# Metabolomic Data Analysis with MetaboAnalyst 5.0

Name: guest11431946729701306757

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 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 Mapping

| | Query | Match | HMDB | PubChem | KEGG | SMILES | Comment |
|---|---|---|---|---|---|---|---|
| 1 | Acetoacetic acid | Acetoacetic acid | HMDB0000060 | 96 | C00164 | CC(=O)CC(=O)O | 1 |
| 2 | Beta-Alanine | Beta-Alanine | HMDB0000056 | 239 | C00099 | C(CN)C(=O)O | 1 |
| 3 | Creatine | Creatine | HMDB0000064 | 586 | C00300 | CN(CC(=O)O)C(=N)N | 1 |
| 4 | Dimethylglycine | Dimethylglycine | HMDB0000092 | 673 | C01026 | CN(C)CC(=O)O | 1 |
| 5 | Fumaric acid | Fumaric acid | HMDB0000134 | 444972 | C00122 | C(=C/C(=O)O)\C(=O)O | 1 |
| 6 | Glycine | Glycine | HMDB0000123 | 750 | C00037 | C(C(=O)O)N | 1 |
| 7 | Homocysteine | Homocysteine | HMDB0000742 | 778 | C00155 | C(CS)C(C(=O)O)N | 1 |
| 8 | L-Cysteine | L-Cysteine | HMDB0000574 | 5862 | C00097 | C([C@@H](C(=O)O)N)S | 1 |
| 9 | L-Isolucine | NA | NA | NA | NA | NA | 0 |
| 10 | L-Phenylalanine | L-Phenylalanine | HMDB0000159 | 6140 | C00079 | C1=CC=C(C=C1)C[C@@H](C(=O)O)N | 1 |
| 11 | L-Serine | L-Serine | HMDB0000187 | 5951 | C00065 | C([C@@H](C(=O)O)N)O | 1 |
| 12 | L-Threonine | L-Threonine | HMDB0000167 | 6288 | C00188 | C[C@H]([C@@H](C(=O)O)N)O | 1 |
| 13 | L-Tyrosine | L-Tyrosine | HMDB0000158 | 6057 | C00082 | C1=CC(=CC=C1C[C@@H](C(=O)O)N)O | 1 |
| 14 | L-Valine | L-Valine | HMDB0000883 | 6287 | C00183 | CC(C)[C@@H](C(=O)O)N | 1 |
| 15 | Phenylpyruvic acid | Phenylpyruvic acid | HMDB0000205 | 997 | C00166 | C1=CC=C(C=C1)CC(=O)C(=O)O | 1 |
| 16 | Propionic acid | Propionic acid | HMDB0000237 | 1032 | C00163 | CCC(=O)O | 1 |
| 17 | Pyruvic acid | Pyruvic acid | HMDB0000243 | 1060 | C00022 | CC(=O)C(=O)O | 1 |
| 18 | Sarcosine | Sarcosine | HMDB0000271 | 1088 | C00213 | CNCC(=O)O | 1 |

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.

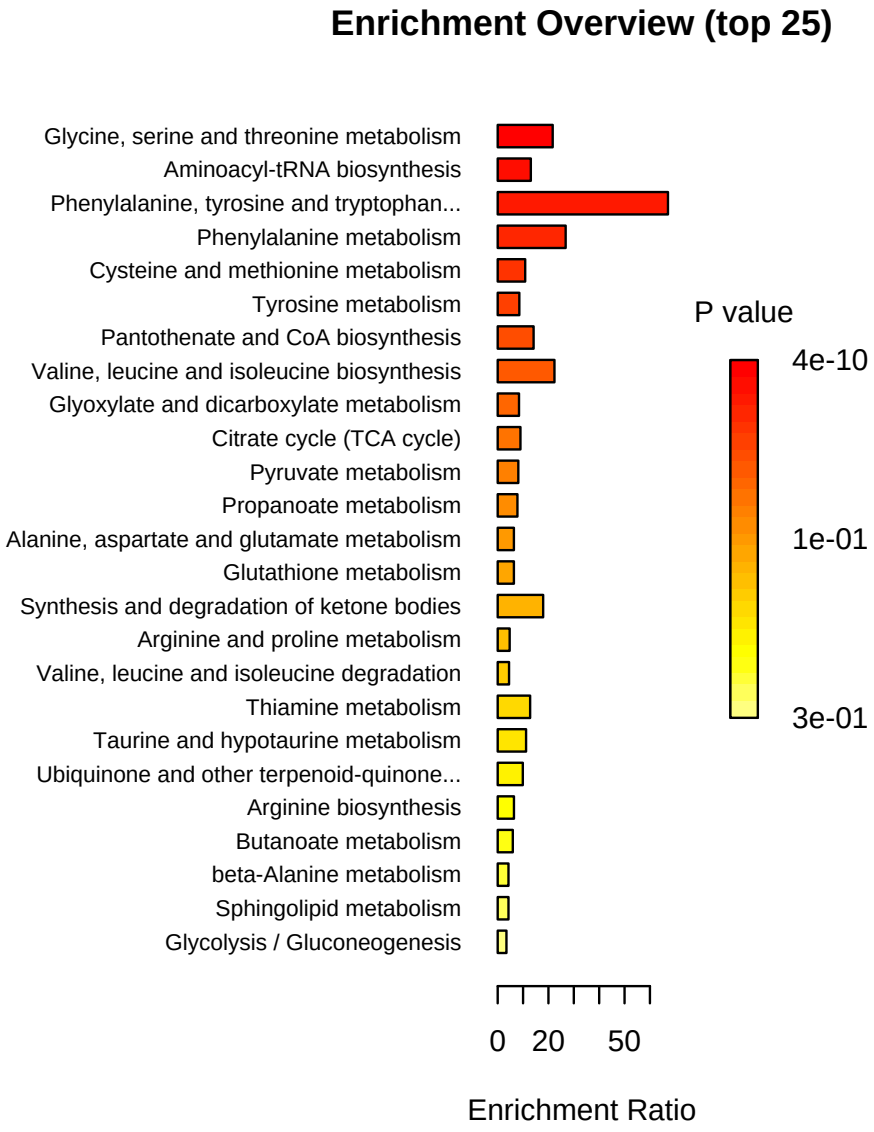


Figure 1: Summary Plot for Over Representation Analysis (ORA)

