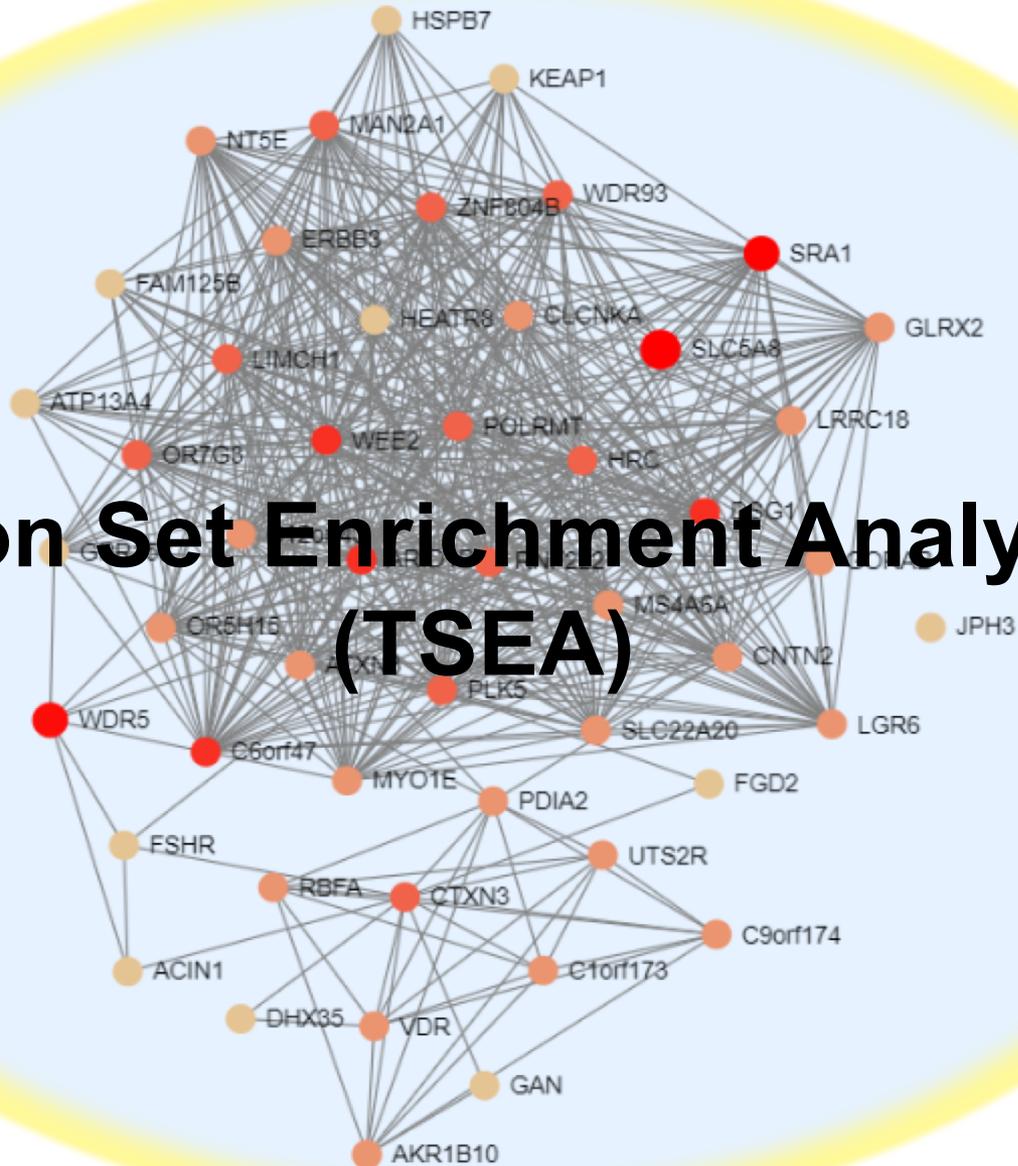


Taxon Set Enrichment Analysis (TSEA)



Goal

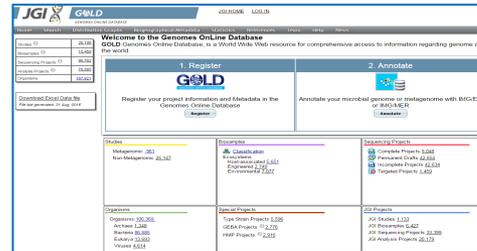
- Starting from a list of microbes of interest to test whether there are enrichments of **taxon sets**
- **Taxon set** as groups of microbes that has something in common. For example,
 - ❖ microbes that share same phenotypical traits, association with host lifestyles or biochemistry. (such as shape, motility, oxygen requirement, diet, food, BMI etc.)
 - ❖ microbes that have been found to be significantly associated particular developmental, physiological or disease conditions
 - ❖ microbes that are altered in composition by a common genetic variation.

