Metabolomic Data Analysis with MetaboAnalyst 5.0

Name: guest2575446555666263837

November 24, 2022

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 Compound Name Mapping

	Query	Match	HMDB	PubChem	KEGG	SMILES
1	HMDB0000244	Riboflavin	HMDB0000244	493570	C00255	CC1=CC2=C(C=C1C)N(C3=NC(=O)NC(
2	HMDB0000446	N-Alpha-acetyllysine	HMDB0000446	192590	C12989	CC(=O)NC(CCCCN)C(=O)O
3	HMDB0000251	Taurine	HMDB0000251	1123	C00245	C(CS(=O)(=O)O)N
4	HMDB0002172	N1,N12-Diacetylspermine	HMDB0002172	132680	C03413	CC(=O)NCCCNCCCCNCCCNC(=O)C
5	HMDB0000893	Suberic acid	HMDB0000893	10457	C08278	C(CCCC(=O)O)CCC(=O)O
6	HMDB0000626	Deoxycholic acid	HMDB0000626	222528	C04483	C[C@H](CCC(=O)O)[C@H]1CC[C@@H]2[C
7	HMDB0000157	Hypoxanthine	HMDB0000157	790	C00262	C1=NC2=C(N1)C(=O)N=CN2
8	HMDB0000752	Methylglutaric acid	HMDB0000752	12284		CC(CC(=O)O)CC(=O)O
9	HMDB0000448	Adipic acid	HMDB0000448	196	C06104	C(CCC(=O)O)CC(=O)O

