# Using web-based tools for integrative analysis of metabolomics and microbiome data for deep functional insights



Jianguo (Jeff) Xia, PhD

Department of Microbiology & Immunology, Institute of Parasitology

McGill University, QC Canada

jeff.xia@mcgill.ca | www.xialab.ca

# Schedule

| Time | Topics | Lab practices |
|---|---|---|
| 12:00 – 12:15 | General introduction | |
| 12:15 – 1:00 | Metabolomics data processing | Live demo 1 & hands on |
| 1:00 – 1:45 | Microbiome data processing | Live demo 2 & hands on |
| 1:45 – 3:00<br>(15 min break) | Microbial community profiling | Live demo 3 & hands on |
| 3:00 – 3:50 | Microbiome-metabolomics integration | Live demo 4 & hands on |
| 3:50 – 4:15 | Summary and discussion | |

# Resources

github.com/xia-lab/Metabolomics_2025

Slides & Materials:

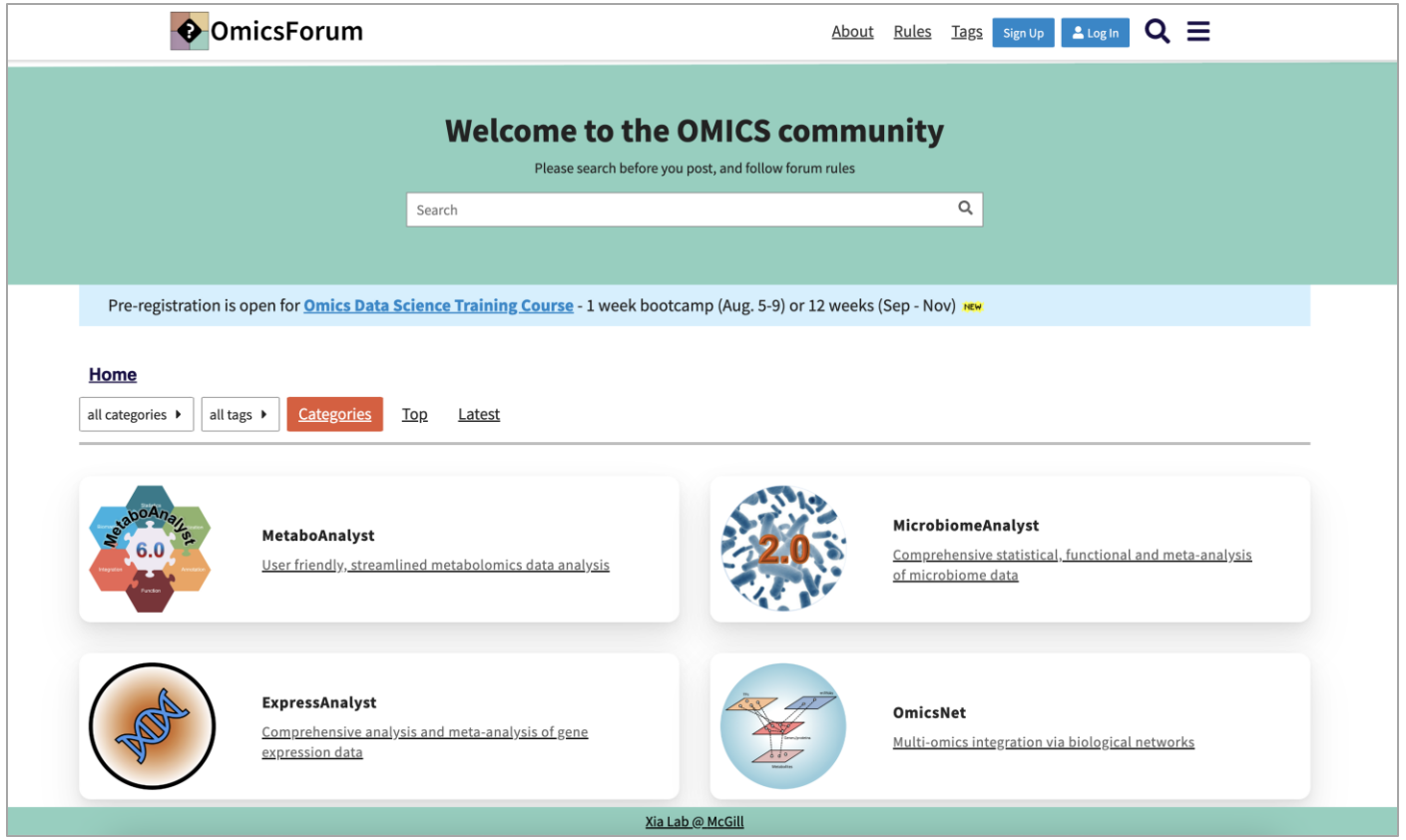
- https://github.com/xia-lab/Metabolomics_2025

Web Sites:

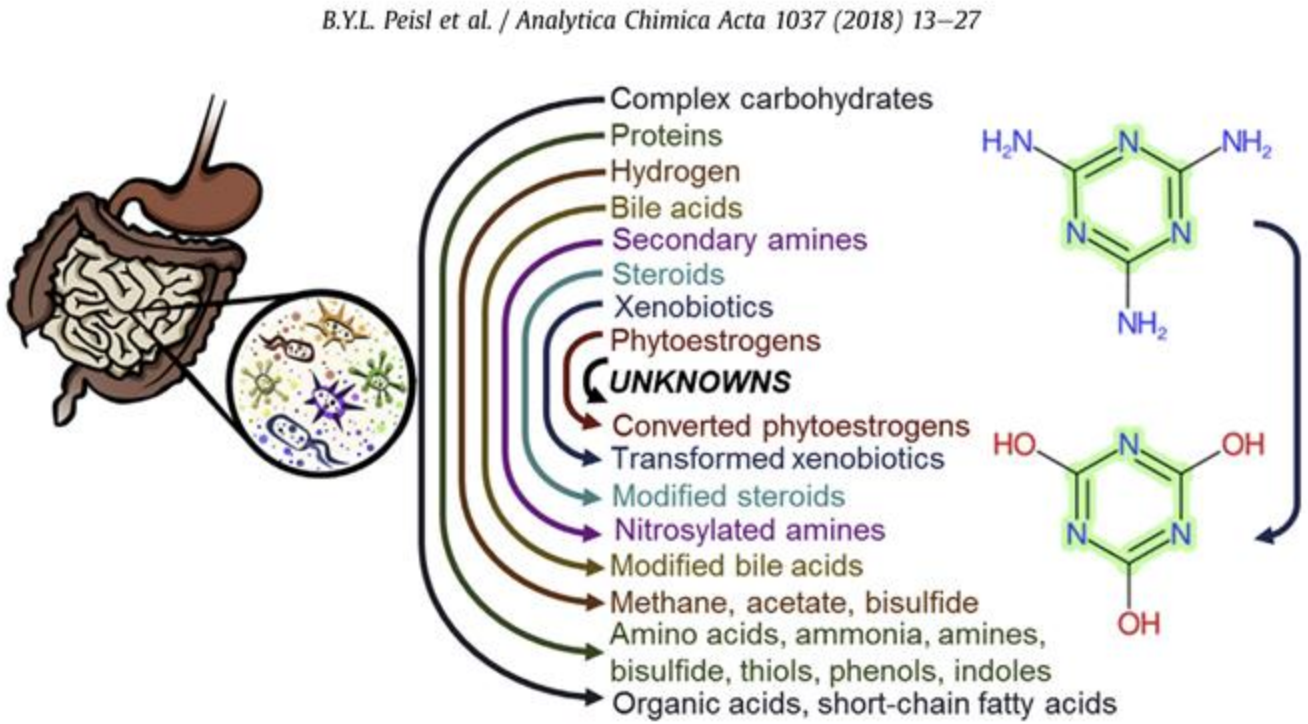
- https://www.metaboanalyst.ca
- https://www.microbiomeanalyst.ca
- https://microbiomenet.com

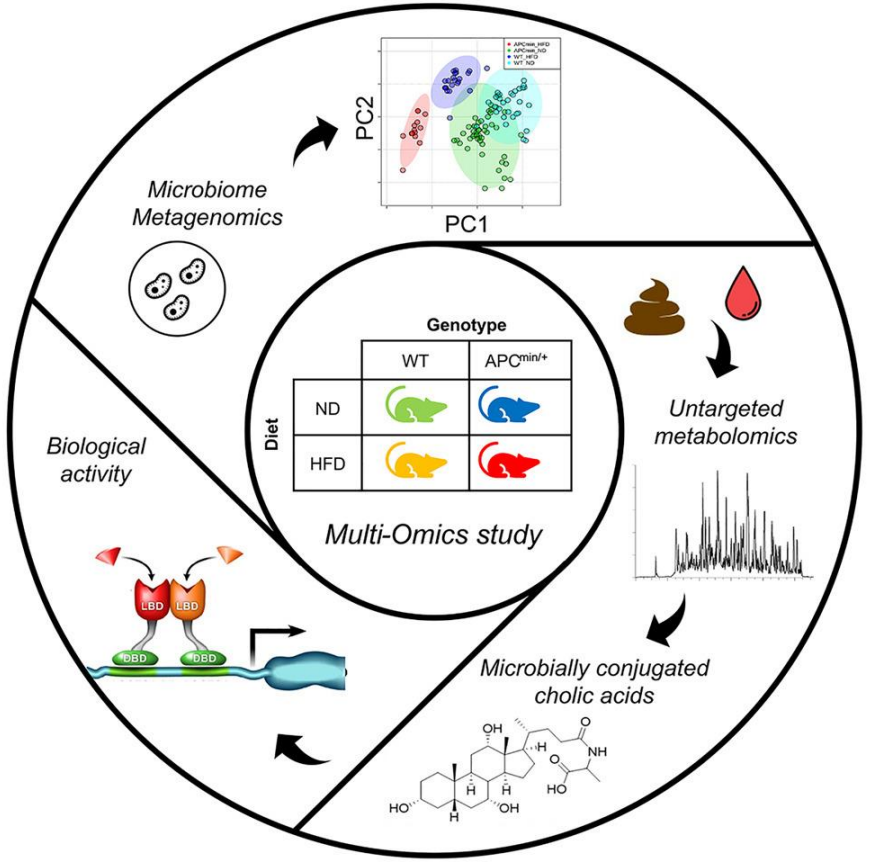
User Forum:

- https://omicsforum.ca



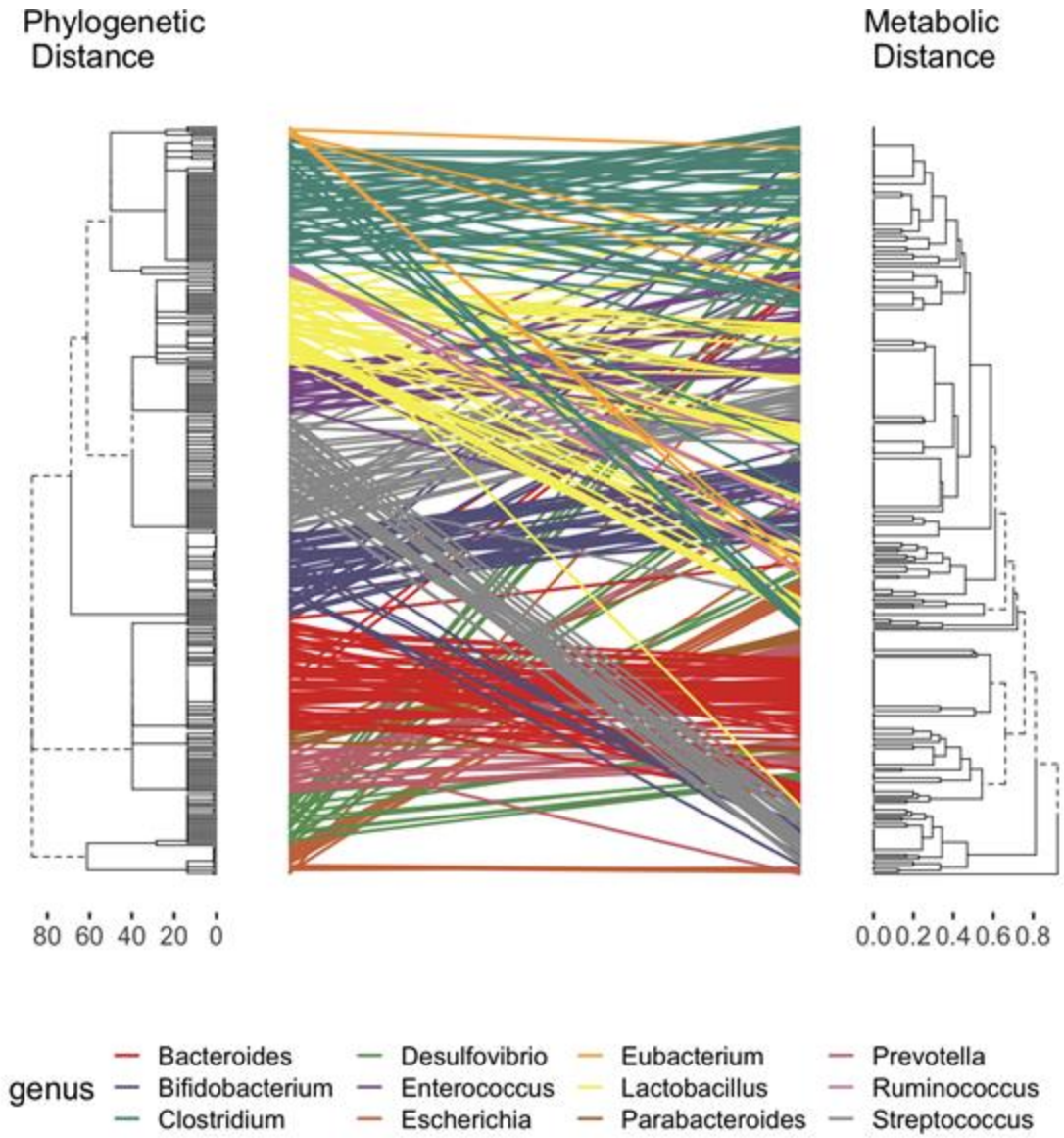
# Metabolomics – functional readout & interactions



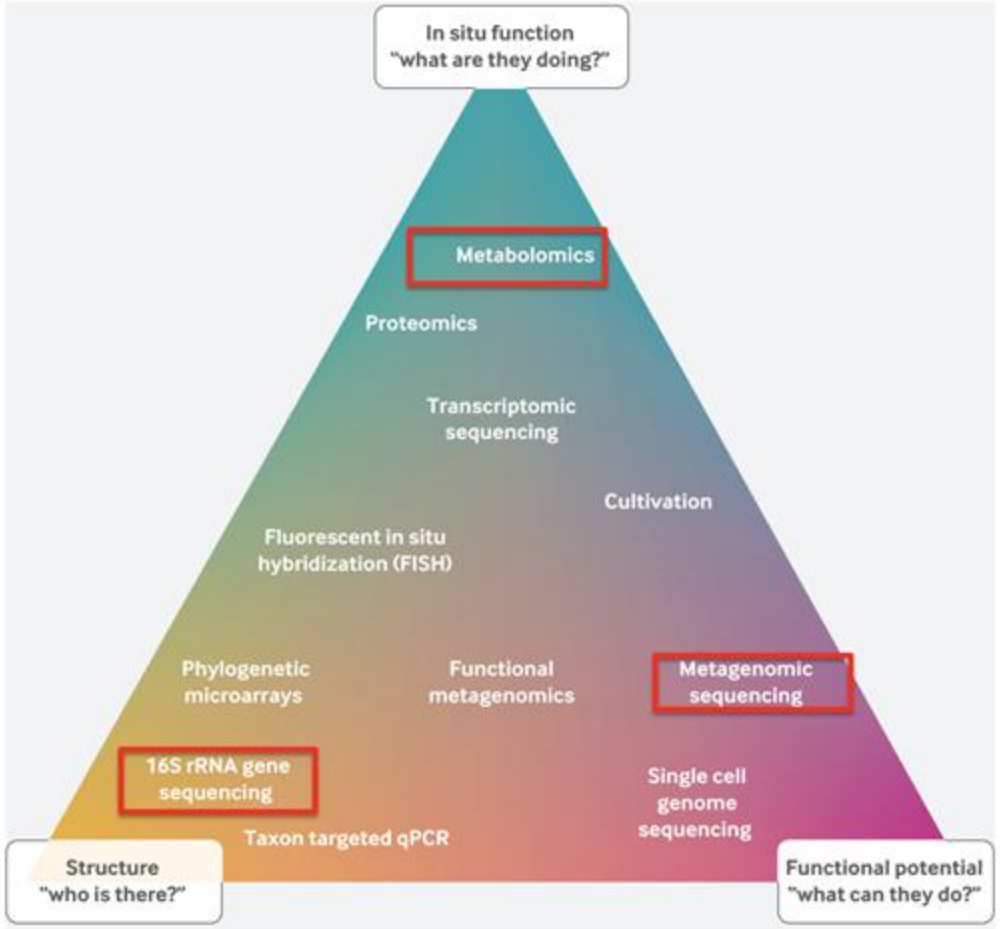

10.1016/j.celrep.2023.112997

# Microbiome composition & functions



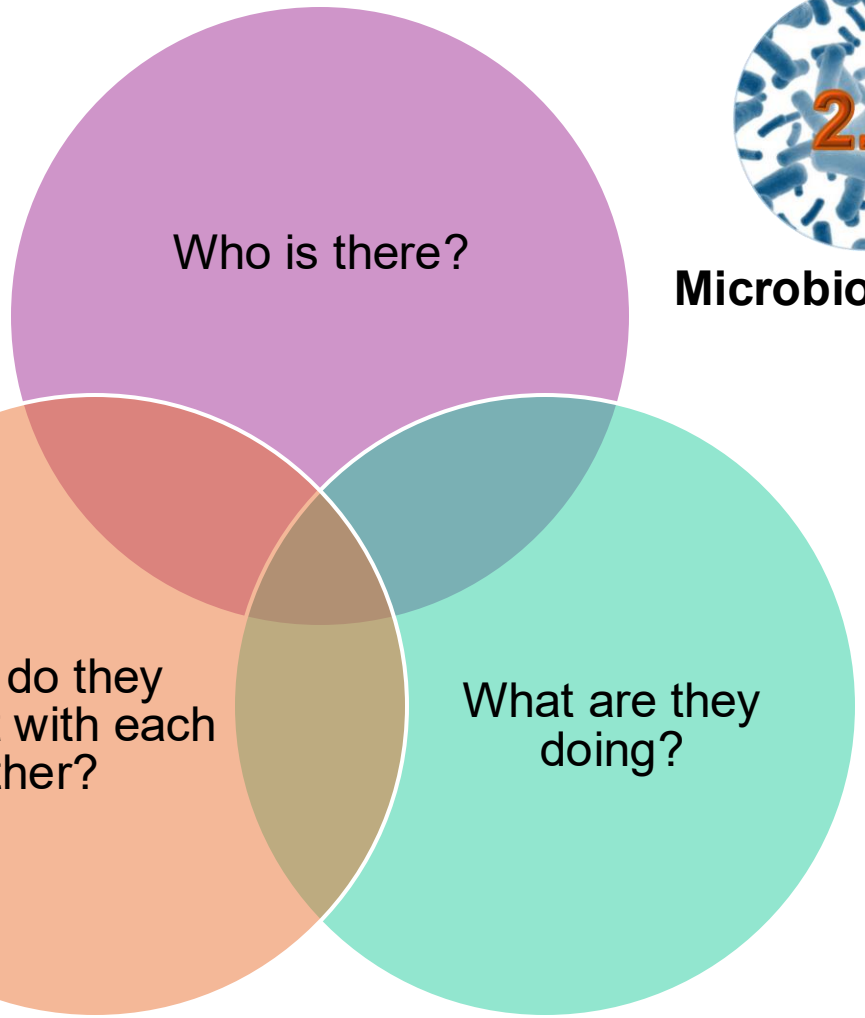
# Microbiomics

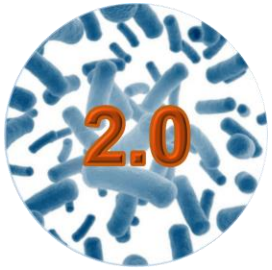
- Marker gene survey
  - Taxonomy
  - Functions (predictive)
- Shotgun sequencing
  - Taxonomy (more accurate)
  - Functions (potential)
- Functions
  - Meta-transcriptomics/proteomics
  - Metabolomics (actual)



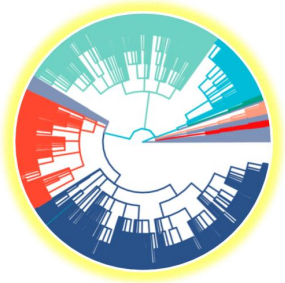

Credit: Lawrence A David

# Overview of our web-based tools




**MicrobiomeAnalyst**

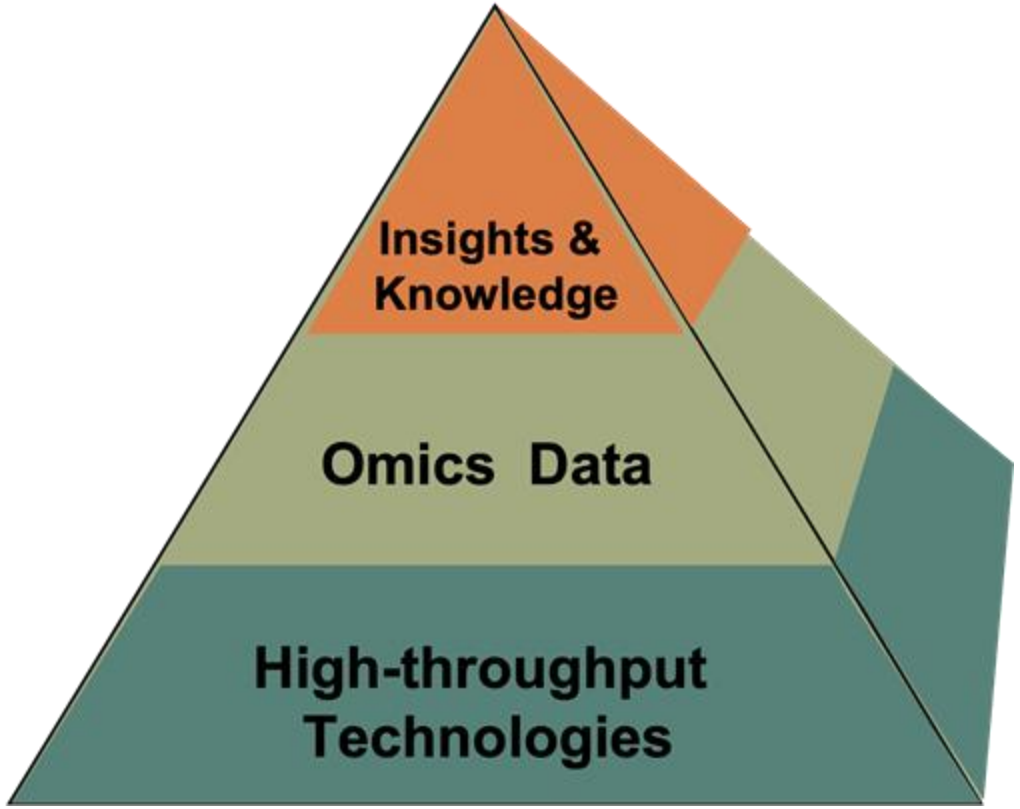


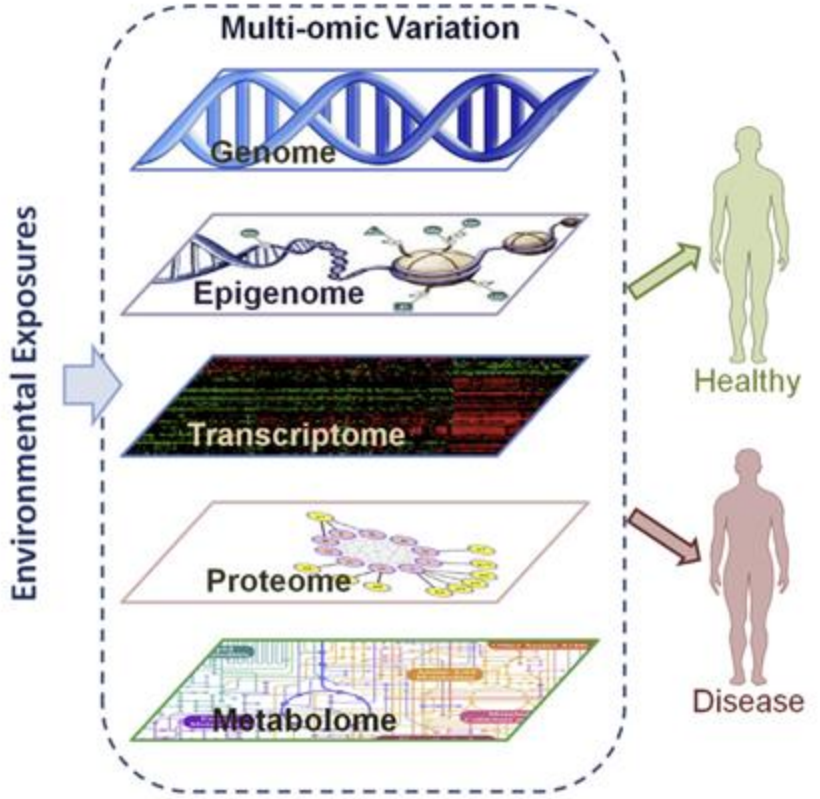
**MicrobiomeNet**


**MetaboAnalyst**

# Intro to Omics Data Science

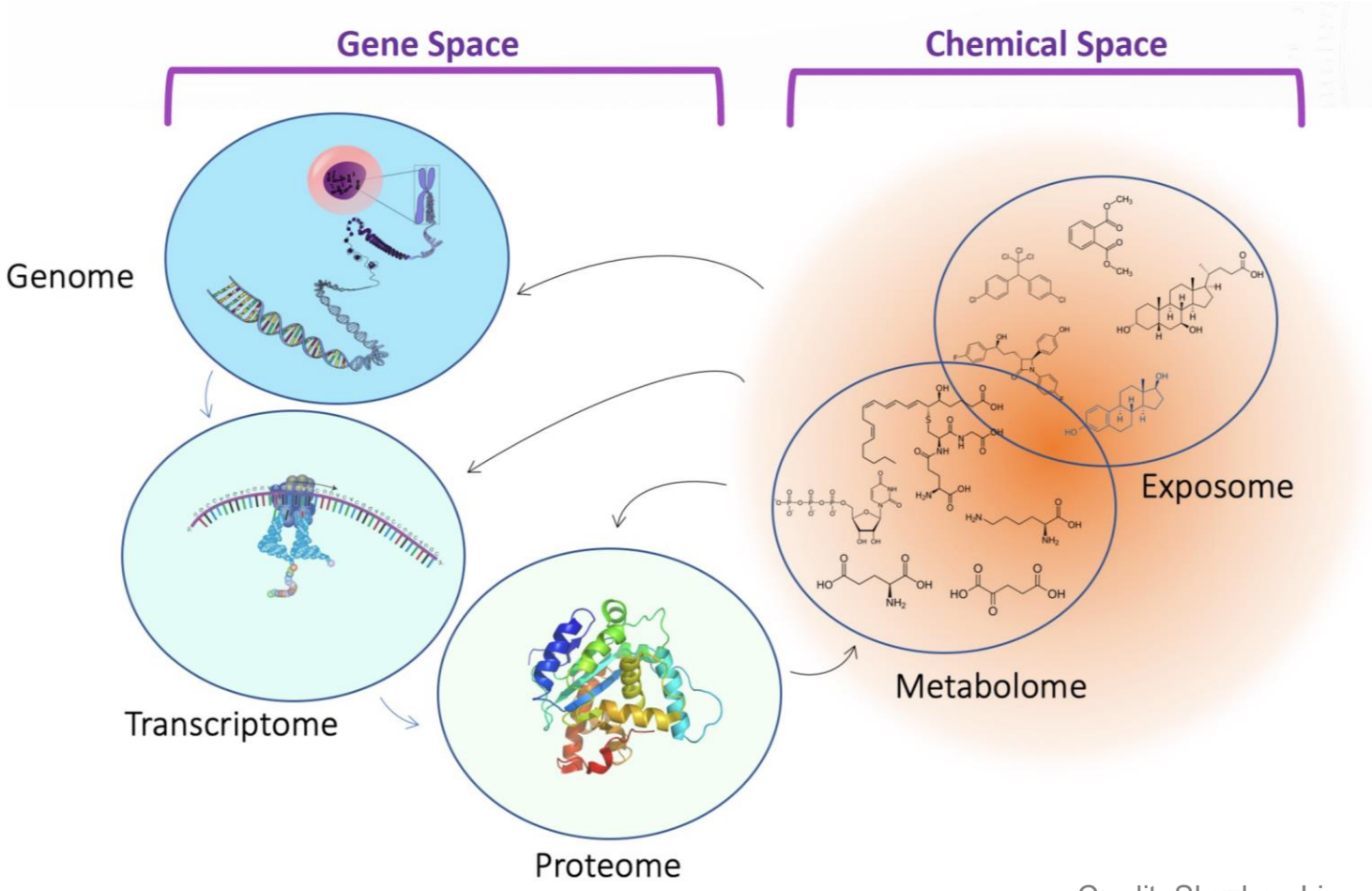
# Omics & multi-omics era



# Two main omics technologies

**NGS Omics**




**MS Omics**

Credit: Shuzhao Li

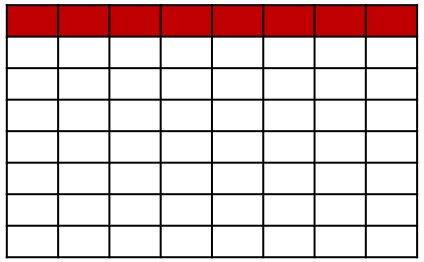
# Two distinct challenges

**Size challenge** (raw data)

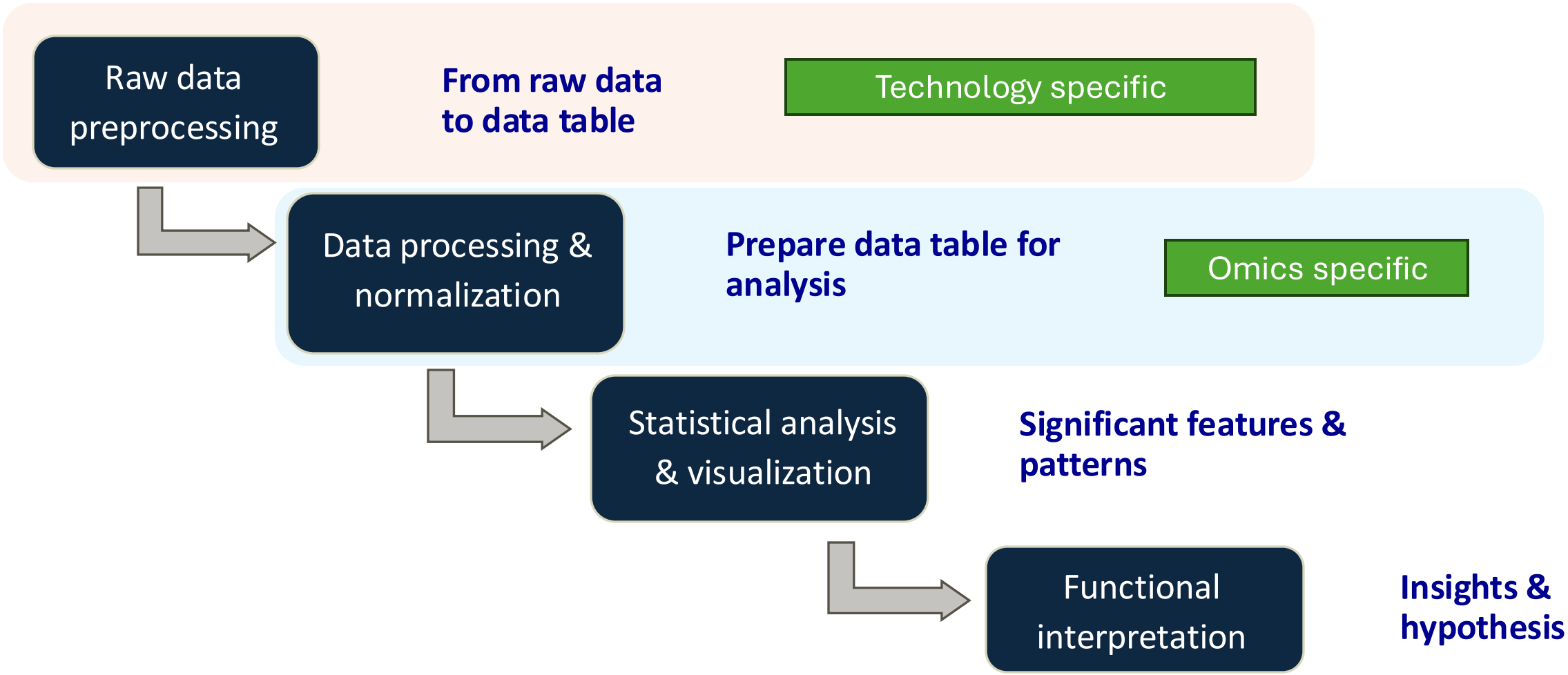
- Raw reads, spectra, images
- Large (GB ~TB)
- Large storage and computing resources

**Complexity challenge** (feature table)

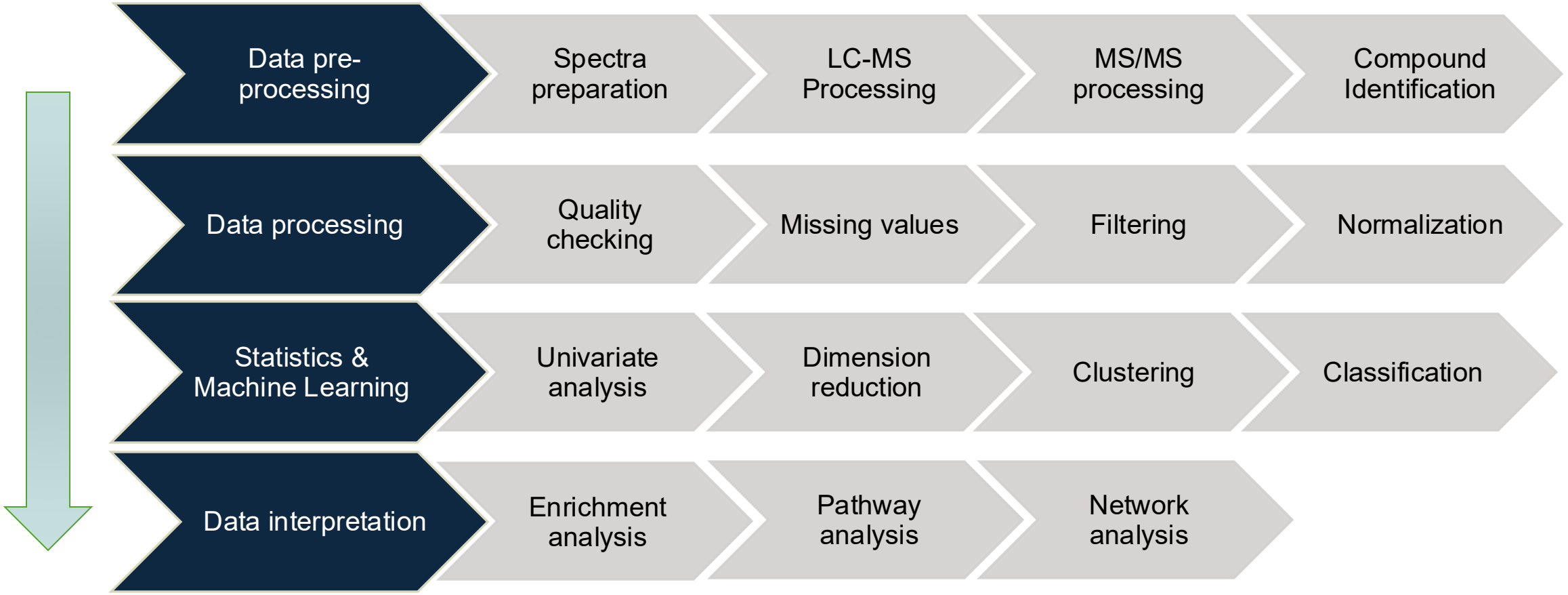
- Feature table (abundance, intensities)
- Small (100s KB ~ MB)
- High-dimensional, missing values
- Data analysis starts here