Taxon Sets

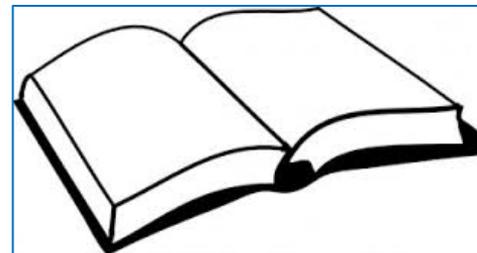
- Currently, MicrobiomeAnalyst supports three type of taxon sets based on the taxonomic resolution of organism or microbes present:
 - ❖ **Mixed-level taxon sets** (for 16S marker sequencing data)
 - Any possible taxonomic level (i.e. phylum to species) can be used
 - Two types of taxon sets: **Human disease** and **Human genetic variations** associated
 - Human disease associated:
 - **39** disease associated taxa sets
 - Derived from public database MicroPattern (<http://www.cuilab.cn/micropattern>)
 - Human genetic variations associated:
 - **1520** genetic variations (genes) associated taxa sets
 - Derived from the a web-based tool HOMINID (<http://blekhman-server1.oit.umn.edu/otugenediagram/>)
 - ❖ **Species-level taxon sets** (for shotgun and 16S marker sequencing data)
 - User can upload list of microbes characterized at species level.
 - **170** host lifestyles, physiology, development and biochemistry associated taxa sets
 - Both 16S as well as marker-gene sequencing can provide taxonomical resolution up to species level.
 - Derived from literate review (<http://science.sciencemag.org/content/352/6285/565>)
 - ❖ **Strain-level taxon sets** (for deep shotgun sequencing data)
 - Lowest most (i.e. stain level of microbes) is used
 - **100** disease and phenotype associated taxa sets
 - High-throughput shotgun sequencing provide such kind of taxonomical resolution for microbes

Strain-level taxa sets

HMP ID	Organism Name	Domain	MCB Superking	HMP Isolation Body Site	Project Status	Current Finish		
1	GI03551	Actinobacteria	defectiva ATCC 49716	BACTERIAL	Bacteria	oral	Complete	Level 2: High
4	GI03553	Actinobacterium	pleuroleus ATCC 43	BACTERIAL	Bacteria	always	Complete	Level 2: High
5	GI03554	Actinobacterium	cytoabundans C54	BACTERIAL	Bacteria	always	Complete	Level 5: Non-
8	GI03556	Actinobacterium	cockii sp. D21	BACTERIAL	Bacteria	gastrointestinal	Complete	Level 5: Para
10	GI03422	Actinobacterium	baumannii ATCC 19608	BACTERIAL	Bacteria	urinary_tract	Complete	Level 2: High
12	GI03421	Actinobacterium	calcoaceticus RH62002	BACTERIAL	Bacteria	skin	Complete	Level 2: High
13	GI03424	Actinobacterium	sp. SH204	BACTERIAL	Bacteria	skin	Complete	Level 2: High
14	GI03423	Actinobacterium	sp. RH2024	BACTERIAL	Bacteria	skin	Complete	Level 2: High
15	GI03597	Actinobacterium	faerholmii ATCC 1...	BACTERIAL	Bacteria	always	Complete	Level 2: High
16	GI03420	Actinobacterium	gibbosus SH346	BACTERIAL	Bacteria	skin	Complete	Level 2: High
17	GI03418	Actinobacterium	leuflin SH145	BACTERIAL	Bacteria	skin	Complete	Level 2: High
18	GI03417	Actinobacterium	radwastriens SH164	BACTERIAL	Bacteria	gastrointestinal	Complete	Level 2: High
19	GI03600	Actinobacterium	radwastriens SH202	BACTERIAL	Bacteria	skin	Complete	Level 2: High
20	GI03494	Actinobacterium	baumannii 6013115	BACTERIAL	Bacteria	skin	Complete	Level 2: High
21	GI03495	Actinobacterium	baumannii 6013150	BACTERIAL	Bacteria	skin	Complete	Level 2: High
22	GI03496	Actinobacterium	baumannii 6014059	BACTERIAL	Bacteria	skin	Complete	Level 2: High
23	GI03599	Actinobacterium	sp. ATCC 23244	BACTERIAL	Bacteria	skin	Complete	Level 2: High
26	GI03419	Actinobacterium	gibbii SH205	BACTERIAL	Bacteria	gastrointestinal	Complete	Level 2: High
27	GI03558	Actinobacterium	ureae ATCC 25976	BACTERIAL	Bacteria	always	Complete	Level 2: High



