metabox: a toolbox for metabolomic data analysis, interpretation and integrative exploration

Kwanjeera Wanichthanarak, Sili Fan, Dmitry Grapov, Dinesh Barupal and Oliver Fiehn

9 January 2017

TUTORIALS

A collection of tutorials for using metabox is provided here to illustrate important features of metabox. These tutorials are corresponding to Results and Discussion section of the main manuscript.

0. Example data sets

The example data sets are based on the published studies about lung adenocarcinoma (transcriptomic data (1) and metabolomic data (2)).

The metabolomic data set containing 39 malignant and adjacent non-malignant lung tissue samples was measured by gas chromatography time-of-flight mass spectrometry (GC-TOF-MS) (2) and pre-processed by the BinBase database (3). 462 compounds were measured, 171 of which have associated PubChem CID. This metabolomic data set was prepared in the input format used in metabox, which is in *metabolomics.xlsx*.

Metabox was used for the statistical analysis comparing tumor and control tissues. For users' convenience, we filtered the list of significantly different compounds comparing tumor and control tissues from the analysis result and provided in *pubchem_stats.txt*.

The transcriptomic data, GEO accession: GSE32863 (1), contain gene expression profiles of 58 lung adenocarcinoma and adjacent non-malignant lung tissues using Illumina HumanWG-6 v3.0 expression BeadChip platform. Differential gene expression analysis comparing lung adenocarcinoma and adjacent non-malignant lung tissues was performed with GEO2R (4). P-values were adjusted with Benjamini and Hochberg false discovery rate at 5% (pFDR < 0.05). Differential gene expression analysis comparing tumor and control tissues reported 171 out of 21,204 genes in which pFDR < 0.05 and |log2FC| > 2. For users' convenience, we filtered the list of significantly different genes comparing tumor and control tissues from the analysis result and provided in *ensembl_stats.txt*.

Table 1. List of example data sets

File	Description	Input of section
metabolomics.xlsx	Metabolic profiles of 39 lung malignant and adjacent non-malignant tissues pre-processed by the BinBase database.	Section 2
pubchem_stats.txt	Multi-column table of significant compounds and associated statistical values computed by metabox.	Section 3
ensembl_stats.txt	Multi-column table of filtered genes and associated statistical values computed by GEO2R.	Section 3

1. Load required libraries and run the metabox using R-terminal

```
## Load library
> library(metabox)
> library(opencpu)

## Run metabox on a web browser
> opencpu$browse('library/metabox/www')
```

2. Using metabox for in-depth analysis of metabolomic data

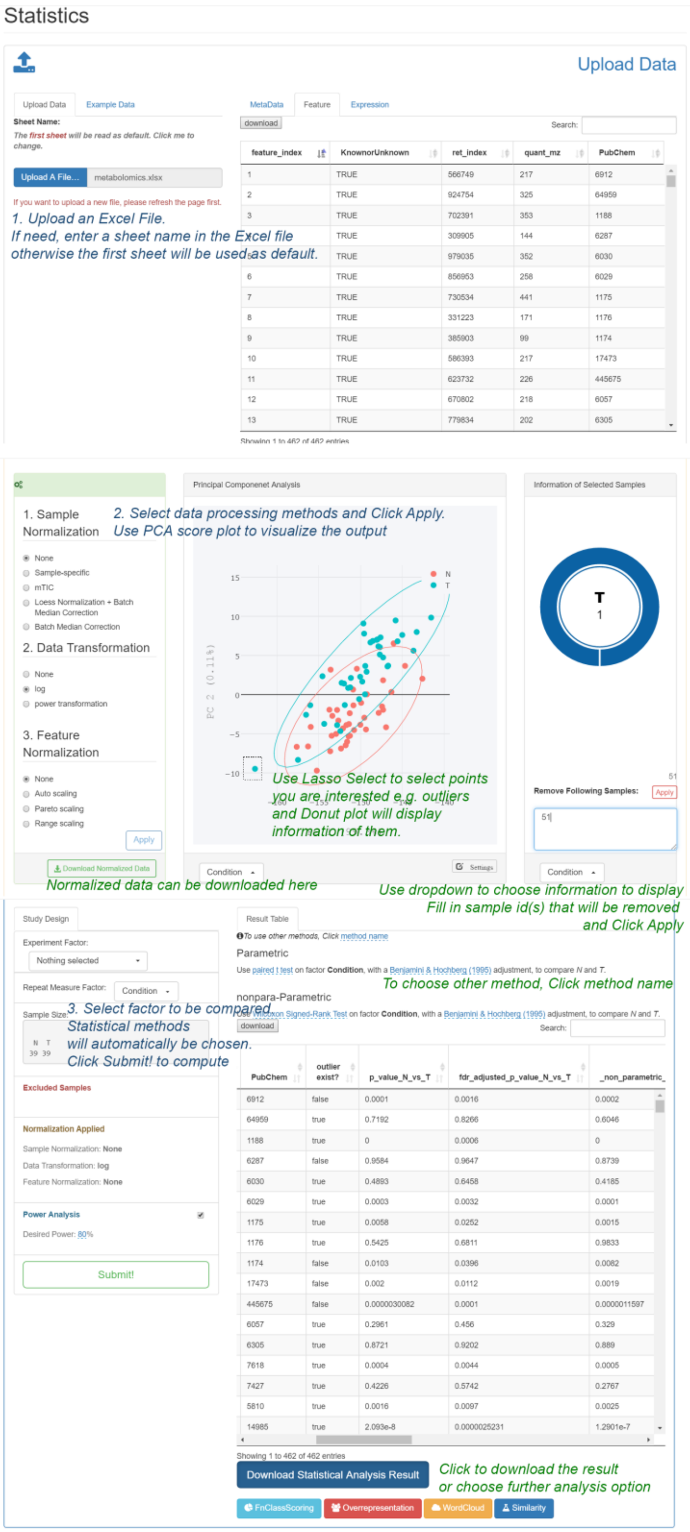
This tutorial shows that metabox is used for deep analysis (data processing, statistical analysis, metabolic network construction and functional interpretation) of metabolomic data. The example metabolomic data set is in *metabolomics.xlsx*. It is uploaded to metabox for log transformation before statistical analysis using paired t-test (Fig 1). The output from the data normalization and the statistical analysis is transferred to the network construction part to calculate the chemical structure similarity network of the PubChem compounds. The default threshold of correlation coefficient > 0.7 was used. The resulting network is enhanced by further mapping with annotation information. In this case, we applied Functional class scoring option to estimate significantly enriched pathways of network nodes by taking metabolic profiles in to account (see Fig 2 for all the steps).

