-CrossReferencing(mSet, \"kegg\");"
[46] "mSet<-CreateMappingResultTable(mSet)"
[47] "mSet<-PreparePrenormData(mSet)"
[48] "mSet<-Normalization(mSet, \"NULL\", \"LogNorm\", \"ParetoNorm\", ratio=FALSE, ratioNum=20)"
[49] "mSet<-PlotNormSummary(mSet, \"norm_1_\", \"png\", 72, width=NA)"
[50] "mSet<-PlotSampleNormSummary(mSet, \"snorm_1_\", \"png\", 72, width=NA)"
[51] "mSet<-Normalization(mSet, \"MedianNorm\", \"LogNorm\", \"ParetoNorm\", ratio=FALSE, ratioNum=2
[52] "mSet<-PlotNormSummary(mSet, \"norm_2_\", \"png\", 72, width=NA)"
[53] "mSet<-PlotSampleNormSummary(mSet, \"snorm_2_\", \"png\", 72, width=NA)"
[54] "mSet<-Normalization(mSet, \"SumNorm\", \"LogNorm\", \"ParetoNorm\", ratio=FALSE, ratioNum=20)"
[55] "mSet<-PlotNormSummary(mSet, \"norm_3_\", \"png\", 72, width=NA)"
[56] "mSet<-PlotSampleNormSummary(mSet, \"snorm_3_\", \"png\", 72, width=NA)"
```

```
[57] "mSet<-Normalization(mSet, \"NULL\", \"LogNorm\", \"MeanCenter\", ratio=FALSE, ratioNum=20)"
[58] "mSet<-PlotNormSummary(mSet, \"norm_4_\", \"png\", 72, width=NA)"
[59] "mSet<-PlotSampleNormSummary(mSet, \"snorm_4_\", \"png\", 72, width=NA)"
[60] "mSet<-Normalization(mSet, \"NULL\", \"LogNorm\", \"ParetoNorm\", ratio=FALSE, ratioNum=20)"
[61] "mSet<-PlotNormSummary(mSet, \"norm_5_\", \"png\", 72, width=NA)"
[62] "mSet<-PlotSampleNormSummary(mSet, \"snorm_5_\", \"png\", 72, width=NA)"
[63] "mSet<-Normalization(mSet, \"NULL\", \"LogNorm\", \"MeanCenter\", ratio=FALSE, ratioNum=20)"
[64] "mSet<-PlotNormSummary(mSet, \"norm_6_\", \"png\", 72, width=NA)"
[65] "mSet<-PlotSampleNormSummary(mSet, \"snorm_6_\", \"png\", 72, width=NA)"
[66] "mSet<-Normalization(mSet, \"NULL\", \"NULL\", \"MeanCenter\", ratio=FALSE, ratioNum=20)"
[67] "mSet<-PlotNormSummary(mSet, \"norm_7_\", \"png\", 72, width=NA)"
[68] "mSet<-PlotSampleNormSummary(mSet, \"snorm_7_\", \"png\", 72, width=NA)"
[69] "mSet<-Normalization(mSet, \"NULL\", \"LogNorm\", \"MeanCenter\", ratio=FALSE, ratioNum=20)"
[70] "mSet<-PlotNormSummary(mSet, \"norm_8_\", \"png\", 72, width=NA)"
[71] "mSet<-PlotSampleNormSummary(mSet, \"snorm_8_\", \"png\", 72, width=NA)"
[72] "mSet<-SetMetabolomeFilter(mSet, F);"
[73] "mSet<-SetCurrentMsetLib(mSet, \"kegg_pathway\", 2);"
[74] "mSet<-CalculateGlobalTestScore(mSet)"
[75] "mSet<-PlotQEA.Overview(mSet, \"qea_0_\", \"net\", \"png\", 72, width=NA)"
[76] "mSet<-PlotEnrichDotPlot(mSet, \"qea\", \"qea_dot_0_\", \"png\", 72, width=NA)"
[77] "mSet<-PrepareSifDownloads(mSet)"
[78] "mSet<-PlotQEA.MetSet(mSet, \"Pentose and glucuronate interconversions\", \"png\", 72, width=NA
[79] "mSet<-PlotQEA.MetSet(mSet, \"Fructose and mannose metabolism\", \"png\", 72, width=NA)"
[80] "mSet<-PlotQEA.MetSet(mSet, \"Fatty acid biosynthesis\", \"png\", 72, width=NA)"
[81] "mSet<-PlotQEA.MetSet(mSet, \"Fatty acid biosynthesis\", \"png\", 72, width=NA)"
[82] "mSet<-PlotQEA.MetSet(mSet, \"Fatty acid biosynthesis\", \"png\", 72, width=NA)"
[83] "mSet<-PlotQEA.MetSet(mSet, \"Fatty acid biosynthesis\", \"png\", 72, width=NA)"
[84] "mSet<-PlotQEA.MetSet(mSet, \"Fatty acid biosynthesis\", \"png\", 72, width=NA)"
[85] "mSet<-SaveTransformedData(mSet)"
[86] "mSet<-PreparePDFReport(mSet, \"guest3628632976710346520\")\n"
```

The report was generated on Sat Apr 27 09:17:28 2024 with R version 4.3.2 (2023-10-31), OS system: Linux, version: -Ubuntu SMP Tue Mar 5 20:16:58 UTC 2024 .