Table 2: Result from Over Representation Analysis

| | total | expected | hits | Raw p | Holm p | FDR |
|---|---|---|---|---|---|---|
| Glycine, serine and threonine metabolism | 33 | 0.37 | 8 | 4.21E-10 | 3.54E-08 | 3.54E-08 |
| Aminoacyl-tRNA biosynthesis | 48 | 0.54 | 7 | 3.04E-07 | 2.53E-05 | 1.28E-05 |
| Phenylalanine, tyrosine and tryptophan biosynthesis | 4 | 0.04 | 3 | 4.61E-06 | 3.78E-04 | 1.29E-04 |
| Phenylalanine metabolism | 10 | 0.11 | 3 | 1.33E-04 | 1.07E-02 | 2.78E-03 |
| Cysteine and methionine metabolism | 33 | 0.37 | 4 | 3.58E-04 | 2.86E-02 | 6.01E-03 |
| Tyrosine metabolism | 42 | 0.47 | 4 | 9.20E-04 | 7.27E-02 | 1.21E-02 |
| Pantothenate and CoA biosynthesis | 19 | 0.21 | 3 | 1.01E-03 | 7.84E-02 | 1.21E-02 |
| Valine, leucine and isoleucine biosynthesis | 8 | 0.09 | 2 | 3.16E-03 | 2.43E-01 | 3.32E-02 |
| Glyoxylate and dicarboxylate metabolism | 32 | 0.36 | 3 | 4.70E-03 | 3.57E-01 | 4.39E-02 |
| Citrate cycle (TCA cycle) | 20 | 0.22 | 2 | 1.98E-02 | 1.00E+00 | 1.67E-01 |
| Pyruvate metabolism | 22 | 0.25 | 2 | 2.38E-02 | 1.00E+00 | 1.81E-01 |
| Propanoate metabolism | 23 | 0.26 | 2 | 2.59E-02 | 1.00E+00 | 1.81E-01 |
| Alanine, aspartate and glutamate metabolism | 28 | 0.31 | 2 | 3.74E-02 | 1.00E+00 | 2.25E-01 |
| Glutathione metabolism | 28 | 0.31 | 2 | 3.74E-02 | 1.00E+00 | 2.25E-01 |
| Synthesis and degradation of ketone bodies | 5 | 0.06 | 1 | 5.47E-02 | 1.00E+00 | 3.06E-01 |
| Arginine and proline metabolism | 38 | 0.42 | 2 | 6.52E-02 | 1.00E+00 | 3.42E-01 |
| Valine, leucine and isoleucine degradation | 40 | 0.45 | 2 | 7.14E-02 | 1.00E+00 | 3.53E-01 |
| Thiamine metabolism | 7 | 0.08 | 1 | 7.58E-02 | 1.00E+00 | 3.54E-01 |
| Taurine and hypotaurine metabolism | 8 | 0.09 | 1 | 8.61E-02 | 1.00E+00 | 3.81E-01 |
| Ubiquinone and other terpenoid-quinone biosynthesis | 9 | 0.10 | 1 | 9.64E-02 | 1.00E+00 | 4.05E-01 |
| Arginine biosynthesis | 14 | 0.16 | 1 | 1.46E-01 | 1.00E+00 | 5.84E-01 |
| Butanoate metabolism | 15 | 0.17 | 1 | 1.56E-01 | 1.00E+00 | 5.95E-01 |
| beta-Alanine metabolism | 21 | 0.23 | 1 | 2.11E-01 | 1.00E+00 | 7.40E-01 |
| Sphingolipid metabolism | 21 | 0.23 | 1 | 2.11E-01 | 1.00E+00 | 7.40E-01 |
| Glycolysis / Gluconeogenesis | 26 | 0.29 | 1 | 2.55E-01 | 1.00E+00 | 8.57E-01 |
| Porphyrin and chlorophyll metabolism | 30 | 0.34 | 1 | 2.88E-01 | 1.00E+00 | 9.32E-01 |
| Pyrimidine metabolism | 39 | 0.44 | 1 | 3.58E-01 | 1.00E+00 | 1.00E+00 |
| Primary bile acid biosynthesis | 46 | 0.51 | 1 | 4.08E-01 | 1.00E+00 | 1.00E+00 |

# 7 Appendix: R Command History

```
 [1] "mSet<-InitDataObjects(\"conc\", \"msetora\", FALSE)"
 [2] "cmpd.vec<-c(\"Acetoacetic acid\",\"Beta-Alanine\",\"Creatine\",\"Dimethylglycine\",\"Fumaric a
 [3] "mSet<-Setup.MapData(mSet, cmpd.vec);"
 [4] "mSet<-CrossReferencing(mSet, \"name\");"
 [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<-PreparePDFReport(mSet, \"guest11431946729701306757\")\n"
```

---

The report was generated on Mon Apr 19 23:26:18 2021 with R version 4.0.2 (2020-06-22).