Alternatively, the output from statistical analysis can be passed to one of the functional analysis method to aid interpretation of measured compounds. In this case, we applied WordCloud option to quickly summarize KEGG pathways of the compounds (see Fig 3 for all the steps). The top ten categories of measured compounds include Central carbon metabolism in cancer, Protein digestion and absorption, Aminoacyl-tRNA biosynthesis, Purine metabolism, Pyrimidine metabolism, Glyoxylate and dicarboxylate metabolism, Arginine and proline metabolism, Pentose and glucuronate interconversions, Galactose metabolism and Fatty acid biosynthesis sorted by the number of compounds in each category. This example shows that after statistical analyses, the result can be interpreted with pathway information.

3. Integrative exploration of significant genes and compounds in biological network context

Metabox is used for joint exploration of the lists of significant genes and compounds from comparisons between tumor and non-tumor tissues in the context of biological networks. The *ensembl_stats.txt* and *pubchem_stats.txt* are used to construct the biological network outlining relationships between genes and compounds as *(:Protein)-[:CONTROL]->(from:Gene)-[:CONVERSION]->(:Protein)-[:CATALYSIS]->(to:Compound)*. The resulting network including all attributes can be downloaded for advance exploration using visualization software such as Cytoscape (5). Fig 4 shows all the steps to construct the network with metabox.

FIGURES



Statistics

Upload Data

Upload Data | Example Data

Sheet Name:

*The **first sheet** will be read as default. Click me to change.*

Upload A File... metabolomics.xlsx

If you want to upload a new file, please refresh the page first.

1. Upload an Excel File.
If need, enter a sheet name in the Excel file otherwise the first sheet will be used as default.

MetaData | Feature | Expression

download

Search:

feature_index	KnownorUnknown	ret_index	quant_mz	PubChem
1	TRUE	566749	217	6912
2	TRUE	924754	325	64959
3	TRUE	702391	353	1188
4	TRUE	309905	144	6287
5	TRUE	979035	352	6030
6	TRUE	856953	258	6029
7	TRUE	730534	441	1175
8	TRUE	331223	171	1176
9	TRUE	385903	99	1174
10	TRUE	586393	217	17473
11	TRUE	623732	226	445675
12	TRUE	670802	218	6057
13	TRUE	779834	202	6305

Showing 1 to 462 of 462 entries

1. Sample Normalization

- ● None
- ○ Sample-specific
- ○ mTIC
- ○ Loess Normalization + Batch Median Correction
- ○ Batch Median Correction

2. Data Transformation

- ○ None
- ● log
- ○ power transformation

3. Feature Normalization

- ● None
- ○ Auto scaling
- ○ Pareto scaling
- ○ Range scaling

Apply

Download Normalized Data

Normalized data can be downloaded here

Principal Componenet Analysis

2. Select data processing methods and Click Apply.
Use PCA score plot to visualize the output

Use Lasso Select to select points you are interested e.g. outliers and Donut plot will display information of them.

Condition

Settings

Information of Selected Samples

Remove Following Samples: Apply

51

Condition

Use dropdown to choose information to display
Fill in sample id(s) that will be removed and Click Apply

Study Design

Experiment Factor: Nothing selected

Repeat Measure Factor: Condition

Sample Size:

N	T
39	39

3. Select factor to be compared.
Statistical methods will automatically be chosen.
Click Submit! to compute

Excluded Samples

Normalization Applied

Sample Normalization: **None**

Data Transformation: **log**

Feature Normalization: **None**

Power Analysis ☑

Desired Power: 80%

Submit!

Result Table

To use other methods, Click method name

Parametric

Use paired t test on factor **Condition**, with a Benjamini & Hochberg (1995) adjustment, to compare *N* and *T*.