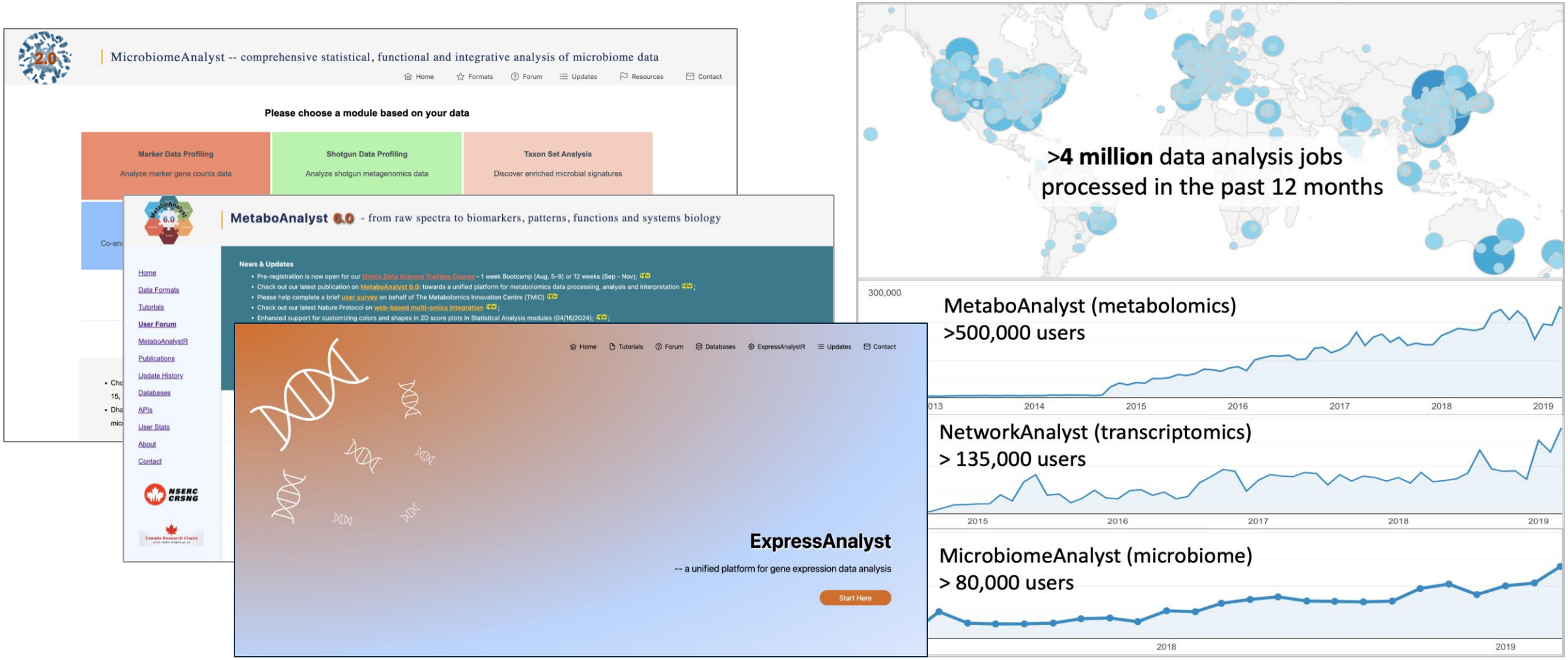
# General workflow for single omics



# Common Tasks (example: metabolomics)

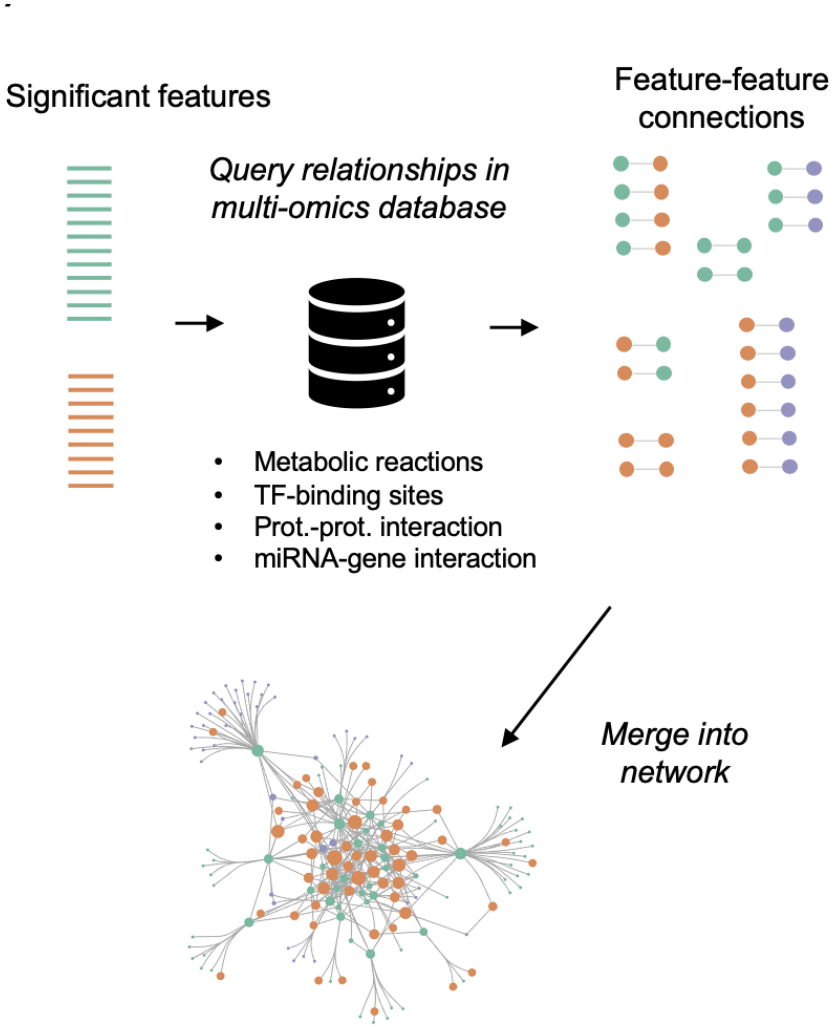


# Towards a unified framework for omics data analysis

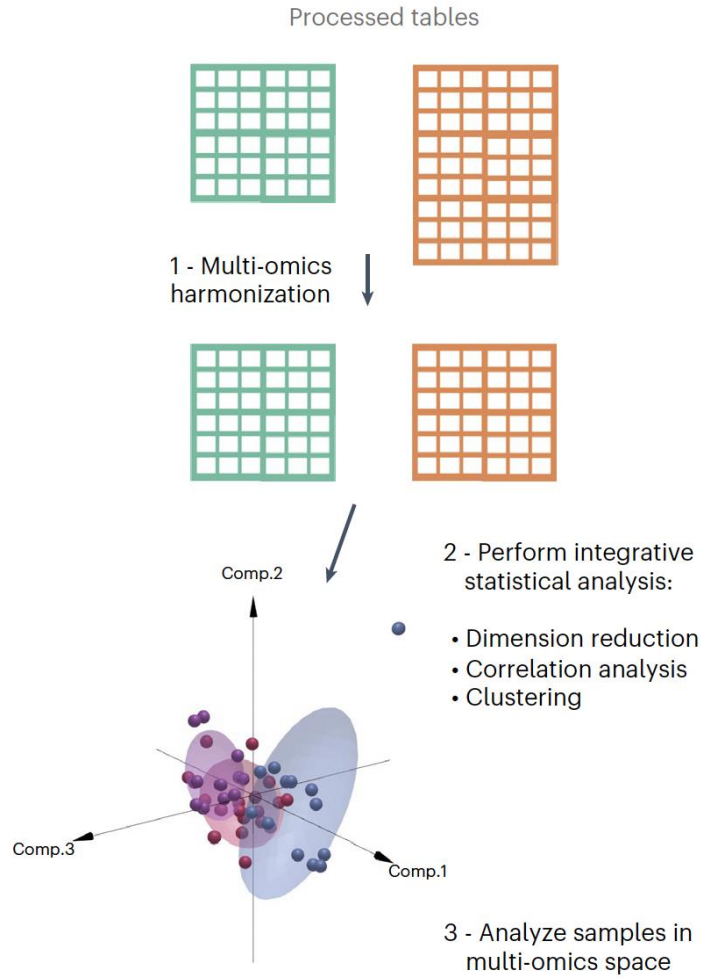



https://www.xialab.ca/tools.xhtml

# General workflow for multi-omics integration




**Qualitative Integration (knowledge-driven)**




**Quantitative Integration (data-driven)**

# Raw data ➔ statistics ➔ patterns ➔ functions

**LC-MS spectra**




MetaboAnalyst

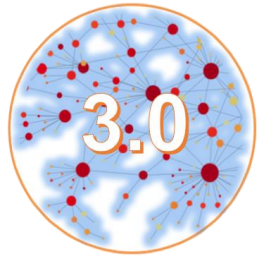

miRNet

**Bulk RNA-seq**

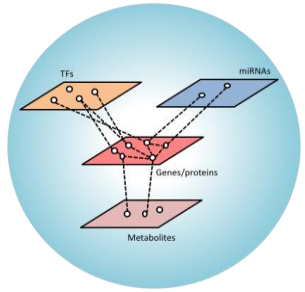


ExpressAnalyst

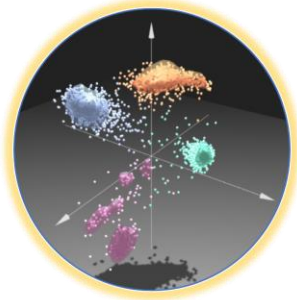





NetworkAnalyst



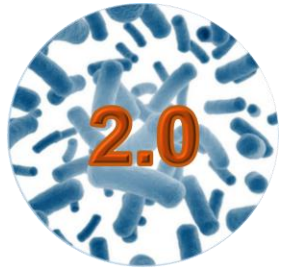

OmicsNet



OmicsAnalyst

**Microbiome**






MicrobiomeAnalyst

# Omics Data Science Training Course

| Topics | Tools | Key Publications |
|---|---|---|
| Transcriptomics | ExpressAnalyst | 1. Nature Communications (2023)<br>2. Current Protocol (2023) |
| Biological Networks | NetworkAnalyst | 3. Methods in Molecular Biology (2022)<br>4. Nucleic Acids Research (2020)<br>5. Nucleic Acids Research (2019) |
| Metabolomics | MetaboAnalyst | |
| Microbiomics | MicrobiomeAnalyst | |
| Multi-omics | OmicsNet, OmicsAnalyst | 12. Nature Protocol (2024)<br>13. Nucleic Acids Research (2022)<br>14. Nucleic Acids Research (2021) |
| Special Topics | Biomarker analysis & meta-analysis | 15. Metabolomics (2013)<br>16. Nature Protocols (2015) |
| | RNAseq processing for non-model species | 17. Genome Research (2021) |
| | LC-MS and MS/MS spectra processing | 18. Nature Communications (2024) |

- Summer Bootcamp (Tokyo time)
  - Aug. 4 - 8, **9:30 - 16:30**
- Regular Session (Montreal time)
  - Saturday morning **9:30 - 12:00**, Sept. - Nov.

**Key Publications**


1. Liu, P., Ewald J., Legrand E., Jeon, Y.S., Sangiovanni, J., Hacariz, O., Pang, Z., Zhou, G., Head, J., Basu, N., and **Xia, J. @** (2023) "A unified platform for RNA-seq analysis in non-model species", Nature Communications 14, 2995.
2. Ewald J., Zhou, G., Lu, Y., and **Xia, J.@** (2023) "Using ExpressAnalyst for Comprehensive Gene Expression Analysis in Model and Non-Model Organisms" Current Protocols 3 (11), e922
3. Chang, L., **Xia, J.@** (2022) "MicroRNA Regulatory Network Analysis Using miRNet 2.0" Chapter Transcription Factor Regulatory Networks, 185-204, Methods in Molecular Biology, Humana Press, New York, NY.
4. Chang, L., Zhou, G., Soufan, O. and **Xia, J.@** (2020) "miRNet 2.0: network-based visual analytics for miRNA functional analysis and systems biology" Nucleic Acids Research (doi: 10.1093/nar/gkaa467)
5. Zhou, G., Soufan, O., Ewald, J., Hancock, REW, Basu, N. and **Xia, J@.** (2019) "NetworkAnalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis" Nucleic Acids Research 47 W234-241

acDonald, P., Wishart, D., Li, S., and Xia, J. @ (2024)
processing, analysis and interpretation" Nucleic Acids

(2022) "Using MetaboAnalyst 5.0 for LC-HRMS spectra
tabolomics data" Nature Protocols (doi: 10.1038/s41596-

for Comprehensive and Integrative Metabolomics Data
36)

oring microbial associations and metabolic profiles for

robiomeAnalyst 2.0: comprehensive statistical, functional
10.1093/nar/gkad407).

mprehensive statistical, functional and meta-analysis of

and **Xia, J. @** (2024) "Web-based Multi-omics Integration
Using the Analyst Software Suite", Nature Protocols (doi: 10.1038/s41596-023-00950-4

13. Zhou, G., Pang Z., Lu, Y., Ewald, J., **Xia, J.@** (2022) "OmicsNet 2.0: a web-based platform for multi-omics integration and network visual analytics" Nucleic Acids Research (doi:10.1093/nar/gkac376).
14. Zhou, G., Ewald, J., and **Xia, J.@** (2021) "OmicsAnalyst: a comprehensive web-based platform for visual analytics of multi-omics data" Nucleic Acids Research (10.1093/nar/gkab394)
15. **Xia, J.,** Broadhurst, D., Wilson, M. and Wishart, D. (2013) "Translational biomarker discovery in clinical metabolomics: an introductory tutorial". Metabolomics (doi: 10.1007/s11306-012-0482-9)
16. **Xia, J@.,** Gill, E., and Hancock, R.E.W@. (2015) "NetworkAnalyst for Statistical, Visual and Network-based Approaches for Meta-analysis of Expression Data" Nature Protocols 10 (6), 823-844
17. Liu, P., Ewald, J., Galvez, J., Head, J., Crump, D., Basu, N. and **Xia, J.@** (2021) "Ultrafast functional profiling of RNA-seq data for nonmodel organisms" Genome Research (doi: 10.1101/gr.269894.120)
18. Pang, Z., Xu, L., Viau, C., Lu Y., Salavati, R., Basu, N., and **Xia, J.@** (2024) "MetaboAnalystR 4.0: a unified LC-MS workflow for global metabolomics" Nature Communications (doi:10.1038/s41467-024-48009-6)


https://www.xialab.ca/translation.xhtml

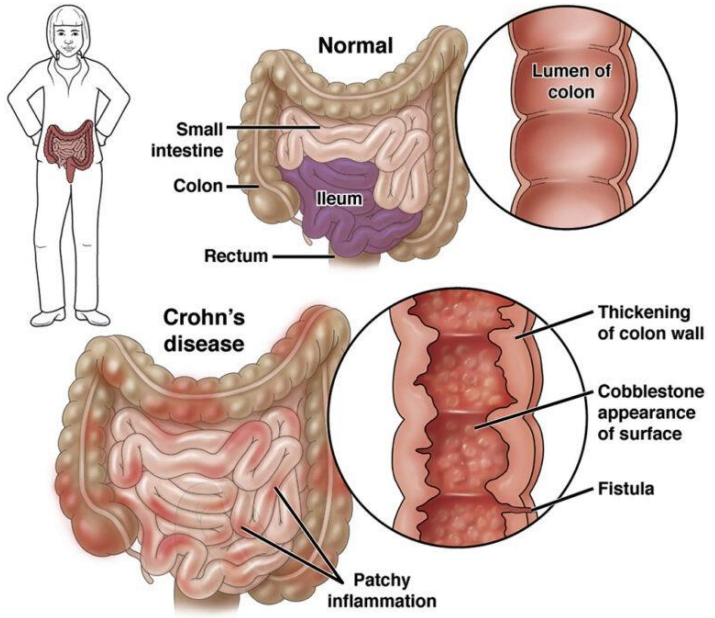
# Tips & recommendations

- **Do not open multiple tabs**
  - Results could overwrite each other!
- (Optional) form a group and share computers
  - Make new friends
  - Help each other & reduce stress
- Page display may be **slow** due to bandwidth limitation
  - Be patient
  - Reduce bandwidth consumption by forming group

# IBD study background

Inflammatory bowel disease (IBD), which includes both Crohn's Disease (CD) and ulcerative colitis (UC), affect several million individuals worldwide and is one of the most-studied imbalances between microbes and the immune system. Achieving a systems-level understanding of the etiology of IBD, particularly the microbiome-driven metabolic consequences, is critical for advancing diagnosis, treatment, and prevention strategies.

- Subjects and samples included:
  - 35 samples (21 CD and 14 non-IBD control);
  - Metabolomics raw spectra data;
  - Microbiome raw sequencing data;
  - Meta-data table, containing diagnosis, age, gender, etc.




References:

https://ibdmdb.org/ (main data resource, part of the Human Microbiome Project)

https://www.nature.com/articles/s41586-019-1237-9

https://genegut.eu/crohns-disease/

# Omics Technology

- Metabolomics:
  - C18 Negative;
  - LC-MS systems comprised of Nexera X2 U-HPLC systems coupled to Q Exactive Plus orbitrap mass spectrometers.
  - LC-MS1 only.
- Microbiome sequencing:
  - 16S rDNA V4 region
  - MiSeq platform (Illumina) 2x250 bp paired-end

The detail protocols can are available at : https://ibdmdb.org/protocols

# Data analysis strategies

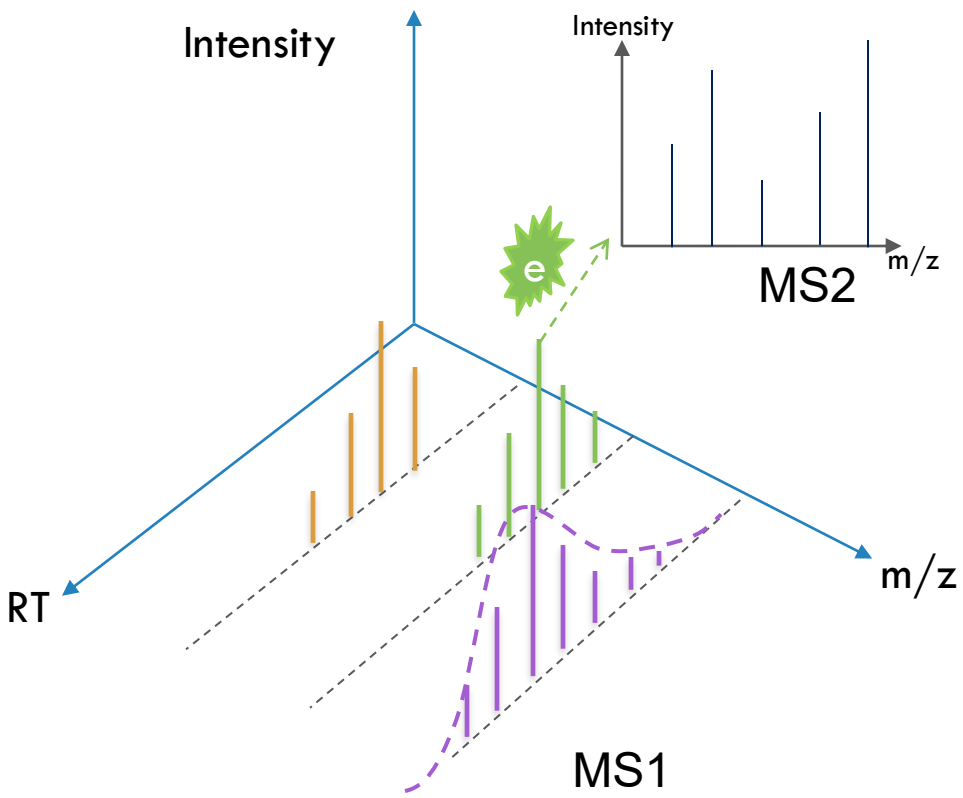
- Metabolomics:
  - Using MetaboAnalyst 6.0 online website – Spectra Processing Module;
  - Explore *Asari* and *centWave-auto* for peak profiling and annotation;
  - Results visualization and downstream analysis – statistics.
- Microbiome:

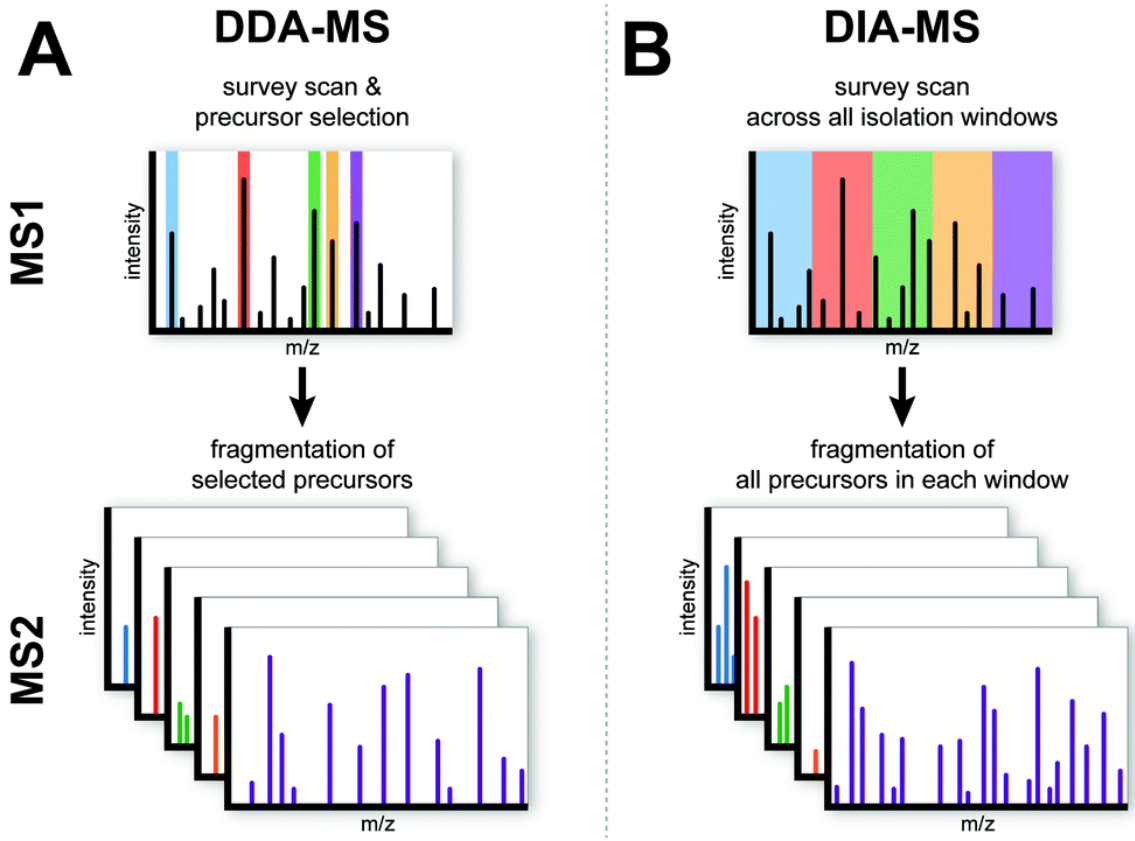
  Using MicrobiomeAnalyst 2.0
  - Raw Data Processing Module for taxonomy annotation and abundance table generation
  - Marker Data Profiling Module for statistical analysis
- Multi-Omics:
  - Using MicrobiomeAnalyst 2.0 for integration analysisi and MicrobiomeNet for interaction exploration

# Schedule

| Time | Topics | Lab practices |
|---|---|---|
| 12:00 – 12:15 | General introduction | |
| 12:15 – 1:00 | Metabolomics data processing | Live demo 1 & hands on |
| 1:00 – 1:45 | Microbiome data processing | Live demo 2 & hands on |
| 1:45 – 3:00<br>(15 min break) | Microbial community profiling | Live demo 3 & hands on |
| 3:00 – 3:50 | Microbiome-metabolomics integration | Live demo 4 & hands on |
| 3:50 – 4:15 | Summary and discussion | |

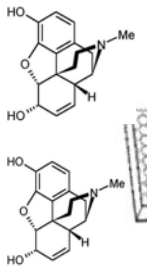
# MS-omics technologies: LC-MS & MS/MS

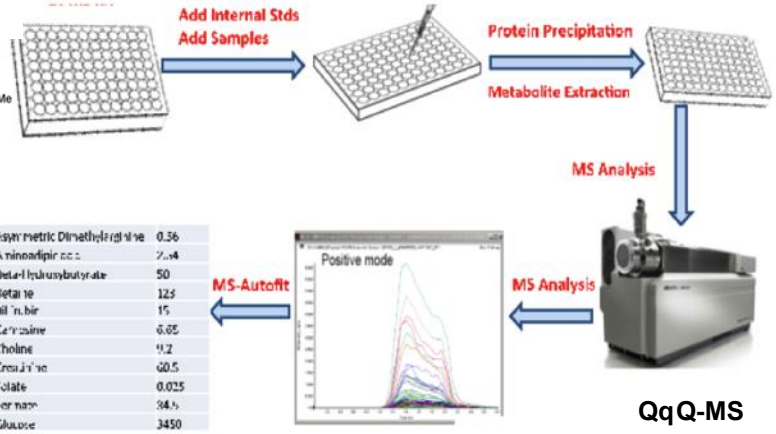


# Two routes to metabolomics (LC-MS)

## Quantitative / Targeted



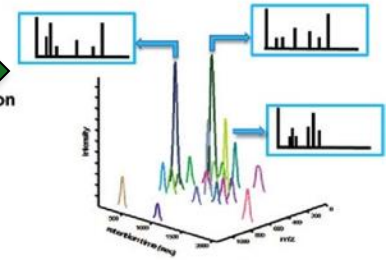



## Global / Untargeted

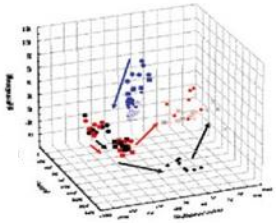
LC-HRMS

Data acquisition

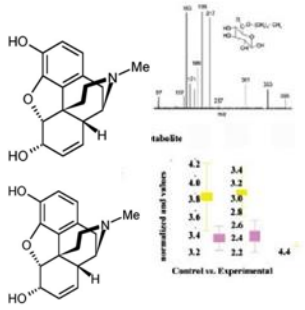
Clustering, Peak detection



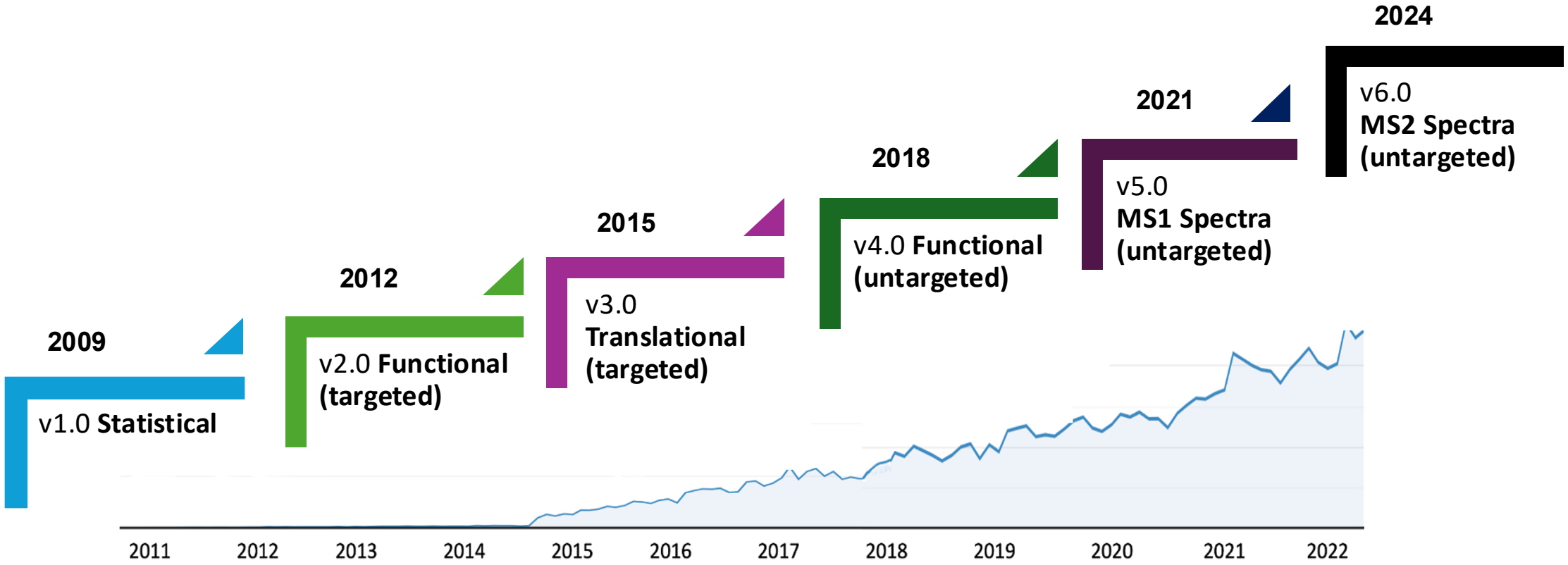
Feature Selection



Metabolite Annotation



# MetaboAnalyst roadmap



# MetaboAnalyst 6.0 Modules

| **Input Data Type** | **Available Modules** (click on a module to proceed, or scroll down to explore a total of 18 modules including **utilities**) | | | | |
|---|---|---|---|---|---|
| LC-MS Spectra<br>(mzML, mzXML or mzData) | | | Spectra Processing<br>[LC-MS w/wo MS2] | | |
| MS Peaks<br>(peak list or intensity table) | | Peak Annotation<br>[MS2-DDA/DIA] | Functional Analysis<br>[LC-MS] | Functional Meta-analysis<br>[LC-MS] | |
| Generic Format<br>(.csv or .txt table files) | Statistical Analysis<br>[one factor] | Statistical Analysis<br>[metadata table] | Biomarker Analysis | Statistical Meta-analysis | Dose Response Analysis |
| Annotated Features<br>(metabolite list or table) | | Enrichment Analysis | Pathway Analysis | Network Analysis | |
| Link to Genomics & Phenotypes<br>(metabolite list) | | | Causal Analysis<br>[Mendelian randomization] | | |

# Streamlined data processing, analysis & interpretation

- **Workflow** should be standardized (SOPs)
- **Parameters** should be minimal & optimized (data specific, not “one size fit all”)
- **Speed** should be fast enough to enable high-throughput analysis

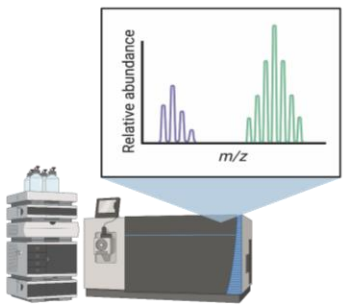
# Study design for LC-MS global metabolomics

## Quantitative analysis (LC-MS)

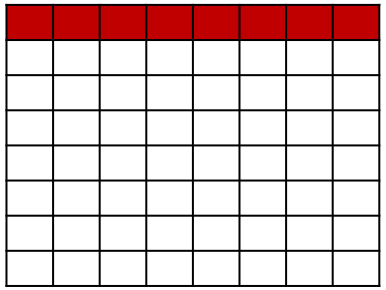
- Biological replicates
  - 10 control vs 10 disease
- Pooled QCs
- Blanks

## Compound identification (MS/MS)

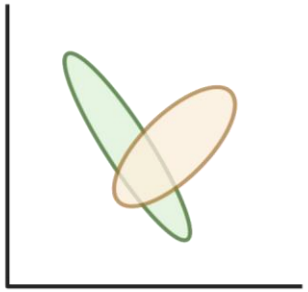
- 3 technical replicates from pooled QCs
  - Aliquots from all samples (better signals & coverage)
  - Spectra consensus to improve MS2 quality



