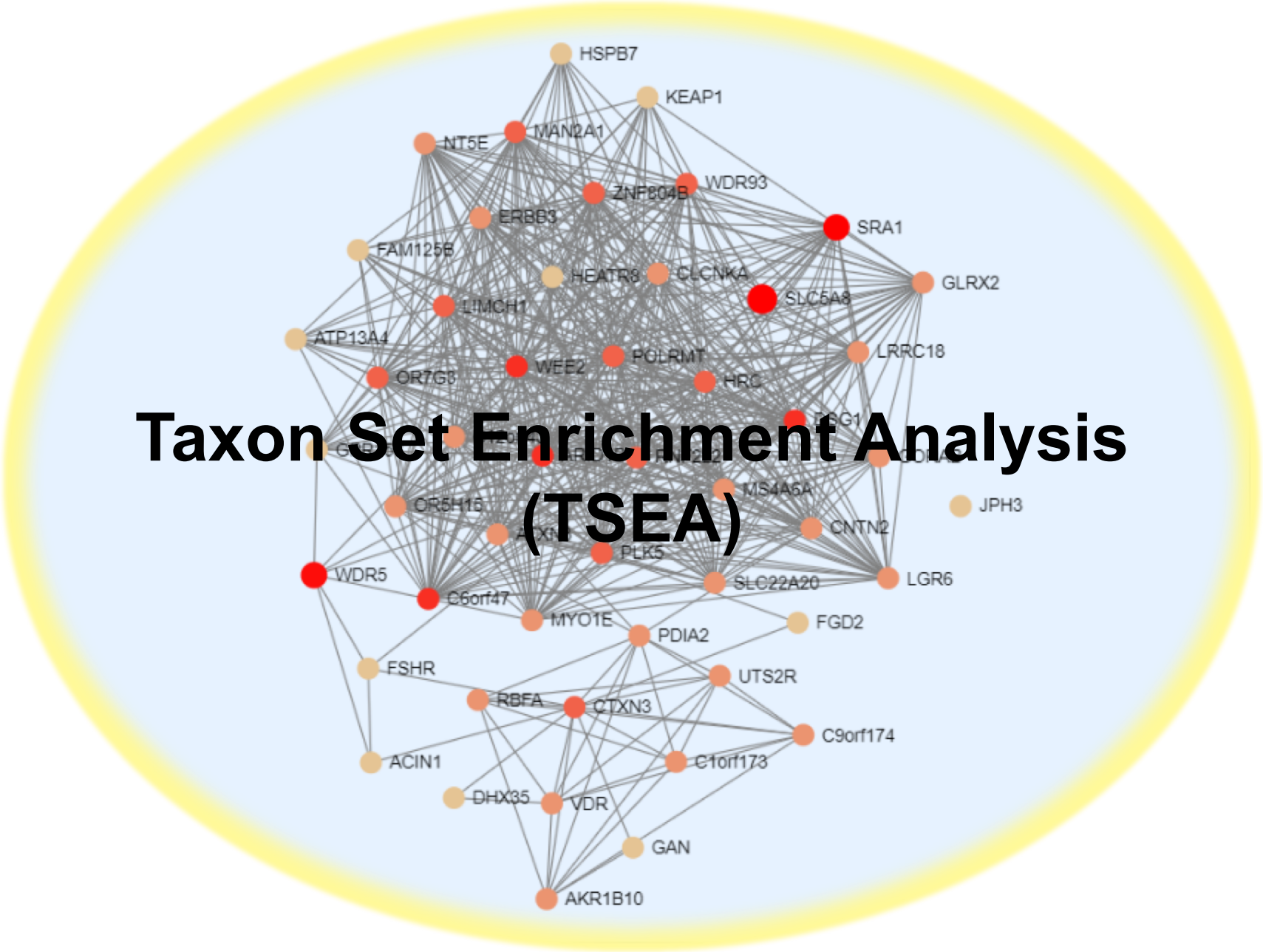



# Taxon Set Enrichment Analysis (TSEA)

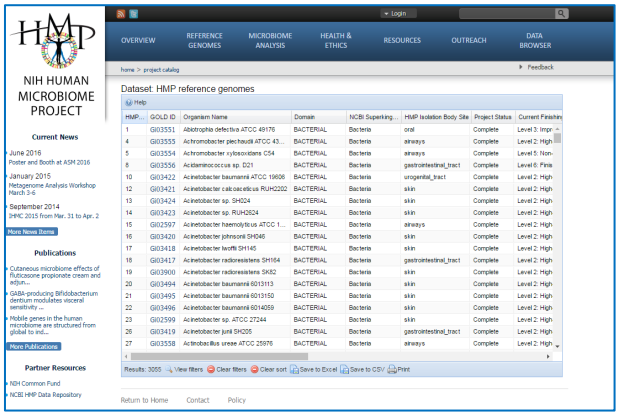
# Goal

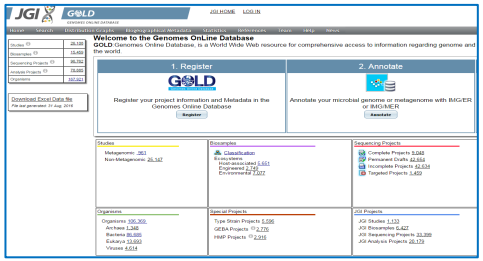
- Starting from a list of microbes of interest to test whether there are enrichments of **taxon sets**
- **Taxon set** as groups of microbes that has something in common. For example,
  - ❖ microbes that share same phenotypical traits, association with host lifestyles or biochemistry. (such as shape, 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