To choose other method, Click method name

nonpara-Parametric

Use Wilcoxon Signed-Rank Test on factor **Condition**, with a Benjamini & Hochberg (1995) adjustment, to compare *N* and *T*.

download

Search:

PubChem	outlier exist?	p_value_N_vs_T	fdr_adjusted_p_value_N_vs_T	_non_parametric_
6912	false	0.0001	0.0016	0.0002
64959	true	0.7192	0.8266	0.6046
1188	true	0	0.0006	0
6287	false	0.9584	0.9647	0.8739
6030	true	0.4893	0.6458	0.4185
6029	true	0.0003	0.0032	0.0001
1175	true	0.0058	0.0252	0.0015
1176	true	0.5425	0.6811	0.9833
1174	false	0.0103	0.0396	0.0082
17473	false	0.002	0.0112	0.0019
445675	false	0.0000030082	0.0001	0.0000011597
6057	true	0.2961	0.456	0.329
6305	true	0.8721	0.9202	0.889
7618	true	0.0004	0.0044	0.0005
7427	true	0.4226	0.5742	0.2767
5810	true	0.0016	0.0097	0.0025
14985	true	2.093e-8	0.0000025231	1.2901e-7

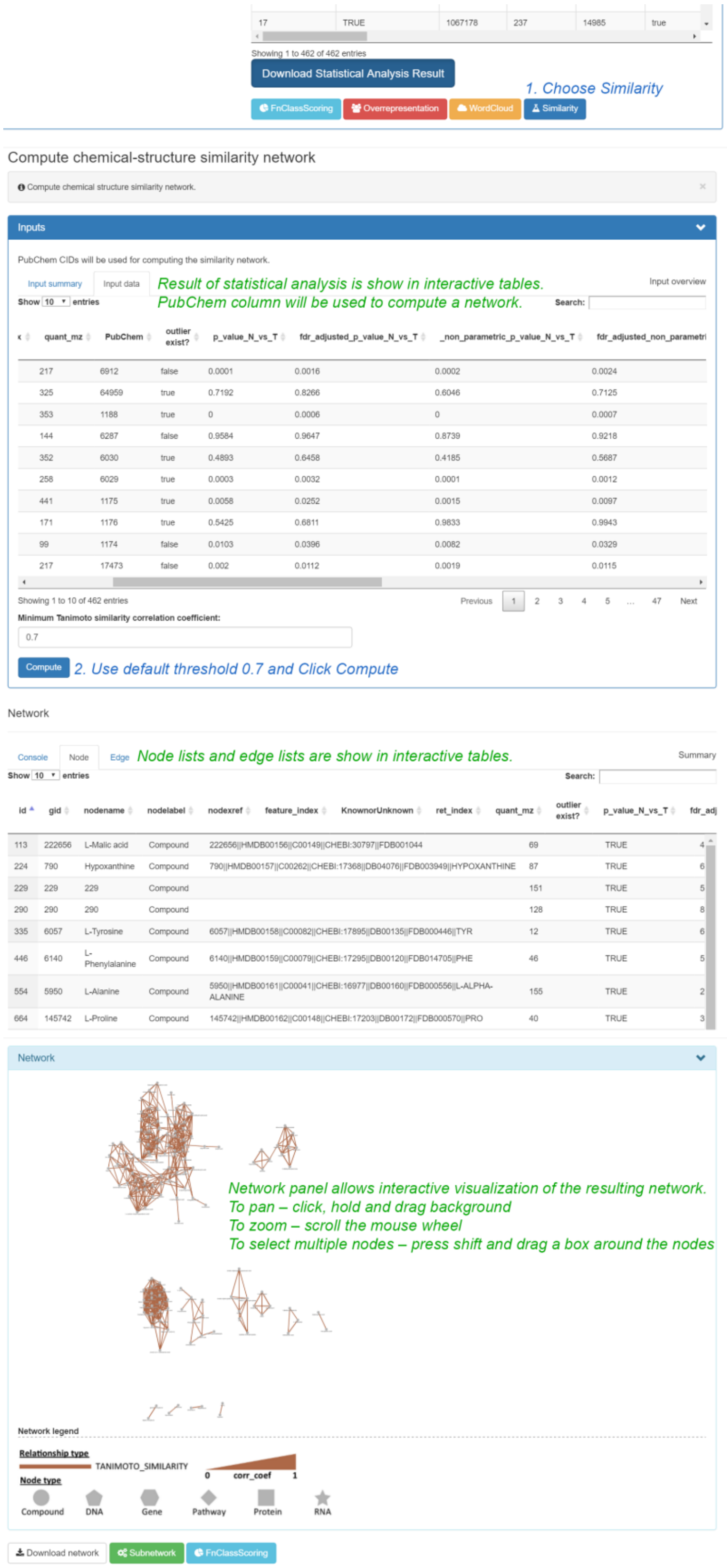
Showing 1 to 462 of 462 entries

Download Statistical Analysis Result

Click to download the result or choose further analysis option

FnClassScoring | Overrepresentation | WordCloud | Similarity

Fig 1. Data normalization and statistical analysis comparing two-paired groups of metabolomic study.



17	TRUE	1067178	237	14985	true

Showing 1 to 462 of 462 entries

1. Choose Similarity

Compute chemical-structure similarity network

Compute chemical structure similarity network.