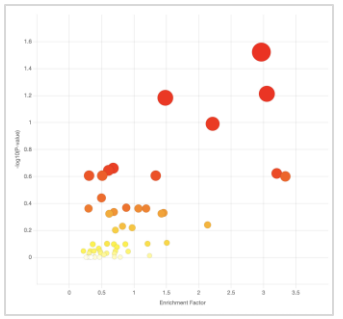

Raw data processing



Statistical analysis



Functional analysis



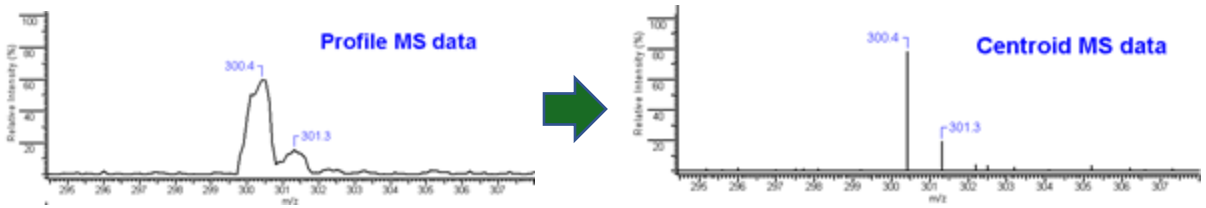
# LC-MS spectra processing

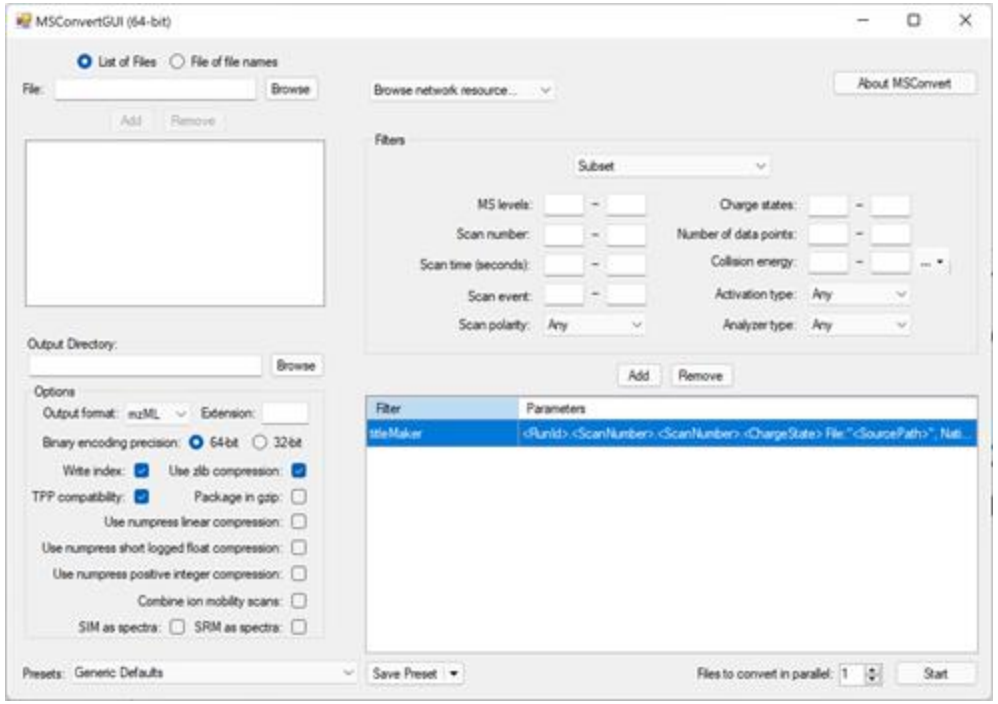
proteowizard → LC-MS1
- MetaboAnalystR
- Asari

→ MS2 - MetaboAnalystR
- DDA
- SWATH-DIA

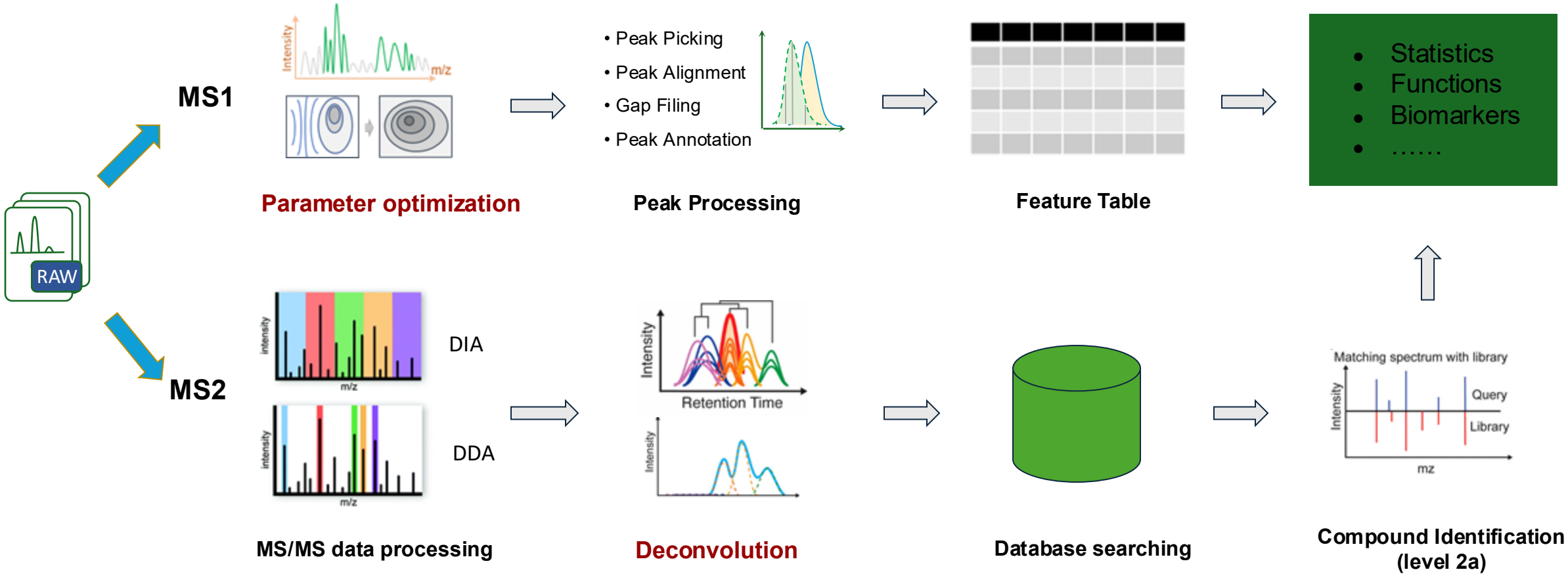
→ MetaboAnalystR
- Mummichog
- GSEA

- The vendor raw spectra data is usually in profile format, which is redundant for regular LC-MS based metabolomics analysis;
- We need to convert the MS data into centroid mode to condense the profile peaks into centroids.
- Open formats (.mzML, etc.)

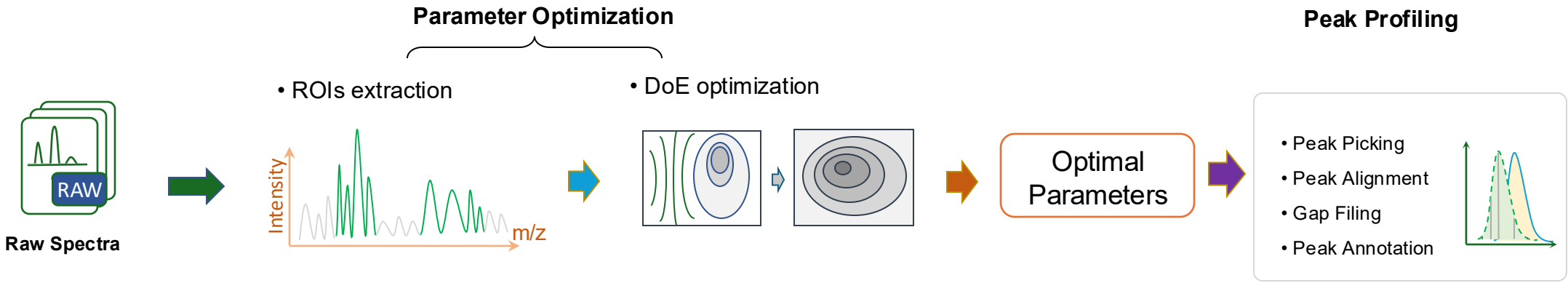


# Under the Hood



# Auto-optimization Workflow




Relative reliable peaks ratio (identified by their isotopes)

A co-efficient describing the stability of a grouped feature

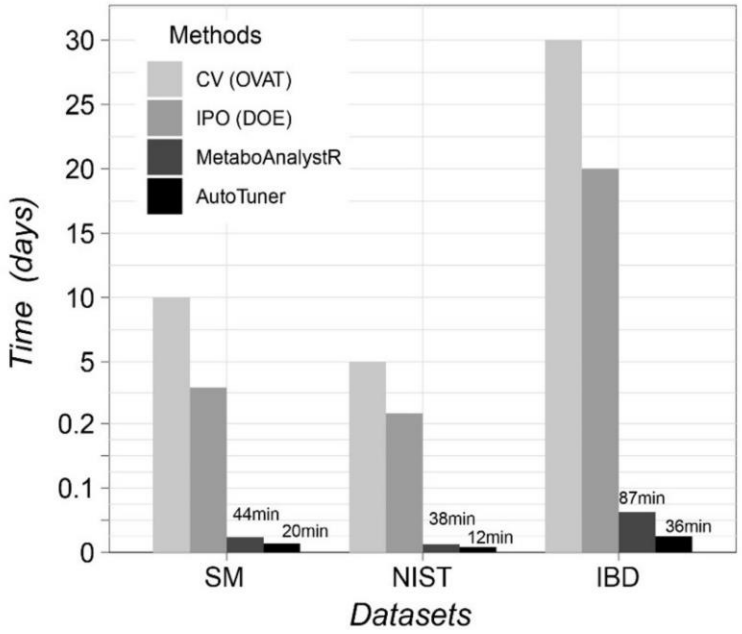
$$QS = \frac{RP^{3/2}}{'all\ peaks' - LIP} * GR^2 * QcoE$$

Gaussian peaks ratio

a “gentle balance” of multiple criteria

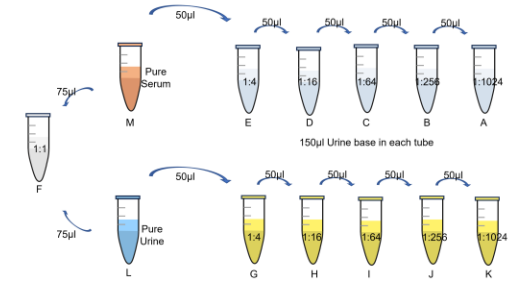
# Benchmarking: better identification

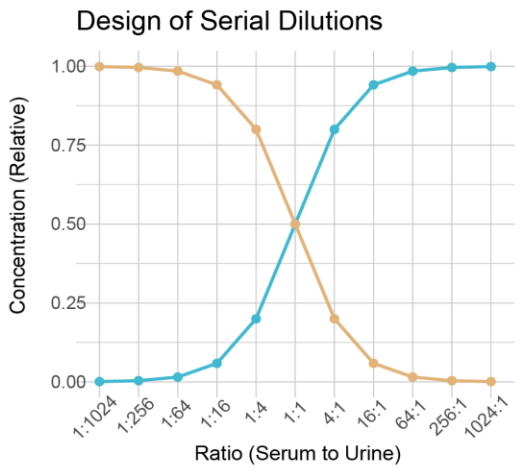
Benchmark with a standard mixture of 838 standards

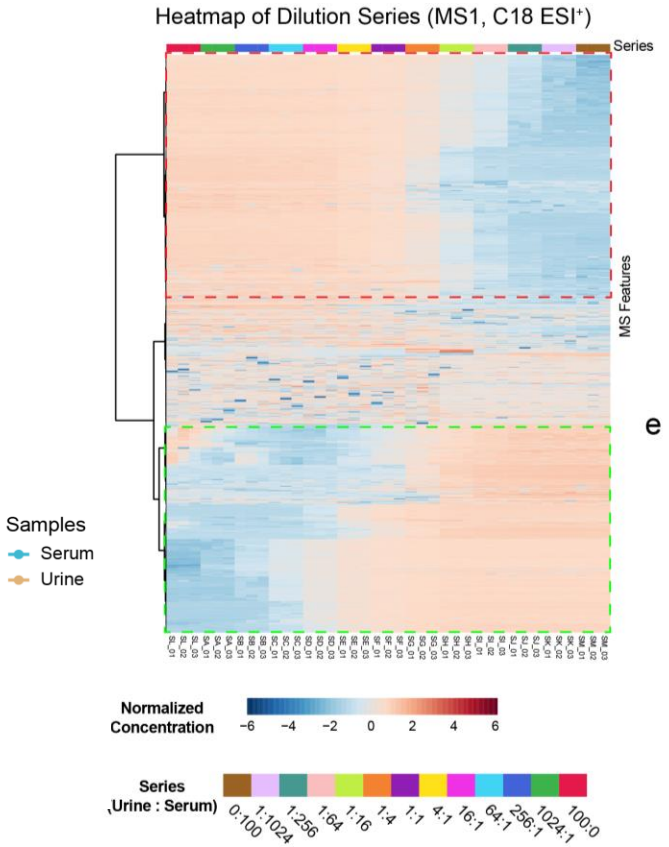


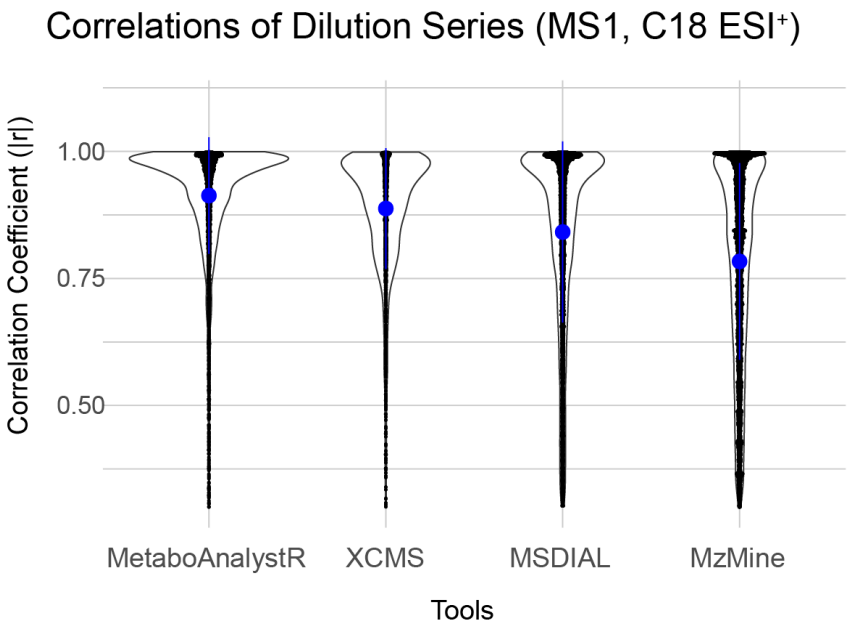

| Methods | Total Peaks | True Peaks | Quantified | Gaussian Peak Ratio |
|---|---|---|---|---|
| Default *centWave* | 16,896 | 382 | 350 | 47.8% |
| IPO | 24,346 | 744 | 663 | 52.0% |
| AutoTuner | 25,517 | 664 | 603 | 40.5% |
| **MetaboAnalystR** | **18,044** | **799** | **754** | **64.4%** |

# Benchmarking: better quantification




Default Settings

# Benchmarking: more biological variance

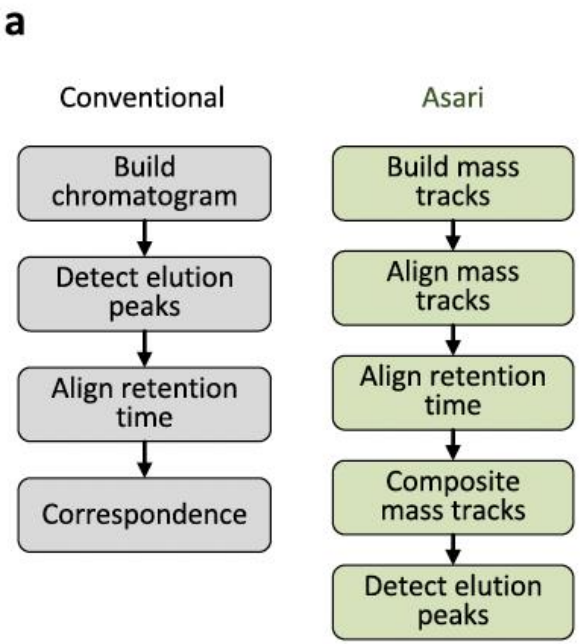
## NIST data

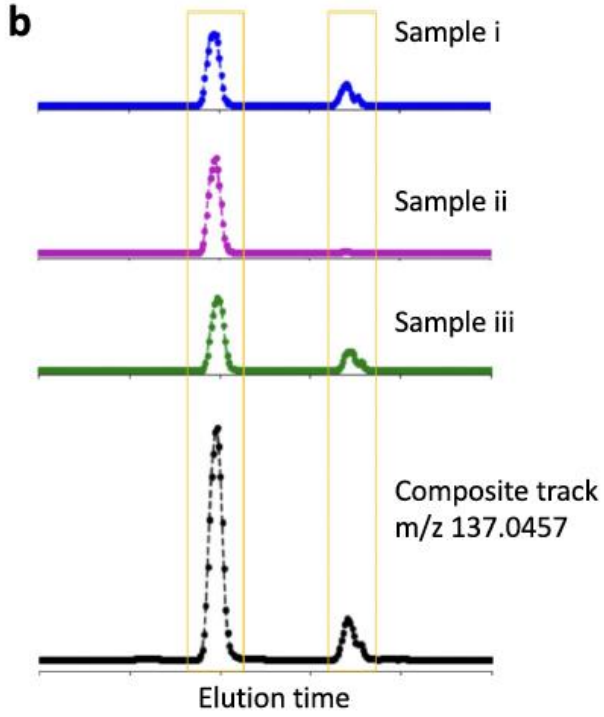
| | Default | Optimized |
|---|---|---|
| **Total peaks** | 2,492 | 2,423 |
| **Isotopes / Adducts** | 667 (26.8%) | 1,112 (45.9%) |
| **Formula Assigned** | 663 | 762 |
| **Potential compounds** | 1,085 | 1,692 |
| **Variance (PC1 + PC2)** | 37% | 50% |
| **Significant peaks** | 855 | 1,091 |

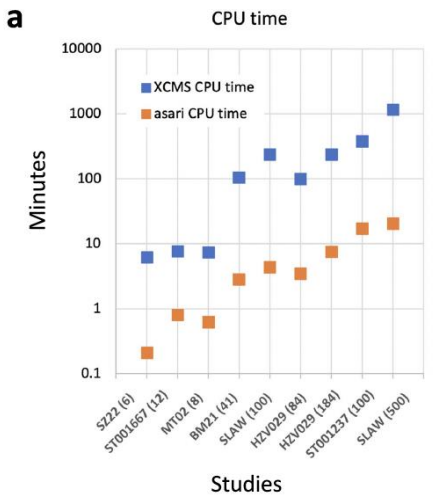
## IBD data

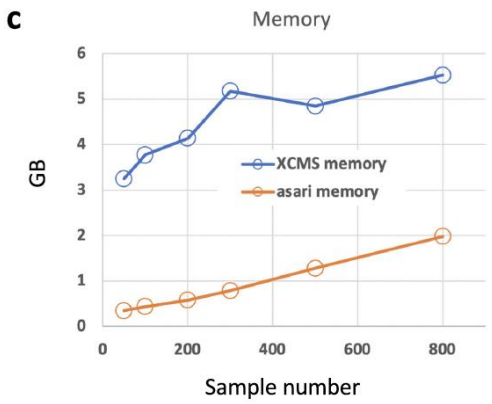
| | Default | Optimized |
|---|---|---|
| **Total peaks** | 4,344 | 5,113 (+ 17.7%) |
| **Isotopes** | 760 | 1,274 (+ 67.6%) |
| **Adducts** | 927 | 1,132 (+ 22.1%) |
| **Formulas assigned** | 632 | 687 (+ 8.7%) |
| **Potential compound matches** | 1,587 | 1,803 (+ 13.6%) |
| **Variance explained (PC1+PC2)** | 76.5% | 81.3% (+ 4.8%) |

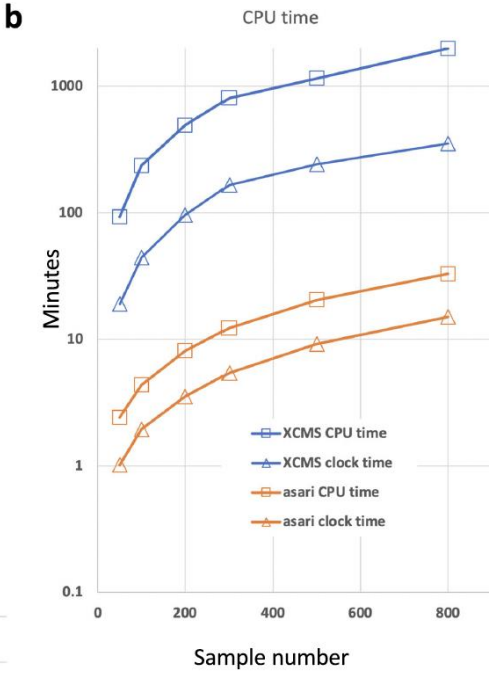
# New algorithm: asari



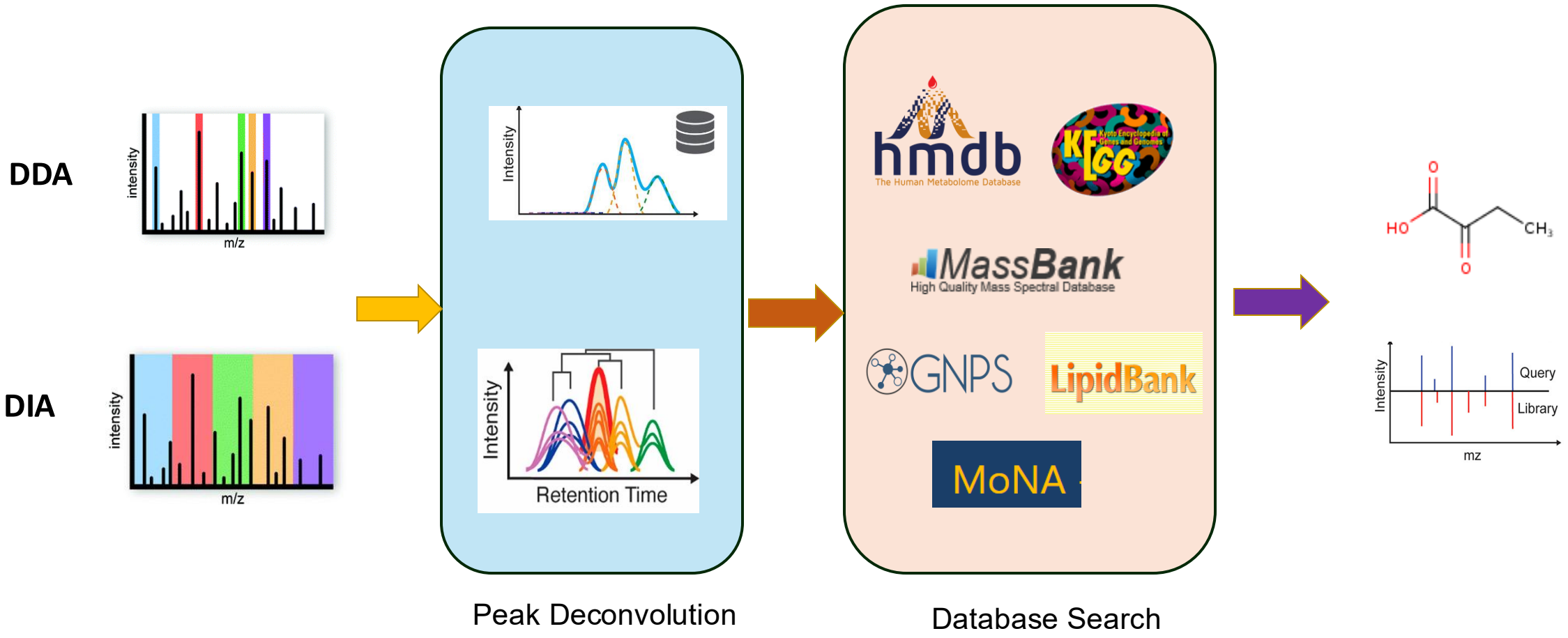



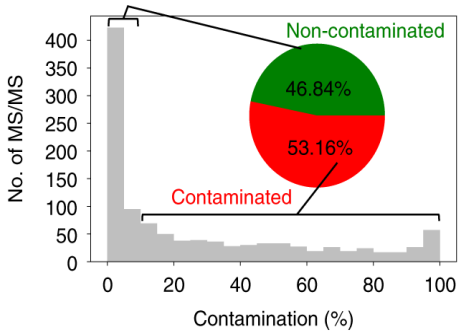




https://www.nature.com/articles/s41467-023-39889-1

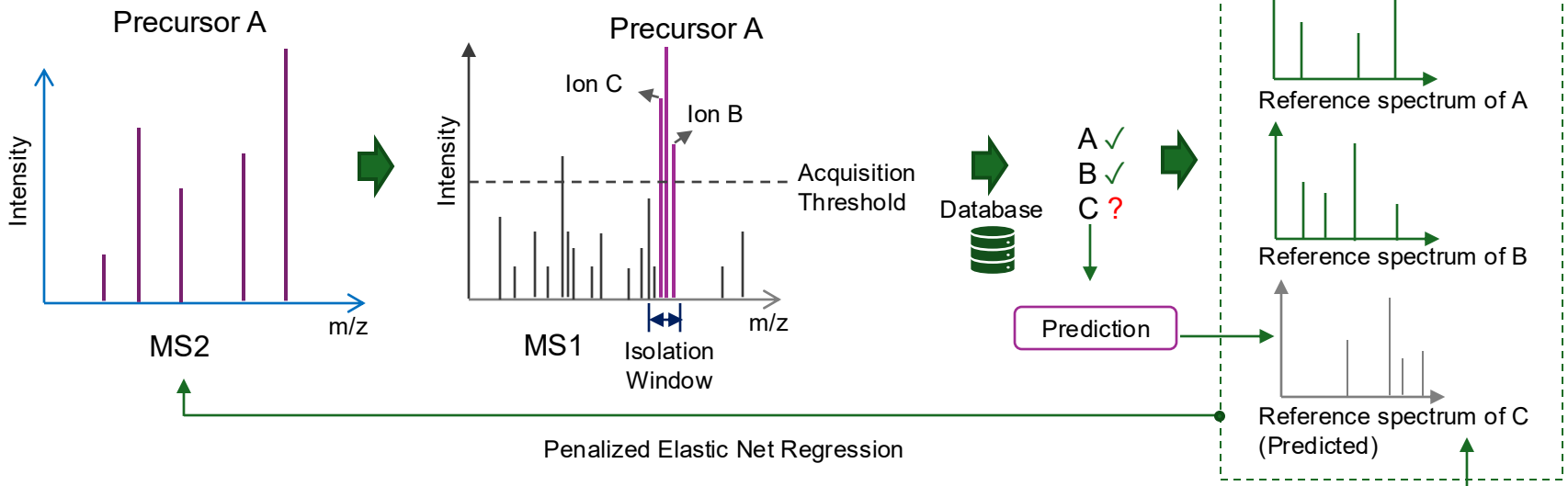
# MS/MS spectra processing in MetaboAnalyst



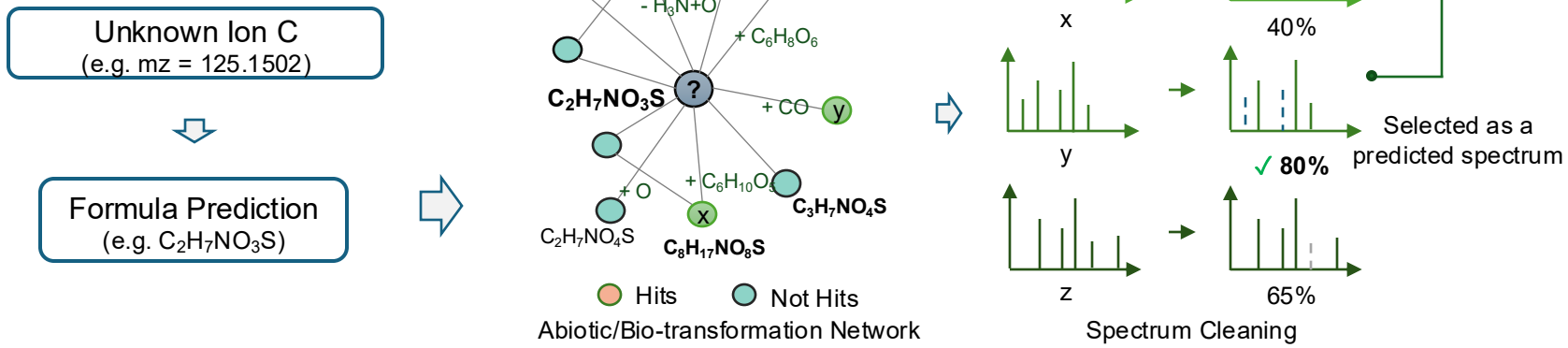
# DDA spectra deconvolution



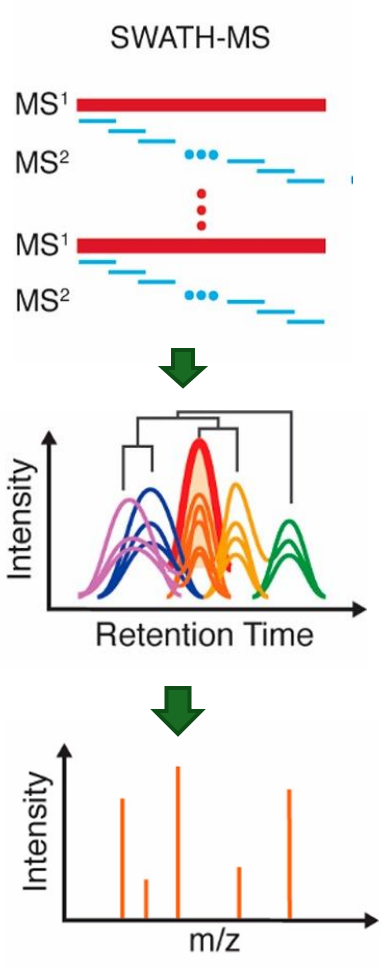

a

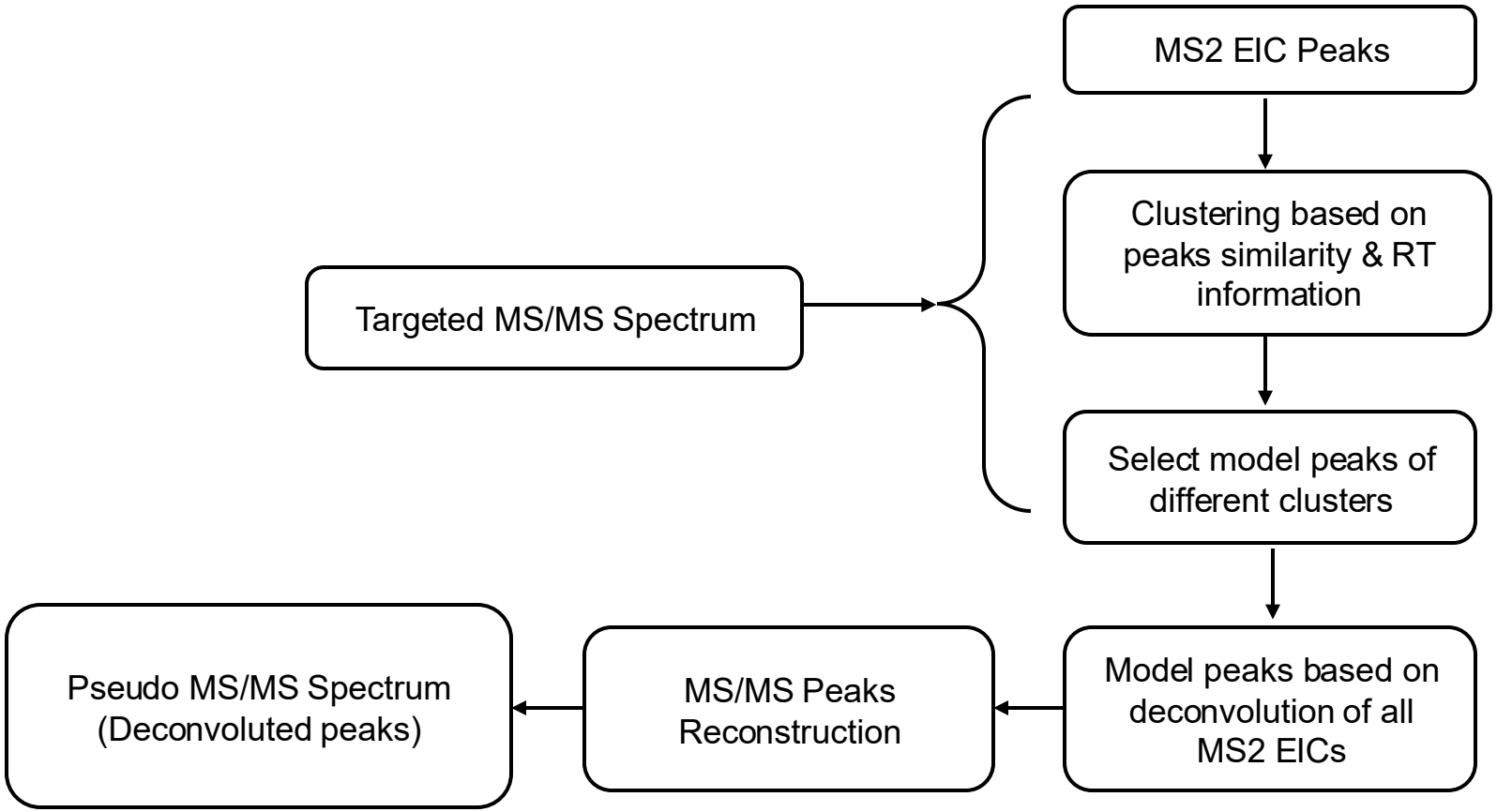



b



# SWATH-DIA spectra deconvolution




*Analytical Chemistry* **2019** *91* (18), 11897-11904

# Comprehensive MS2 databases

| Name | Records | Unique compounds | Size (MS2 \| neutral loss) |
|---|---|---|---|
| Complete library | 10,420,215 | 1,551,012 | 7.2 \| 6.4 GB |
| Pathway library | 172,370 | 3456 | 138.2 \| 94.1 MB |
| Biology library | 864,386 | 49,055 | 744.0 \| 491.0 MB |
| Exposome library | 1,883,828 | 106,387 | 1.5 \| 1.1 GB |
| Lipid library | 3,221,409 | 878,220 | 1.9 \| 1.1 GB |