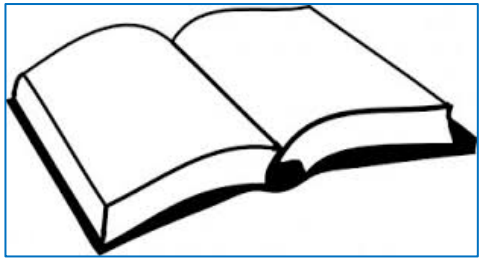
- Currently, MicrobiomeAnalyst supports three type of taxon sets based on the taxonomic resolution of organism or microbes present:
  - ❖ **Mixed-level taxon sets** (for 16S marker sequencing data)
    - Any possible taxonomic level (i.e. phylum to species) can be used
    - Two types of taxon sets**: 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 and16S 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




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).

MicrobiomeAnalyst -- comprehensive statistical, visual and meta-analysis of microbiome data

Home | Data Format | FAQs | Tutorials | Updates | About

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

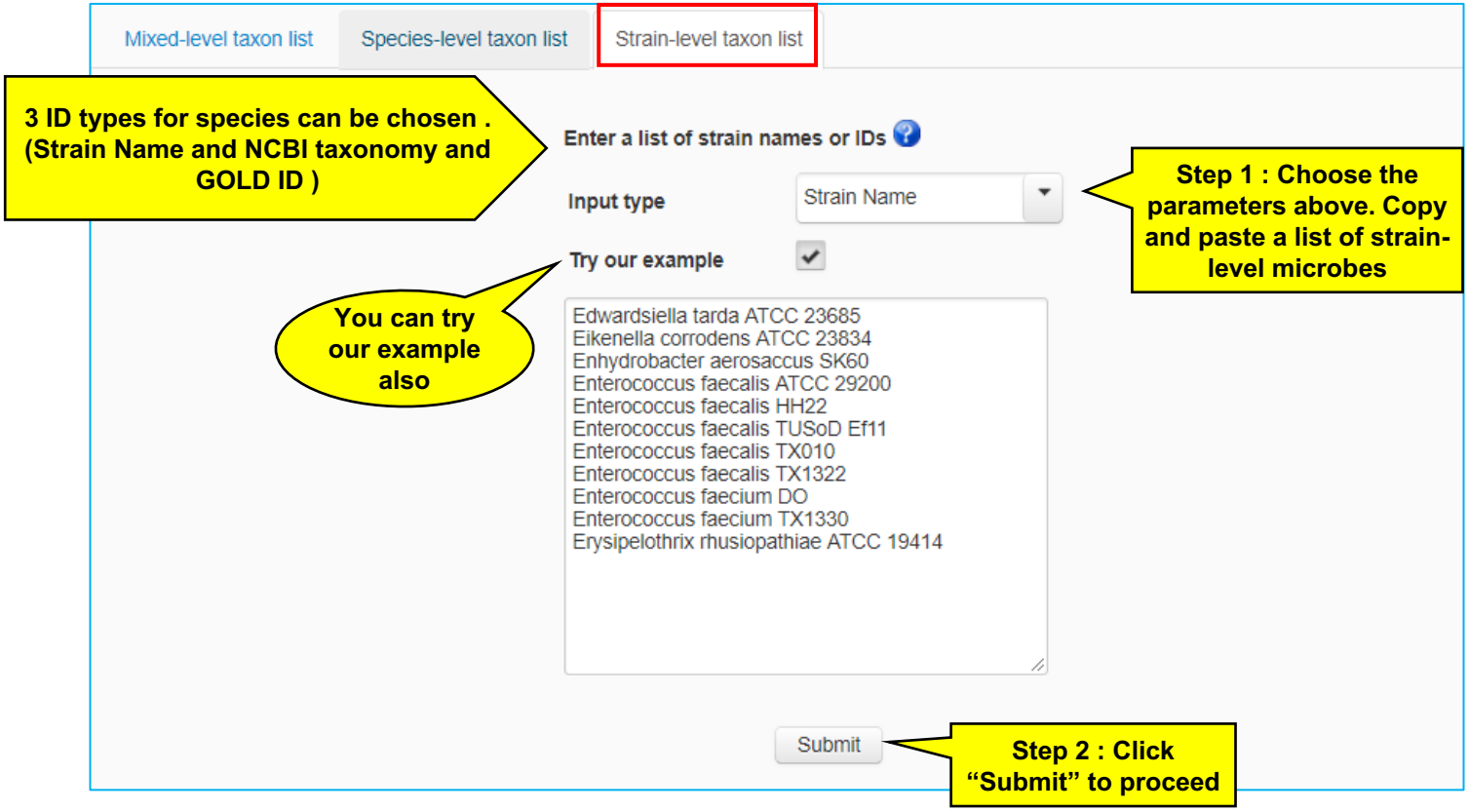
**Projection with Public Data (PPD)**

**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



# 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 |

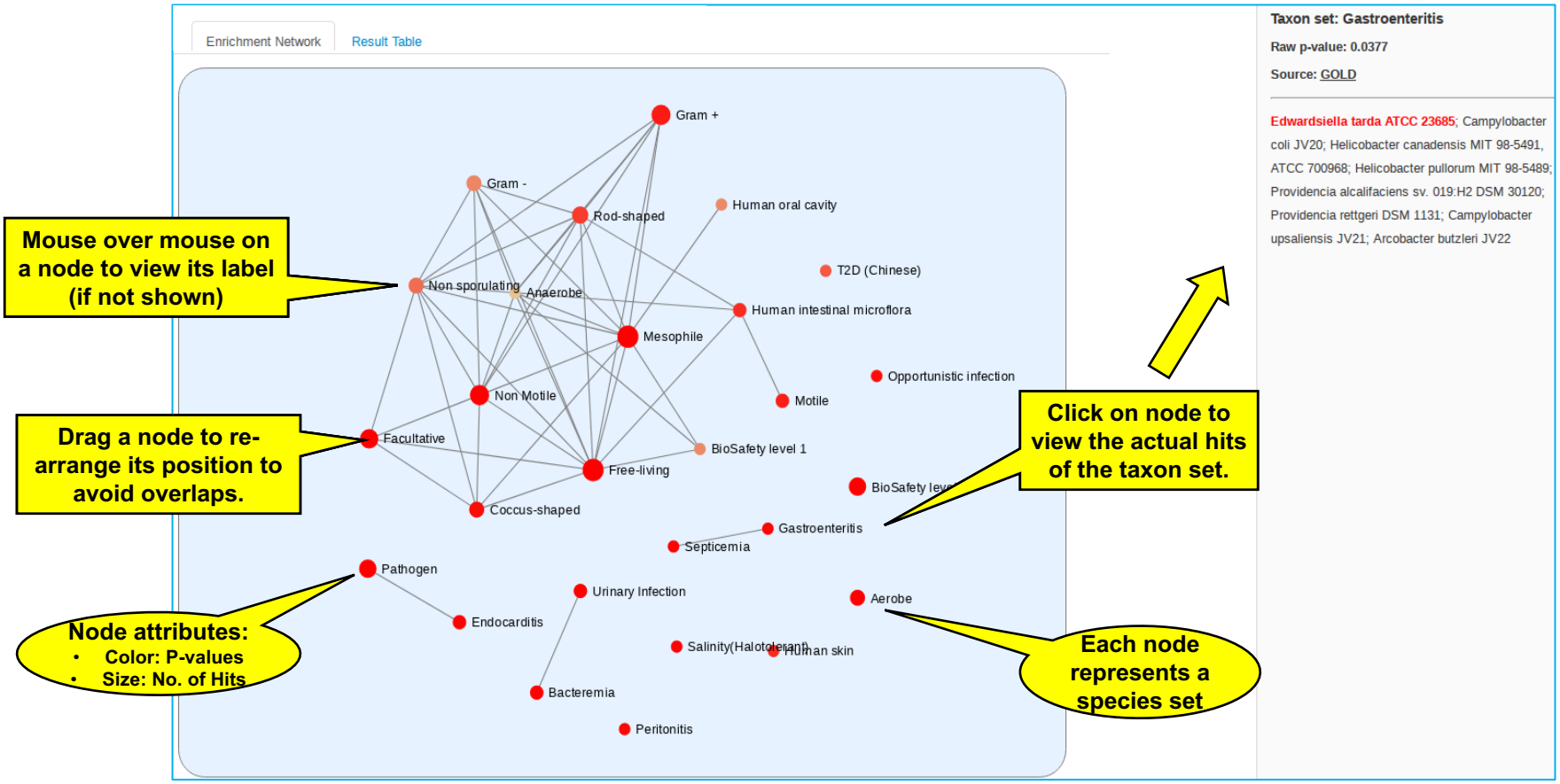