Organism phenotypes



Disease phenotypes



Phenotype database

- We derived our list of species from HMP (Human Microbiome Project) reference genome. (~2270 species)
- Then, we collected all their **phenotypes, ecological niches and disease associations** through **GOLD** database and **literature review**, for creating **species set**.
- Currently, **MicrobiomeAnalyst** have **100** such bacterial species set (~40 disease association sets).



Starting from marker gene abundance data (OTU table, BIOM file, mothur output)

Marker Data Profiling (MDP)

Shotgun Data Profiling (SDP)

Starting from gene list or gene abundance data annotated by KO, EC or COG

Visually exploring your 16S rRNA data with a public data in a 3D PCoA plot

Projection with Public Data (PPD)

Taxon Set Enrichment Analysis (TSEA)

Starting with a list of taxa of interest (strains, species or higher level taxa)

Click here to start

C. Upload a list of strain-level microbes

Mixed-level taxon list Species-level taxon list **Strain-level taxon list**

Enter a list of strain names or IDs 

Input type Strain Name 

Try our example

You can try our example also

Step 1 : Choose the parameters above. Copy and paste a list of strain-level microbes

```
Edwardsiella tarda ATCC 23685
Eikenella corrodens ATCC 23834
Enhydrobacter aerosaccus SK60
Enterococcus faecalis ATCC 29200
Enterococcus faecalis HH22
Enterococcus faecalis TUSoD Ef11
Enterococcus faecalis TX010
Enterococcus faecalis TX1322
Enterococcus faecium DO
Enterococcus faecium TX1330
Erysipelothrix rhusiopathiae ATCC 19414
```

Submit

Step 2 : Click "Submit" to proceed

2. Name mapping for microbes

Taxonomic Name/ID Mapping

- NCBI Taxonomy and GOLD database IDs are only applicable for strain-level ID mapping;
- Greek alphabets are not recognized, they should be replaced by English names (i.e. alpha, beta);
- Query names in normal white indicate exact match;
- Query names highlighted indicate **no exact or unique match**;
- Unmatched query names will be removed from further analysis;

Query	Database Hit	Species	Genus	NCBI Taxonomy	GOLD STAMP ID
Edwardsiella tarda ATCC 23685	Edwardsiella tarda ATCC 23685	Edwardsiella tarda	Edwardsiella	500638	Gp0003503
Eikenella corrodens ATCC 23834	Eikenella corrodens ATCC 23834	Eikenella corrodens	Eikenella	546274	Gp0003877
Enhydrobacter aerosaccus SK60	Enhydrobacter aerosaccus SK60	Enhydrobacter aerosaccus	Enhydrobacter	553217	Gp0004124
Enterococcus faecalis ATCC 29200	Enterococcus faecalis ATCC 29200	Enterococcus faecalis	Enterococcus	525271	Gp0003505
Enterococcus faecalis HH22	Enterococcus faecalis HH22	Enterococcus faecalis	Enterococcus	491075	Gp0001011
Enterococcus faecalis TUSoD Ef11	Enterococcus faecalis TUSoD Ef11	Enterococcus faecalis	Enterococcus	553209	Gp0005537
Enterococcus faecalis TX010	-	-	-	-	-
Enterococcus faecalis TX1322	Enterococcus faecalis TX1322	Enterococcus faecalis	Enterococcus	525278	Gp0003506
Enterococcus faecium DO	Enterococcus faecium DO	Enterococcus faecium	Enterococcus	333849	Gp0003507
Enterococcus faecium TX1330	Enterococcus faecium TX1330	Enterococcus faecium	Enterococcus	525279	Gp0003508
Erysipelothrix rhusiopathiae ATCC 19414	Erysipelothrix rhusiopathiae ATCC 19414	Erysipelothrix rhusiopathiae	Erysipelothrix	525280	Gp0003509