Database SQLite files: https://metaboanalyst.ca/docs/Databases.xhtml

# Open schema

| ID | CompoundName | DBID | PrecursorMZ | PrecursorType | Formula | Smiles | InchiKey | InstrumentType | CollisionEnergy | RetentionTime | Ontology | NumberOfPeak | MS2Peaks |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Filter | Filter | Filter | Filter | Filter | Filter | Filter | Filter | Filter | Filter | Filter | Filter | Filter | Filter |
| 1 | Pyruvic acid | BMDM... | 89.02 | [M+H]+ | C3H4O3 | O=C(O)C... | LCTONWC... | Orbitrap | 10.0 | 2.2 | Alpha-keto... | 7 | 68.952 2... |
| 2 | Pyruvic acid | BMDM... | 111.01 | [M+Na]+ | C3H4O3 | O=C(O)C... | LCTONWC... | Orbitrap | 10.0 | 2.2 | Alpha-keto... | 4 | 110.975 69... |
| 3 | Pyruvic acid | BMDM... | 89.02 | [M+H]+ | C3H4O3 | O=C(O)C... | LCTONWC... | Orbitrap | 10.0 | 2.2 | Alpha-keto... | 18 | 68.452 21... |
| 4 | Pyruvic acid | BMDM... | 111.01 | [M+Na]+ | C3H4O3 | O=C(O)C... | LCTONWC... | Orbitrap | 10.0 | 2.2 | Alpha-keto... | 4 | 110.975 64... |
| 5 | Pyruvic acid | BMDM... | 89.02 | [M+H]+ | C3H4O3 | O=C(O)C... | LCTONWC... | Orbitrap | 10.0 | 2.2 | Alpha-keto... | 21 | 67.951 7... |
| 6 | Pyruvic acid | BMDM... | 111.01 | [M+Na]+ | C3H4O3 | O=C(O)C... | LCTONWC... | Orbitrap | 10.0 | 2.2 | Alpha-keto... | 5 | 110.975 76... |
| 7 | Pyruvic acid | BMDM... | 89.02 | [M+H]+ | C3H4O3 | O=C(O)C... | LCTONWC... | Orbitrap | 10.0 | 2.2 | Alpha-keto... | 20 | 68.452 14... |
| 8 | Pyruvic acid | BMDM... | 111.01 | [M+Na]+ | C3H4O3 | O=C(O)C... | LCTONWC... | Orbitrap | 10.0 | 2.2 | Alpha-keto... | 7 | 83.049 10... |
| 9 | Pyruvic acid | BMDM... | 89.02 | [M+H]+ | C3H4O3 | O=C(O)C... | LCTONWC... | Orbitrap | 10.0 | 2.2 | Alpha-keto... | 19 | 68.452 77... |
| 10 | Pyruvic acid | BMDM... | 111.01 | [M+Na]+ | C3H4O3 | O=C(O)C... | LCTONWC... | Orbitrap | 10.0 | 2.2 | Alpha-keto... | 8 | 83.049 15... |
| 11 | Pyruvic acid | BMDM... | 89.02 | [M+H]+ | C3H4O3 | O=C(O)C... | LCTONWC... | Orbitrap | 10.0 | 2.2 | Alpha-keto... | 19 | 68.452 28... |
| 12 | Pyruvic acid | BMDM... | 111.01 | [M+Na]+ | C3H4O3 | O=C(O)C... | LCTONWC... | Orbitrap | 10.0 | 2.2 | Alpha-keto... | 6 | 110.975 94... |
| 13 | Pyruvic acid | BMDM... | 89.02 | [M+H]+ | C3H4O3 | O=C(O)C... | LCTONWC... | Orbitrap | 10.0 | 2.2 | Alpha-keto... | 21 | 67.951 12... |
| 14 | Pyruvic acid | BMDM... | 111.01 | [M+Na]+ | C3H4O3 | O=C(O)C... | LCTONWC... | Orbitrap | 10.0 | 2.2 | Alpha-keto... | 4 | 110.975 74... |
| 15 | Pyruvic acid | BMDM... | 89.02 | [M+H]+ | C3H4O3 | O=C(O)C... | LCTONWC... | Orbitrap | 10.0 | 2.2 | Alpha-keto... | 20 | 67.951 10... |
| 16 | Pyruvic acid | BMDM... | 111.01 | [M+Na]+ | C3H4O3 | O=C(O)C... | LCTONWC... | Orbitrap | 10.0 | 2.2 | Alpha-keto... | 5 | 110.975 31... |
| 17 | Pyruvic acid | BMDM... | 89.02 | [M+H]+ | C3H4O3 | O=C(O)C... | LCTONWC... | Orbitrap | 10.0 | 2.2 | Alpha-keto... | 18 | 67.951 6... |
| 18 | Pyruvic acid | BMDM... | 111.01 | [M+Na]+ | C3H4O3 | O=C(O)C... | LCTONWC... | Orbitrap | 10.0 | 2.2 | Alpha-keto... | 7 | 110.975 29... |
| 19 | Pyruvic acid | BMDM... | 89.02 | [M+H]+ | C3H4O3 | O=C(O)C... | LCTONWC... | Orbitrap | 10.0 | 2.2 | Alpha-keto... | 20 | 68.452 25... |
| 20 | Pyruvic acid | BMDM... | 111.01 | [M+Na]+ | C3H4O3 | O=C(O)C... | LCTONWC... | Orbitrap | 10.0 | 2.2 | Alpha-keto... | 6 | 110.975 65... |
| 21 | Pyruvic acid | BMDM... | 89.02 | [M+H]+ | C3H4O3 | O=C(O)C... | LCTONWC... | Orbitrap | 10.0 | 2.2 | Alpha-keto... | 22 | 68.452 48... |
| 22 | Pyruvic acid | BMDM... | 111.01 | [M+Na]+ | C3H4O3 | O=C(O)C... | LCTONWC... | Orbitrap | 10.0 | 2.2 | Alpha-keto... | 3 | 110.975 88... |

# DDA benchmark – IROA mixture (ESI+)

**406** Compounds – ESI+, Isolation window: 1Da

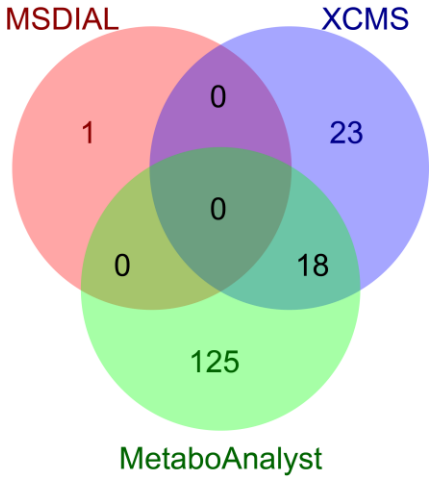
| Tools | Number of detected standards (MS1) | Compounds correctly annotated (MS2) | Percentage | Time elapsed (1 CPU core) |
|---|---|---|---|---|
| MS-DIAL + MS-FINDER | 165 | 82 | 20.2% | 32 min |
| MZmine + SIRIUS | 221 | 77 | 19.0% | 4 hours |
| **MetaboAnalyst*** | **239** | **159** | **39.1%** | **22 min** |
| MetaboAnalyst (nonDeco) | 239 | 146 | 36.0% | 12 min |

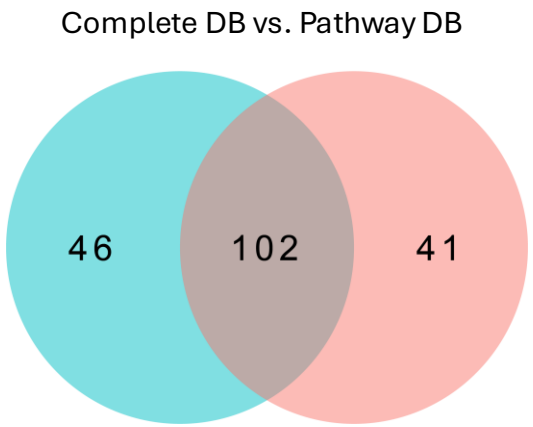
* Based on Complete Database with deconvolution enabled

# SWATH-DIA benchmark (ESI$^+$)

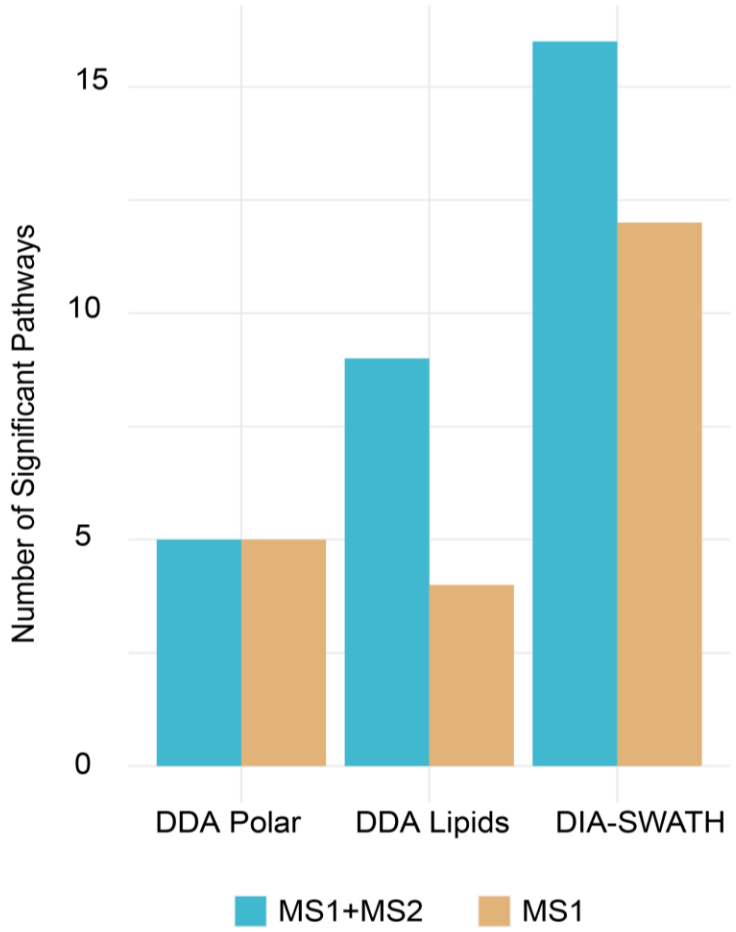
**406** Compounds.

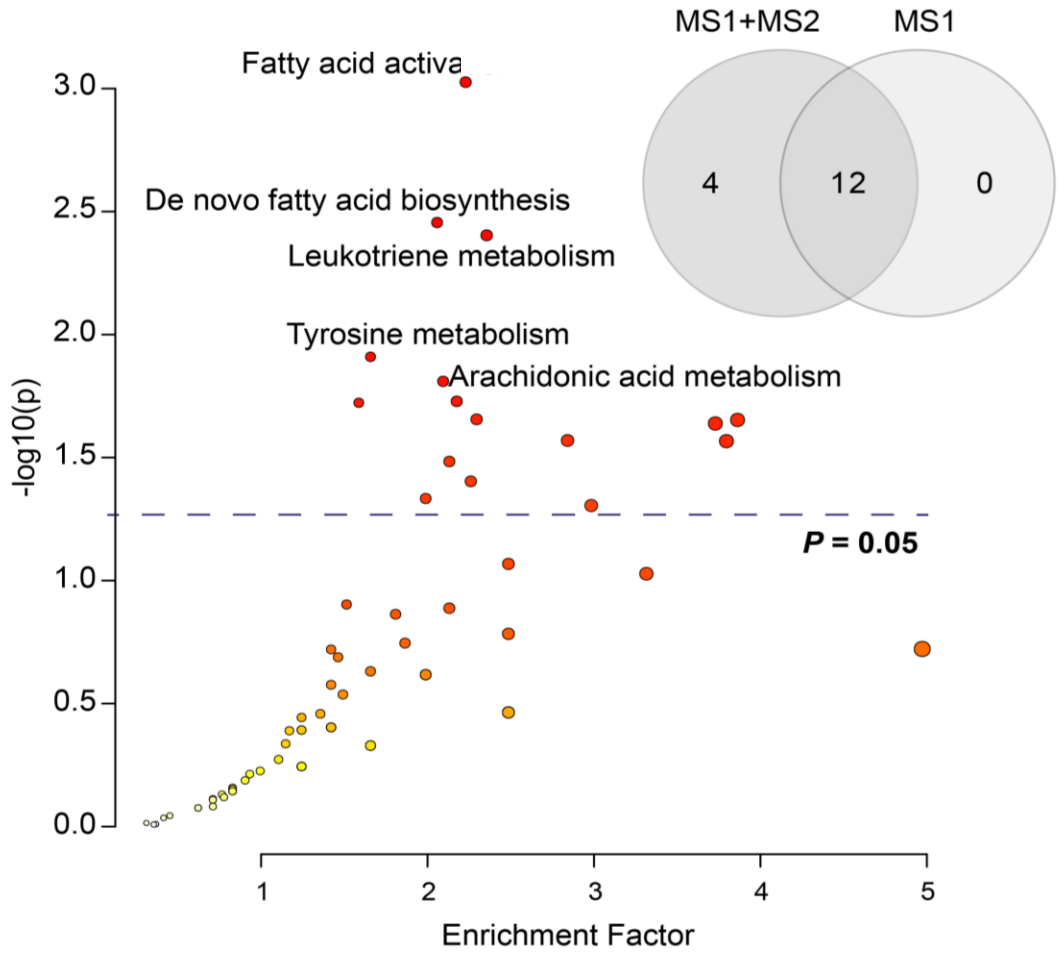
| Tools | Number of detected standards (MS1) | Compounds correctly annotated (MS2) | Percentage | Time elapsed (1 CPU core) |
|---|---|---|---|---|
| MS-DIAL + MS-FINDER | 5 | 1 | 0.25% | 3 min |
| XCMS + SIRIUS | 108 | 42 | 10.3% | ~ 12 h |
| **MetaboAnalyst** | **324** | **143** | **35.22%** | **14 min** |
| **MetaboAnalyst** (PathwayDB) | **324** | **148** | **36.45%** | **5 min** |



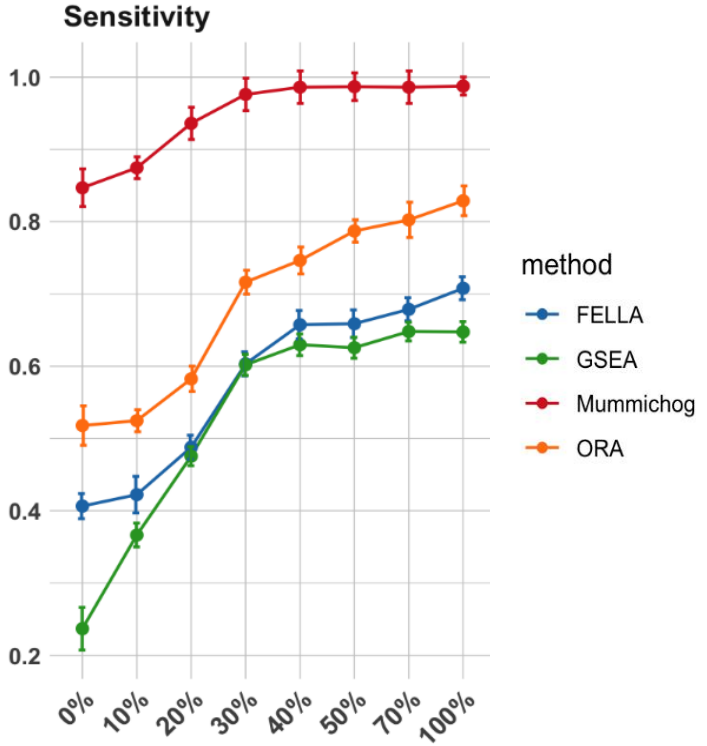
# Coupling LC-MS and MS/MS improves biology



# “Approximately correct” is sufficient for functional analysis

- Biological systems showing coordinated changes or group behaviors
- Group behavior can tolerate the random errors/inaccuracies associated with individual features
- Wide application of high-res MS greatly improves annotation



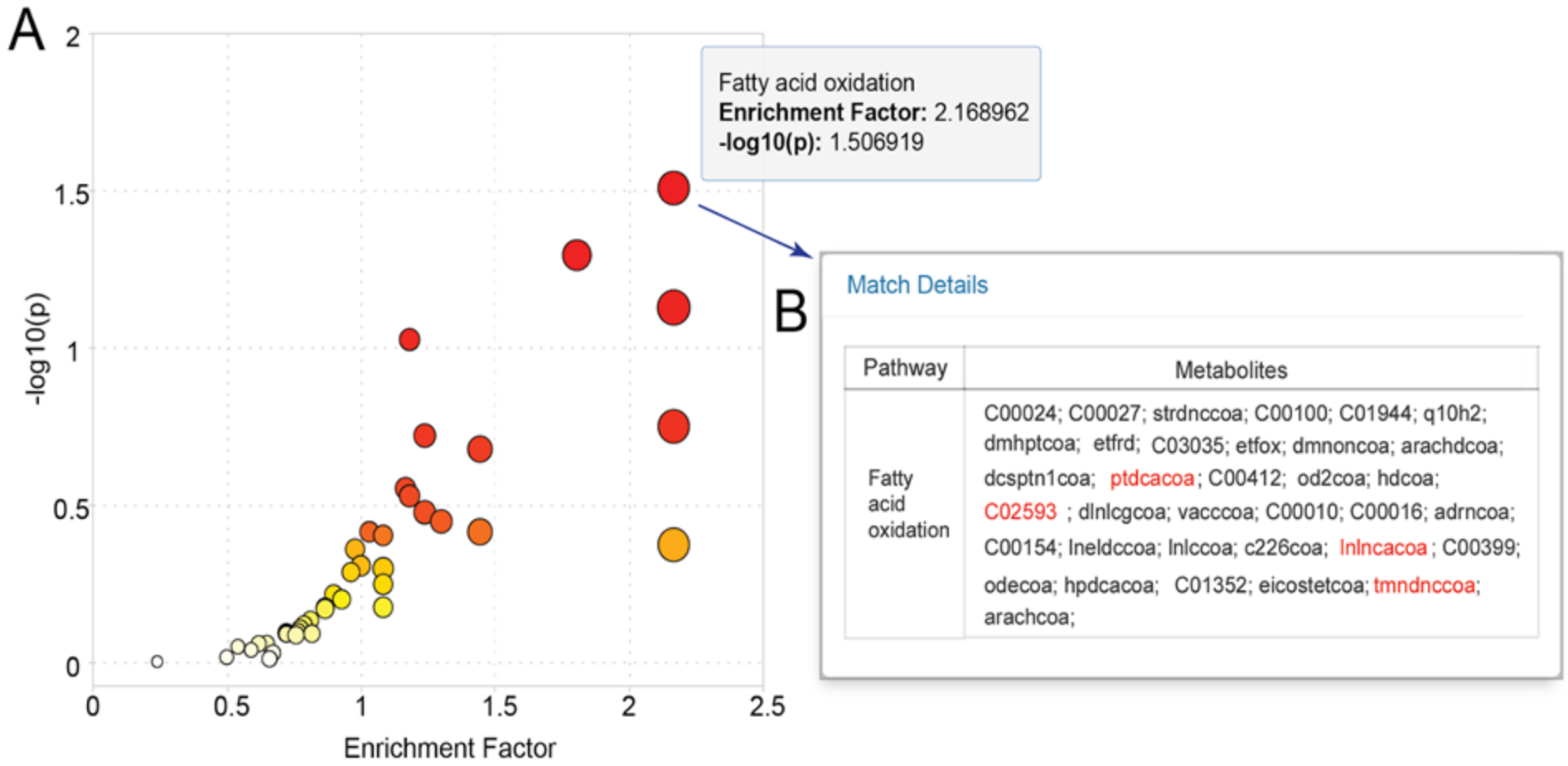

~30% correct annotation rate (remaining random assignment) enable ~100% pathway detection

https://doi.org/10.1093/bib/bbac553

# Critical: input preparation

- LC - **high-resolution** MS (LCHR-MS)
  - Orbitrap, Q-TOF
  - Reason: putative annotation needs to be approximately correct (better guess leads to more accurate functional analysis)
- Needs to be **complete** peak list or peak intensity table
  - Not just significant peaks
  - Reason: mummichog using permutation to estimate the null/background distribution
- In general, the algorithm works well for **> 3000** peaks (assuming human plasma samples).

# From ordered peak lists to functions

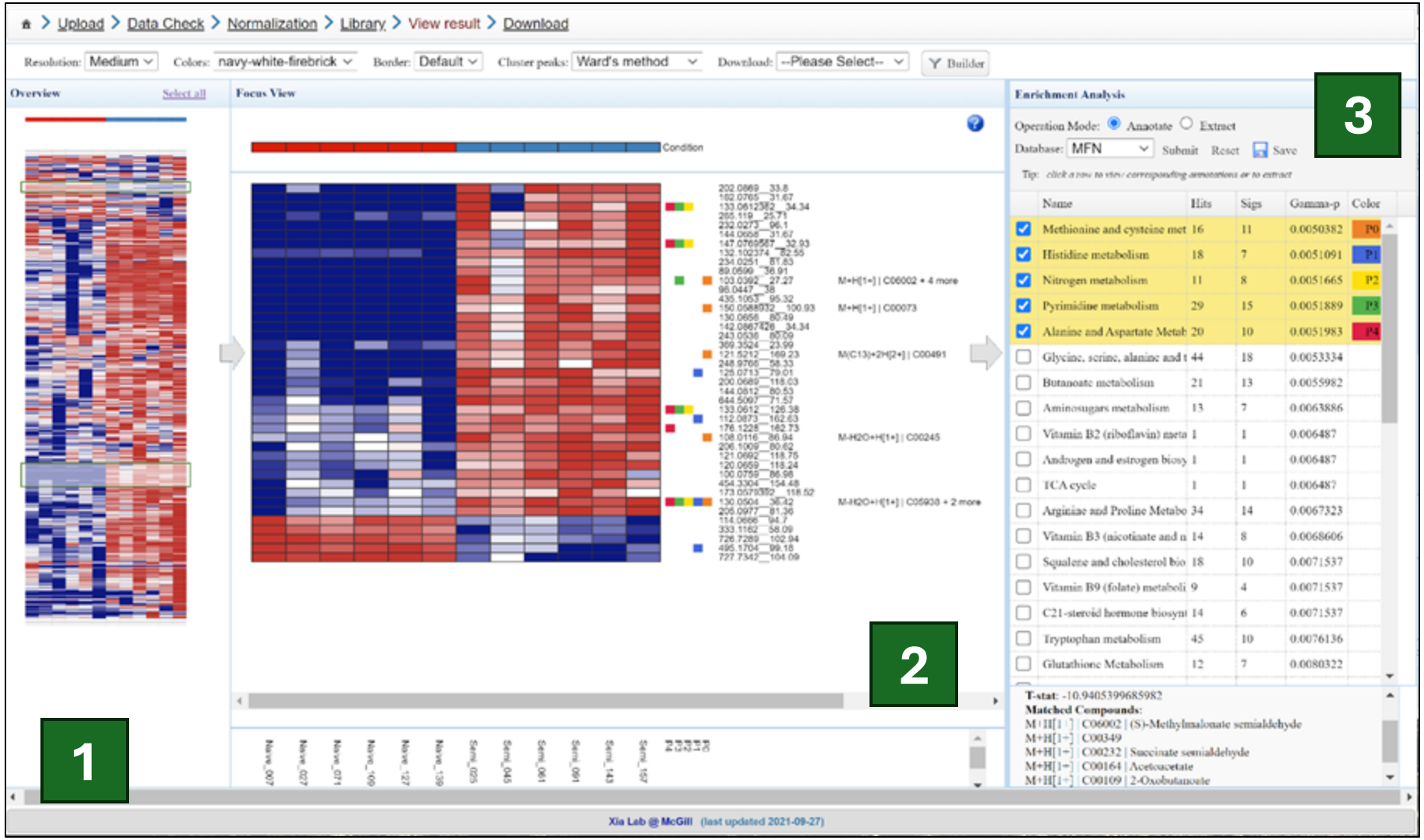


# How to interpret the results

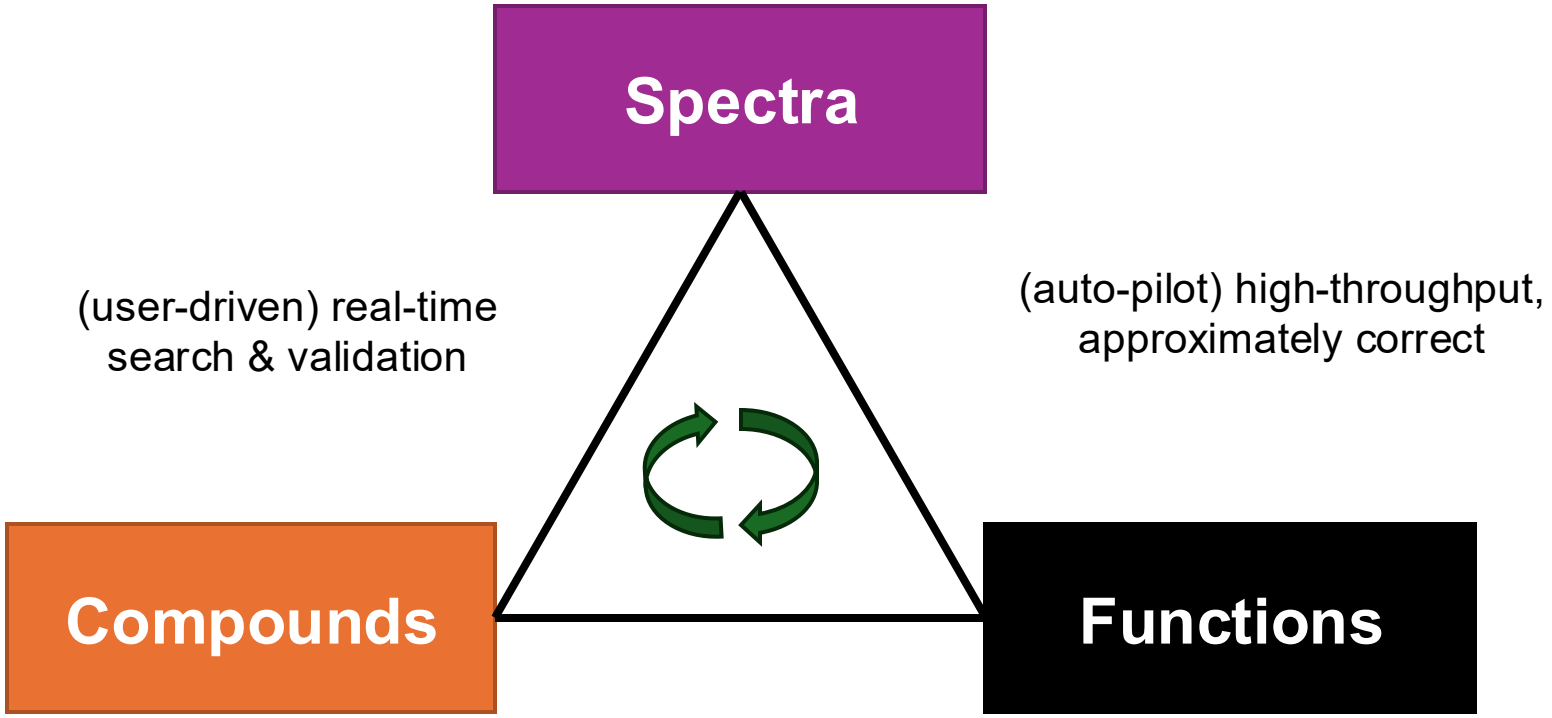
| Pathway Name | Total | Hits (all) | Hits (sig.) | Expected | P(Fisher) | P(Gamma) | Details |
|---|---|---|---|---|---|---|---|
| Vitamin E metabolism | 54 | 38 | 15 | 5.0563 | 0.030024 | 0.025523 | View |
| Carnitine shuttle | 72 | 25 | 10 | 6.7418 | 0.06554 | 0.028334 | View |

- **Total**: the total number of the given pathway
- **Hits (all)**: all the peaks mapped to the pathway
- **Hits (sig)**: all the significant peaks mapped to the pathway
- **Expected:** The expected number of metabolite hits in the pathway.
- **P(Fischer)**: The Fisher's exact p-value for the pathway
- **P(Gamma)**: P-values derived from Gamma distribution based on permutation tests for the pathway.

# Not just significant peaks



# Iterative refinement & understanding



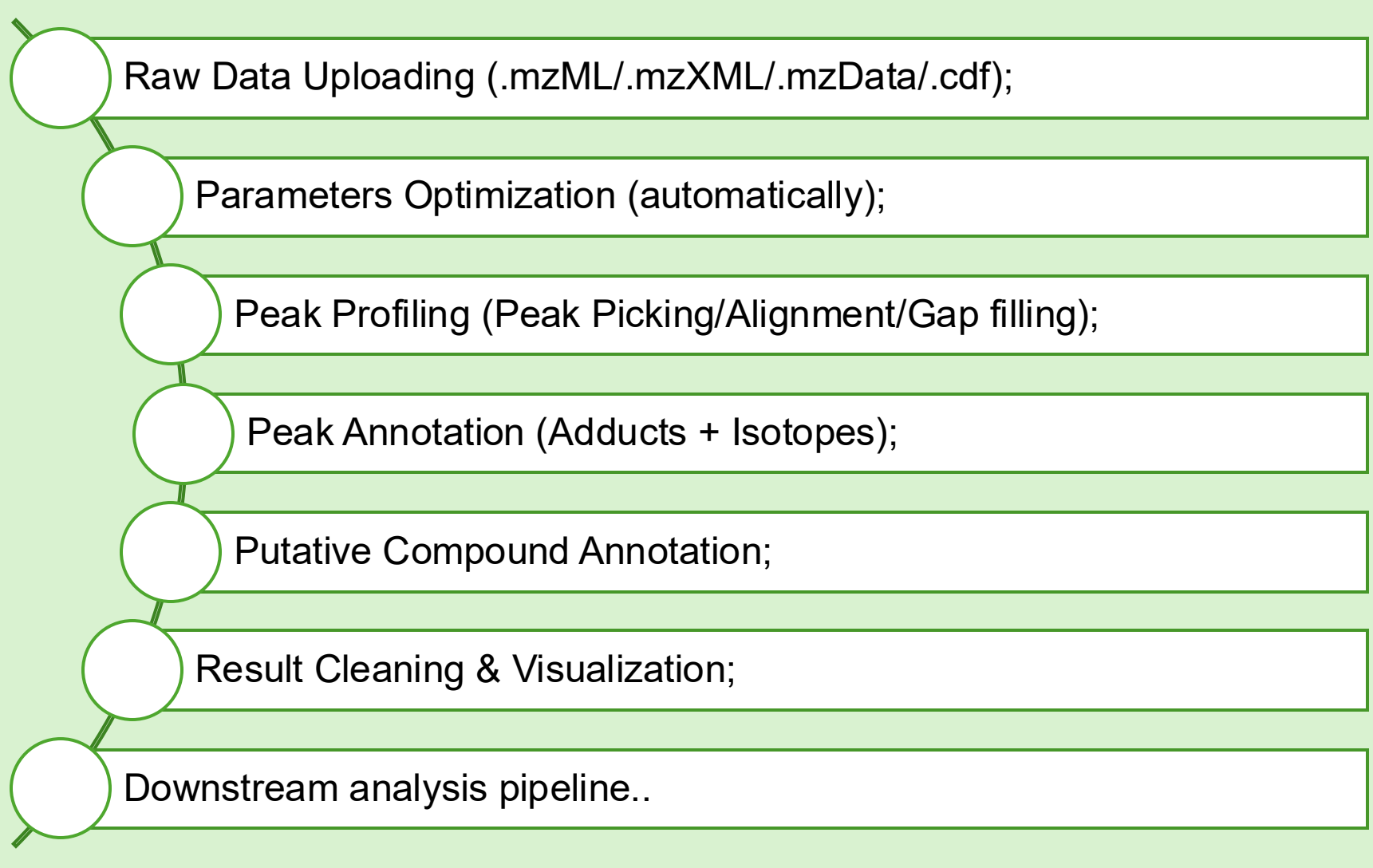

Developer friendly (open-source R package);
researcher friendly (web-based tools)

# Main updates since last year

1. Improved raw spectral analysis pipeline
2. Support exposomics annotation
3. Enrichment network
4. Improved missing values estimation
5. Better graphics
6. Dose response analysis
   - Supporting both continuous & categorical dosage
7. Causal analysis
   - Fast database

# Live Demo 1 - LC-MS spectra processing

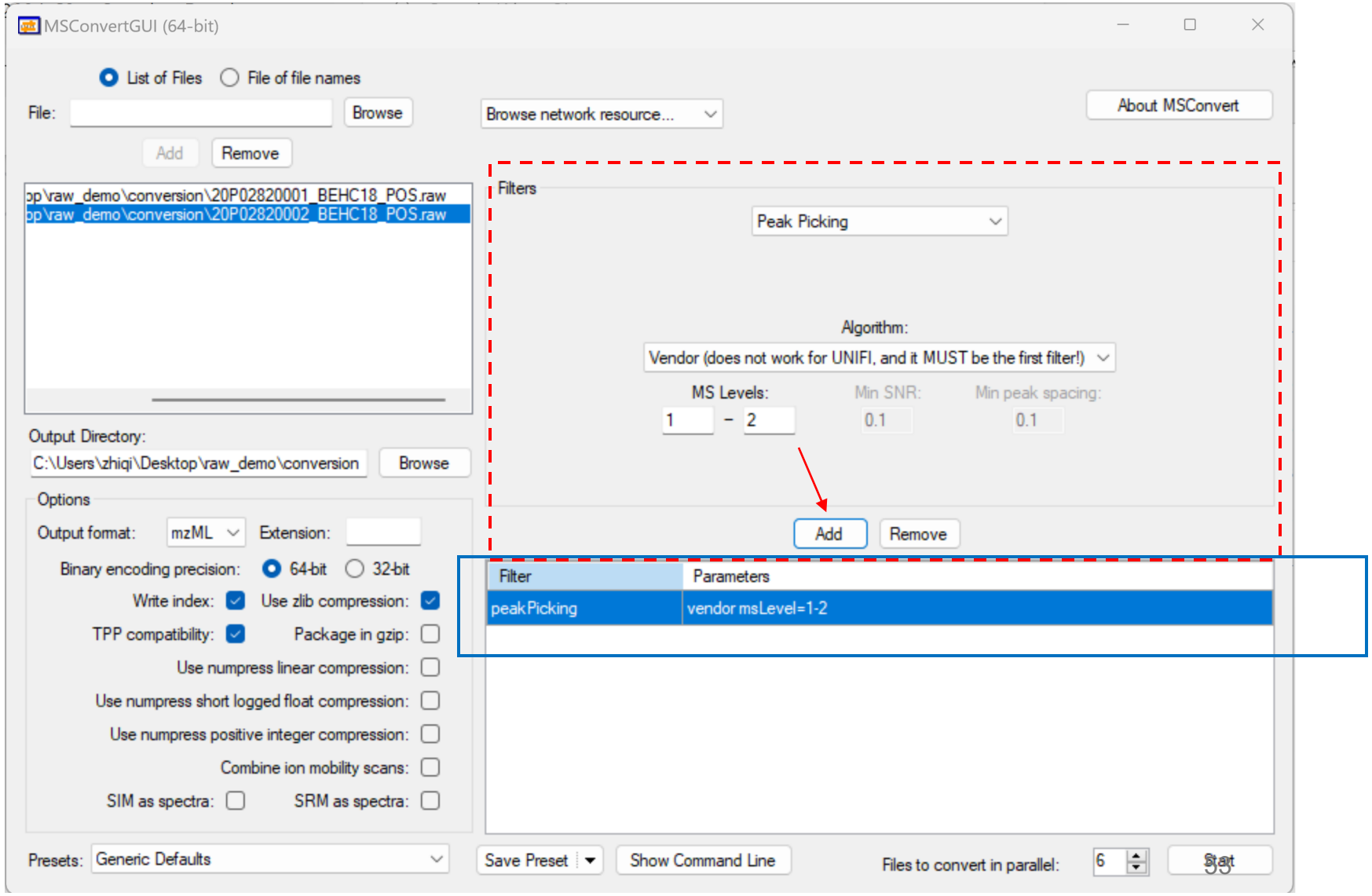
# Overview of raw spectra processing



# Spectral Centroiding - Windows

MSConvert GUI (Windows Only)

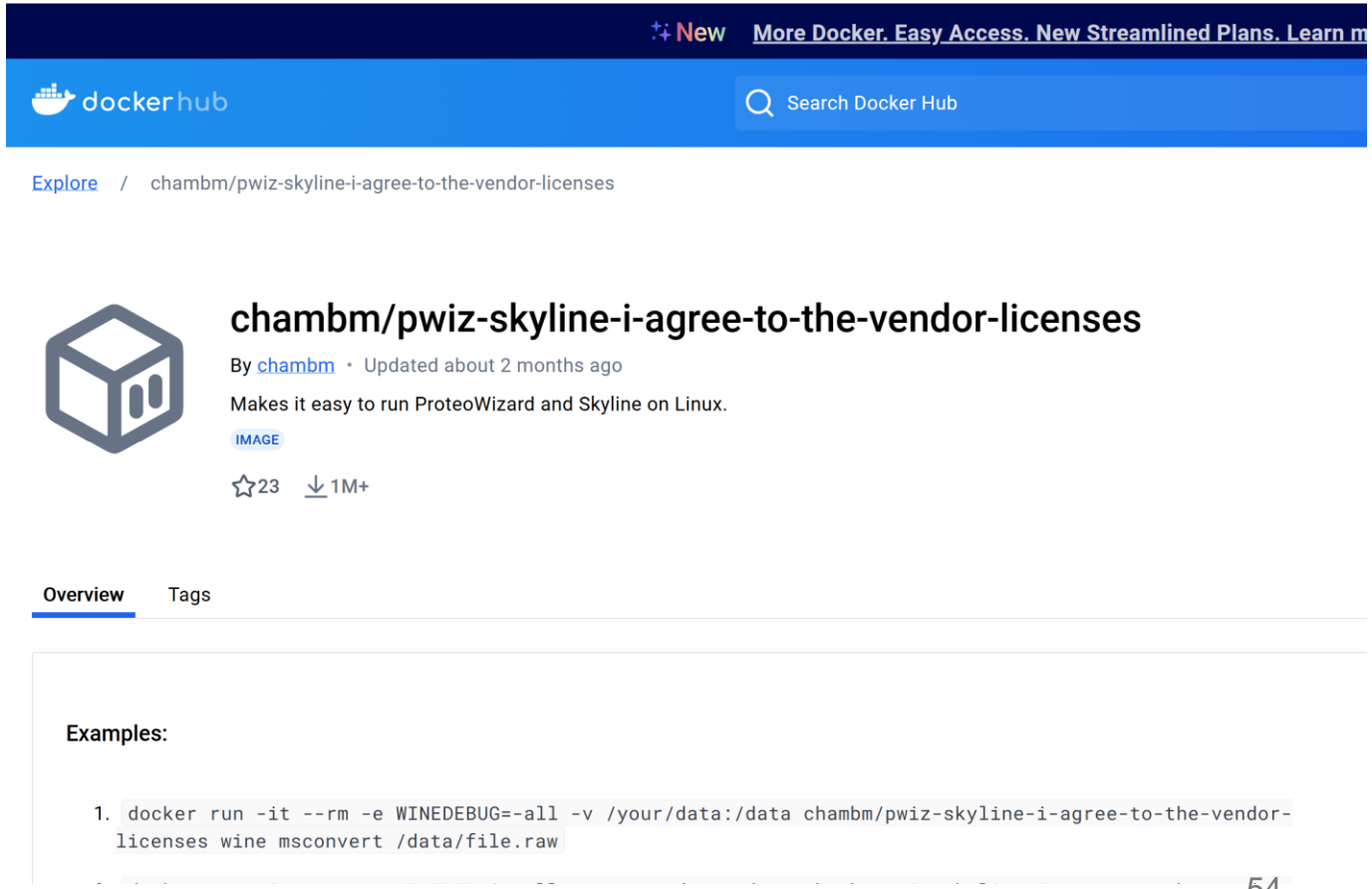
How to centroid MS spectral data ?