PubChem CIDs will be used for computing the similarity network.

Result of statistical analysis is show in interactive tables.
PubChem column will be used to compute a network.

Show 10 entries

Search:

x	quant_mz	PubChem	outlier exist?	p_value_N_vs_T	fdr_adjusted_p_value_N_vs_T	_non_parametric_p_value_N_vs_T	fdr_adjusted_non_parametri
	217	6912	false	0.0001	0.0016	0.0002	0.0024
	325	64959	true	0.7192	0.8266	0.6046	0.7125
	353	1188	true	0	0.0006	0	0.0007
	144	6287	false	0.9584	0.9647	0.8739	0.9218
	352	6030	true	0.4893	0.6458	0.4185	0.5687
	258	6029	true	0.0003	0.0032	0.0001	0.0012
	441	1175	true	0.0058	0.0252	0.0015	0.0097
	171	1176	true	0.5425	0.6811	0.9833	0.9943
	99	1174	false	0.0103	0.0396	0.0082	0.0329
	217	17473	false	0.002	0.0112	0.0019	0.0115

Showing 1 to 10 of 462 entries

Minimum Tanimoto similarity correlation coefficient:

0.7

2. Use default threshold 0.7 and Click Compute

Network

Node lists and edge lists are show in interactive tables.

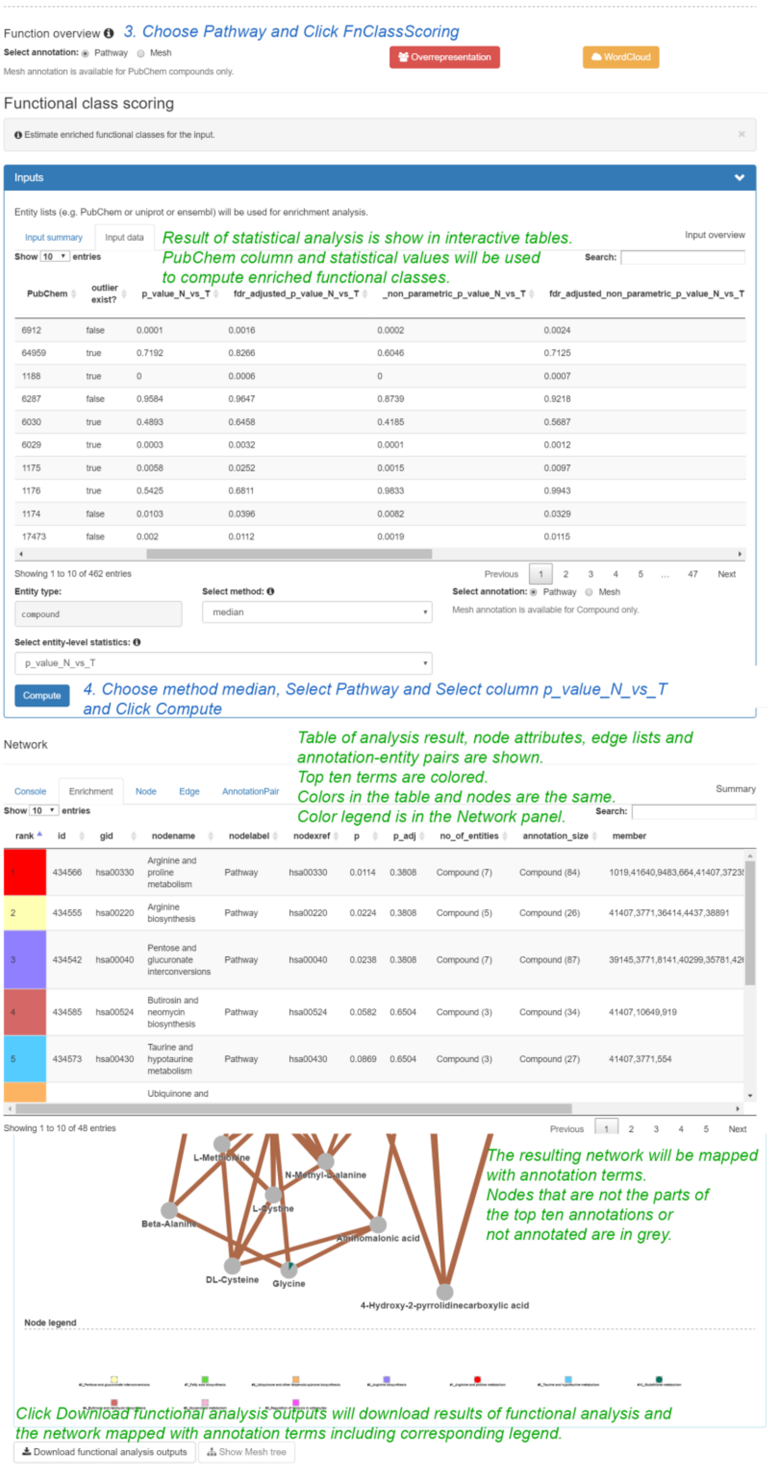
Show 10 entries

Search:

