Metabolomic Data Analysis with MetaboAnalyst 5.0

Name: guest15047850574617834199

May 21, 2021

1 Background

The Pathway Analysis module combines results from powerful pathway enrichment analysis with pathway topology analysis to help researchers identify the most relevant pathways involved in the conditions under study.

There are many commercial pathway analysis software tools such as Pathway Studio, MetaCore, or Ingenuity Pathway Analysis (IPA), etc. Compared to these commercial tools, the pathway analysis module was specifically developed for metabolomics studies. It uses high-quality KEGG metabolic pathways as the backend knowledgebase. This module integrates many well-established (i.e. univariate analysis, over-representation analysis) methods, as well as novel algorithms and concepts (i.e. Global Test, GlobalAncova, network topology analysis) into pathway analysis. Another feature is a Google-Map style interactive visualization system to deliver the analysis results in an intuitive manner.

2 Data Input

The Pathway Analysis module accepts either a list of compound labels (common names, HMDB IDs or KEGG IDs) with one compound per row, or a compound concentration table with samples in rows and compounds in columns. The second column must be phenotype labels (binary, multi-group, or continuous). The table is uploaded as comma separated values (.csv).

3 Compound Name Matching

The first step is to standardize the compound labels used in user uploaded data. This is a necessary step since these compounds will be subsequently compared with compounds contained in the pathway library. There are three outcomes from the step - exact match, approximate match (for common names only), and no match. Users should click the textbfView button from the approximate matched results to manually select the correct one. Compounds without match will be excluded from the subsequently pathway analysis.

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 Comp

	Query	Match	HMDB	PubChem	KEGG
1	HMDB0038791	NA	NA	NA	NA
2	HMDB0033584	Diethyl tartrate	HMDB0033584	62333	
3	HMDB0059767	NA	NA	NA	NA
4	HMDB0157322	NA	NA	NA	NA
5	HMDB0157321	NA	NA	NA	NA
6	HMDB0157320	NA	NA	NA	NA
7	HMDB0157323	NA	NA	NA	NA
8	HMDB0157324	NA	NA	NA	NA
9	HMDB0029416	L-Targinine	HMDB0029416	132862	C0388

10	HMDB0000670	Homo-L-arginine	HMDB0000670	9085	C0192
11	HMDB0015454	Rasagiline	HMDB0015454	3052776	
12	HMDB0032926	Polycartine B	HMDB0032926	77623	
13	HMDB0005973	Dimethyltryptamine	HMDB0005973	6089	C0830
14	HMDB0033961	2-(3-Methylbutyl)-1H-pyrrolo[2,3-b]pyridine	HMDB0033961	55272645	
15	HMDB0014861	Methoxamine	HMDB0014861	6082	C0751
16	HMDB0167331	NA	NA	NA	NA
17	HMDB0014954	Orciprenaline	HMDB0014954	4086	C0714
18	HMDB0015197	Isoproterenol	HMDB0015197	3779	C0705
19	HMDB0011175	Leucylproline	HMDB0011175	44369311	
20	HMDB0028937	NA	NA	NA	NA
21	HMDB0011174	Isoleucylproline	HMDB0011174	342734	
22	HMDB0172369	NA	NA	NA	NA
23	HMDB0028915	NA	NA	NA	NA
24	HMDB0240365	NA	NA	NA	NA
25	HMDB0013816	2,4-Di-tert-butylphenol	HMDB0013816	14339290	
26	HMDB0013825	4-(1,1,3,3-Tetramethylbutyl)-phenol	HMDB0013825	0	C1420
27	HMDB0032528	alpha-(p-(1,1,3,3-Tetramethylbutyl)phenyl)-omega-hydroxypoly(oxyethylene)	HMDB0032528	8814	C1420
28	HMDB0032397	Methyl-delta-ionone	HMDB0032397	5463913	
29	HMDB0061831	NA	NA	NA	NA
30	HMDB0031738	3-Methyl-4-(2,6,6-trimethyl-2-cyclohexen-1-yl)-3-buten-2-one	HMDB0031738	5372174	
31	HMDB0038130	NA	NA	NA	NA
32	HMDB0031737	delta-Methylionone	HMDB0031737	5372195	
33	HMDB0035245	1-(2,6,6-Trimethyl-2-cyclohexen-1-yl)-1-penten-3-one	HMDB0035245	5371084	
34	HMDB0036024	10-Isopropyl-2,7-dimethyl-1-oxaspiro[4.5]deca-3,6-diene	HMDB0036024	20268659	
35	HMDB0036023	Etaspirene	HMDB0036023	22219099	
36	HMDB0133179	NA	NA	NA	NA
37	HMDB0035631	alpha-Irone	HMDB0035631	5371002	C0969
38	HMDB0060537	NA	NA	NA	NA
39	HMDB0060826	NA	NA	NA	NA
40	HMDB0004185	5-Hydroxyindoleacetylglycine	HMDB0004185	440806	C0583
41	HMDB0003066	Chalcone	HMDB0003066	637760	C1558
42	HMDB0135598	NA	NA	NA	NA
43	HMDB0060323	NA	NA	NA	NA
44	HMDB0032357	N-Lactoyl ethanolamine phosphate	HMDB0032357	11550267	
45	HMDB0155289	NA	NA	NA	NA
46	HMDB0060757	NA	NA	NA	NA
47	HMDB0000684	L-Kynurenine	HMDB0000684	161166	C0032
48	HMDB0012948	Formyl-5-hydroxykynurenamine	HMDB0012948	440743	C0564
49	HMDB0014350	Pyrimethamine	HMDB0014350	4993	C0739
50	HMDB0060321	NA	NA	NA	NA
51	HMDB0155347	NA	NA	NA	NA
52	HMDB0155348	NA	NA	NA	NA
53	HMDB0034064	2-Angeloyl-9-(3-methyl-2E-pentenoyl)-2b,9a-dihydroxy-4Z,10(14)-oplopadien-3-one	HMDB0034064	131751519	
54	HMDB0012936	Dynorphin B (10-13)	HMDB0012936	53481556	