Submit

**User query**

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

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

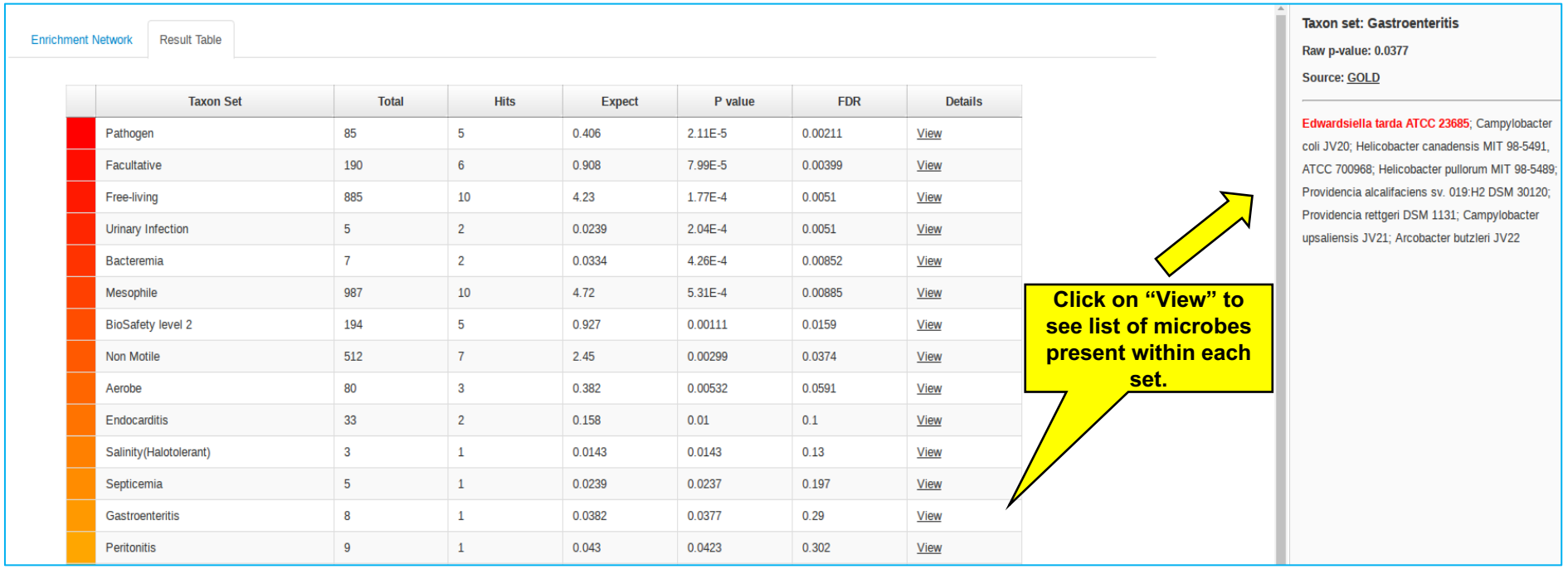
# 3. Enrichment Analysis Summary



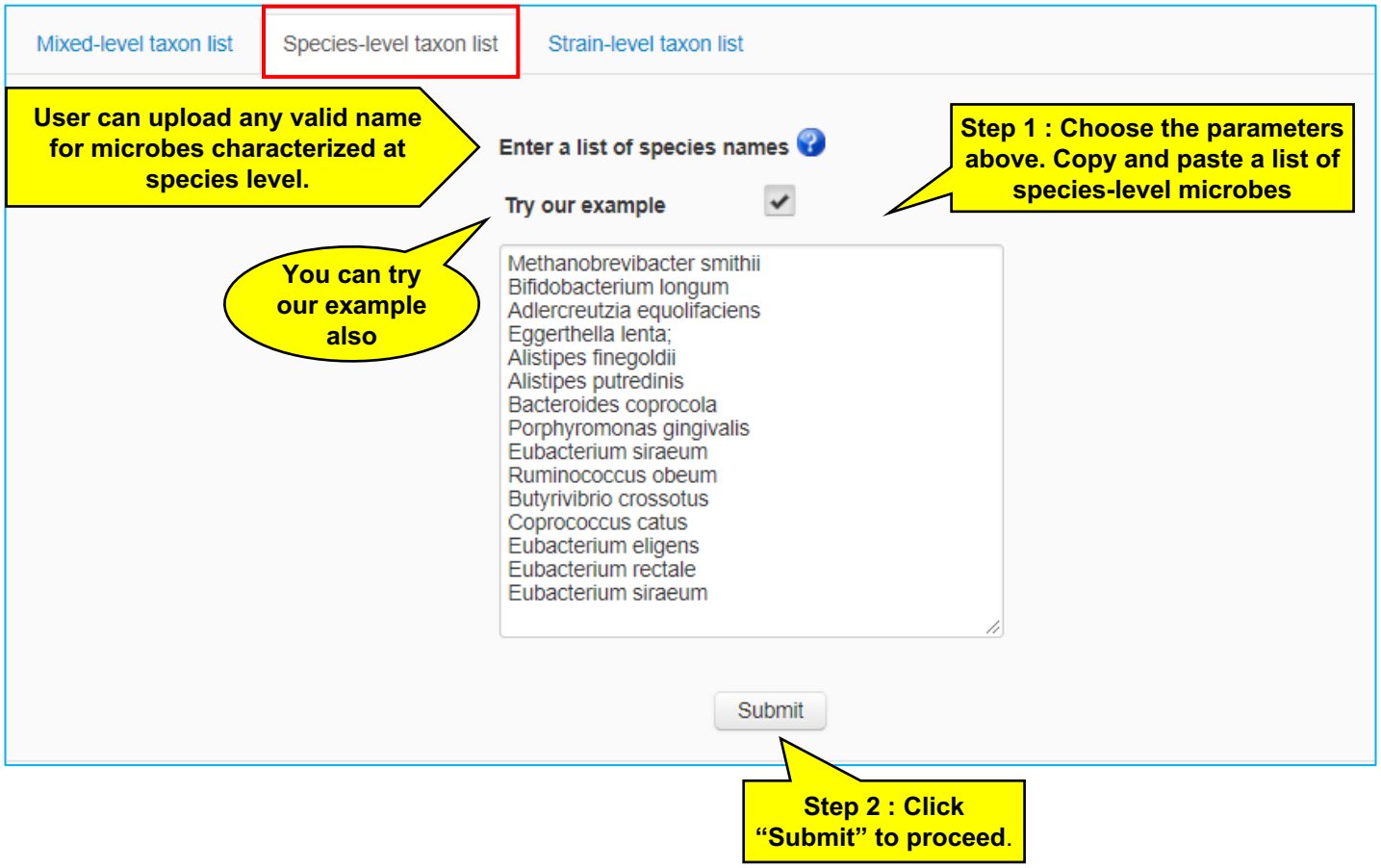