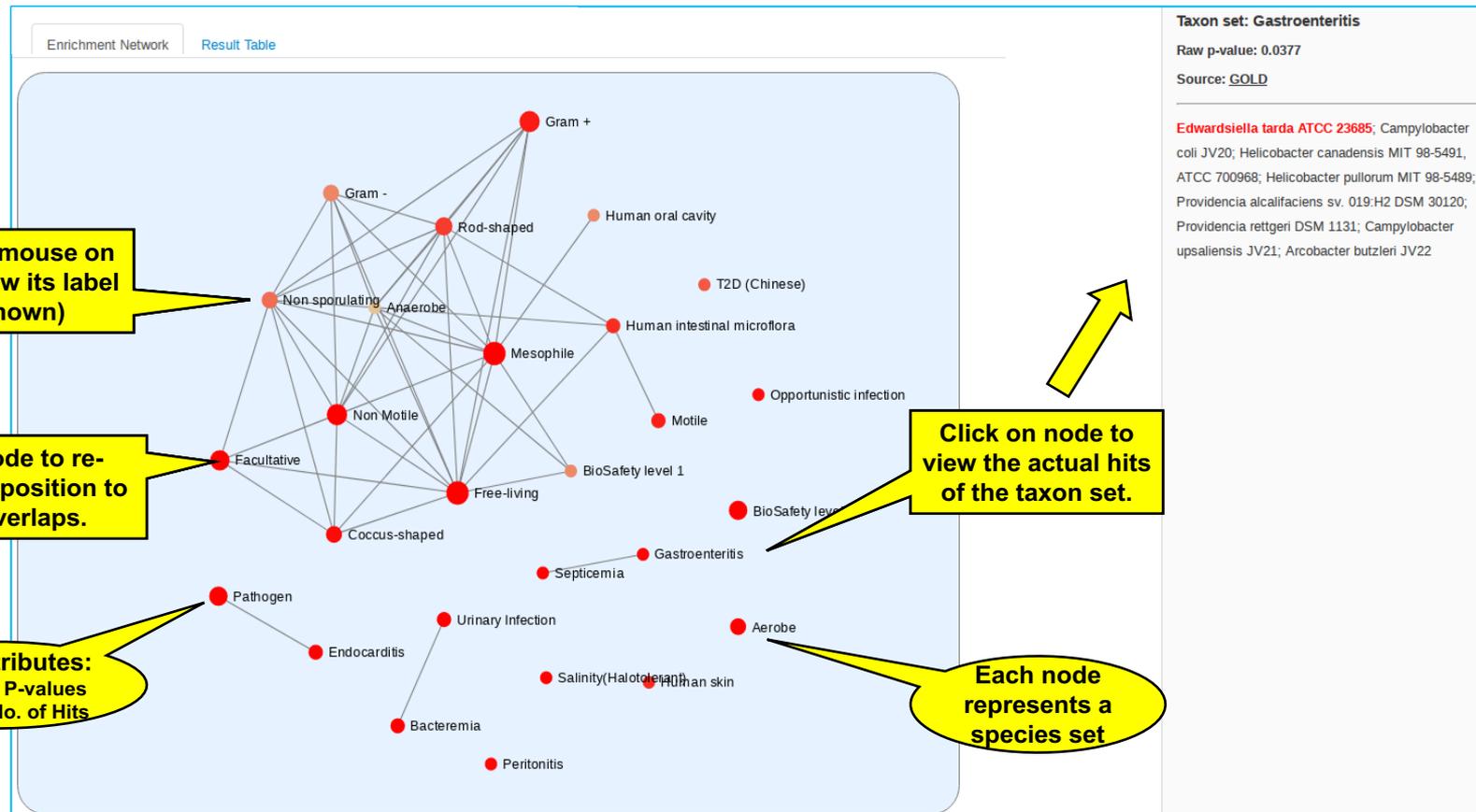
Submit

Click "Submit" to perform Enrichment analysis on strain-level taxon set.

User query

- We map the user upload organisms list with our list of organisms derived from HMP reference genome.