10	HMDB0000929	L-Tryptophan	HMDB0000929	6305	C00078	C1=CC=C2C(=C1)C(=CN2)C[C@@H](C(=
11	HMDB0000500	4-Hydroxybenzoic acid	HMDB0000500	135	C00156	C1=CC(=CC=C1C(=O)O)O
12	HMDB0000422	2-Methylglutaric acid	HMDB0000422	12046		CC(CCC(=O)O)C(=O)O
13	HMDB0003331	1-Methyladenosine	HMDB0003331	27476	C02494	CN1C=NC2=C(C1=N)N=CN2[C@H]3[C@
14	HMDB0000159	L-Phenylalanine	HMDB0000159	6140	C00079	C1=CC=C(C=C1)C[C@@H](C(=O)O)N
15	HMDB0000254	Succinic acid	HMDB0000254	1110	C00042	C(CC(=O)O)C(=O)O
16	HMDB0001138	N-Acetylglutamic acid	HMDB0001138	185	C00624	CC(=O)NC(CCC(=O)O)C(=O)O
17	HMDB0000661	Glutaric acid	HMDB0000661	743	C00489	C(CC(=O)O)CC(=O)O
18	HMDB0000631	Deoxycholic acid glycine conjugate	HMDB0000631	3035026	C05464	C[C@H](CCC(=O)NCC(=O)O)[C@H]1CC
19	HMDB0000784	Azelaic acid	HMDB0000784	2266	C08261	C(CCCC(=O)O)CCCC(=O)O
20	HMDB0004620	N-a-Acetyl-L-arginine	HMDB0004620	67427		CC(=O)N[C@@H](CCCN=C(N)N)C(=O)O
21	HMDB0000678	Isovalerylglycine	HMDB0000678	546304		CC(C)CC(=O)NCC(=O)O
22	HMDB0001032	Dehydroepiandrosterone sulfate	HMDB0001032	12594	C04555	C[C@]12CC[C@H]3[C@H]([C@@H]1CCC2=
23	HMDB0011103	1,7-Dimethyluric acid	HMDB0011103	91611	C16356	CN1C2=C(NC1=O)NC(=O)N(C2=O)C
24	HMDB0000708	Glycoursodeoxycholic acid	HMDB0000708	12310288		C[C@H](CCC(=O)NCC(=O)O)[C@H]1CC[
25	HMDB0002123	1,3,7-Trimethyluric acid	HMDB0002123	79437	C16361	CN1C2=C(NC1=O)N(C(=O)N(C2=O)C)C
26	HMDB0003334	Symmetric dimethylarginine	HMDB0003334	169148		CNC(=NC)NCCC[C@@H](C(=O)O)N
27	HMDB0013677	3,5-Dihydroxybenzoic acid	HMDB0013677	7424	C00180	C1=C(C=C(C=C1O)O)C(=O)O
28	HMDB0028942	Leucyl-Valine	HMDB0028942	6993116		CC(C)CC(N)C(=O)NC(C(C)C)C(O)=O
29	HMDB0001844	Methylsuccinic acid	HMDB0001844	10349	C08645	CC(CC(=O)O)C(=O)O
30	HMDB0244966	NA	NA	NA	NA	NA
31	HMDB0255727	NA	NA	NA	NA	NA
32	HMDB0000730	Isobutyrylglycine	HMDB0000730	10855600		CC(C)C(=O)NCC(=O)O
33	HMDB0005807	Gallic acid	HMDB0005807	370	C01424	C1=C(C=C(C(=C1O)O)O)C(=O)O
34	HMDB0000687	L-Leucine	HMDB0000687	6106	C00123	CC(C)C[C@@H](C(=O)O)N
35	HMDB0000881	Xanthurenic acid	HMDB0000881	5699	C02470	C1=CC2=C(C(=C1)O)NC(=CC2=O)C(=C
36	HMDB0000956	Tartaric acid	HMDB0000956	444305	C00898	O[C@H]([C@@H](O)C(O)=O)C(O)=O
37	HMDB0000729	Alpha-Hydroxyisobutyric acid	HMDB0000729	11671		CC(C)(C(=O)O)O
38	HMDB0062640	3-hydroxy-2-isobutyrate	HMDB0062640	87	C01188	CC(CO)C(O)=O
39	HMDB0001991	7-Methylxanthine	HMDB0001991	68374	C16353	CN1C=NC2=C1C(=O)NC(=O)N2
40	HMDB0061384	NA	NA	NA	NA	NA
41	HMDB0013713	N-acetyltryptophan	HMDB0013713	700653		[H][C@@](CC1=CNC2=CC=CC=C12)(N=
42	HMDB0000152	Gentisic acid	HMDB0000152	3469	C00628	C1=CC(=C(C=C1O)C(=O)O)O
43	HMDB0000301	Urocanic acid	HMDB0000301	736715	C00785	C1=C(NC=N1)/C=C/C(=O)O
44	HMDB0001847	Caffeine	HMDB0001847	2519	C07481	CN1C=NC2=C1C(=O)N(C(=O)N2C)C
45	HMDB0000822	p-Hydroxymandelic acid	HMDB0000822	7721	C11527	C1=CC(=CC=C1C(C(=O)O)O)O
46	HMDB0001406	Niacinamide	HMDB0001406	936	C00153	C1=CC(=CN=C1)C(=O)N
47	HMDB0012275	Phenylethylamine	HMDB0012275	1001	C05332	C1=CC=C(C=C1)CCN
48	HMDB0000226	Orotic acid	HMDB0000226	967	C00295	C1=C(NC(=O)NC1=O)C(=O)O
49	HMDB0006029	N-Acetylglutamine	HMDB0006029	25561		CC(=O)NC(CCC(=O)N)C(=O)O
50	HMDB0001325	N6,N6,N6-Trimethyl-L-lysine	HMDB0001325	440120	C03793	C[N+](C)(C)CCCC[C@@H](C(=O)[O-])N
51	HMDB0000235	Thiamine	HMDB0000235	1130	C00378	CC1=C(SC=[N+]1CC2=CN=C(N=C2N)C
52	HMDB0013676	2,6-Dihydroxybenzoic acid	HMDB0013676	9338	C21298	C1=CC(=C(C(=C1)O)C(=O)O)O
53	HMDB0000721	Glycylproline	HMDB0000721	79101		C1CC(N(C1)C(=O)CN)C(=O)O
54	HMDB0000641	L-Glutamine	HMDB0000641	5961	C00064	C(CC(=O)N)[C@@H](C(=O)O)N
55	HMDB0000158	L-Tyrosine	HMDB0000158	6057	C00082	C1=CC(=CC=C1C[C@@H](C(=O)O)N)O
56	HMDB0000355	3-Hydroxymethylglutaric acid	HMDB0000355	1662	C03761	CC(CC(=O)O)(CC(=O)O)O