**Network view**

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

# 4. Enrichment Analysis Results



# B. Upload a list of species-level microbes



# 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 | - | - |

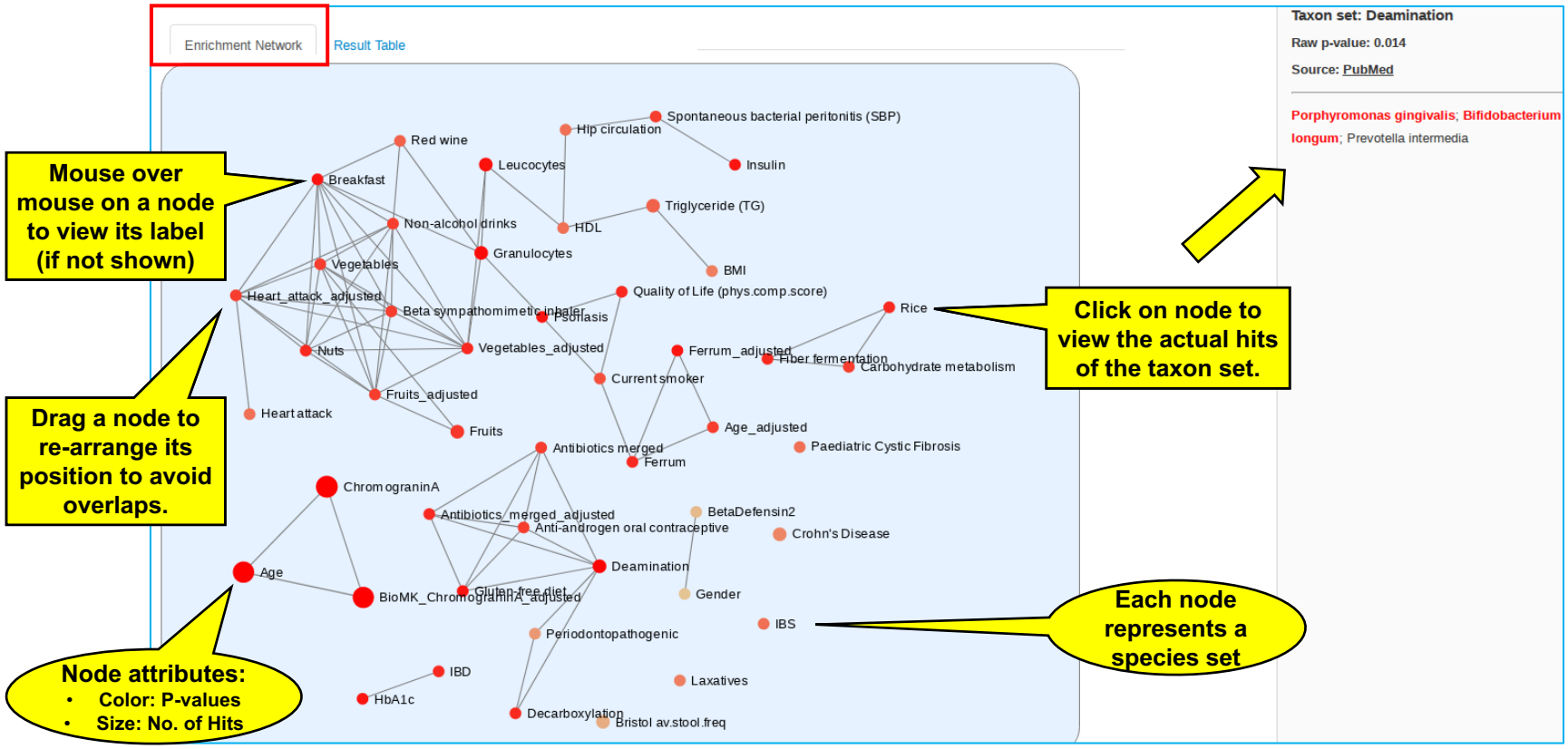
Submit

**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.

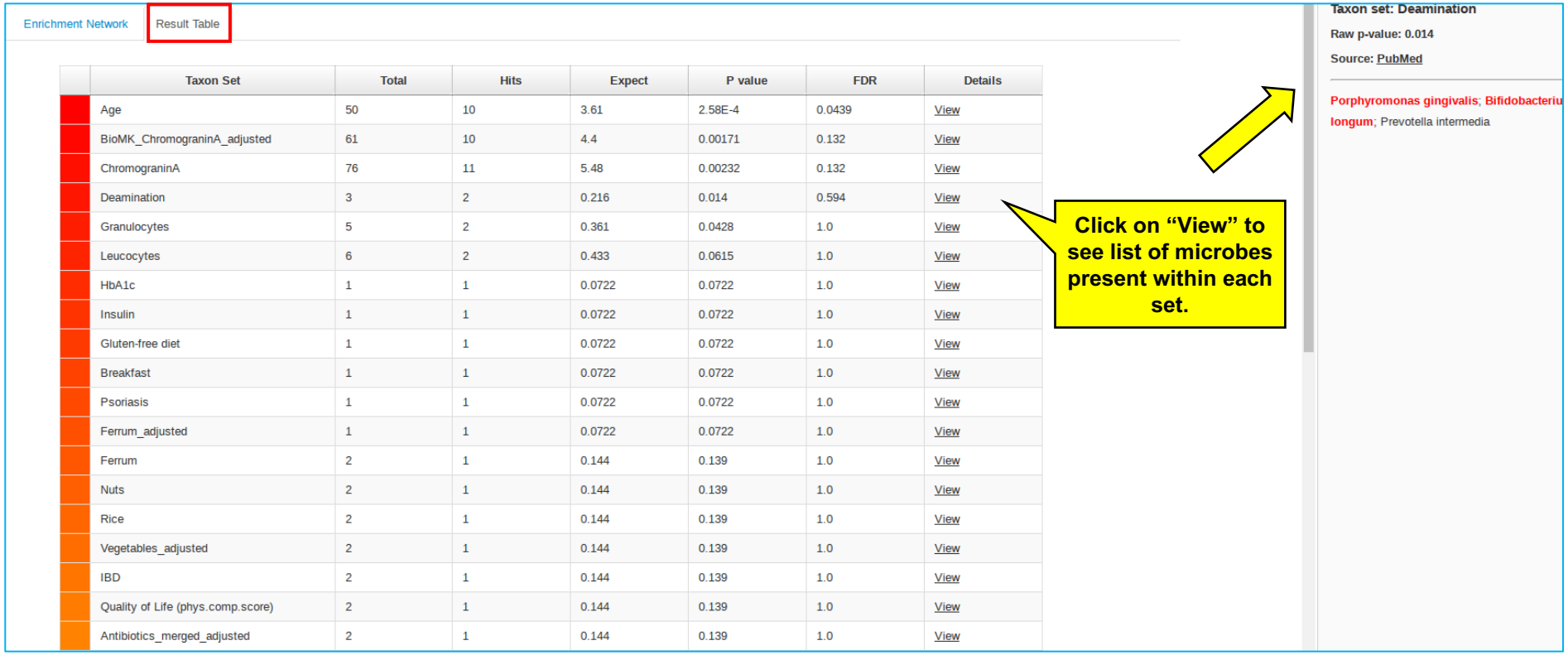