3. Enrichment Analysis Summary



Network view

Hypergeometric test is used to test if certain species sets are represented more often than expected by chance

4. Enrichment Analysis Results

Enrichment Network **Result Table**

	Taxon Set	Total	Hits	Expect	P value	FDR	Details
	Pathogen	85	5	0.406	2.11E-5	0.00211	View
	Facultative	190	6	0.908	7.99E-5	0.00399	View
	Free-living	885	10	4.23	1.77E-4	0.0051	View
	Urinary Infection	5	2	0.0239	2.04E-4	0.0051	View
	Bacteremia	7	2	0.0334	4.26E-4	0.00852	View
	Mesophile	987	10	4.72	5.31E-4	0.00885	View
	BioSafety level 2	194	5	0.927	0.00111	0.0159	View
	Non Motile	512	7	2.45	0.00299	0.0374	View
	Aerobe	80	3	0.382	0.00532	0.0591	View
	Endocarditis	33	2	0.158	0.01	0.1	View
	Salinity(Halotolerant)	3	1	0.0143	0.0143	0.13	View
	Septicemia	5	1	0.0239	0.0237	0.197	View
	Gastroenteritis	8	1	0.0382	0.0377	0.29	View
	Peritonitis	9	1	0.043	0.0423	0.302	View

Click on "View" to see list of microbes present within each set.

Taxon set: Gastroenteritis

Raw p-value: 0.0377

Source: [GOLD](#)

Edwardsiella tarda ATCC 23685; Campylobacter coli JV20; Helicobacter canadensis MIT 98-5491, ATCC 700968; Helicobacter pullorum MIT 98-5489; Providencia alcalifaciens sv. 019:H2 DSM 30120; Providencia rettgeri DSM 1131; Campylobacter upsaliensis JV21; Arcobacter butzleri JV22

B. Upload a list of species-level microbes

Mixed-level taxon list **Species-level taxon list** Strain-level taxon list

User can upload any valid name for microbes characterized at species level.

Enter a list of species names 

Try our example

You can try our example also

Step 1 : Choose the parameters above. Copy and paste a list of species-level microbes

Methanobrevibacter smithii
Bifidobacterium longum
Adlercreutzia equolifaciens
Eggerthella lenta;
Alistipes finegoldii
Alistipes putredinis
Bacteroides coprocola
Porphyromonas gingivalis
Eubacterium siraeum
Ruminococcus obeum
Butyrivibrio crossotus
Coprococcus catus
Eubacterium eligens
Eubacterium rectale
Eubacterium siraeum

Submit

Step 2 : Click "Submit" to proceed.

2. Name mapping for microbes

Taxonomic Name/ID Mapping

- NCBI Taxonomy and GOLD database IDs are only applicable for strain-level ID mapping;
- Greek alphabets are not recognized, they should be replaced by English names (i.e. alpha, beta);
- Query names in normal white indicate exact match;
- Query names highlighted indicate no exact or unique match;
- Unmatched query names will be removed from further analysis;

Query	Database Hit	Genus	Family	NCBI Taxonomy	GOLD STAMP ID
Methanobrevibacter smithii	Methanobrevibacter smithii	Methanobrevibacter	Methanobacteriaceae	-	-
Bifidobacterium longum	Bifidobacterium longum	Bifidobacterium	Bifidobacteriaceae	-	-
Adlercreutzia equolifaciens	Adlercreutzia equolifaciens	Adlercreutzia	Eggerthellaceae	-	-
Eggerthella lenta	Eggerthella lenta	Eggerthella	Eggerthellaceae	-	-
Alistipes finegoldii	Alistipes finegoldii	Alistipes	Rikenellaceae	-	-
Alistipes putredinis	Alistipes putredinis	Alistipes	Rikenellaceae	-	-
Bacteroides coprocola	Bacteroides coprocola	Bacteroides	Bacteroidaceae	-	-
Porphyromonas gingivalis	Porphyromonas gingivalis	Porphyromonas	Porphyromonadaceae	-	-
Eubacterium siraeum	Eubacterium siraeum	Eubacterium	Eubacteriaceae	-	-
Ruminococcus obeum	Ruminococcus obeum	Ruminococcus	Ruminococcaceae	-	-
Butyrivibrio crossotus	Butyrivibrio crossotus	Butyrivibrio	Lachnospiraceae	-	-
Coprococcus catus	Coprococcus catus	Coprococcus	Lachnospiraceae	-	-
Eubacterium eligens	Eubacterium eligens	Eubacterium	Eubacteriaceae	-	-
Eubacterium rectale	Eubacterium rectale	Eubacterium	Eubacteriaceae	-	-
Eubacterium siraeum	Eubacterium siraeum	Eubacterium	Eubacteriaceae	-	-