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

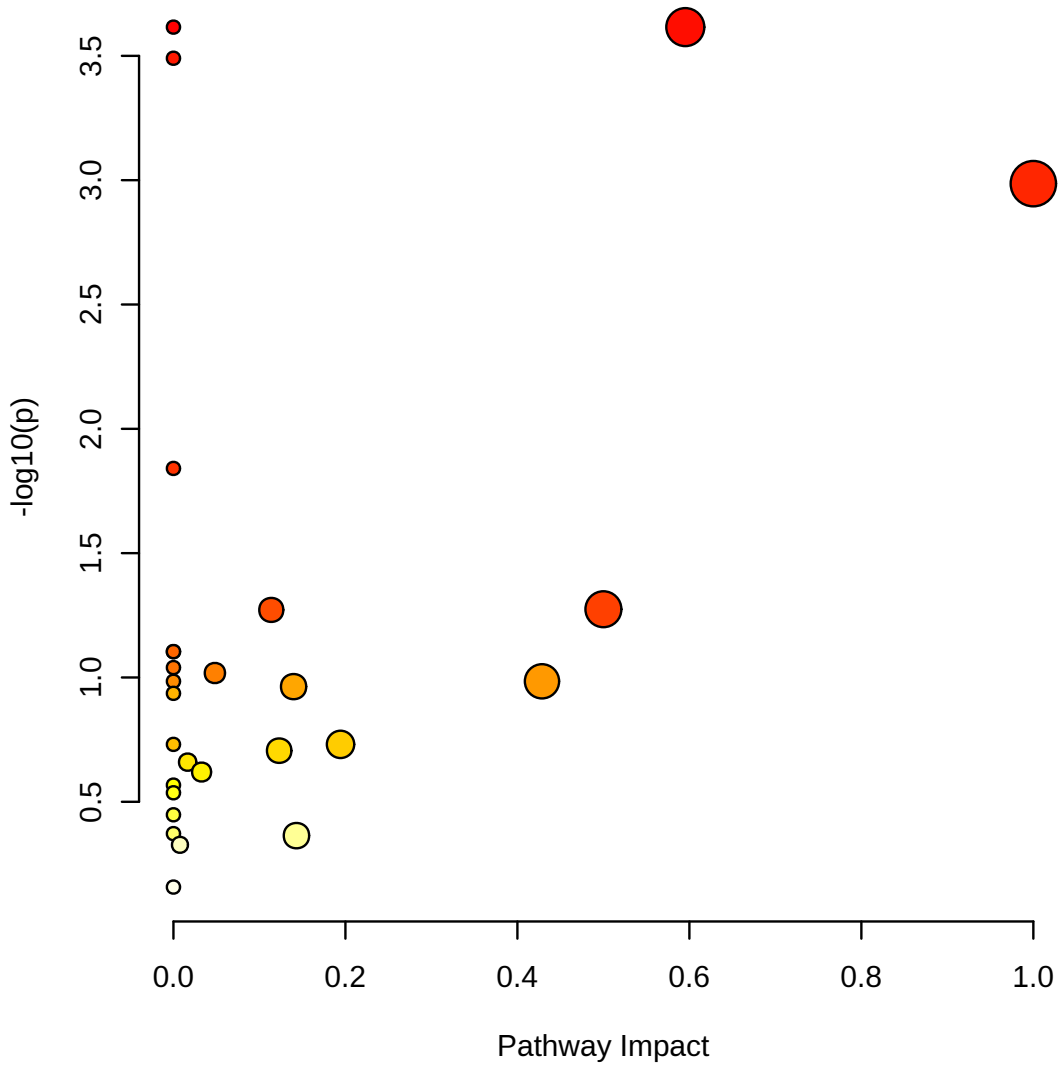

Figure 1: Summary of Pathway Analysis

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
Caffeine metabolism	10	0.14	3	2.42E-04	3.62E+00	2.04E-02	9.05E-03	0.00
Phenylalanine metabolism	10	0.14	3	2.42E-04	3.62E+00	2.04E-02	9.05E-03	0.60
Aminoacyl-tRNA biosynthesis	48	0.65	5	3.23E-04	3.49E+00	2.65E-02	9.05E-03	0.00
Phenylalanine, tyrosine and tryptophan biosynthesis	4	0.05	2	1.03E-03	2.99E+00	8.36E-02	2.17E-02	1.00
Arginine biosynthesis	14	0.19	2	1.44E-02	1.84E+00	1.00E+00	2.42E-01	0.00
Riboflavin metabolism	4	0.05	1	5.32E-02	1.27E+00	1.00E+00	6.42E-01	0.50
Alanine, aspartate and glutamate metabolism	28	0.38	2	5.35E-02	1.27E+00	1.00E+00	6.42E-01	0.11
D-Glutamine and D-glutamate metabolism	6	0.08	1	7.87E-02	1.10E+00	1.00E+00	6.49E-01	0.00
Nitrogen metabolism	6	0.08	1	7.87E-02	1.10E+00	1.00E+00	6.49E-01	0.00
Thiamine metabolism	7	0.09	1	9.12E-02	1.04E+00	1.00E+00	6.49E-01	0.00
Pyrimidine metabolism	39	0.53	2	9.60E-02	1.02E+00	1.00E+00	6.49E-01	0.05
Valine, leucine and isoleucine biosynthesis	8	0.11	1	1.04E-01	9.85E-01	1.00E+00	6.49E-01	0.00