# Spectral Centroiding - MacOS/Linux

docker pull chambm/pwiz-skyline-i-agree-to-the-vendor-licenses

docker run -it --rm -e WINEDEBUG=-all -v **/project/E-waste_Metabolomics_data/HILIC_NEGATIVE**:/data chambm/pwiz-skyline-i-agree-to-the-vendor-licenses wine msconvert ***.raw** -o mzML **--mzML** --filter "**peakPicking true 1-2**" --filter "zeroSamples removeExtra" --64 --zlib --singleThreaded



# MetaboAnalyst 6.0 Modules

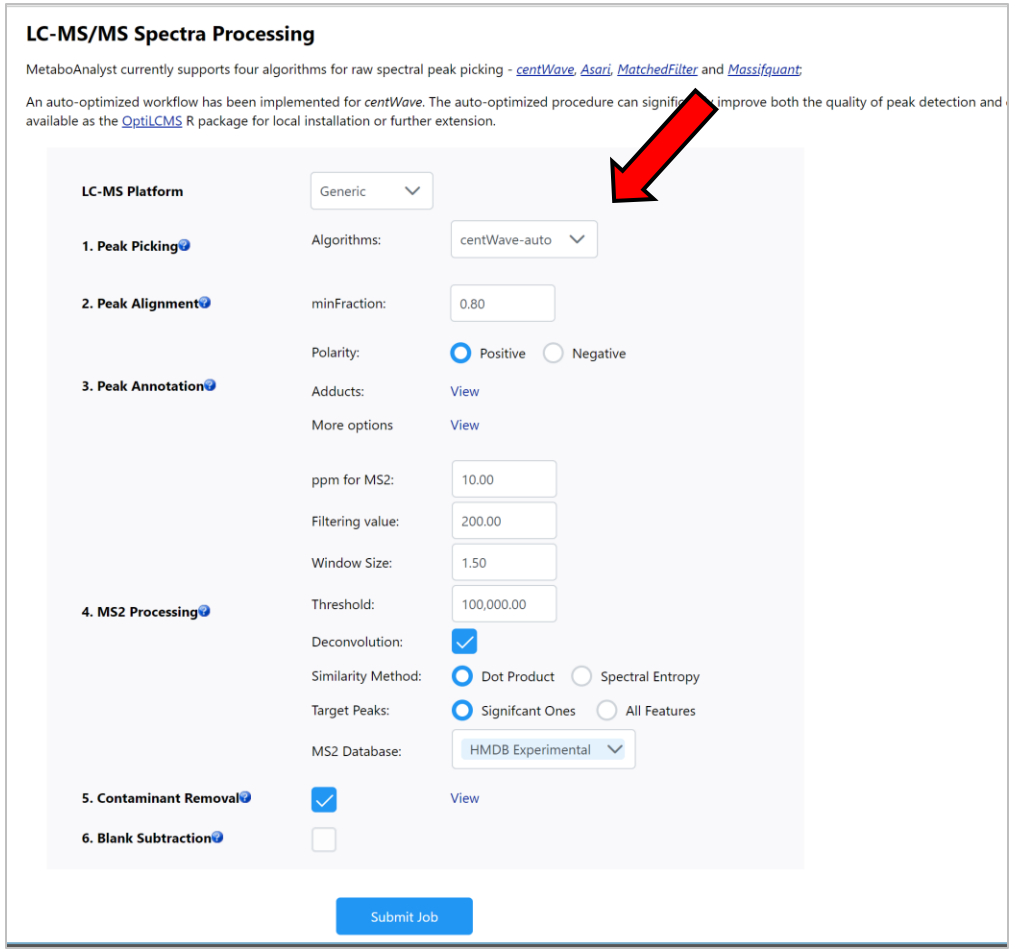
| Input Data Type | Available Modules (click on a module to pr[illegible]d, or scroll down to explore a total of 18 modules including utilities) | | | | |
|---|---|---|---|---|---|
| LC-MS Spectra (mzML, mzXML or mzData) | | | Spectra Processing [LC-MS w/wo MS2] | | |
| MS Peaks (peak list or intensity table) | | Peak Annotation [MS2-DDA/DIA] | Functional Analysis [LC-MS] | Functional Meta-analysis [LC-MS] | |
| Generic Format (.csv or .txt table files) | Statistical Analysis [one factor] | Statistical Analysis [metadata table] | Biomarker Analysis | Statistical Meta-analysis | Dose Response Analysis |
| Annotated Features (metabolite list or table) | | Enrichment Analysis | Pathway Analysis | Network Analysis | |
| Link to Genomics & Phenotypes (metabolite list) | | | Causal Analysis [Mendelian randomization] | | |

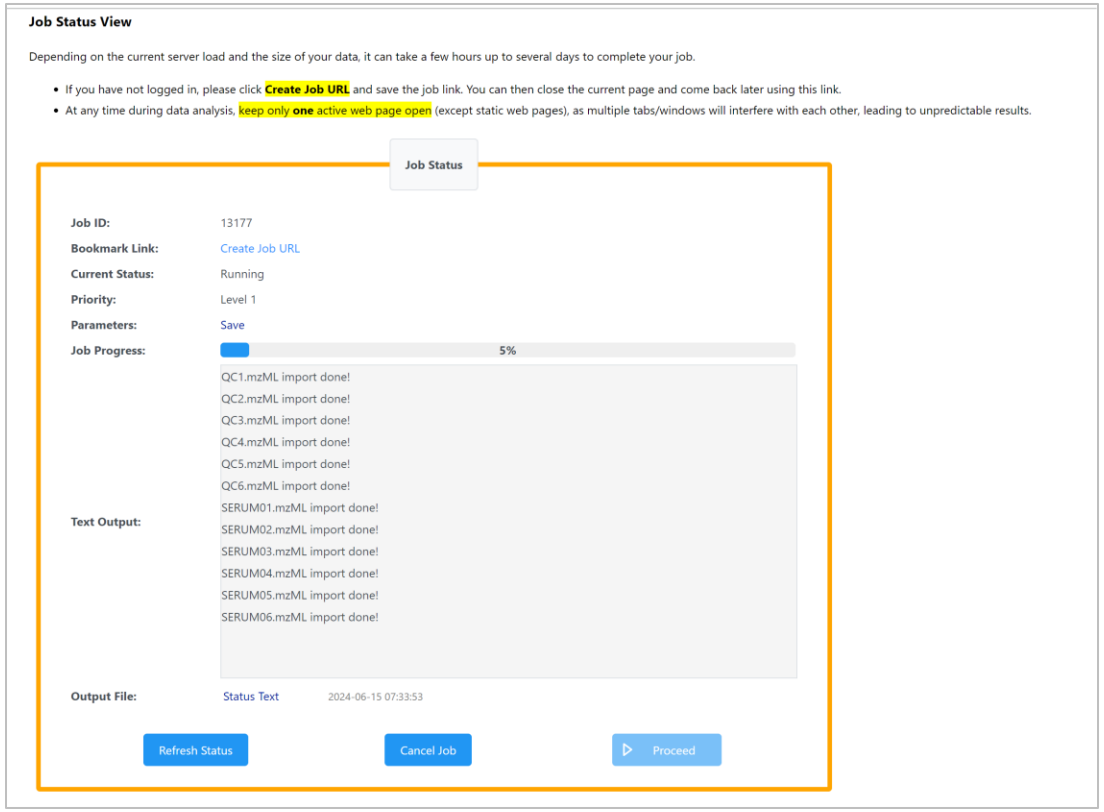
# LC-MS Spectra Preparation

MetaboAnalyst currently supports mzML, mzXML, CDF or mzData formats in centroid mode. For MS2 data, spectra should be acquired in either **DDA** or **SWATH-DIA** mode for each job. Mixed mode is not supported.

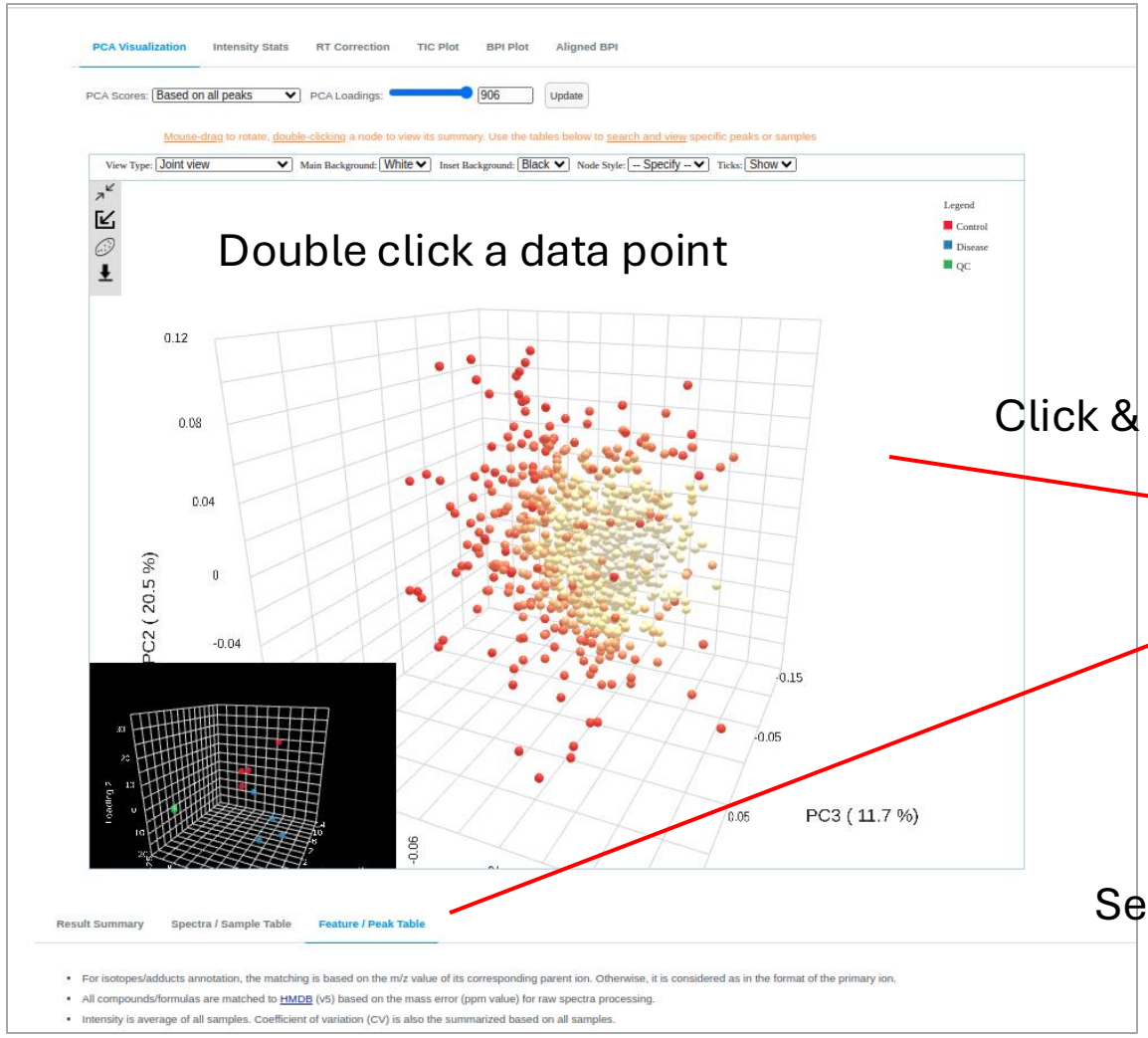
1. [Required] MS1 Spectra uploaded as individual zip files - one zip (.zip) per spectrum [max: 200 spectra].
2. [Optional] Either **DDA**- or **SWATH-DIA**-based LC-MS/MS Spectra should be uploaded as individual zip files (same as MS1) [max: 50 spectra]. MS2 data must start with **"MS2_"** or marked as "MS2" in meta data file.
3. [Optional] Meta data uploaded as a plain text (.txt) file containing two columns - spectral names and group labels
4. [Optional] Quality control (QC) spectra should start with **"QC_"** or marked as "QC" in meta data. BLANK should be marked as "BLANK" in meta data for subtraction.

# Parameter setting & job submission



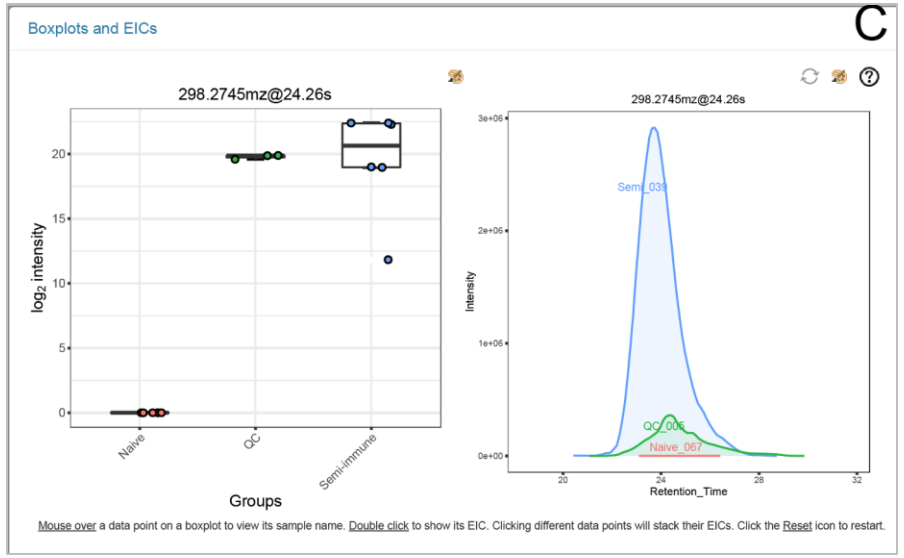



# Result Exploration




Click & view

Search & view



# Output: peak abundance table

- Uniquely identify each peak: retention time and m/z value
- Calculate the relative intensity in each sample

Retention time

m/z value

| Sample | X1014 | X1049 | X1068 | X1070 | X1071 | X1073 | X1074 | X1075 | X1076 | X1078 |
|---|---|---|---|---|---|---|---|---|---|---|
| 85.02798773_398.845656 | 91281.129 | 295971.19 | 244257.92 | 82883.828 | 357387.91 | 314793.29 | 296933.07 | 259134.23 | 316398.3 | 298981.38 |
| 85.03918591_540.7198895 | 20368705 | 23645645 | 27541993 | 20197810 | 20698441 | 27700133 | 18903295 | 21151136 | 22135283 | 23551889 |
| 85.03934182_206.8491361 | 100801.73 | 147630.84 | 128838.32 | 48201.572 | 14503.911 | 94388.175 | 147840.04 | 94226.848 | 47368.725 | 86117.51 |
| 85.05850153_553.5489174 | 28578.672 | NA | 42871.286 | 45854.92 | 31862.665 | 42511.683 | 16638.517 | 21645.293 | 42802.335 | 47630.422 |
| 85.06447722_552.8676506 | 64506.008 | 36993.153 | 64365.242 | 21970.254 | 22431.698 | 42717.702 | 49608.002 | 61113.878 | 45457.694 | 31242.437 |
| 85.07557123_503.1977875 | 5185552 | 6545664.8 | 4849575.1 | 7455068.2 | 4687812.5 | 8568037.4 | 5092330.2 | 3961282.2 | 6480194.6 | 7331818.4 |
| 85.07616337_141.9029172 | 82899.952 | 207861.36 | 50610.657 | 79208.885 | 225161.43 | NA | 347408.98 | 236485.2 | 776251.79 | 164112 |
| 85.0838011_198.0411769 | 85303.336 | 123532.16 | 91254.97 | 66497.463 | 172721.72 | 236255.05 | 47396.288 | 78663.557 | 189683.64 | 245493.04 |
| 85.50950642_172.3411474 | 339908.68 | 321187.16 | 322001.53 | 255557.48 | 330914.06 | 254245.84 | NA | NA | 290287.6 | 298955.37 |
| 85.51517772_50.65023803 | 118159.94 | 112972.04 | 114059.62 | 113950.95 | 167858.69 | 103292.57 | 86749.39 | 82707.461 | 119298.44 | 107657.2 |
| 85.5363475_41.45434989 | 53482.821 | 17514.179 | 35163.947 | 36411.914 | 59951.47 | 51123.602 | 41371.083 | 30019.615 | 22520.943 | 47343.966 |
| 85.96264165_42.73935005 | 81788.089 | 78215.738 | 50882.903 | 65819.686 | 73752.586 | 57479.55 | 71399.888 | 42905.115 | 49373.813 | 68847.43 |
| 86.00545485_545.3171583 | 46468.886 | 40671.699 | 23324.775 | 36142.339 | 31310.553 | 56563.276 | 26034.229 | NA | NA | 29480.762 |
| 86.01779309_54.25356378 | 57728.236 | 36204.919 | 31645.834 | 63374.773 | 42848.297 | 70339.755 | NA | 46788.918 | 78406.509 | 49801.696 |
| 86.03613685_568.0578201 | 120163.19 | 121293.45 | 137159.94 | 118697.36 | 114696.1 | 147598.85 | 95348.512 | 97339.544 | 120371.54 | 117616.77 |
| 86.04255464_546.3279646 | 773051.95 | 675716.91 | 764306.84 | 716529.31 | 614985.95 | 775433.46 | 527588.69 | 666915.52 | 719938.04 | 659466.81 |
| 86.05953662_575.4776799 | 1305749 | 986112.65 | 1107787.1 | 896955.61 | 623282.13 | 622941.45 | 627053.74 | 507228.47 | 1017792.5 | 491771.67 |
| 86.05955314_395.2147633 | 1151506.4 | 827450.26 | 484189.22 | 252791.25 | 1586988.1 | 522492.9 | 1083396.6 | 410343.24 | 291013.34 | 591663.48 |
| 86.0596265_321.9552286 | 2306641.6 | 2636648.8 | 2057971 | 2244866.3 | 2813936.3 | 2650464.4 | 2521397.2 | 2291594.2 | 2794708.7 | 2986888.4 |
| 86.07101485_545.5074342 | 18024.92 | 48694.834 | 39266.12 | NA | 21814.652 | 14367.843 | NA | 16065.358 | 11001.248 | 26206.676 |
| 86.07888004_507.2294891 | 186762.52 | 274866.34 | 292333 | 168433.73 | 130364.59 | 257889.76 | 129553.62 | 137593.85 | 315715.95 | 134660.58 |
| 86.08334857_524.9644006 | NA | NA | 51854.327 | NA | 55064.237 | 84586.362 | 38654.123 | 45651.322 | 54524.784 | 40857.812 |

Relative intensities

# MetaboAnalyst 6.0 Modules

| Input Data Type | Available Modules (click on a module to proceed, or scroll down to explore a total of 18 modules including utilities) | | | | |
|---|---|---|---|---|---|
| LC-MS Spectra (mzML, mzXML or mzData) | | | Spectra Processing [LC-MS w/wo MS2] | | |
| MS Peaks (peak list or intensity table) | | Peak Annotation [MS2-DDA/DIA] | Functional Analysis [LC-MS] | Functional Meta-analysis [LC-MS] | |
| Generic Format (.csv or .txt table files) | Statistical Analysis [one factor] | Statistical Analysis [metadata table] | Biomarker Analysis | Statistical Meta-analysis | Dose Response Analysis |
| Annotated Features (metabolite list or table) | | Enrichment Analysis | Pathway Analysis | Network Analysis | |
| Link to Genomics & Phenotypes (metabolite list) | | | Causal Analysis [Mendelian randomization] | | |

# Accepted formats and expected results

To accommodate application scenario and offer compatibility with MS2 spectra results from other popular tools. There are three formats supported:

1. Simple text file (m/z and intensity separated by tab);
2. MGF format (standard);
3. MSP format (MS-DIAL);

The MS2 spectra/spectrum searching provides results including comprehensive compound identification summary and visualization of the matching pattern:

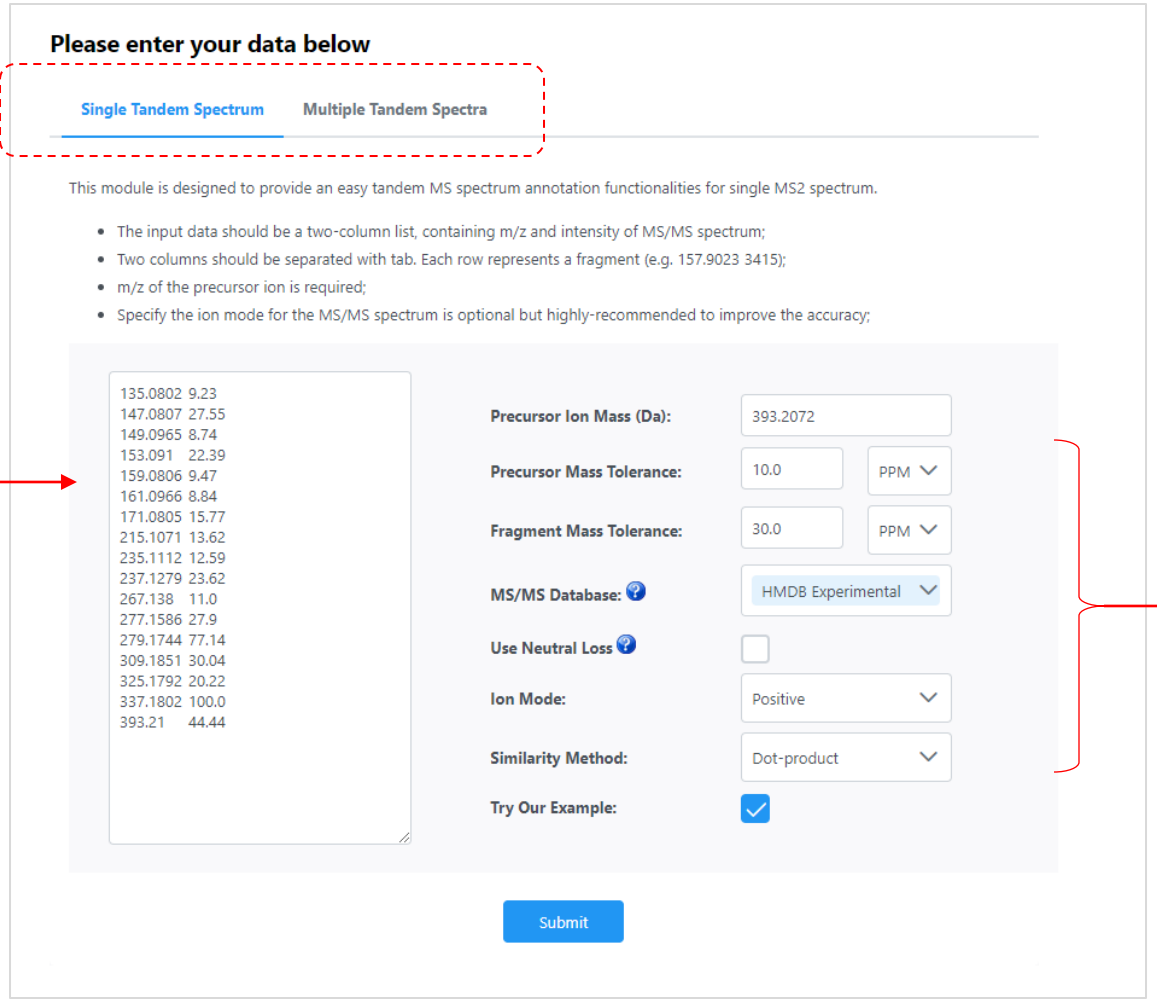
1. Compound identification summary table;
2. Visualization on MS2 matching pattern and annotation of fragments;

# Single spectrum upload

At the first page, user can upload single spectrum or multiple spectra. We used the "***Single Tandem Spectrum***" at this stage.

**For single spectrum uploading,**

1. It should be a text containing two columns. The first column is *m/z* values, while the second column is intensity values.
2. The two columns must be separated by tab (not space).
3. It is unnecessary to normalize the intensity values, we will automatically do it.




**Parameters for searching,**

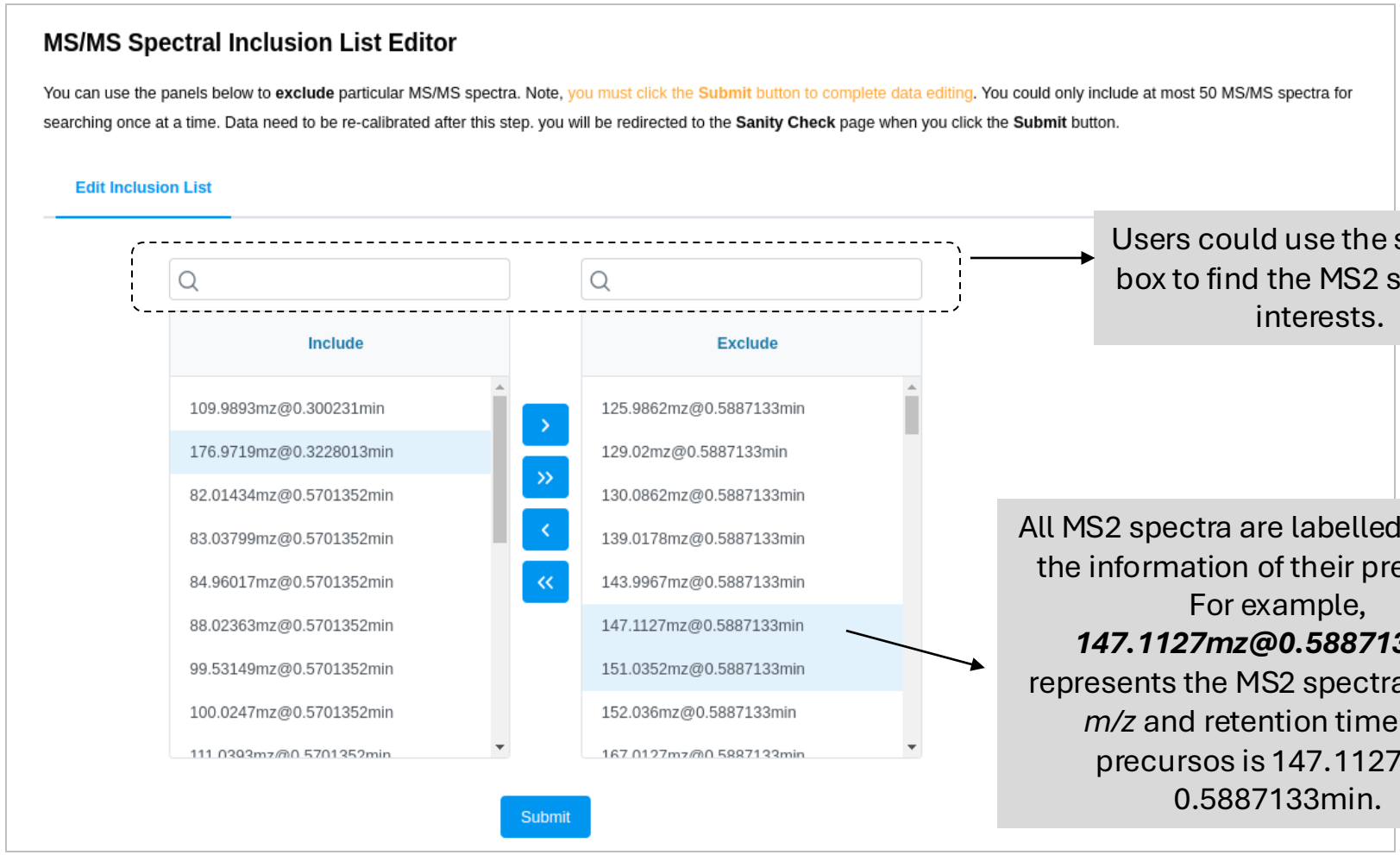
1. **Precursor Ion Mass** is required, please input the value as precisely as possible;
2. **Tolerance**: both tolerance values are recommended to be optimized based on MS instrument'
3. **MS/MS Databases**: user could customize their database option (see 3.2 for more information);
4. **Use Neutral Loss:** user could optionally use Neutral Loss for database search by use the option. Please note, this is only encouraged for unknown new compound discovery;
5. **Similarity Method**: User could choose traditional way (dot-product) or a new strategy (spectral entropy).

**Mouse hover the help tip to view the detailed information on different database options.**

# MS/MS spectral batch processing

**Spectra inclusion editor,**

1. Since MetaboAnalyst only support at most 20 spectra searching once at a time, user could manually customize the inclusion list for MS2 database search;
2. By default, the first 20 spectra will be listed into “Include” list to be included for searching;
3. User could move MS2 spectra features between two lists by using the blue moving arrows;
4. Once the editing is done, Click “***Submit***” button to confirm.




Users could use the searching box to find the MS2 spectra of interests.

All MS2 spectra are labelled based on the information of their precursors. For example, ***147.1127mz@0.5887133min*** represents the MS2 spectra, and the *m/z* and retention time of its precursos is 147.1127 and 0.5887133min.

# MS2 spectra searching results

- Database search results would be summarized as a table;
- User could expand a row to visually explore the matching results of a MS2 spectrum;
- Information of fragments will be automatically displayed when mouse hover the fragment;
- The top (blue) part are users' input, while the bottom (red) parts are from the reference library;
- All matched fragments will be marked with red diamond at the top.

# Demo Datasets

- **IBD dataset**
  - Clinical metabolomics dataset;
  - CD and non-IBD patients: 35 samples are included.
  - LC-MS1 only.

- **DDA dataset (homework)**
  - Whole blood metabolomics dataset
  - A total of 30 samples (serum, plasma and whole blood samples (n=6 for each), 6 QCs and 6 DDA MS/MS samples)

# Results summary

Result Summary | Spectra / Sample Table | Feature / Peak Table | MS/MS Results

**Raw Spectra Processing Result Summary:**

MetaboAnalyst has finished raw spectra processing with OptiLCMS (1.2.0):

There are 24 samples of 4 groups (Plasma, QC, Serum, whole_blood) included for processing!

Total of 4910 features have been detected and aligned across the whole sample list.

The mass deviation of this study was estimated/set as 5 ppm.

2413 features (49.13%) have been annotated as isotopes.

2312 features (47.08%) have been annotated as adducts.

189 unique formulas have been matched to HMDB database.

958 potential compounds have been matched to HMDB database.