id	gid	nodename	nodelabel	nodexref	feature_index	KnownorUnknown	ret_index	quant_mz	outlier exist?	p_value_N_vs_T	fdr_adj
113	222656	L-Malic acid	Compound	222656||HMDB00156||C00149||CHEBI:30797||FDB001044				69		TRUE	4
224	790	Hypoxanthine	Compound	790||HMDB00157||C00262||CHEBI:17368||DB04076||FDB003949||HYPOXANTHINE				87		TRUE	6
229	229	229	Compound					151		TRUE	5
290	290	290	Compound					128		TRUE	8
335	6057	L-Tyrosine	Compound	6057||HMDB00158||C00082||CHEBI:17895||DB00135||FDB000446||TYR				12		TRUE	6
446	6140	L-Phenylalanine	Compound	6140||HMDB00159||C00079||CHEBI:17295||DB00120||FDB014705||PHE				46		TRUE	5
554	5950	L-Alanine	Compound	5950||HMDB00161||C00041||CHEBI:16977||DB00160||FDB000556||L-ALPHA-ALANINE				155		TRUE	2
664	145742	L-Proline	Compound	145742||HMDB00162||C00148||CHEBI:17203||DB00172||FDB000570||PRO				40		TRUE	3

Network panel allows interactive visualization of the resulting network.
To pan – click, hold and drag background
To zoom – scroll the mouse wheel
To select multiple nodes – press shift and drag a box around the nodes

Fig 2. Computing chemical similarity network and performing FnClassScoring on network nodes.



Function overview 🛈 *3. Choose Pathway and Click FnClassScoring*

Select annotation: ● Pathway ○ Mesh

Mesh annotation is available for PubChem compounds only.

Functional class scoring

🛈 Estimate enriched functional classes for the input.

Inputs

Entity lists (e.g. PubChem or uniprot or ensembl) will be used for enrichment analysis.

Result of statistical analysis is show in interactive tables. PubChem column and statistical values will be used to compute enriched functional classes.

Show 10 ▾ **entries** **Search:**

PubChem	outlier exist?	p_value_N_vs_T	fdr_adjusted_p_value_N_vs_T	_non_parametric_p_value_N_vs_T	fdr_adjusted_non_parametric_p_value_N_vs_T
6912	false	0.0001	0.0016	0.0002	0.0024
64959	true	0.7192	0.8266	0.6046	0.7125
1188	true	0	0.0006	0	0.0007
6287	false	0.9584	0.9647	0.8739	0.9218
6030	true	0.4893	0.6458	0.4185	0.5687
6029	true	0.0003	0.0032	0.0001	0.0012
1175	true	0.0058	0.0252	0.0015	0.0097
1176	true	0.5425	0.6811	0.9833	0.9943
1174	false	0.0103	0.0396	0.0082	0.0329
17473	false	0.002	0.0112	0.0019	0.0115

Showing 1 to 10 of 462 entries

Entity type: compound

Select method: 🛈 median

Select annotation: ● Pathway ○ Mesh

Mesh annotation is available for Compound only.

Select entity-level statistics: 🛈 p_value_N_vs_T

4. Choose method median, Select Pathway and Select column p_value_N_vs_T and Click Compute

Network

Table of analysis result, node attributes, edge lists and annotation-entity pairs are shown.
Top ten terms are colored.
Colors in the table and nodes are the same.
Color legend is in the Network panel.