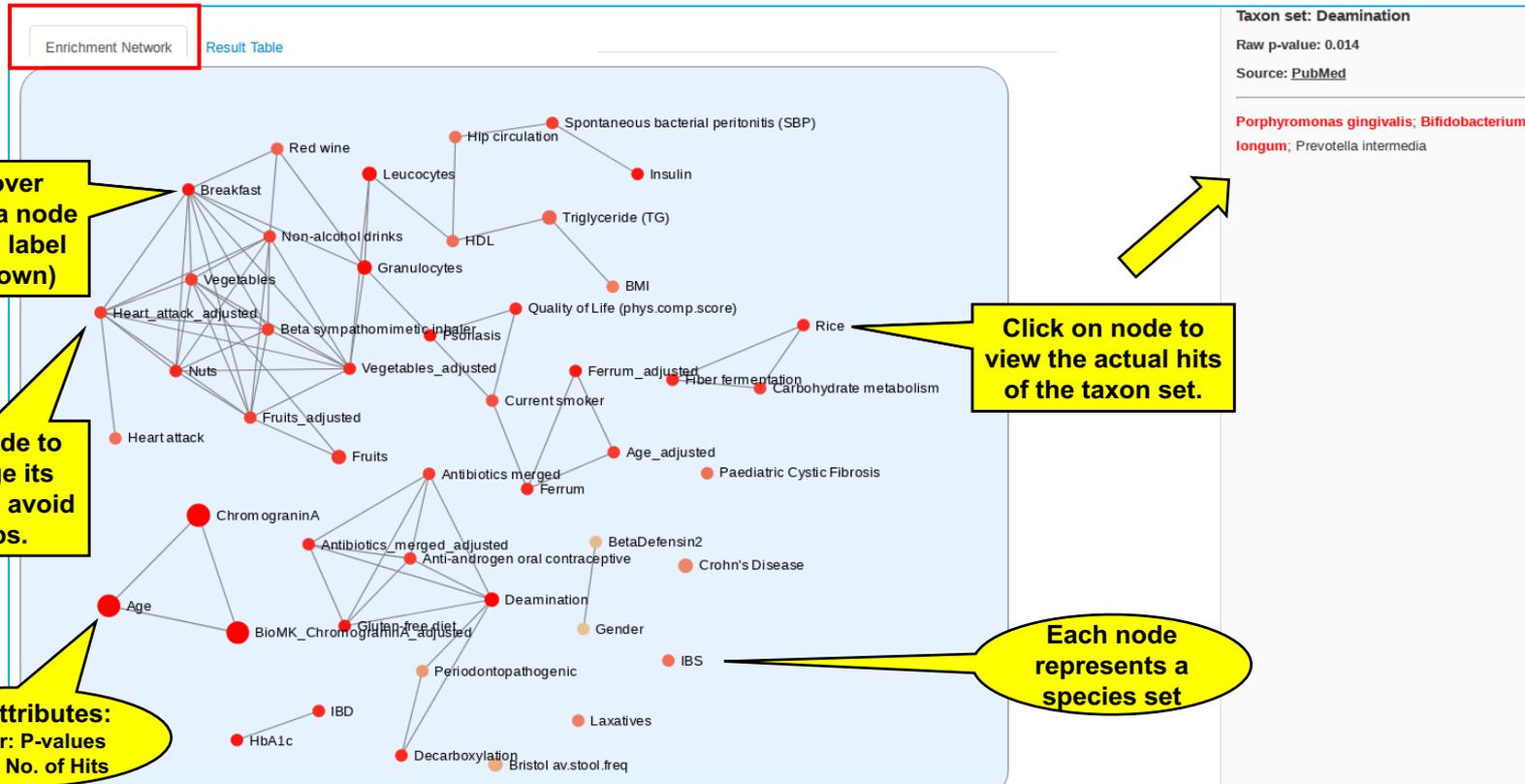
Submit

User query

Click "Submit" to perform Enrichment analysis on Species set.

- We map the user upload organisms list with list of organisms derived from literature review by other researchers.