4 Pathway Analysis

In this step, users are asked to select a pathway library, as well as specify the algorithms for pathway enrichment analysis and pathway topology analysis.

4.1 Pathway Library

There are 15 pathway libraries currently supported, with a total of 1173 pathways :

- Homo sapiens (human) [80]
- Mus musculus (mouse) [82]
- Rattus norvegicus (rat) [81]
- Bos taurus (cow) [81]
- Danio rerio (zebrafish) [81]
- Drosophila melanogaster (fruit fly) [79]
- Caenorhabditis elegans (nematode) [78]
- Saccharomyces cerevisiae (yeast) [65]
- Oryza sativa japonica (Japanese rice) [83]
- Arabidopsis thaliana (thale cress) [87]
- Escherichia coli K-12 MG1655 [87]
- Bacillus subtilis [80]
- Pseudomonas putida KT2440 [89]
- Staphylococcus aureus N315 (MRSA/VSSA)[73]
- Thermotoga maritima [57]

Your selected pathway library code is **hsa** (KEGG organisms abbreviation).

4.2 Over Representation Analysis

Over-representation analysis tests if a particular group of compounds is represented more than expected by chance within the user uploaded compound list. In the context of pathway analysis, we are testing if compounds involved in a particular pathway are enriched compared to random hits. MetPA offers two of the most commonly used methods for over-representation analysis:

- Fishers'Exact test
- Hypergeometric Test

Please note, MetPA uses one-tailed Fisher's exact test which will give essentially the same result as the result calculated by the hypergeometric test.

The selected over-representation analysis method is **Hypergeometric test**.

4.3 Pathway Topology Analysis

The structure of biological pathways represent our knowledge about the complex relationships among molecules within a cell or a living organism. However, most pathway analysis algorithms fail to take structural information into consideration when estimating which pathways are significantly changed under conditions of study. It is well-known that changes in more important positions of a network will trigger a more severe impact on the pathway than changes occurred in marginal or relatively isolated positions.

The pathway topology analysis uses two well-established node centrality measures to estimate node importance - **degree centrality** and **betweenness centrality**. Degree centrality is defined as the number of links occurred upon a node. For a directed graph there are two types of degree: in-degree for links come from other nodes, and out-degree for links initiated from the current node. Metabolic networks are directed graph. Here we only consider the out-degree for node importance measure. It is assumed that nodes upstream will have regulatory roles for the downstream nodes, not vice versa. The betweenness centrality measures the number of shortest paths going through the node. Since the metabolic network is directed, we use the relative betweenness centrality for a metabolite as the importance measure. The degree centrality measure focuses more on local connectivities, while the betweenness centrality measure focuses more on global network topology. For more detailed discussions on various graph-based methods for analyzing biological networks, please refer to the article by Tero Aittokallio, T. et al. [1]

Please note, for comparison among different pathways, the node importance values calculated from centrality measures are further normalized by the sum of the importance of the pathway. Therefore, the total/maximum importance of each pathway is 1; the importance measure of each metabolite node is actually the percentage w.r.t the total pathway importance, and the pathway impact value is the cumulative percentage from the matched metabolite nodes.

Your selected node importance measure for topological analysis is **relative betweenness centrality**.

5 Pathway Analysis Result

The results from pathway analysis are presented graphically as well as in a detailed table.

A Google-map style interactive visualization system was implemented to facilitate data exploration. The graphical output contains three levels of view: **metabolome view**, **pathway view**, and **compound view**. Only the metabolome view is shown below. Pathway views and compound views are generated dynamically based on your interactions with the visualization system. They are available in your downloaded files.