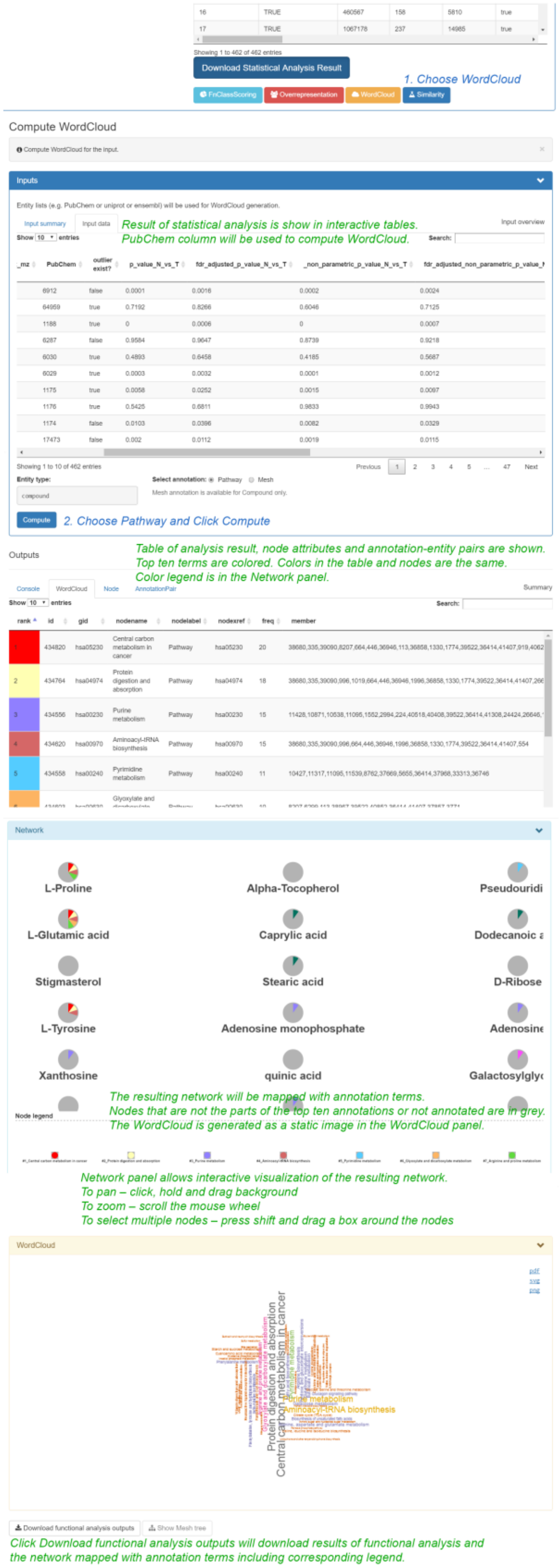
Show 10 ▾ **entries** **Search:**

rank	id	gid	nodename	nodelabel	nodexref	p	p_adj	no_of_entities	annotation_size	member
1	434566	hsa00330	Arginine and proline metabolism	Pathway	hsa00330	0.0114	0.3808	Compound (7)	Compound (84)	1019,41640,9483,664,41407,3723
2	434555	hsa00220	Arginine biosynthesis	Pathway	hsa00220	0.0224	0.3808	Compound (5)	Compound (26)	41407,3771,36414,4437,38891
3	434542	hsa00040	Pentose and glucuronate interconversions	Pathway	hsa00040	0.0238	0.3808	Compound (7)	Compound (87)	39145,3771,8141,40299,35781,42
4	434585	hsa00524	Butirosin and neomycin biosynthesis	Pathway	hsa00524	0.0582	0.6504	Compound (3)	Compound (34)	41407,10649,919
5	434573	hsa00430	Taurine and hypotaurine metabolism	Pathway	hsa00430	0.0869	0.6504	Compound (3)	Compound (27)	41407,3771,554
			Ubiquinone and							

Showing 1 to 10 of 48 entries

The resulting network will be mapped with annotation terms.
Nodes that are not the parts of the top ten annotations or not annotated are in grey.

Click Download functional analysis outputs will download results of functional analysis and the network mapped with annotation terms including corresponding legend.



16	TRUE	460567	158	5810	true
17	TRUE	1067178	237	14985	true

Showing 1 to 462 of 462 entries

1. Choose WordCloud

Compute WordCloud

Compute WordCloud for the input.

Entity lists (e.g. PubChem or uniprot or ensembl) will be used for WordCloud generation.

Result of statistical analysis is show in interactive tables.
PubChem column will be used to compute WordCloud.

Show 10 entries Search:

_mz	PubChem	outlier exist?	p_value_N_vs_T	fdr_adjusted_p_value_N_vs_T	_non_parametric_p_value_N_vs_T	fdr_adjusted_non_parametric_p_value_N
	6912	false	0.0001	0.0016	0.0002	0.0024
	64959	true	0.7192	0.8266	0.6046	0.7125
	1188	true	0	0.0006	0	0.0007
	6287	false	0.9584	0.9647	0.8739	0.9218
	6030	true	0.4893	0.6458	0.4185	0.5687
	6029	true	0.0003	0.0032	0.0001	0.0012
	1175	true	0.0058	0.0252	0.0015	0.0097
	1176	true	0.5425	0.6811	0.9833	0.9943
	1174	false	0.0103	0.0396	0.0082	0.0329
	17473	false	0.002	0.0112	0.0019	0.0115

Showing 1 to 10 of 462 entries Previous 1 2 3 4 5 … 47 Next

Entity type: compound

Select annotation: ● Pathway ○ Mesh

Mesh annotation is available for Compound only.

2. Choose Pathway and Click Compute

Outputs

Table of analysis result, node attributes and annotation-entity pairs are shown.
Top ten terms are colored. Colors in the table and nodes are the same.
Color legend is in the Network panel.

Show 10 entries Search:

rank	id	gid	nodename	nodelabel	nodexref	freq	member
1	434820	hsa05230	Central carbon metabolism in cancer	Pathway	hsa05230	20	38680,335,39090,8207,664,446,36946,113,36858,1330,1774,39522,36414,41407,919,4062
2	434764	hsa04974	Protein digestion and absorption	Pathway	hsa04974	18	38680,335,39090,996,1019,664,446,36946,1996,36858,1330,1774,39522,36414,41407,266
3	434556	hsa00230	Purine metabolism	Pathway	hsa00230	15	11428,10871,10538,11095,1552,2994,224,40518,40408,39522,36414,41308,24424,26646,
4	434620	hsa00970	Aminoacyl-tRNA biosynthesis	Pathway	hsa00970	15	38680,335,39090,996,664,446,36946,1996,36858,1330,1774,39522,36414,41407,554
5	434558	hsa00240	Pyrimidine metabolism	Pathway	hsa00240	11	10427,11317,11095,11539,8762,37669,5655,36414,37968,33313,36746
6	434603	hsa00630	Glyoxylate and dicarboxylate	Pathway	hsa00630	10	8207,6299,113,38867,39522,40662,36414,41407,37857,3771