3. Enrichment Analysis Summary



Network view

Hypergeometric test is used to test if certain species sets are represented more often than expected by chance

4. Enrichment Analysis Results

Enrichment Network **Result Table**

	Taxon Set	Total	Hits	Expect	P value	FDR	Details
	Age	50	10	3.61	2.58E-4	0.0439	View
	BioMK_ChromograninA_adjusted	61	10	4.4	0.00171	0.132	View
	ChromograninA	76	11	5.48	0.00232	0.132	View
	Deamination	3	2	0.216	0.014	0.594	View
	Granulocytes	5	2	0.361	0.0428	1.0	View
	Leucocytes	6	2	0.433	0.0615	1.0	View
	HbA1c	1	1	0.0722	0.0722	1.0	View
	Insulin	1	1	0.0722	0.0722	1.0	View
	Gluten-free diet	1	1	0.0722	0.0722	1.0	View
	Breakfast	1	1	0.0722	0.0722	1.0	View
	Psoriasis	1	1	0.0722	0.0722	1.0	View
	Ferrum_adjusted	1	1	0.0722	0.0722	1.0	View
	Ferrum	2	1	0.144	0.139	1.0	View
	Nuts	2	1	0.144	0.139	1.0	View
	Rice	2	1	0.144	0.139	1.0	View
	Vegetables_adjusted	2	1	0.144	0.139	1.0	View
	IBD	2	1	0.144	0.139	1.0	View
	Quality of Life (phys.comp.score)	2	1	0.144	0.139	1.0	View
	Antibiotics_merged_adjusted	2	1	0.144	0.139	1.0	View

Click on "View" to see list of microbes present within each set.

Taxon set: Deamination

Raw p-value: 0.014

Source: [PubMed](#)

Porphyromonas gingivalis; Bifidobacterium longum; Prevotella intermedia

A. Upload a list of mixed-level of microbes

Mixed-level taxon list Species-level taxon list Strain-level taxon list

User can upload any valid name for microbes and can be at any possible taxonomical level (phylum to species)

Enter a list of taxon names ?

Choose a taxon-set library Taxon sets associated with genetic variations

Try our example

You can try our example also

Actinobacillus porcinus
Alistipes finegoldii
Alistipes putredinis
Bacteroides coprocola
Bacteroides fragilis
Lachnospiraceae
Fusobacterium
Porphyromonas gingivalis
Bacteroides
Prevotella
Treponema

Submit

Step 1 : Choose the suitable taxon-set library from either disease-associated or genetic variations associated and paste a list of microbes

Step 2 : Copy and paste a list of microbes

Step 3: Click "Submit" to proceed.

2. Name mapping for microbes

Taxonomic Name/ID Mapping

- NCBI Taxonomy and GOLD database IDs are only applicable for strain-level ID mapping;
- Greek alphabets are not recognized, they should be replaced by English names (i.e. alpha, beta);
- Query names in normal white indicate exact match;
- Query names highlighted indicate **no exact or unique match**;
- Unmatched query names will be removed from further analysis;

Query	Database Hit	Species	Genus	NCBI Taxonomy	GOLD STAMP ID
Actinobacillus porcinus		-	-	-	-
Alistipes finegoldii		-	-	-	-
Alistipes putredinis		-	-	-	-
Bacteroides coprocola		-	-	-	-
Bacteroides fragilis		-	-	-	-
Lachnospiraceae	Lachnospiraceae	-	-	-	-
Fusobacterium	Fusobacterium	-	-	-	-
Porphyromonas gingivalis		-	-	-	-
Bacteroides	Bacteroides	-	-	-	-
Prevotella	Prevotella	-	-	-	-
Treponema		-	-	-	-