# 3. Enrichment Analysis Summary



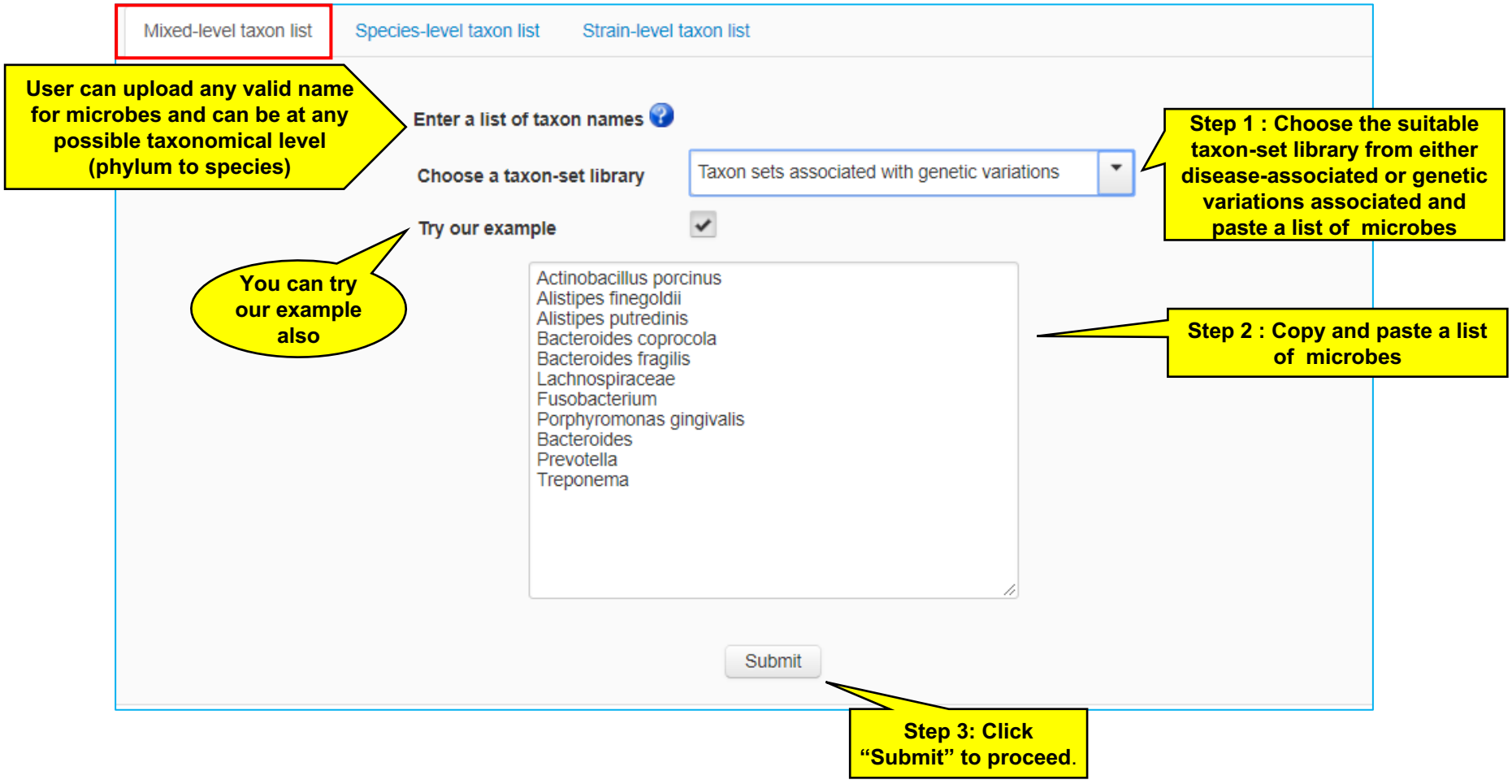

**Network view**

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

# 4. Enrichment Analysis Results



# A. Upload a list of mixed-level of microbes



# 