Download Page

# MS/MS Results

Result Summary | Spectra / Sample Table | Feature / Peak Table | **MS/MS Results**

- MS/MS-based compounds identification results are displayed below.
- Similarity of MS/MS are evaluated based on dot-product or spectral entropy methods. Top 5 compounds are listed from high to low (100, perfect match; 0, not matched).
- User could click View button below to view the MS/MS pattern matching results.

| Compound | Formula | Matching Score ↑↓ | InchiKey | Database | View |
|---|---|---|---|---|---|
| ⌄ mz137.0459@42.46min | | | | | |
| Hypoxanthine | C5H4N4O | 84.79 | FDGQSTZJBFJUBT-UHFFFAOYSA-N | HMDB_experimental | 🖼 |
| Allopurinol | C5H4N4O | 83.53 | OFCNXPDARWKPPY-UHFFFAOYSA-N | HMDB_experimental | 🖼 |
| › mz98.9830@51.97min | | | | | |
| › mz83.0597@45.80min | | | | | |
| › mz147.1130@35.24min | | | | | |
| › mz150.0585@43.66min | | | | | |

# Hands on Practices (15 min)

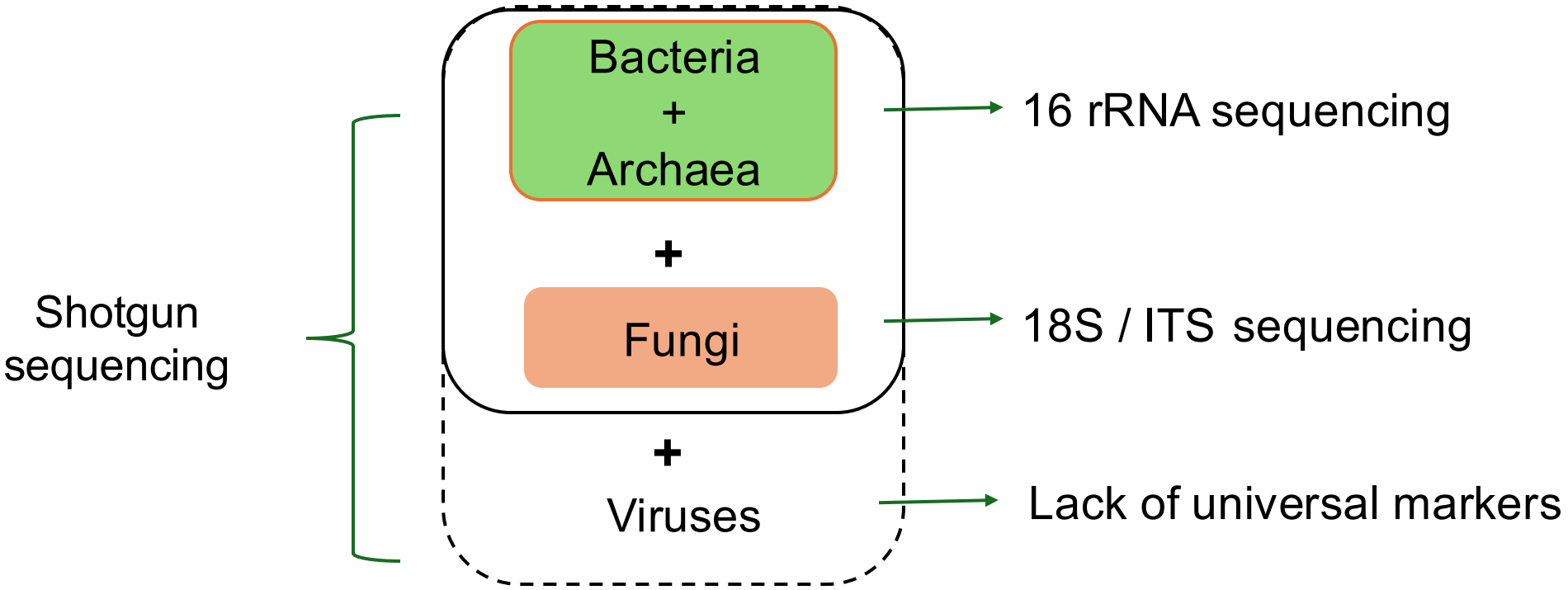
## Tips

1. For raw data processing, demo data can be found in https://drive.google.com/drive/folders/1n5DD2RJmrY_UbXcV5zcCBULSYaSWyQ4B?usp=sharing ;
2. Avoid downloading and uploading any example raw spectra data due to the limited bandwidth.
3. Default MetaboAnalyst includes www.metaboanalyst.ca , new.metaboanalyst.ca and dev.metaboanalyst.ca. please use either of them.

# Schedule

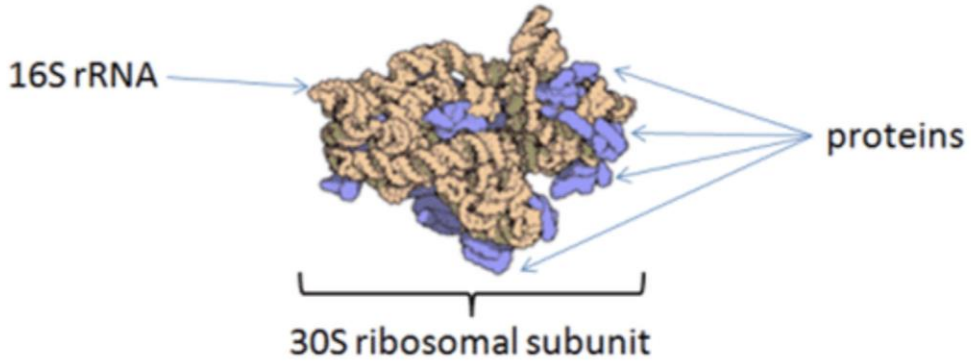
| Time | Topics | Lab practices |
|---|---|---|
| 12:00 – 12:15 | General introduction | |
| 12:15 – 1:00 | Metabolomics data processing | Live demo 1 & hands on |
| 1:00 – 1:45 | Microbiome data processing | Live demo 2 & hands on |
| 1:45 – 3:00<br>(15 min break) | Microbial community profiling | Live demo 3 & hands on |
| 3:00 – 3:50 | Microbiome-metabolomics integration | Live demo 4 & hands on |
| 3:50 – 4:15 | Summary and discussion | |

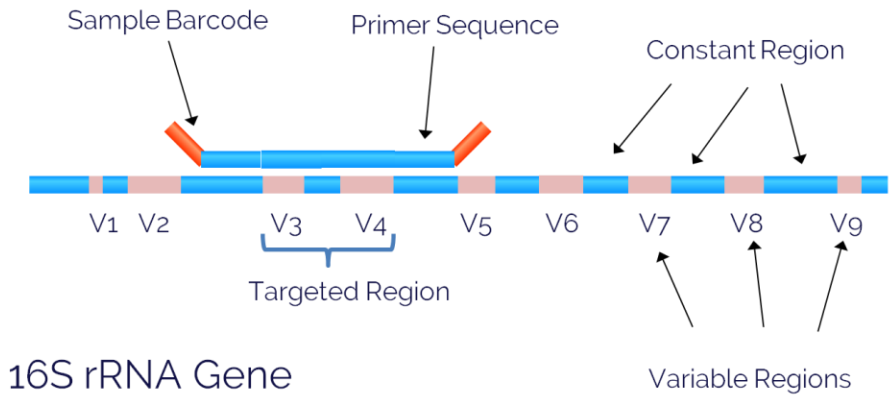
# Sequencing microbiome



# 16S RNA marker gene (bacteria and archaea)

- Containing regions that are highly conserved between different species of bacteria and archaea
- Whereas the rest of genetic content varies greatly across species
- Predict functional profiles for well-studies communities with reference genomes

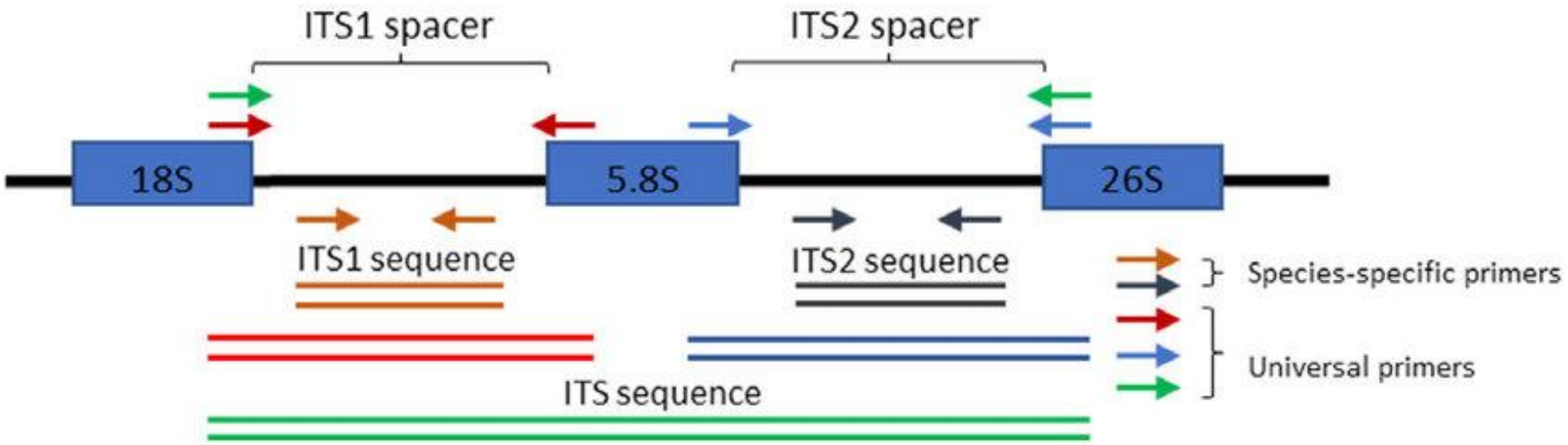


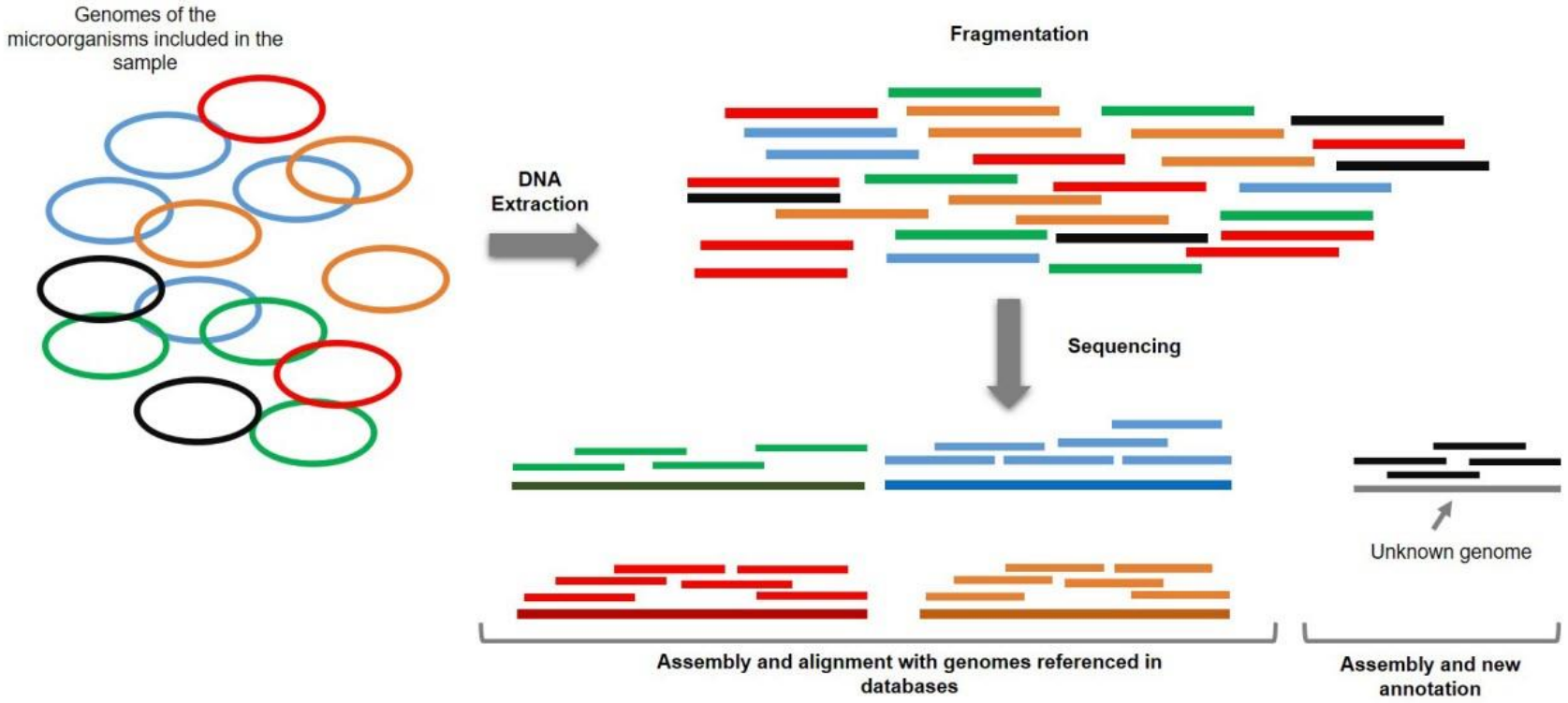

Chen et al. 2012 *Bioinformatics.*

# 18S / ITS region gene sequencing (fungi)

- (Conserved) 18S is considered a potential biomarker for fungi classification above the species level for wide phylogenetic analyses and environmental biodiversity screenings
- (Less conserved) ITS region tends to be hypervariable between fungal species. Good for survey for genetic diversity at the species level



# Shotgun metagenomics (multiple markers and functions)




france-genomique.org

Many researchers choose 16S sequencing for large studies, followed by deep shotgun sequencing on a subset of targeted samples

# Processing microbiome sequencing data

- DADA2 (single marker)
  - Marker gene sequencing data => Taxonomy
- MetaPhlan (multiple markers)
  - Shotgun metagenomics data => Taxonomy
- HUMAnN
  - Shotgun metagenomics data => **Function**
- **FunTaxSeq**
  - Shotgun metagenomics data => Taxonomy **& Function**

# From Raw Reads to Unit of Analysis

**Phylotypes**

- Sequences are directly assigned to taxa on the basis on known reference sequences

**Operational Taxonomic Unit (OTU)**

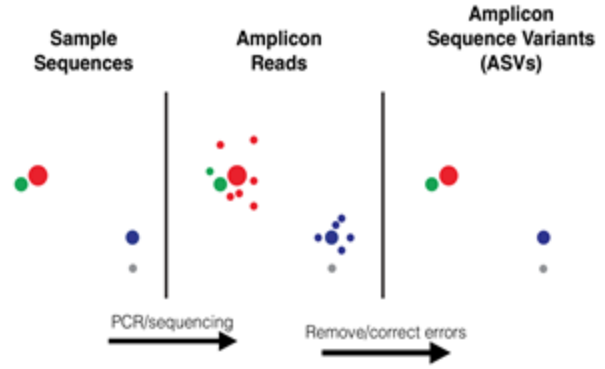
- Based on clustering (similarity to other sequences in the same dataset)
- Usually OTUs (~species) are delineated with a 3% sequence dissimilarity, and higher taxa with increasingly larger dissimilarity
- Data specific & computing intensive

**Amplicon Sequence Variant (ASV)**

- Exact sequences without clustering
- Universal
- Fast

# Amplicon Sequence Variants (ASV)

- Exact sequences, generated without clustering or reference databases
- Comparison of similar reads to determine the probability that a given read at a given frequency is not due to sequencer error using an error model for sequencing run
- A given target gene sequence (i.e. based on the primers) should always generate the same ASV
- ASV can be compared to a reference database at a much higher resolution
- DADA2 is among the best performers for ASV




| | ASVs | De novo | Closed-ref |
|---|---|---|---|
| Precise | ✓ | ~ | ~ |
| Tractable | ✓ | ~ | ✓ |
| Reproducible | ✓ | X | ✓ |
| Comprehensive | ✓ | ✓ | X |

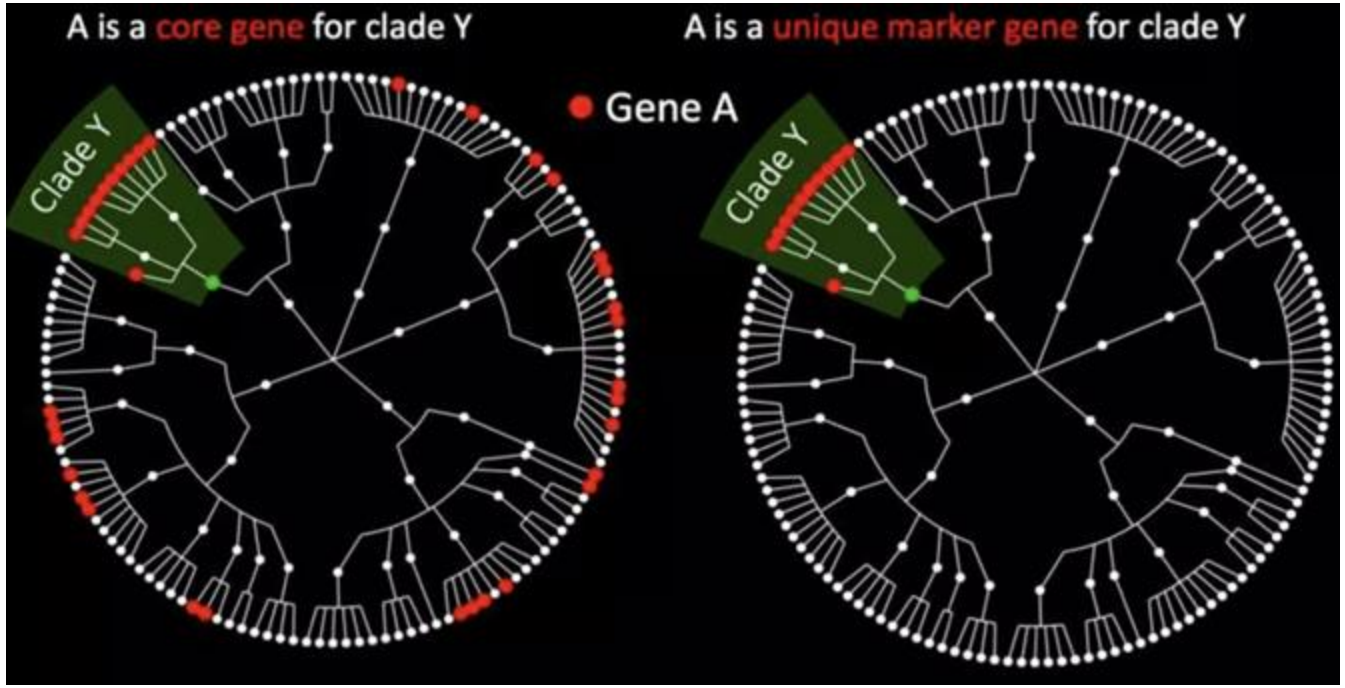
DADA2: Fast and accurate sample inference from amplicon data with single-nucleotide resolution

DADA2

Amplicon Sequencing. **Exactly.**

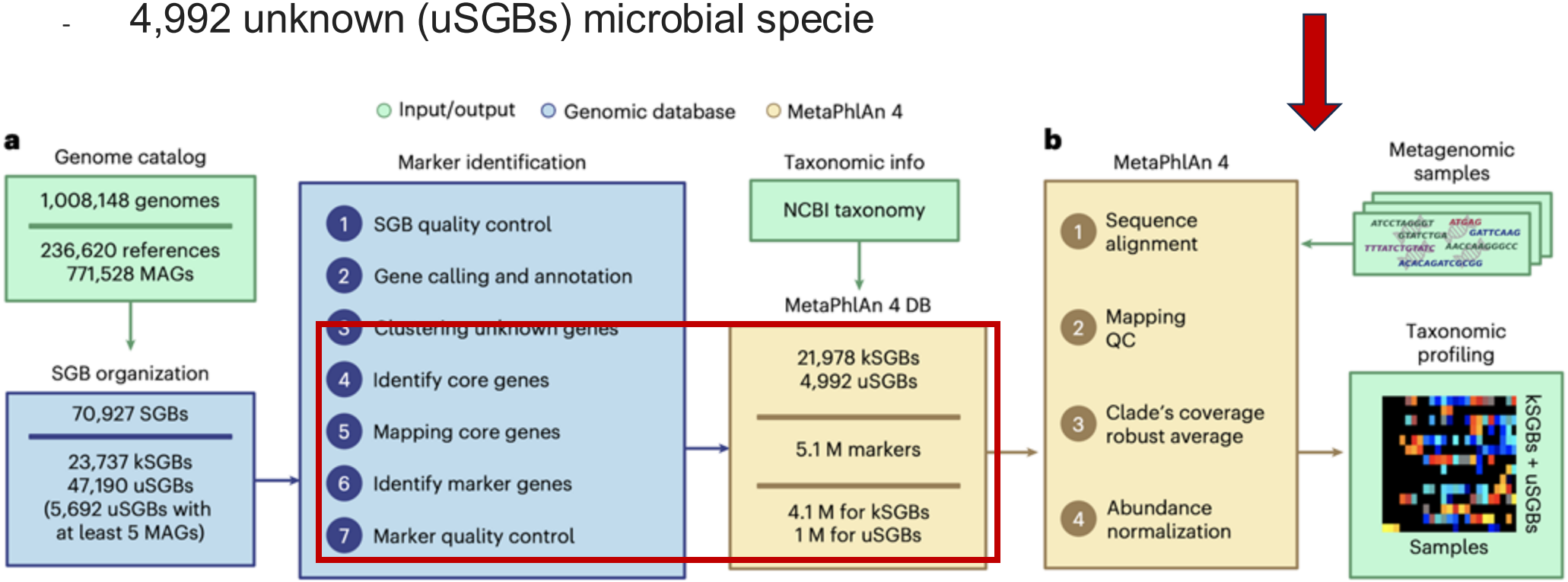
# From single to multiple marker genes

- Singe gene (universal markers)
  - 16S / 18S / ITS
- Multiple genes (clade-specific markers)



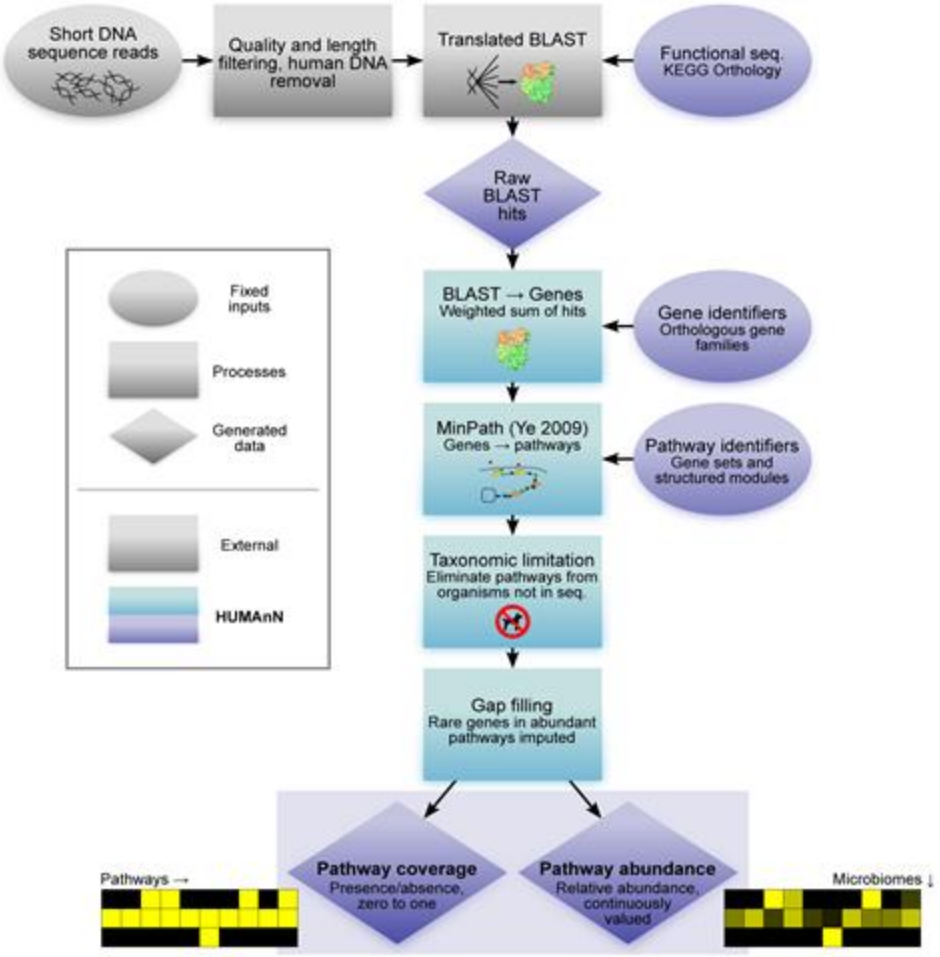
# Tool: MetaPhlan

- Clade-specific marker genes
- Species-level genome bins (SGBs) based on assembly
  - 21,978 known (kSGBs)
  - 4,992 unknown (uSGBs) microbial specie




https://www.nature.com/articles/s41587-023-01688-w

# Tool: HUMAnN Pipeline




## Major updates in HUMAnN 3.0

- Designed in tandem with MetaPhlAn 3.0.
- Based on UniProt/UniRef 2019_01 sequences and annotations.
- Contains 2x more species pangenomes and 3x more gene families (vs. HUMAnN 2.0).
- Incorporates MetaCyc v24.0 pathway definitions.
- Removed version number from executables (call metaphlan and humann).
- Pangenome sequences must be covered at >50% of sites to be reported (tunable).
- Additional accuracy and performance re-tuning across all search steps.

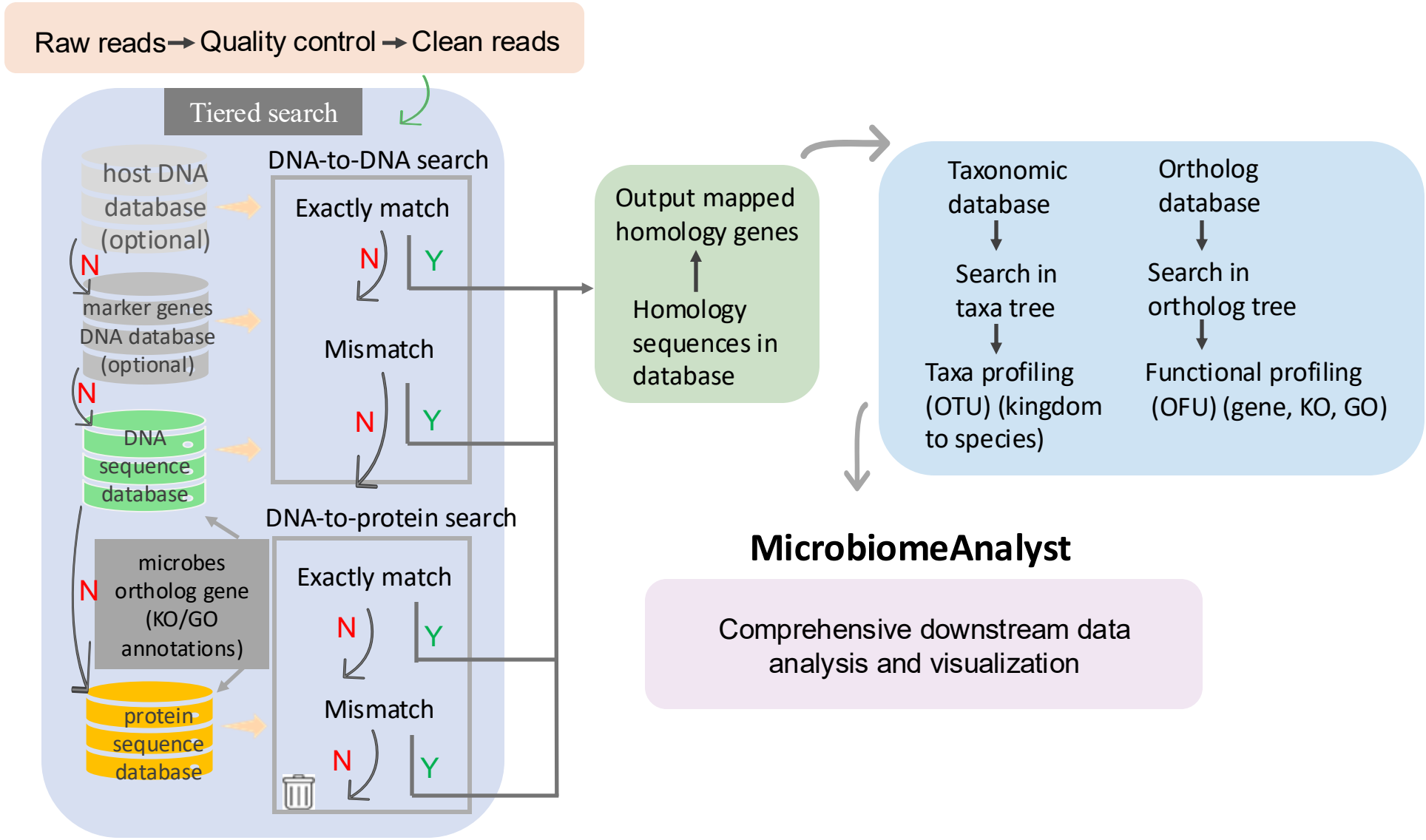
https://github.com/biobakery/humann

# FunTaxSeq: simultaneous taxonomic & functional profiling of shotgun metagenomics

- A unified and versatile platform for microbiome data
  - metagenomics, metatranscriptomics, amplicon
- All-in-one, standalone software (high I/O efficiency)
  - raw reads to quality control to taxa & function tables
- Phylogenetic/hierarchy tree-based assignment
  - taxonomic classification and functional profiling

Dr. Peng Liu, Environment and Climate Change Canada (ECCC)

# FunTaxSeq workflow



# FunTaxSeq Databases

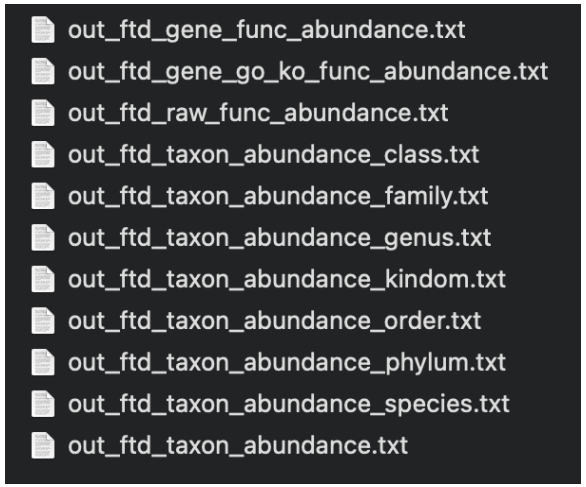
- ~70 M gene and protein sequences from ~18 K prokaryotes from orthoDB 12
- Taxa tree: ~17k; ortholog tree: ~3M

**GitHub:**

- https://github.com/rocpengliu/FunTaxSeq
- Under benchmarking
- Will be released late this summer