Taurine and hypotaurine metabolism	8	0.11	1	1.04E-01	9.85E-01	1.00E+00	6.49E-01	0.43
Tyrosine metabolism	42	0.57	2	1.09E-01	9.63E-01	1.00E+00	6.49E-01	0.14
Ubiquinone and other terpenoid-quinone biosynthesis	9	0.12	1	1.16E-01	9.36E-01	1.00E+00	6.49E-01	0.00
Butanoate metabolism	15	0.20	1	1.86E-01	7.31E-01	1.00E+00	9.18E-01	0.00
Nicotinate and nicotinamide metabolism	15	0.20	1	1.86E-01	7.31E-01	1.00E+00	9.18E-01	0.19
Histidine metabolism	16	0.22	1	1.97E-01	7.06E-01	1.00E+00	9.19E-01	0.12
Purine metabolism	65	0.88	2	2.19E-01	6.60E-01	1.00E+00	9.68E-01	0.02
Citrate cycle (TCA cycle)	20	0.27	1	2.40E-01	6.20E-01	1.00E+00	1.00E+00	0.03
Propanoate metabolism	23	0.31	1	2.71E-01	5.67E-01	1.00E+00	1.00E+00	0.00
Lysine degradation	25	0.34	1	2.91E-01	5.36E-01	1.00E+00	1.00E+00	0.00
Glyoxylate and dicarboxylate metabolism	32	0.43	1	3.57E-01	4.48E-01	1.00E+00	1.00E+00	0.00
Valine, leucine and isoleucine degradation	40	0.54	1	4.25E-01	3.72E-01	1.00E+00	1.00E+00	0.00
Tryptophan metabolism	41	0.56	1	4.33E-01	3.64E-01	1.00E+00	1.00E+00	0.14
Primary bile acid biosynthesis	46	0.62	1	4.71E-01	3.27E-01	1.00E+00	1.00E+00	0.01
Steroid hormone biosynthesis	85	1.15	1	6.96E-01	1.57E-01	1.00E+00	1.00E+00	0.00