[1] Tero Aittokallio and Benno Schwikowski. *Graph-based methods for analyzing networks in cell biology*, Briefings in Bioinformatics 2006 7(3):243-255

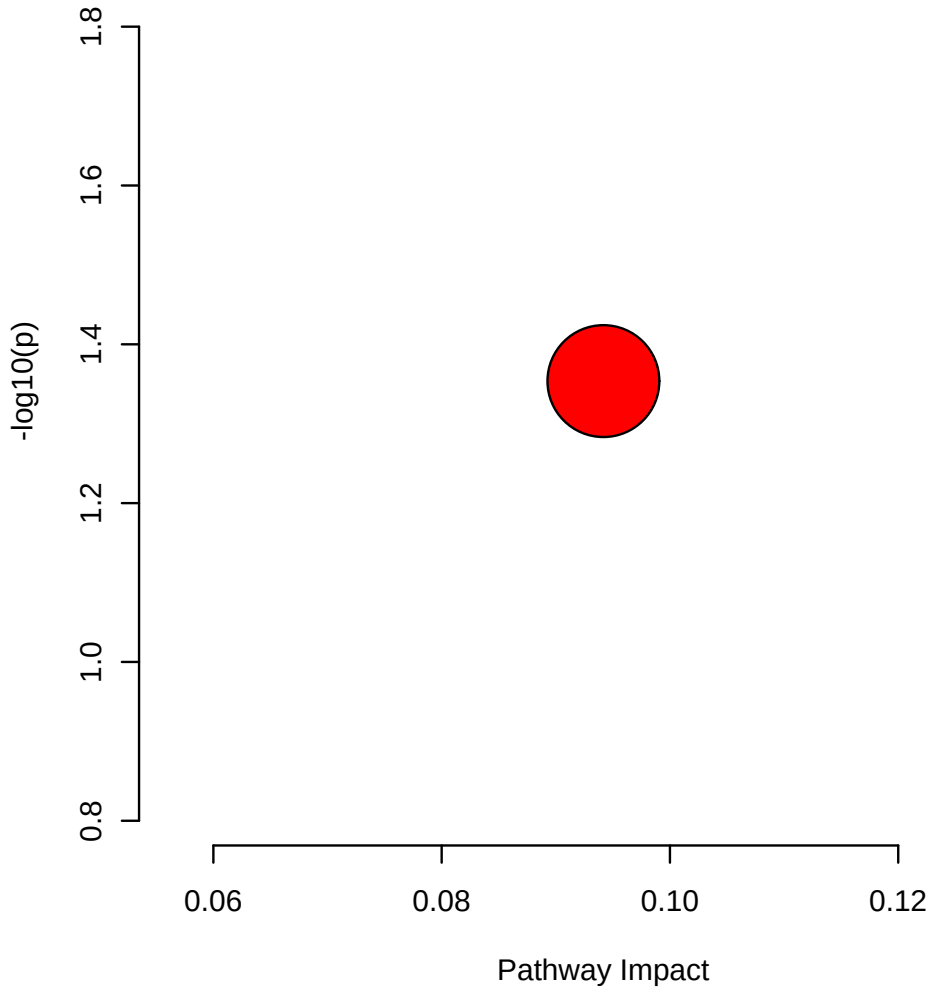

Figure 1: Summary of Pathway Analysis

The table below shows the detailed results from the pathway analysis. Since we are testing many pathways at the same time, the statistical p values from enrichment analysis are further adjusted for multiple testings. In particular, the **Total** is the total number of compounds in the pathway; the **Hits** is the actually matched number from the user uploaded data; the **Raw p** is the original p value calculated from the enrichment analysis; the **Holm p** is the p value adjusted by Holm-Bonferroni method; the **FDR p** is the p value adjusted using False Discovery Rate; the **Impact** is the pathway impact value calculated from pathway topology analysis.

Table 2: Result from Pathway Analysis

	Total	Expected	Hits	Raw p	-log10(p)	Holm adjust	FDR	Impact
Tryptophan metabolism	41	0.34	2	4.43E-02	1.35E+00	1.00E+00	1.00E+00	0.09

6 Appendix: R Command History

```
 [1] "mSet<-InitDataObjects(\"conc\", \"pathora\", FALSE)"
 [2] "cmpd.vec<-c(\"HMDB0038791\",\"HMDB0033584\",\"HMDB0059767\",\"HMDB0157322\",\"HMDB0157321\",\"
 [3] "mSet<-Setup.MapData(mSet, cmpd.vec);"
 [4] "mSet<-CrossReferencing(mSet, \"hmdb\");"
 [5] "mSet<-CreateMappingResultTable(mSet)"
 [6] "mSet<-SetKEGG.PathLib(mSet, \"hsa\", \"current\")"
 [7] "mSet<-SetMetabolomeFilter(mSet, F);"
 [8] "mSet<-CalculateOraScore(mSet, \"rbc\", \"hyperg\")"
 [9] "mSet<-PlotPathSummary(mSet, F, \"path_view_0_\", \"png\", 72, width=NA)"
[10] "mSet<-SaveTransformedData(mSet)"
[11] "mSet<-PreparePDFReport(mSet, \"guest15047850574617834199\")\n"
```

The report was generated on Fri May 21 16:17:52 2021 with R version 4.0.2 (2020-06-22).