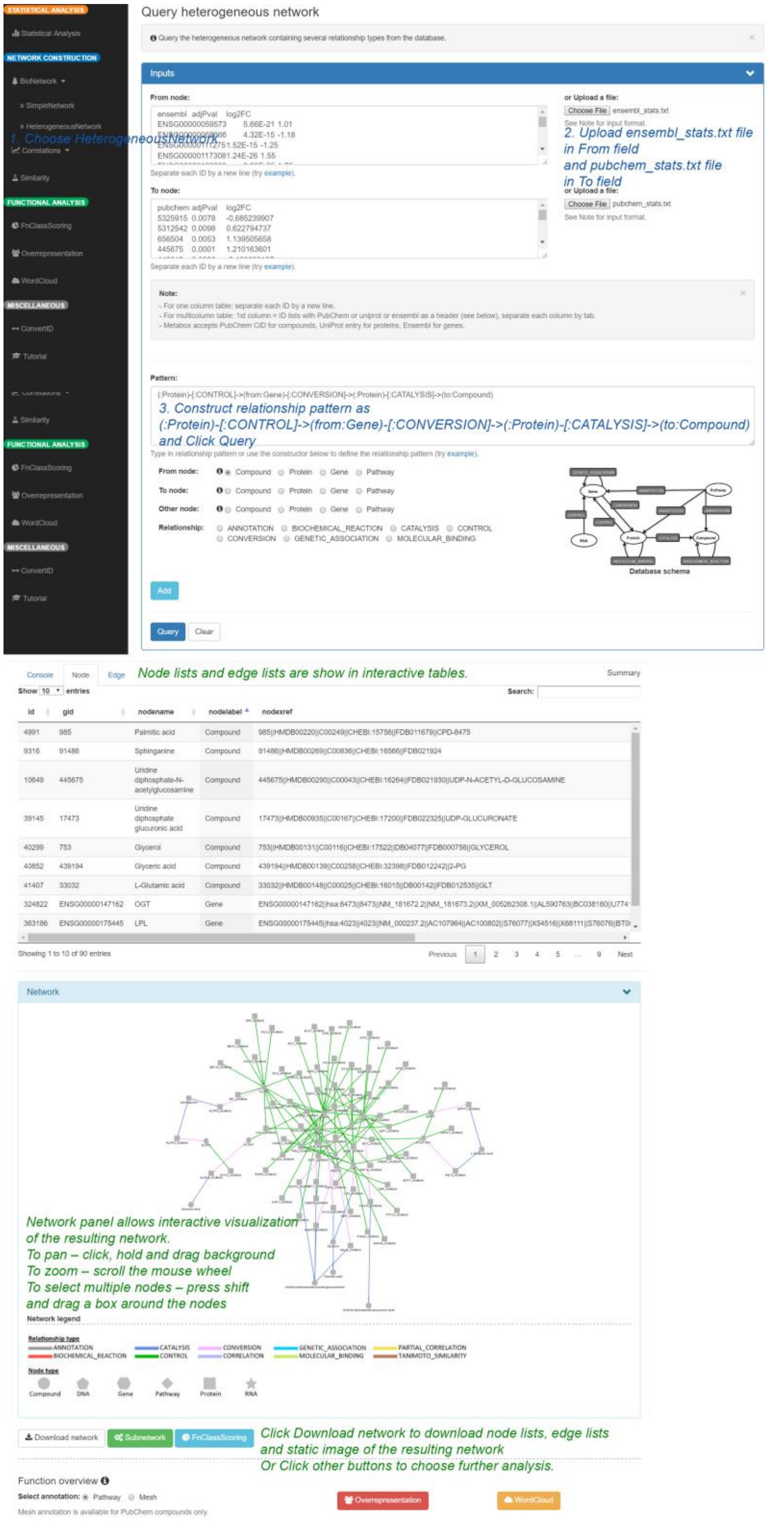
The resulting network will be mapped with annotation terms.
Nodes that are not the parts of the top ten annotations or not annotated are in grey.
The WordCloud is generated as a static image in the WordCloud panel.

Network panel allows interactive visualization of the resulting network.
To pan – click, hold and drag background
To zoom – scroll the mouse wheel
To select multiple nodes – press shift and drag a box around the nodes

WordCloud
pdf
svg
png
Central carbon metabolism in cancer
Protein digestion and absorption
Purine metabolism
Aminoacyl-tRNA biosynthesis
Citrate cycle (TCA cycle)
Biosynthesis of unsaturated fatty acids
Alanine, aspartate and glutamate metabolism
Valine, leucine and isoleucine biosynthesis
Download functional analysis outputs
Show Mesh tree

Click Download functional analysis outputs will download results of functional analysis and the network mapped with annotation terms including corresponding legend.

Fig 3. Computing WordCloud for measured compounds.



Query heterogeneous network

Query the heterogeneous network containing several relationship types from the database.

1. Choose HeterogeneousNetwork

From node:

```
ensembl  adjPval   log2FC
ENSG00000059573      5.66E-21 1.01
ENSG00000068066      4.32E-15 -1.18
ENSG000001112751.52E-15 -1.25
ENSG000001173081.24E-26 1.55
```

Separate each ID by a new line (try example).

or Upload a file: Choose File ensembl_stats.txt

See Note for input format.