# Outputs of FunTaxSeq

- Taxonomic ID
- Gene id, gene name,
- GO, KO



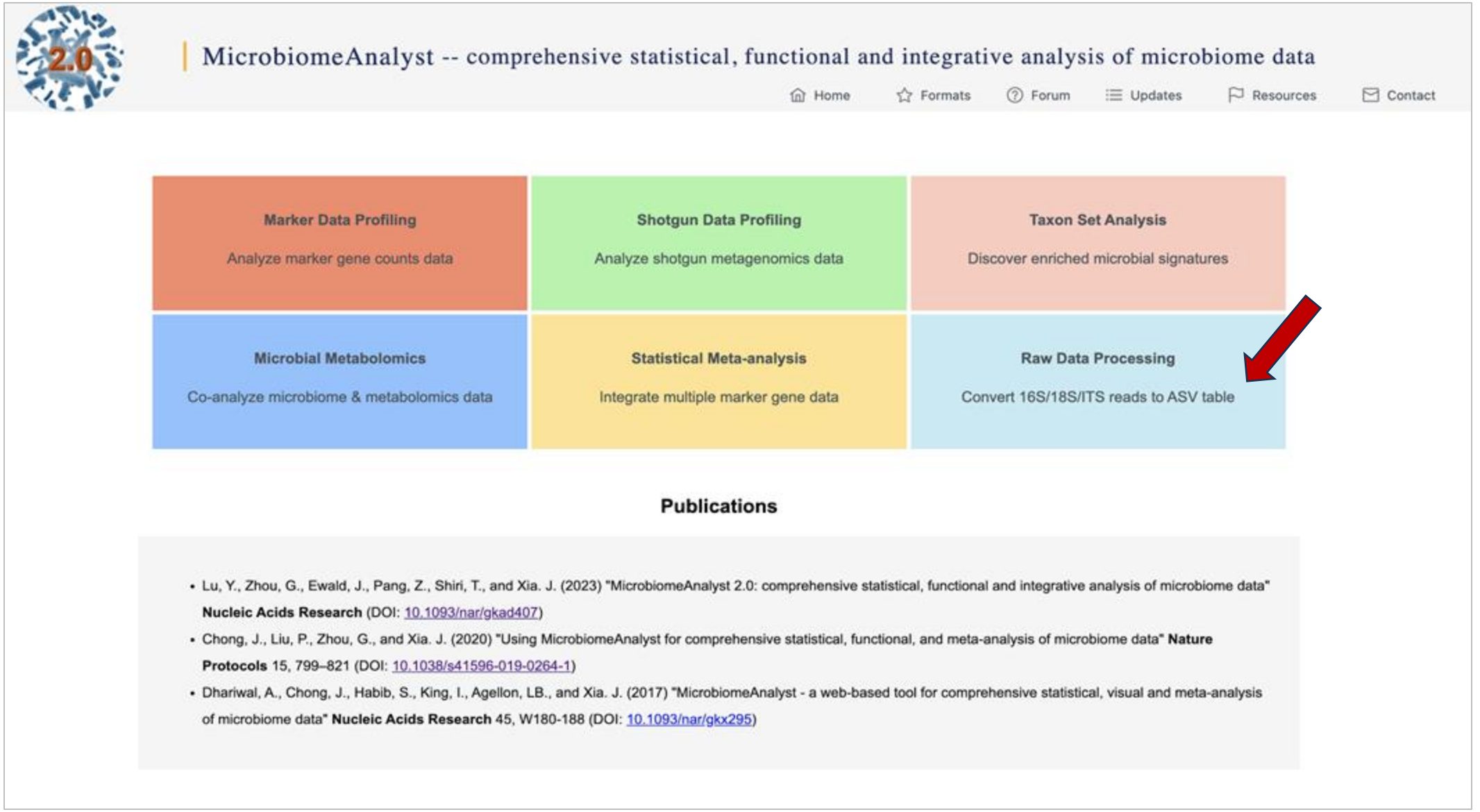

out_ftd_gene_func_abundance

| #ortholog | GO | KO | gene_size | sim_rep1 | sim_rep2 | sim_rep3 |
|---|---|---|---|---|---|---|
| (2e,6e)-farnesyl diphosphate synthase | GO:0004659;GO:0008299;GO:0016740 | K00795 | 894 | 0 | 3 | 1 |
| (2fe-2s)-binding protein | 0 | 0 | 315 | 0 | 1 | 0 |
| (2fe-2s)-binding protein | 0 | K03518 | 507 | 0 | 0 | 1 |
| (2fe-2s)-binding protein | GO:0005829;GO:0051536;GO:0051537;GO:0140647 | K04755 | 321 | 0 | 1 | 0 |
| (2fe-2s)-binding protein | GO:0016491;GO:0046872;GO:0051536 | K03518 | 487 | 2 | 4 | 1 |
| (2fe-2s)-binding protein | GO:0016491;GO:0046872;GO:0051536 | K03518;K03519 | 492 | 0 | 0 | 2 |
| (2fe-2s)-binding protein | GO:0016491;GO:0046872;GO:0051536;GO:0051537 | K03518 | 491 | 3 | 2 | 0 |
| (2fe-2s)-binding protein | GO:0016491;GO:0046872;GO:0051536;GO:0051537 | K07302 | 492 | 1 | 2 | 0 |
| (2fe-2s)-binding protein | GO:0016491;GO:0051536 | K22550 | 297 | 1 | 0 | 0 |
| (2fe-2s)-binding protein | GO:0016491;GO:0051536;GO:0051537 | 0 | 306 | 0 | 0 | 1 |
| (3r)-hydroxyacyl-acp dehydratase subunit hadb | 0 | 0 | 429 | 0 | 2 | 0 |
| (4fe-4s)-binding protein | 0 | 0 | 1032 | 1 | 0 | 0 |
| (d)cmp kinase | 0 | K00945 | 672 | 5 | 2 | 3 |
| (e)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase | 0 | K03526 | 1998 | 2 | 0 | 1 |
| (fe-s)-binding protein | 0 | 0 | 1620 | 1 | 0 | 0 |
| (fe-s)-binding protein | 0 | K08264 | 790 | 1 | 1 | 0 |
| (fe-s)-binding protein | 0 | K18928 | 717 | 1 | 1 | 4 |
| (fe-s)-binding protein | GO:0016020;GO:0046872;GO:0051536 | 0 | 1326 | 1 | 2 | 2 |
| (fe-s)-binding protein | GO:0016020;GO:0046872;GO:0051536 | K08264 | 3039 | 3 | 0 | 0 |
| (fe-s)-binding protein | GO:0016020;GO:0046872;GO:0051536 | K21834 | 1924 | 1 | 2 | 1 |
| (fe-s)-binding protein | GO:0016491 | K08264 | 792 | 1 | 0 | 0 |
| (fe-s)-binding protein | GO:0051536 | K08264 | 1152 | 2 | 0 | 1 |
| (r)-citramalate synthase | GO:0003824;GO:0003852;GO:0009082;GO:0009097;GO:000909 | K01649 | 1587 | 3 | 3 | 2 |

out_ftd_taxon_abundance

| #taxon | genome_size | sim_rep1 | sim_rep2 | sim_rep3 |
|---|---|---|---|---|
| k__Archaea;p__Methanobacteriota;c__Halobacteria;o__Halobacteriales | 3665414 | 7 | 2 | 5 |
| k__Archaea;p__Methanobacteriota;c__Halobacteria;o__Halobacteriales;f__Haladaptataceae;g__Haladaptatus;s__Haladaptatus cibarius | 3926724 | 0 | 1 | 0 |
| k__Archaea;p__Methanobacteriota;c__Halobacteria;o__Halobacteriales;f__Haladaptataceae;g__Halorussus;s__Halorussus amylolyticus | 3660772 | 0 | 1 | 0 |
| k__Archaea;p__Methanobacteriota;c__Halobacteria;o__Halobacteriales;f__Haladaptataceae;g__Halorussus;s__Halorussus halobius | 3933954 | 1 | 0 | 0 |
| k__Archaea;p__Methanobacteriota;c__Halobacteria;o__Halobacteriales;f__Haladaptataceae;g__Halorussus;s__Halorussus pelagicus | 3490924 | 1 | 3 | 1 |
| k__Archaea;p__Methanobacteriota;c__Halobacteria;o__Halobacteriales;f__Haladaptataceae;g__Halorussus;s__Halorussus ruber | 4318307 | 2 | 2 | 0 |
| k__Archaea;p__Methanobacteriota;c__Halobacteria;o__Halobacteriales;f__Haladaptataceae;g__Halorussus;s__Halorussus salinus | 4991811 | 0 | 1 | 0 |
| k__Archaea;p__Methanobacteriota;c__Halobacteria;o__Halobacteriales;f__Haladaptataceae;g__Halorussus;s__Halorussus sp. MSC15.2 | 4408556 | 1 | 0 | 0 |
| k__Archaea;p__Methanobacteriota;c__Halobacteria;o__Halobacteriales;f__Haloarculaceae;g__Halapricum;s__Halapricum salinum | 3451774 | 0 | 1 | 0 |
| k__Archaea;p__Methanobacteriota;c__Halobacteria;o__Halobacteriales;f__Haloarculaceae;g__Haloarcula;s__Haloarcula amylovorans | 5038561 | 1 | 1 | 0 |
| k__Archaea;p__Methanobacteriota;c__Halobacteria;o__Halobacteriales;f__Haloarculaceae;g__Haloarcula;s__Haloarcula hispanica | 3890005 | 1 | 0 | 0 |
| k__Archaea;p__Methanobacteriota;c__Halobacteria;o__Halobacteriales;f__Haloarculaceae;g__Haloarcula;s__Haloarcula japonica | 4280359 | 1 | 0 | 0 |
| k__Archaea;p__Methanobacteriota;c__Halobacteria;o__Halobacteriales;f__Haloarculaceae;g__Haloarcula;s__Haloarcula rubripromontorii | 4008530 | 1 | 3 | 1 |
| k__Archaea;p__Methanobacteriota;c__Halobacteria;o__Halobacteriales;f__Haloarculaceae;g__Haloarcula;s__Haloarcula salina | 3793362 | 0 | 1 | 0 |
| k__Archaea;p__Methanobacteriota;c__Halobacteria;o__Halobacteriales;f__Haloarculaceae;g__Haloarcula;s__Haloarcula sp. CBA1127 | 4359264 | 1 | 4 | 3 |
| k__Archaea;p__Methanobacteriota;c__Halobacteria;o__Halobacteriales;f__Haloarculaceae;g__Haloarcula;s__Haloarcula sp. CBA1129 | 4048546 | 0 | 3 | 1 |
| k__Archaea;p__Methanobacteriota;c__Halobacteria;o__Halobacteriales;f__Haloarculaceae;g__Haloarcula;s__Haloarcula vallismortis | 3923205 | 2 | 5 | 0 |
| k__Archaea;p__Methanobacteriota;c__Halobacteria;o__Halobacteriales;f__Haloarculaceae;g__Halomarina;s__Halomarina oriensis | 4069707 | 0 | 1 | 1 |
| k__Archaea;p__Methanobacteriota;c__Halobacteria;o__Halobacteriales;f__Haloarculaceae;g__Halomicrobium;s__Halomicrobium mukohata | 3337602 | 3 | 0 | 0 |
| k__Archaea;p__Methanobacteriota;c__Halobacteria;o__Halobacteriales;f__Haloarculaceae;g__Halorientalis;s__Halorientalis pallida | 4087668 | 0 | 1 | 0 |
| k__Archaea;p__Methanobacteriota;c__Halobacteria;o__Halobacteriales;f__Haloarculaceae;g__Halorientalis;s__Halorientalis regularis | 4032076 | 0 | 0 | 1 |

# Demo 2 – marker data processing



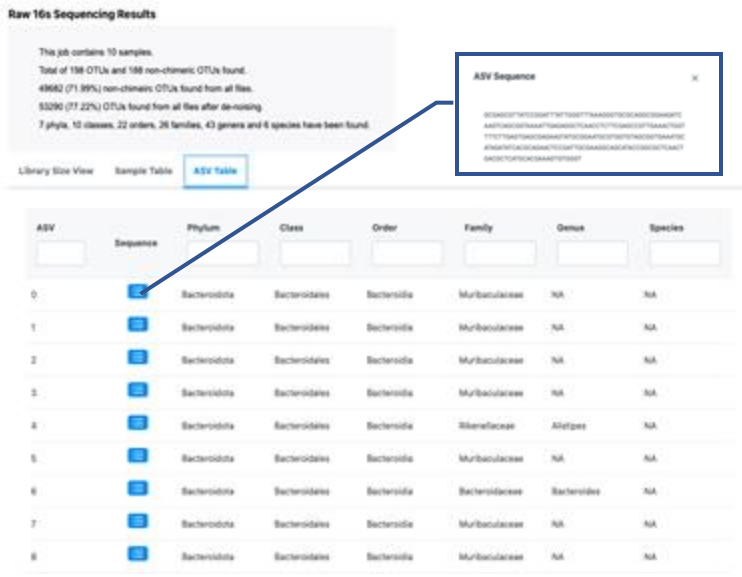

**https://www.microbiomeanalyst.ca**

# Raw data processing (16S)

Raw data input → Trimming and quality control → Processing and Annotation → MDP

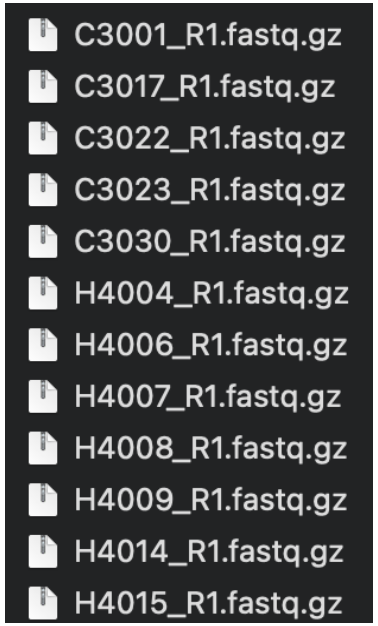
**Raw data**





# Input file for raw data processing:

- Sequencing data uploaded as individual zip/fastq.gz files - one zip per data (no larger than **100 MB**) [max: 200 files].
- The original fastq files (before compress) must end with R1.fastq/R2.fastq.
- Metadata uploaded as a plain text (.txt) file containing multiple columns - files names, group labels and other experiment factors




| sample | diagnosis | consent_age | sex |
|---|---|---|---|
| M2008_R1.fastq.gz | CD | 30 | Female |
| M2025_R1.fastq.gz | CD | 43 | Female |
| M2027_R1.fastq.gz | CD | 41 | Male |
| M2028_R1.fastq.gz | CD | 24 | Female |
| C3001_R1.fastq.gz | CD | 43 | Female |
| C3017_R1.fastq.gz | CD | 45 | Male |
| C3022_R1.fastq.gz | nonIBD | 69 | Male |
| C3023_R1.fastq.gz | CD | 60 | Male |
| C3030_R1.fastq.gz | CD | 44 | Male |
| H4004_R1.fastq.gz | CD | 14 | Male |
| H4006_R1.fastq.gz | CD | 8 | Male |
| H4008_R1.fastq.gz | nonIBD | 13 | Female |
| H4007_R1.fastq.gz | CD | 15 | Female |
| H4009_R1.fastq.gz | nonIBD | 6 | Female |
| H4014_R1.fastq.gz | CD | 10 | Female |
| H4015_R1.fastq.gz | CD | 15 | Male |
| H4016_R1.fastq.gz | nonIBD | 10 | Female |

**Example data:** Single-end FASTQ files derived from a 2×250 bp Illumina MiSeq paired-end run targeting the V4 region of the 16S rRNA gene.

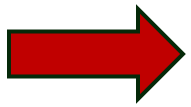
# Hands on Practices (10 min)

Demo sequence data can be found:

https://drive.google.com/drive/folders/1HbCeMeh1CIYWyCaz9QeIxQdupwMMLOyP?usp=drive_link

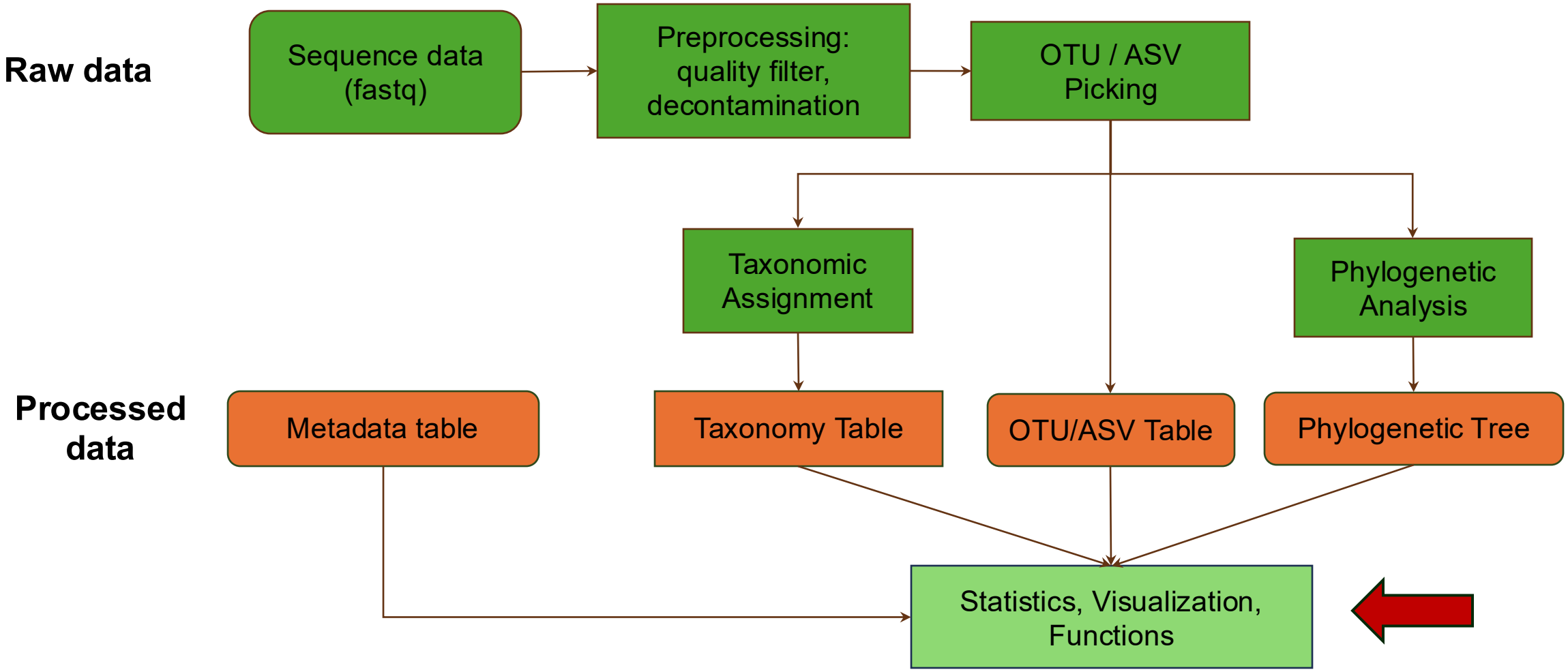
Try other example data on our website for learning purpose.

# Schedule



| Time | Topics | Lab practices |
|---|---|---|
| 12:00 – 12:15 | General introduction | |
| 12:15 – 1:00 | Metabolomics data processing | Live demo 1 & hands on |
| 1:00 – 1:45 | Microbiome data processing | Live demo 2 & hands on |
| 1:45 – 3:00<br>(15 min break) | Microbial community profiling | Live demo 3 & hands on |
| 3:00 – 3:50 | Microbiome-metabolomics integration | Live demo 4 & hands on |
| 3:50 – 4:15 | Summary and discussion | |

# Bioinformatics workflow for marker data

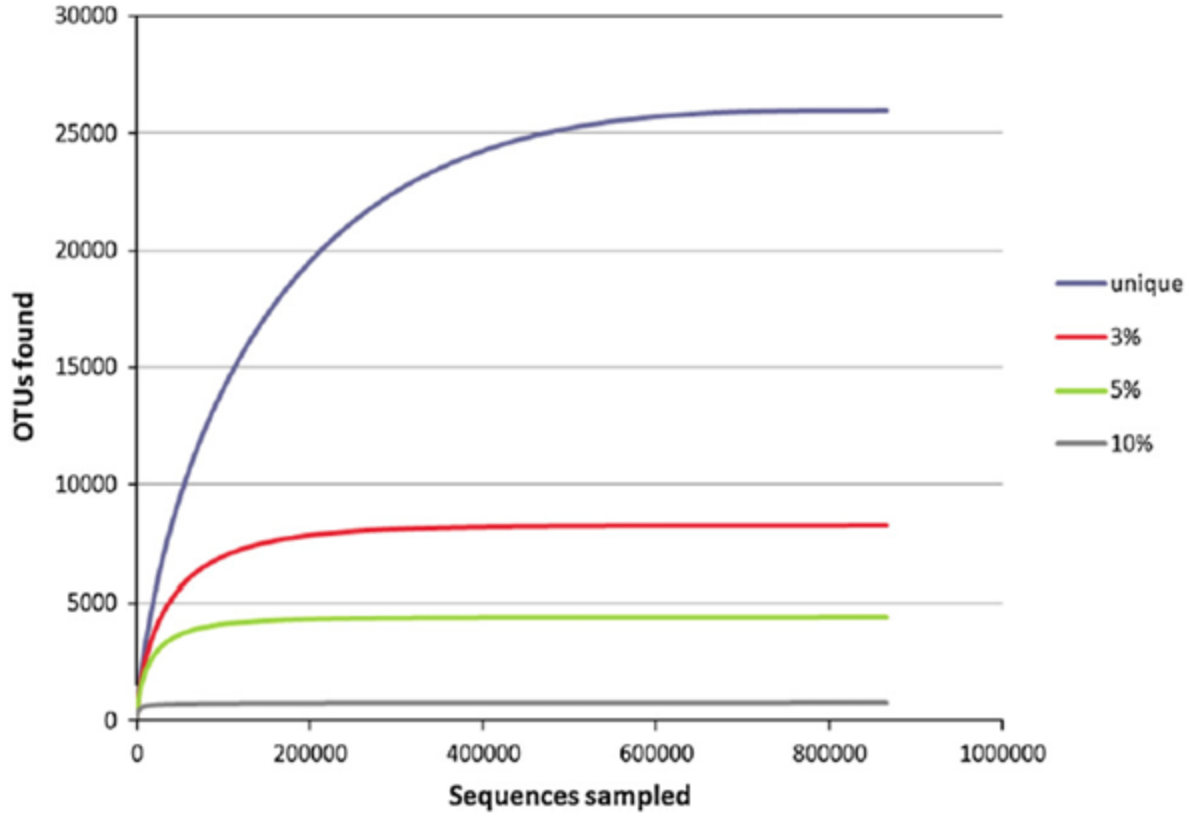


# Community Composition & Diversities

1. Overview of community composition
   - Descriptive visualization methods
2. Alpha diversity (within sample)
   - Number (richness) and distribution (evenness) of taxa expected within a single sample.
3. Beta diversity (between samples)
   - Describes taxonomic variations that are shared between samples

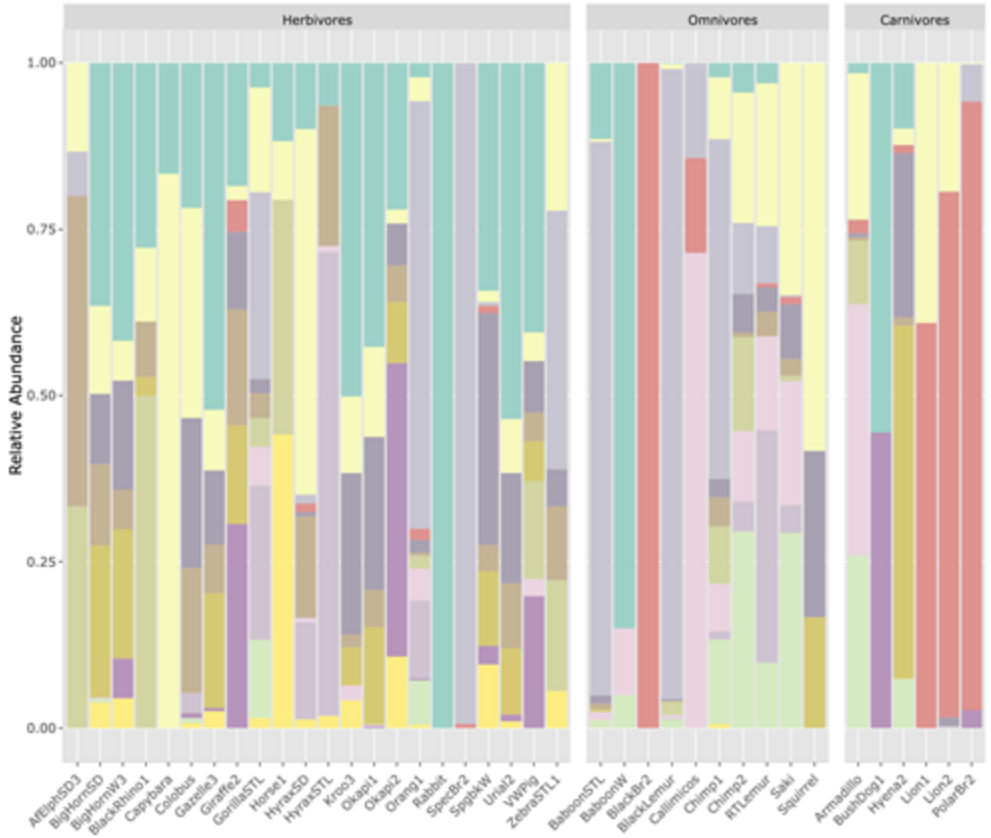
# Richness - Rarefaction Curve

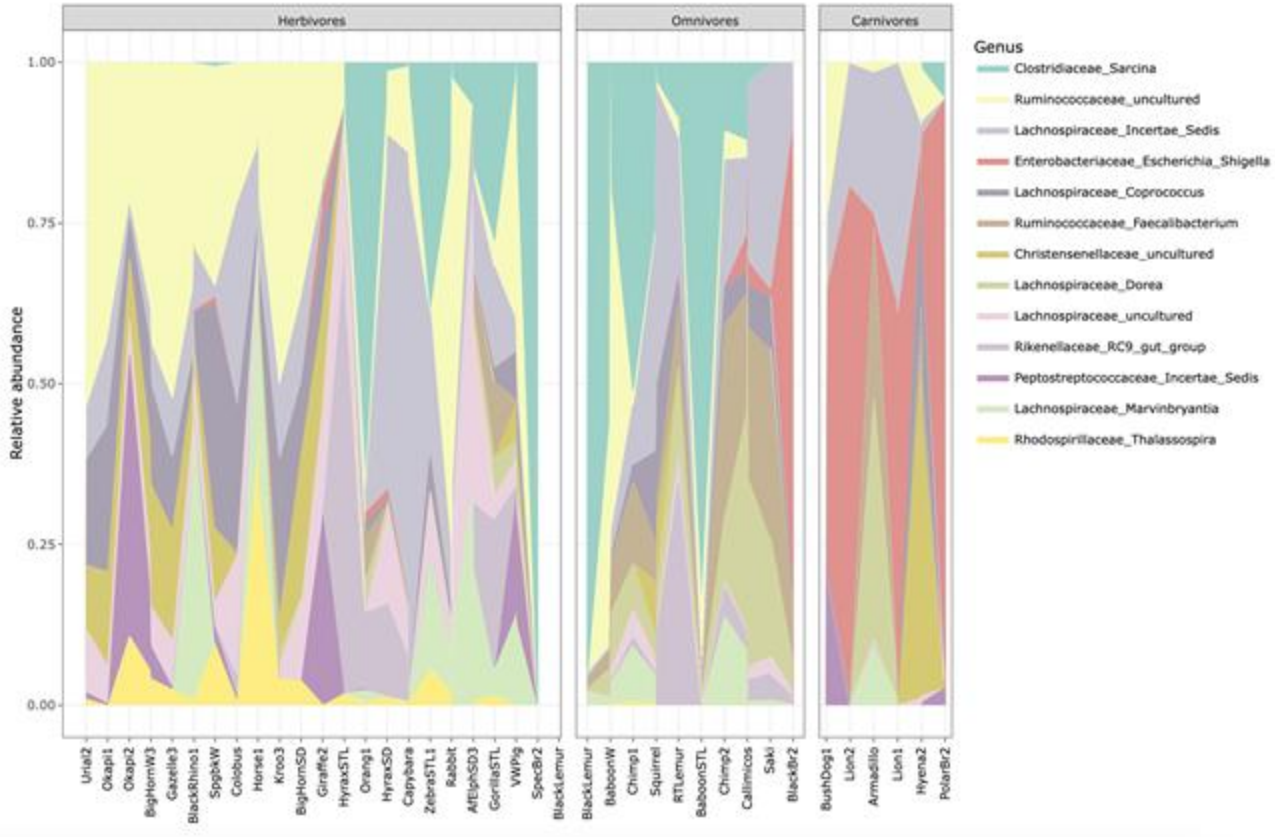
- Assess species richness from the results of sampling
- Evaluate the completeness of a sample and explore biodiversity
  - Increasing numbers of sequenced taxa allow increasingly precise estimates of total population diversity



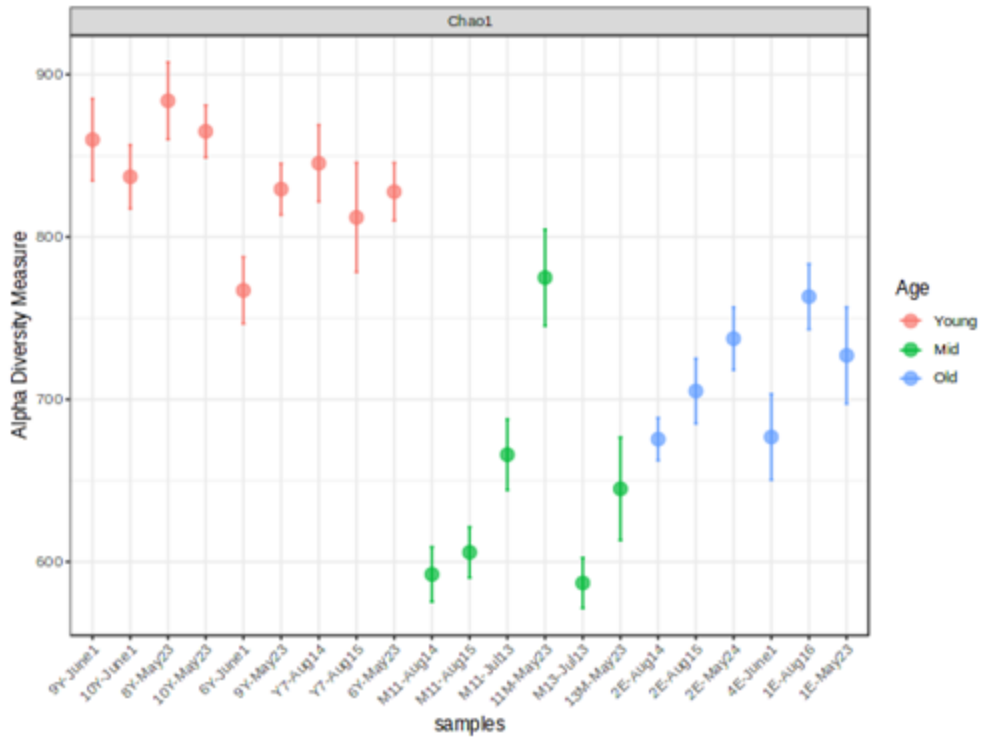
# Stacked bar / area plot

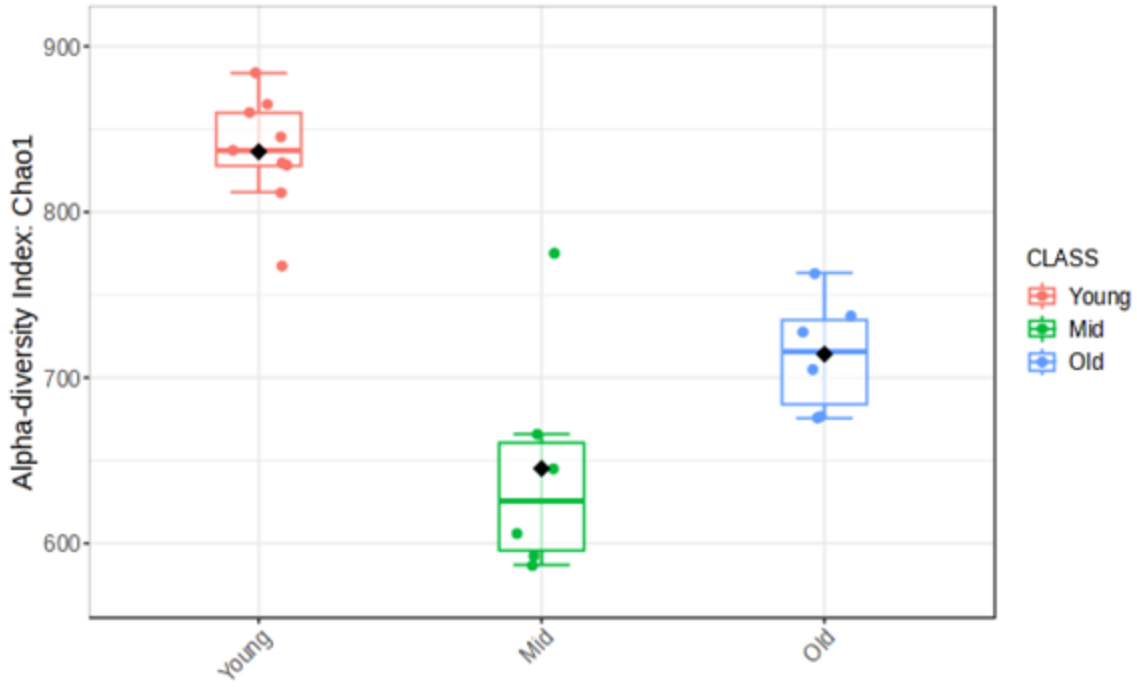
- Visualizing composition across individual samples



# Alpha diversity visualization (chao1)




Comparing different alpha diversity measures across samples & groups

# Alpha diversity - richness & evenness

Same richness, yet different evenness

## Chao1 Estimator

$$S_{chao1} = S_{obs} + \frac{n_1 (n_1 - 1)}{2 (n_2 + 1)}$$

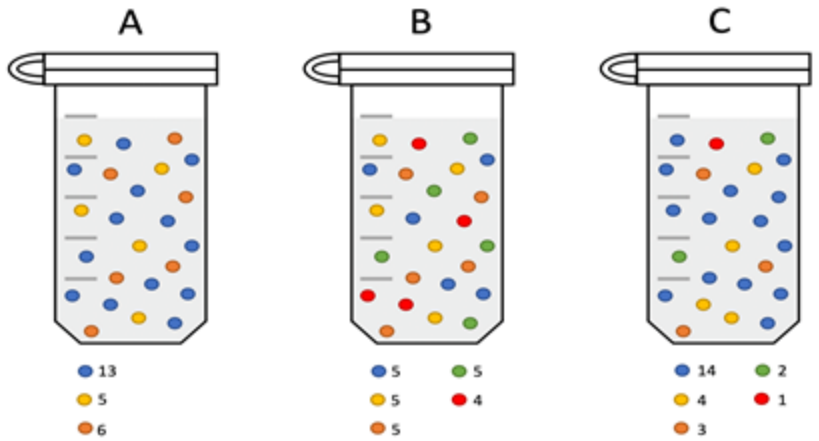
## Shannon Diversity

$$H = -\sum_{j=1}^{S} p_i \ln p_i$$

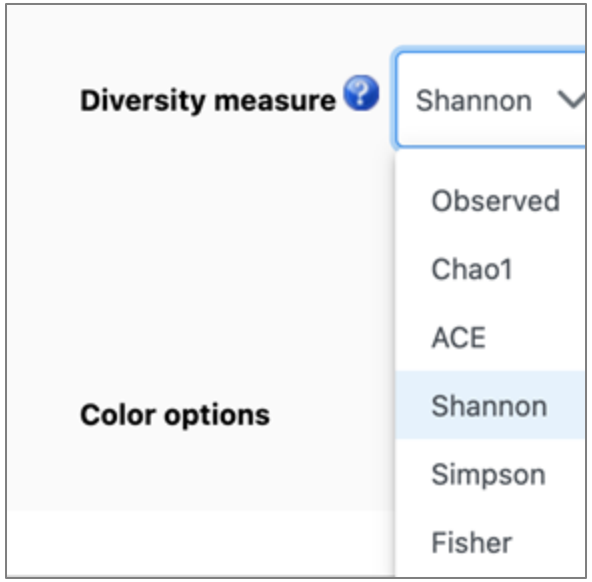
## Simpson Diversity

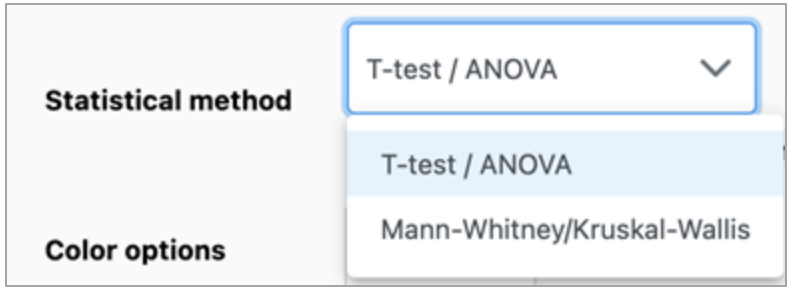
$$D = \frac{N(N-1)}{\sum n(n-1)}$$

# Test differences in alpha diversity




| | A | B | C |
|---|---|---|---|
| **Chao1** | 3 | 5 | 5 |
| **Shannon** | 1.005 | **1.606** | 1.212 |
| **Simpson** | 0.601 | **0.799** | 0.608 |




https://docs.onecodex.com/en/articles/4136553-alpha-diversity

# Beta Diversity measures (I)

Describe shared taxa between samples (relative to total counts)

- Jaccard index
- Bray-Curtis distance (BC)

$$Jaccard_{AB} = 1 - \frac{|A \cap B|}{|A \cup B|}$$

$$BC_{ij} = 1 - \frac{2C_{ij}}{S_i + S_j}$$

where $C_{ij}$ is the sum of the lesser values for species in common between the two samples, and $S_i$ and $S_j$ are the total number of microbes found in samples *i* and *j*, respectively.