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** | | - | - | - | - |

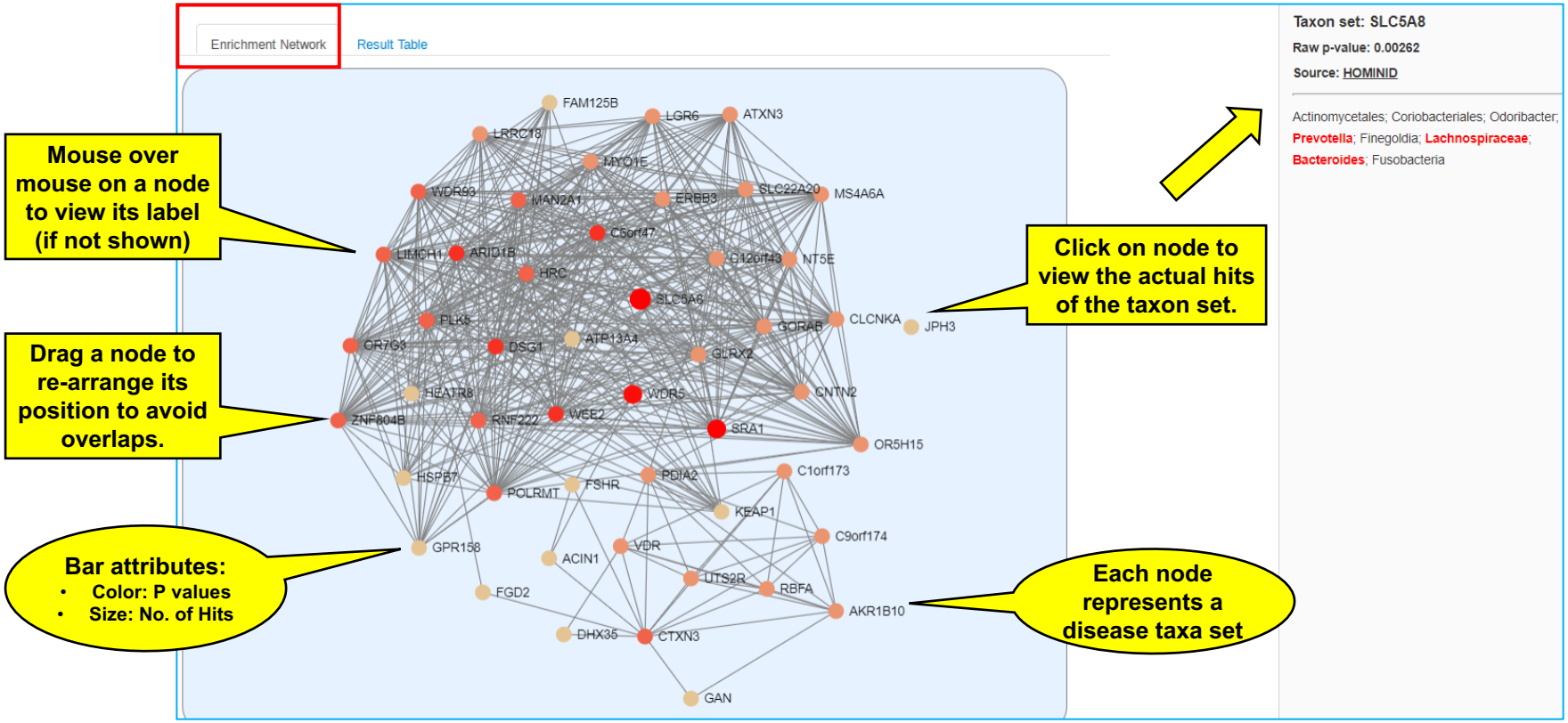
Submit

**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).

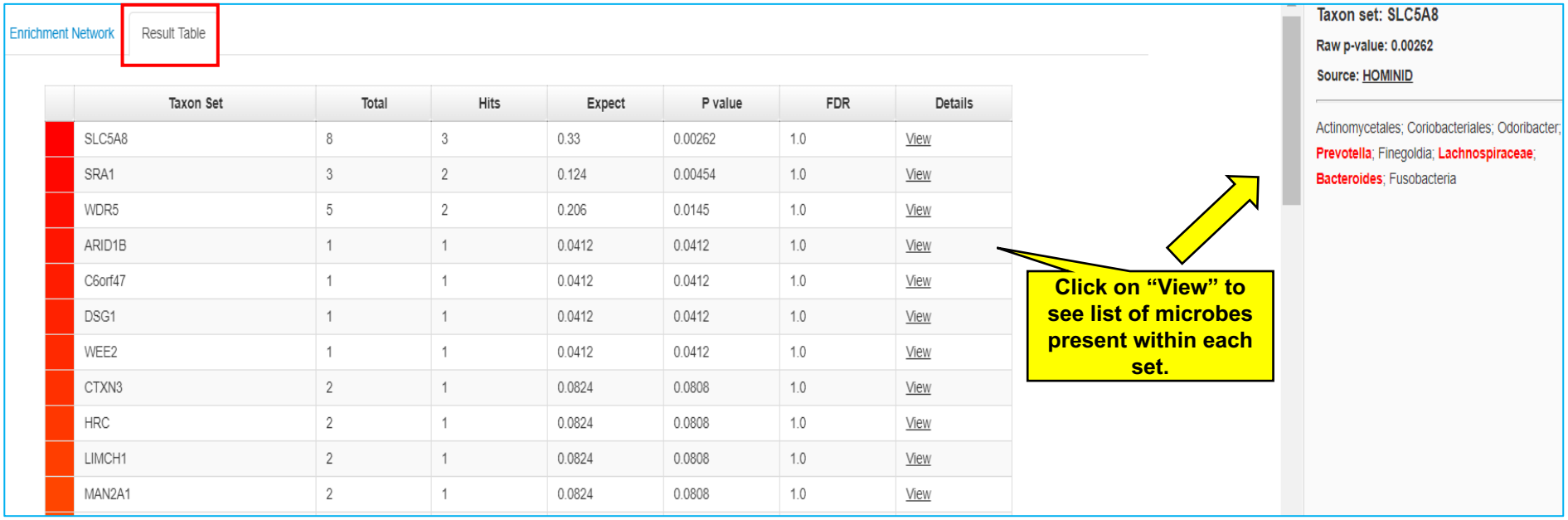
# 3. Enrichment Analysis Summary




**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



=== END ===