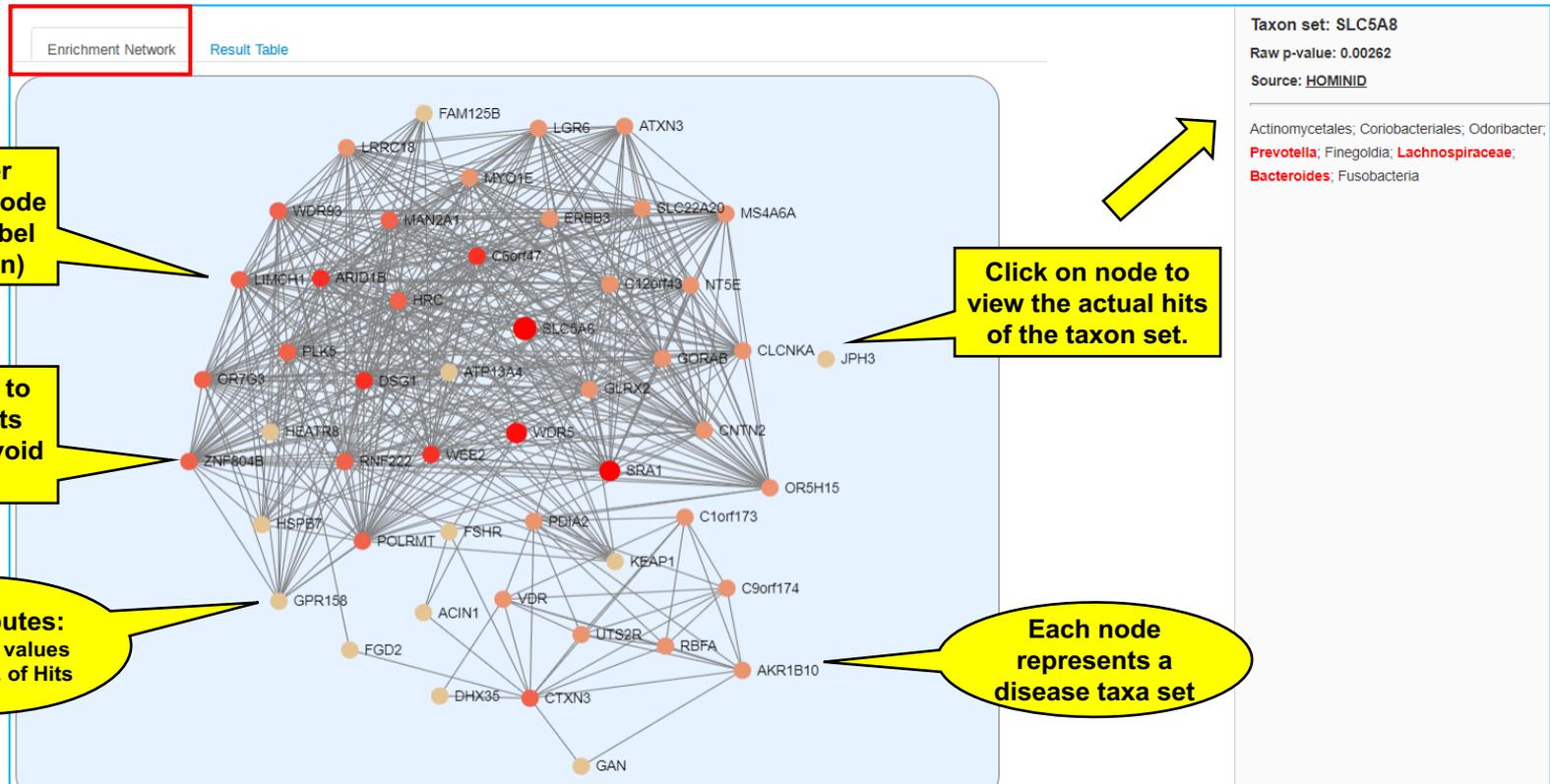
Submit

User query

Click "Submit" to perform Enrichment analysis on Species set.

- We map the user upload species list with our list of species derived from disease association database (MicroPattern).

3. Enrichment Analysis Summary



Network view

Hypergeometric test is used to test if certain disease-associated taxa sets are represented more often than expected by chance

4. Enrichment Analysis Results

Enrichment Network Result Table

Taxon Set	Total	Hits	Expect	P value	FDR	Details
SLC5A8	8	3	0.33	0.00262	1.0	View
SRA1	3	2	0.124	0.00454	1.0	View
WDR5	5	2	0.206	0.0145	1.0	View
ARID1B	1	1	0.0412	0.0412	1.0	View
C6orf47	1	1	0.0412	0.0412	1.0	View
DSG1	1	1	0.0412	0.0412	1.0	View
WEE2	1	1	0.0412	0.0412	1.0	View
CTXN3	2	1	0.0824	0.0808	1.0	View
HRC	2	1	0.0824	0.0808	1.0	View
LIMCH1	2	1	0.0824	0.0808	1.0	View
MAN2A1	2	1	0.0824	0.0808	1.0	View

Click on "View" to see list of microbes present within each set.

Taxon set: SLC5A8
Raw p-value: 0.00262
Source: [HOMINID](#)

Actinomycetales; Coriobacteriales; Odontobacter;
Prevotella; Finegoldia; **Lachnospiraceae**;
Bacteroides; Fusobacteria

=== END ===