2. Upload ensembl_stats.txt file in From field and pubchem_stats.txt file in To field

To node:

```
pubchem adjPval   log2FC
5325915  0.0078    -0.685239907
5312542  0.0098    0.622794737
656504   0.0053    1.139505658
445675   0.0001    1.210163601
```

Separate each ID by a new line (try example).

or Upload a file: Choose File pubchem_stats.txt

See Note for input format.

Note:
- For one column table: separate each ID by a new line.
- For multicolumn table: 1st column = ID lists with PubChem or uniprot or ensembl as a header (see below), separate each column by tab.
- Metabox accepts PubChem CID for compounds, UniProt entry for proteins, Ensembl for genes.

Pattern:

(:Protein)-[:CONTROL]->(from:Gene)-[:CONVERSION]->(:Protein)-[:CATALYSIS]->(to:Compound)

3. Construct relationship pattern as (:Protein)-[:CONTROL]->(from:Gene)-[:CONVERSION]->(:Protein)-[:CATALYSIS]->(to:Compound) and Click Query

Type in relationship pattern or use the constructor below to define the relationship pattern (try example).

From node: ● Compound ○ Protein ○ Gene ○ Pathway

To node: ○ Compound ○ Protein ○ Gene ○ Pathway

Other node: ○ Compound ○ Protein ○ Gene ○ Pathway

Relationship: ○ ANNOTATION ○ BIOCHEMICAL_REACTION ○ CATALYSIS ○ CONTROL ○ CONVERSION ○ GENETIC_ASSOCIATION ○ MOLECULAR_BINDING

Node lists and edge lists are show in interactive tables.

Show 10 entries Search:

id	gid	nodename	nodelabel	nodexref
4991	985	Palmitic acid	Compound	985||HMDB00220||C00249||CHEBI:15756||FDB011679||CPD-8475
9316	91486	Sphinganine	Compound	91486||HMDB00269||C00836||CHEBI:16566||FDB021924
10649	445675	Uridine diphosphate-N-acetylglucosamine	Compound	445675||HMDB00290||C00043||CHEBI:16264||FDB021930||UDP-N-ACETYL-D-GLUCOSAMINE
39145	17473	Uridine diphosphate glucuronic acid	Compound	17473||HMDB00935||C00167||CHEBI:17200||FDB022325||UDP-GLUCURONATE
40299	753	Glycerol	Compound	753||HMDB00131||C00116||CHEBI:17522||DB04077||FDB000756||GLYCEROL
40852	439194	Glyceric acid	Compound	439194||HMDB00139||C00258||CHEBI:32398||FDB012242||2-PG
41407	33032	L-Glutamic acid	Compound	33032||HMDB00148||C00025||CHEBI:16015||DB00142||FDB012535||GLT
324822	ENSG00000147162	OGT	Gene	ENSG00000147162||hsa:8473||8473||NM_181672.2||NM_181673.2||XM_005262308.1||AL590763||BC038180||U774
363186	ENSG00000175445	LPL	Gene	ENSG00000175445||hsa:4023||4023||NM_000237.2||AC107964||AC100802||S76077||X54516||X68111||S76076||BT0

Showing 1 to 10 of 90 entries

Previous 1 2 3 4 5 … 9 Next

Network panel allows interactive visualization of the resulting network.
To pan – click, hold and drag background
To zoom – scroll the mouse wheel
To select multiple nodes – press shift and drag a box around the nodes

Network legend

Relationship type

▬ ANNOTATION	▬ CATALYSIS	▬ CONVERSION	▬ GENETIC_ASSOCIATION	▬ PARTIAL_CORRELATION
▬ BIOCHEMICAL_REACTION	▬ CONTROL	▬ CORRELATION	▬ MOLECULAR_BINDING	▬ TANIMOTO_SIMILARITY

Node type

● Compound ⬟ DNA ⬢ Gene ◆ Pathway ■ Protein ★ RNA

Click Download network to download node lists, edge lists and static image of the resulting network
Or Click other buttons to choose further analysis.

Function overview

Select annotation: ● Pathway ○ Mesh

Mesh annotation is available for PubChem compounds only.

Fig 4. Constructing biological network from gene lists and compound lists. Relationship pattern between genes and compounds is:

(:Protein)-[:CONTROL]->(from:Gene)-[:CONVERSION]->(:Protein)-[:CATALYSIS]->(to:Compound)

REFERENCES

1. Selamat SA, Chung BS, Girard L, Zhang W, Zhang Y, Campan M, et al. Genome-scale analysis of DNA methylation in lung adenocarcinoma and integration with mRNA expression. Genome research. 2012;22(7):1197-211.

2. Wikoff WR, Grapov D, Fahrmann JF, DeFelice B, Rom WN, Pass HI, et al. Metabolomic markers of altered nucleotide metabolism in early stage adenocarcinoma. Cancer prevention research. 2015;8(5):410-8.

3. Fiehn O, Wohlgemuth G, Scholz M. Setup and annotation of metabolomic experiments by integrating biological and mass spectrometric metadata. Lect Notes Comput Sc. 2005;3615:224-39.

4. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets--update. Nucleic acids research. 2013;41(Database issue):D991-5.

5. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research. 2003;13(11):2498-504.