6 Appendix: R Command History

```
 [1] "mSet<-InitDataObjects(\"conc\", \"pathora\", FALSE)"
 [2] "cmpd.vec<-c(\"HMDB0000244\",\"HMDB0000446\",\"HMDB0000251\",\"HMDB0002172\",\"HMDB0000893\",\"
 [3] "mSet<-Setup.MapData(mSet, cmpd.vec);"
 [4] "mSet<-CrossReferencing(mSet, \"hmdb\");"
 [5] "mSet<-CreateMappingResultTable(mSet)"
 [6] "mSet<-PerformDetailMatch(mSet, \"HMDB0244966\");"
 [7] "mSet<-GetCandidateList(mSet);"
 [8] "mSet<-SetKEGG.PathLib(mSet, \"hsa\", \"current\")"
 [9] "mSet<-SetMetabolomeFilter(mSet, F);"
[10] "mSet<-CalculateOraScore(mSet, \"rbc\", \"hyperg\")"
[11] "mSet<-PlotPathSummary(mSet, F, \"path_view_0_\", \"png\", 72, width=NA, NA, NA )"
[12] "mSet<-PlotKEGGPath(mSet, \"Phenylalanine, tyrosine and tryptophan biosynthesis\",576, 480, \"p
[13] "mSet<-RerenderMetPAGraph(mSet, \"zoom1669282612181.png\",576.0, 480.0, 100.0)"
[14] "mSet<-PlotKEGGPath(mSet, \"Phenylalanine, tyrosine and tryptophan biosynthesis\",576, 480, \"p
[15] "mSet<-PlotKEGGPath(mSet, \"Phenylalanine, tyrosine and tryptophan biosynthesis\",576, 480, \"p
[16] "mSet<-RerenderMetPAGraph(mSet, \"zoom1669282640153.png\",576.0, 480.0, 100.0)"
[17] "mSet<-RerenderMetPAGraph(mSet, \"zoom1669282640816.png\",576.0, 480.0, 100.0)"
[18] "mSet<-RerenderMetPAGraph(mSet, \"zoom1669282641999.png\",576.0, 480.0, 100.0)"
[19] "mSet<-RerenderMetPAGraph(mSet, \"zoom1669282642327.png\",576.0, 480.0, 100.0)"
[20] "mSet<-RerenderMetPAGraph(mSet, \"zoom1669282642905.png\",576.0, 480.0, 100.0)"
[21] "mSet<-RerenderMetPAGraph(mSet, \"zoom1669282644076.png\",576.0, 480.0, 100.0)"
[22] "mSet<-RerenderMetPAGraph(mSet, \"zoom1669282644737.png\",576.0, 480.0, 100.0)"
[23] "mSet<-PlotKEGGPath(mSet, \"Phenylalanine metabolism\",576, 480, \"png\", NULL)"
[24] "mSet<-PlotKEGGPath(mSet, \"Phenylalanine metabolism\",576, 480, \"png\", NULL)"
[25] "mSet<-PlotKEGGPath(mSet, \"Phenylalanine metabolism\",576, 480, \"png\", NULL)"
[26] "mSet<-PlotKEGGPath(mSet, \"Riboflavin metabolism\",576, 480, \"png\", NULL)"
[27] "mSet<-SaveTransformedData(mSet)"
[28] "mSet<-PlotKEGGPath(mSet, \"Phenylalanine, tyrosine and tryptophan biosynthesis\",576, 480, \"p
[29] "mSet<-RerenderMetPAGraph(mSet, \"zoom1669282839481.png\",576.0, 480.0, 100.0)"
[30] "mSet<-SaveTransformedData(mSet)"
[31] "mSet<-PreparePDFReport(mSet, \"guest2575446555666263837\")\n"
```

The report was generated on Thu Nov 24 04:41:02 2022 with R version 4.2.2 (2022-10-31), OS system: Linux, version: -Ubuntu SMP Thu Oct 13 08:03:55 UTC 2022 .