# Beta Diversity measures (II) - UniFrac

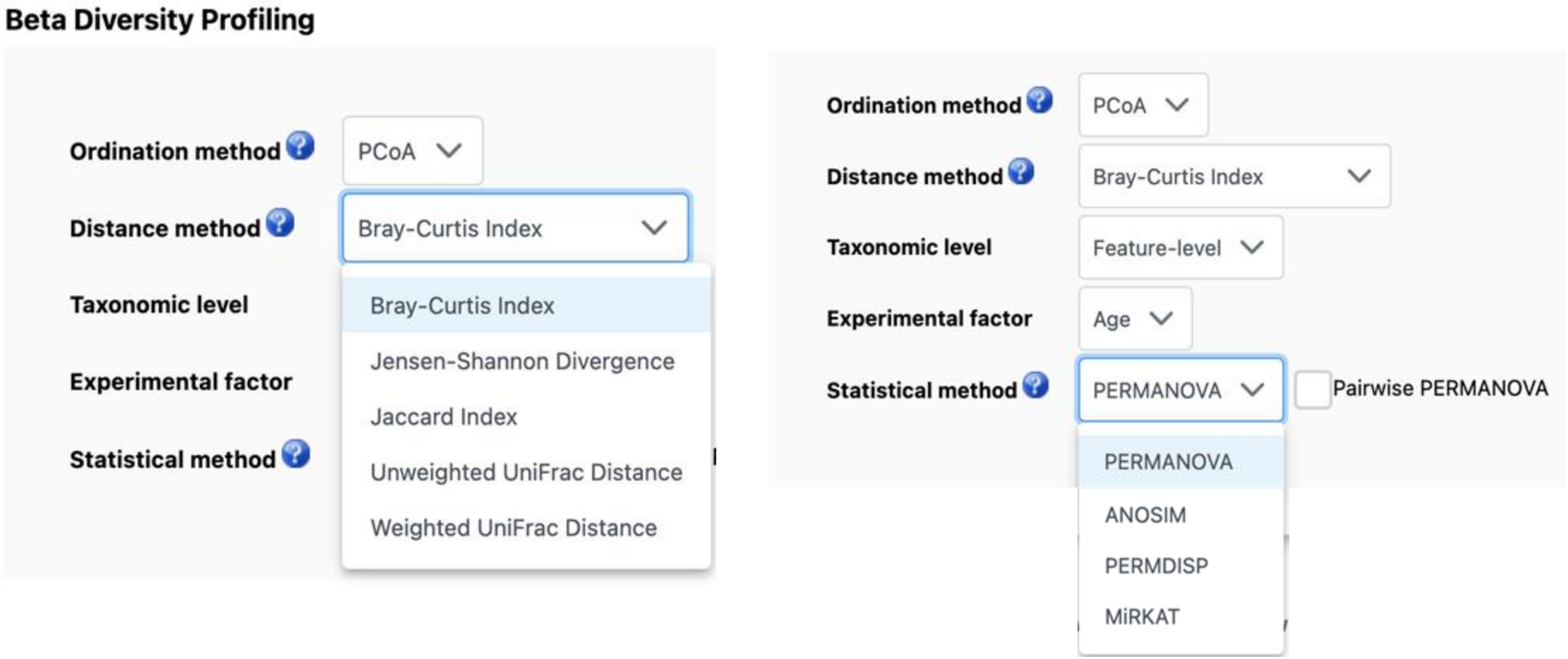
Incorporating phylogenetic distances between observed organisms in the computation

- Unweighted UniFrac only consider the presence or absence of the species,
- Weighted UniFrac consider the actual abundance.

$$\left(\frac{sum\ of\ unshared\ branch\ lengths}{sum\ of\ all\ tree\ branch\ lengths}\right) = fraction\ of\ total\ unshared\ branch\ lengths$$

*Bioinformatics*. **28** (16): 2106–2113.

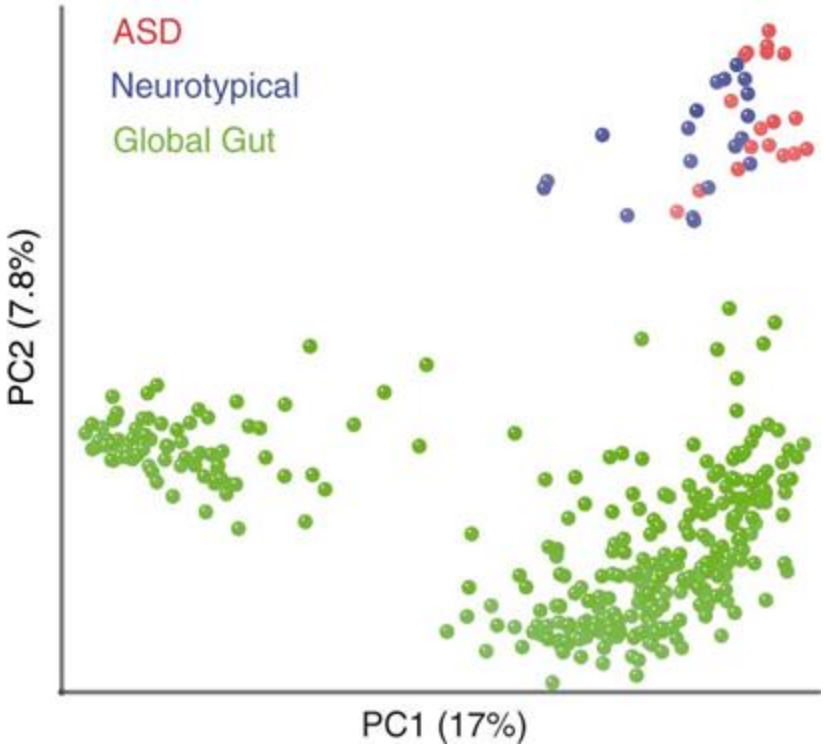
# Common beta diversity measures & tests



# Visualizing Beta Diversity

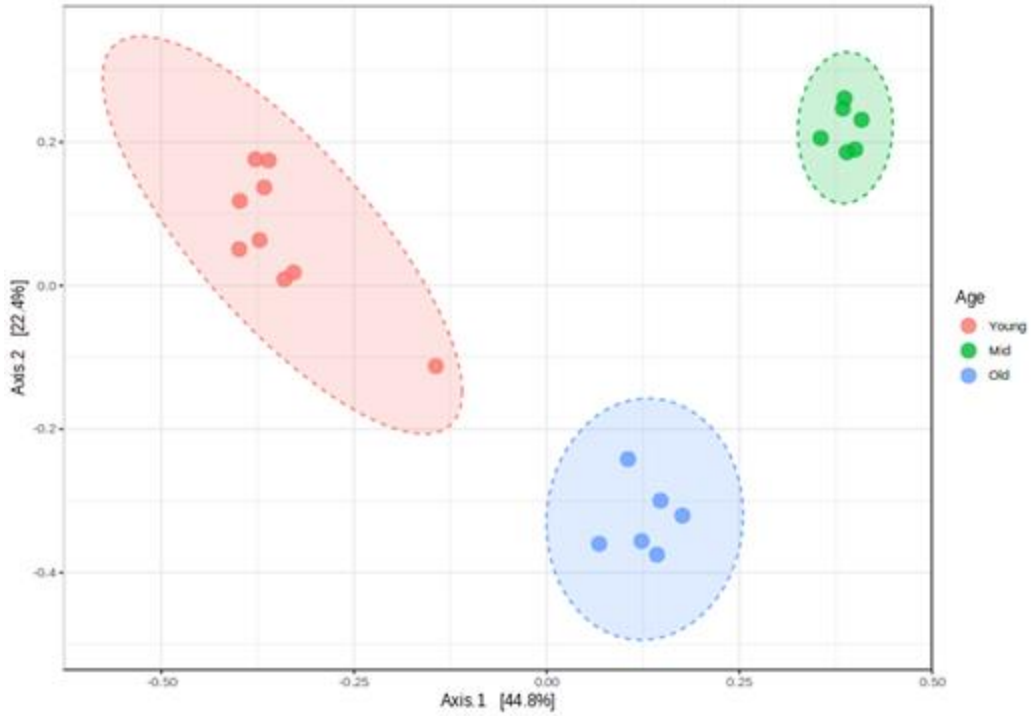
## Principal Coordinates Analysis (PCoA)

- Similar to principal component analysis (PCA), but use ecological distance measures
- If we use Euclidean distance matrix to a PCoA, PCoA and PCA are essentially identical

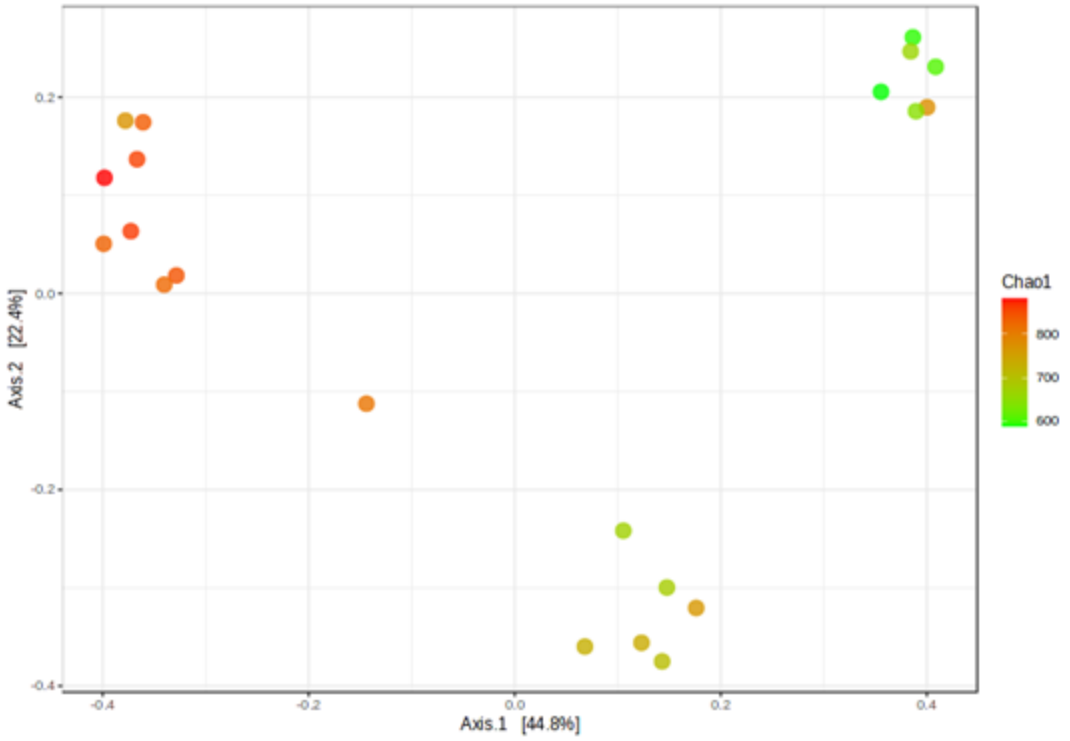



Microbial Ecology in Health & Disease 2015, 26: 26914

# View alpha & beta-diversity together

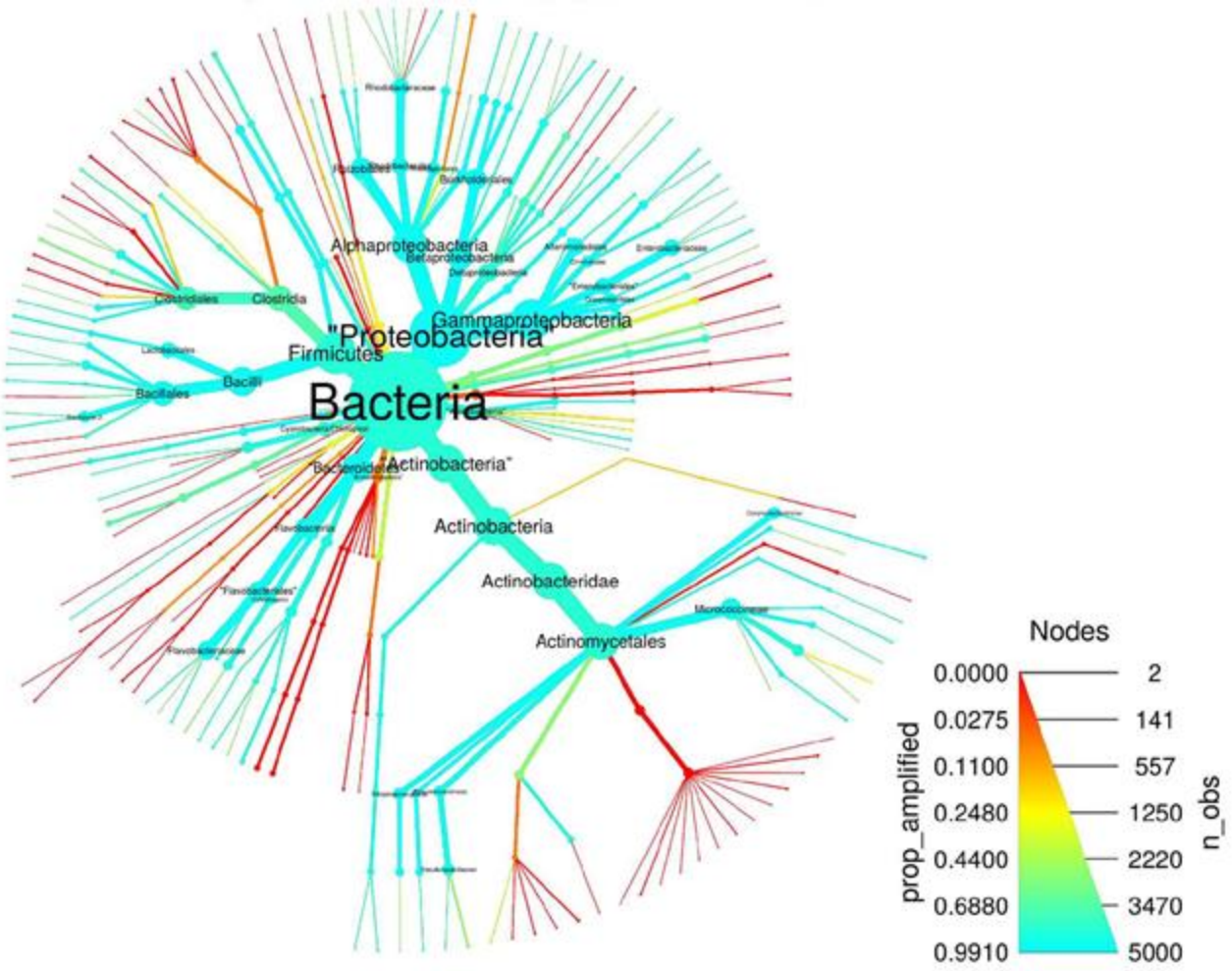



Coloring based on groups



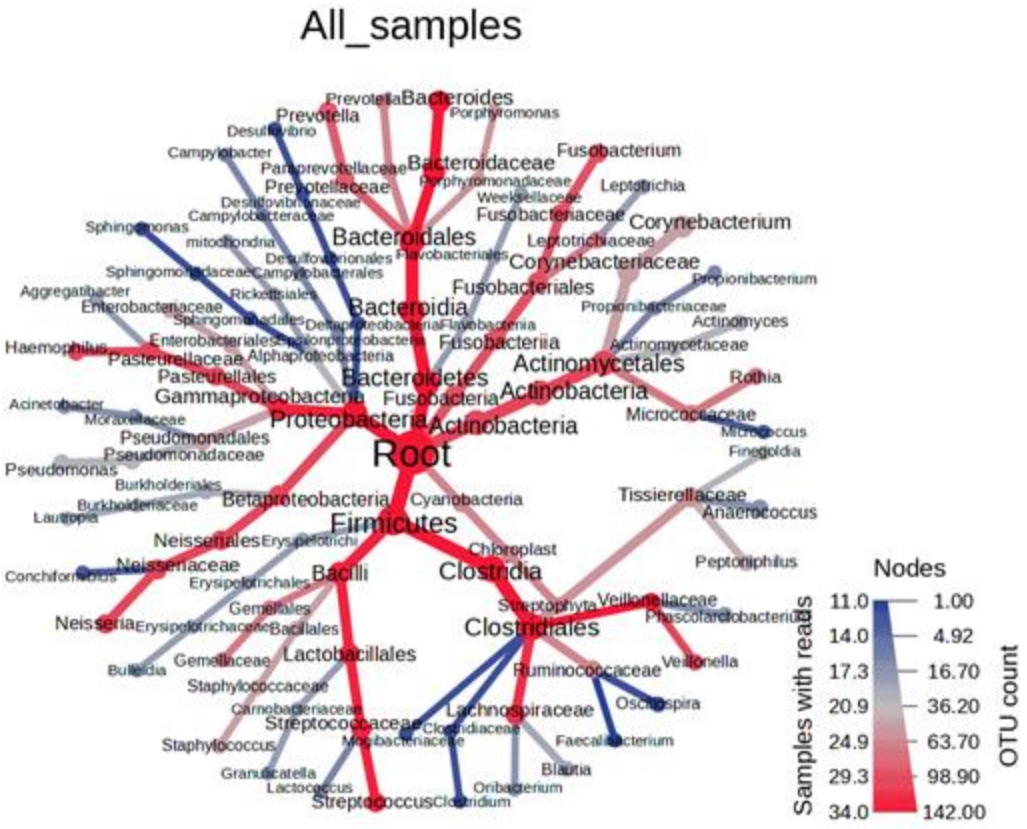

Overlay alpha diversity

# Heat tree visualization of community diversity

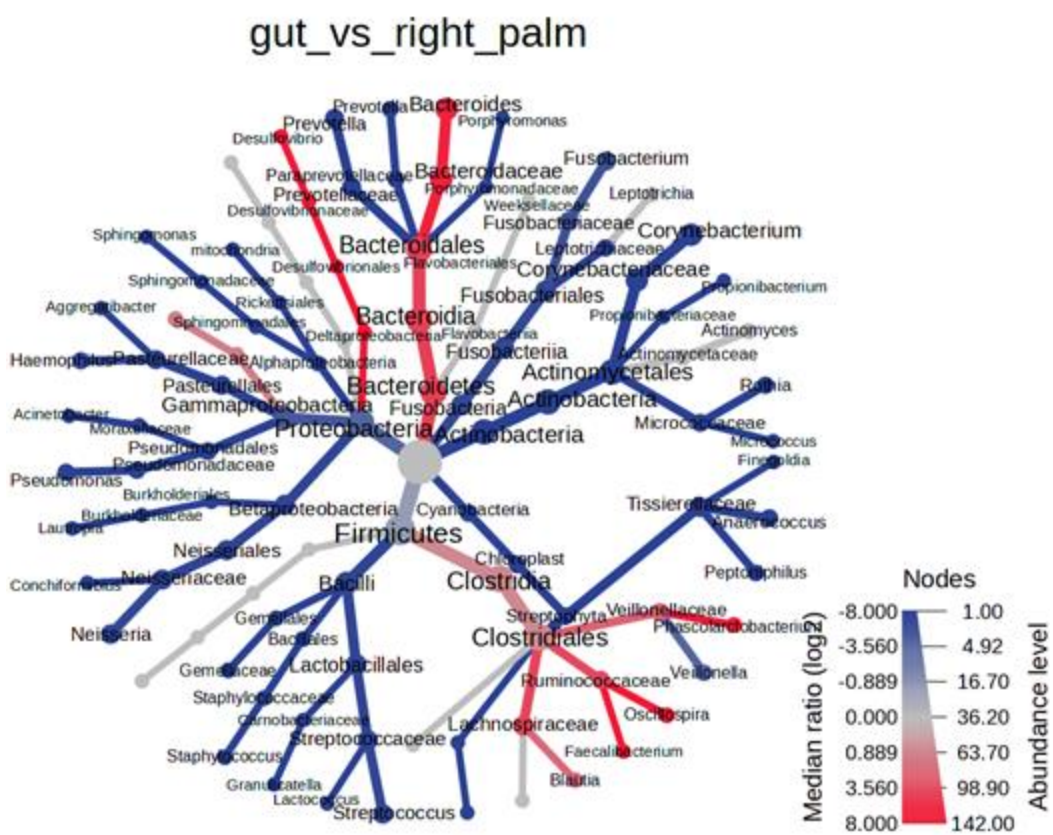



- Using tree to convey a taxonomic hierarchy
- Displaying how statistics are distributed throughout the tree, including internal taxa
- Size and color of nodes and edges are correlated with the abundance of organisms in each community
  - n_obs: number of observations for each taxon
  - prop_amplified: proportion of sequences amplified for each taxo

# Heat tree for community abundance

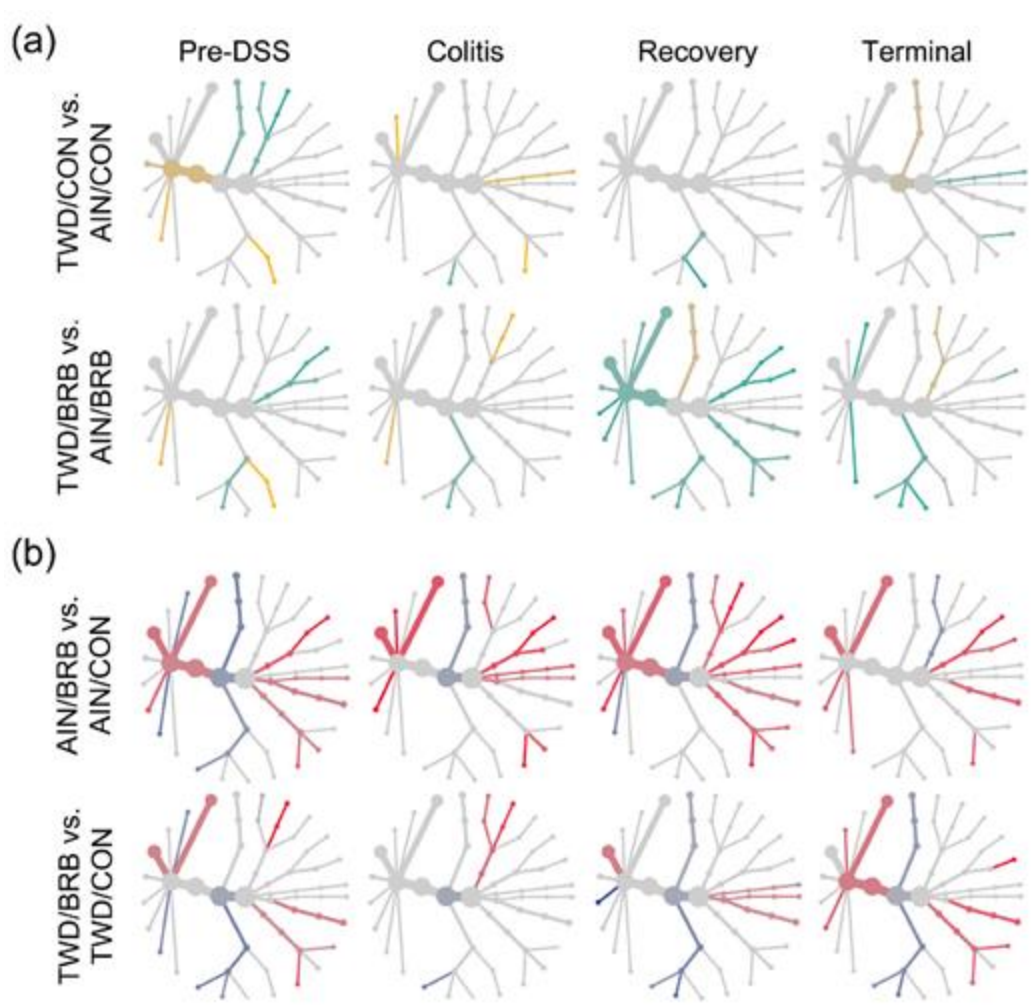


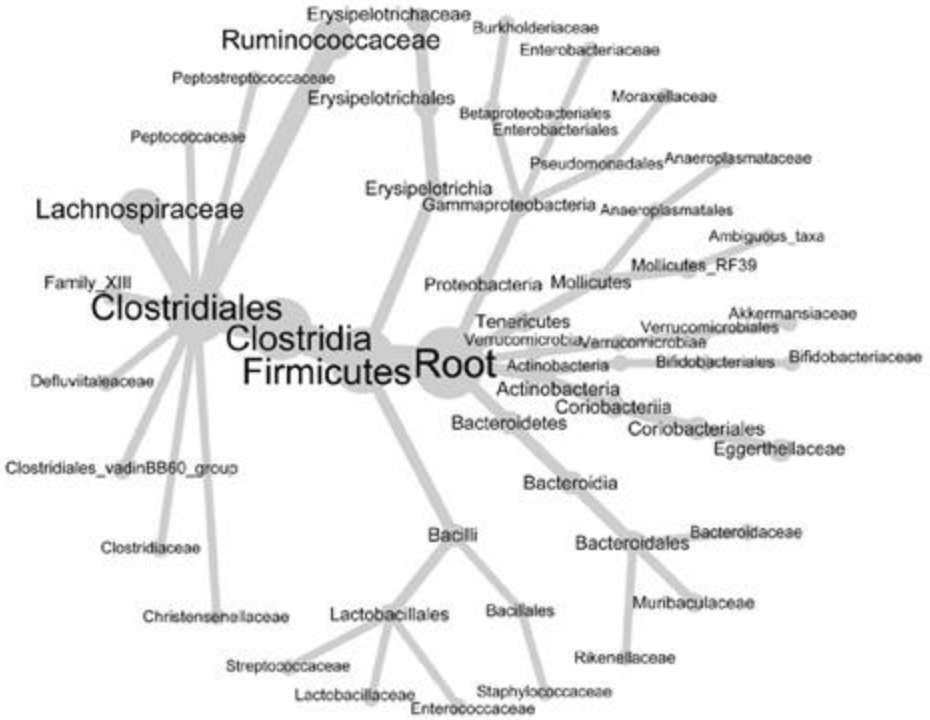

Abundance View

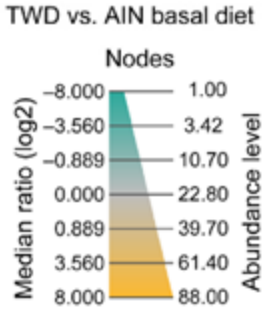


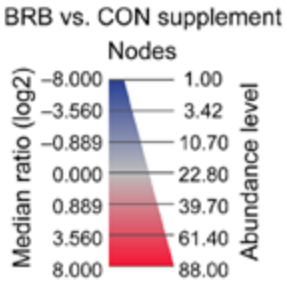

Comparison View

# Heat tree visualization of temporal changes



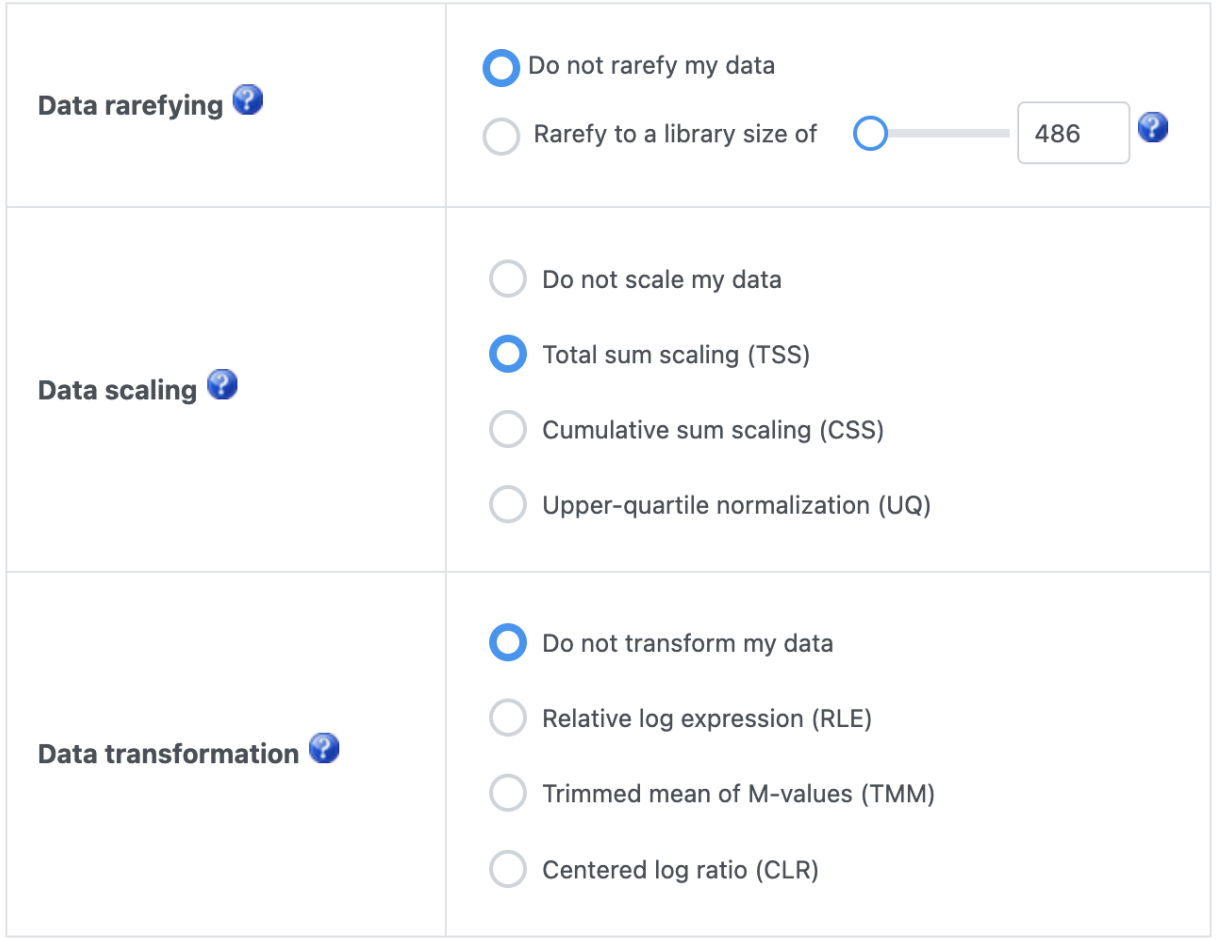
# Statistical modeling of microbiome data

Microbiome abundance data presents unique challenges:

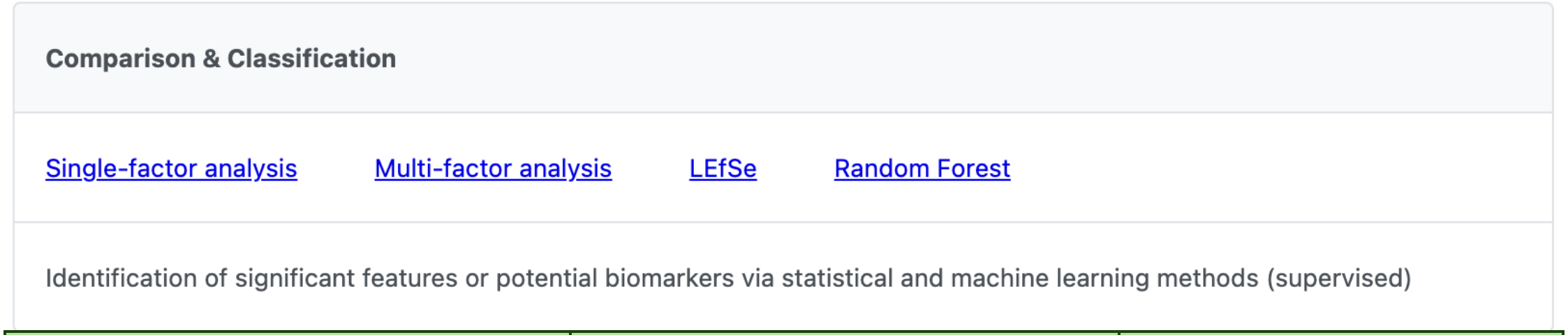
- Sparsity (containing too many zeros)
  - Filtering
- Vast differences in sequencing depth
  - Rarefaction
  - Normalization
- Large variance in distributions (over-dispersion)
  - Non-parametric methods
  - RNAseq methods

# Normalization for comparative analysis

- Normalization aims to address the variability in sampling depth and the sparsity of the data to enable more biologically meaningful comparisons.
- When the library sizes are very different (i.e. > 10 times), rarefying is also recommended.

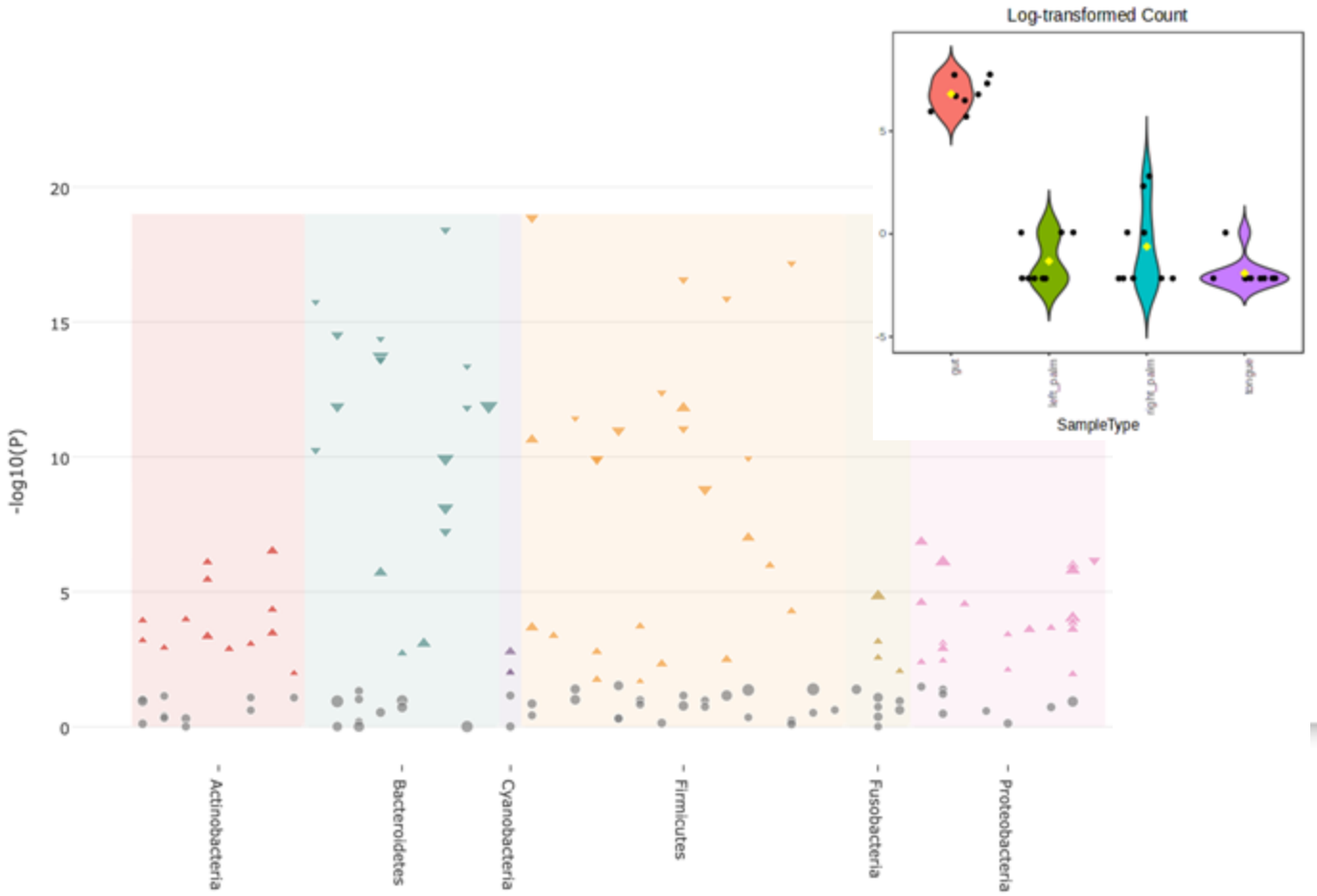


# Comparative analysis in MicrobiomeAnalyst




| General methods | Microbiome specific | RNAseq methods |
| --- | --- | --- |
| ○ T-tests / ANOVA<br>○ Non-parametric tests | ○ LEfSe<br>○ MetagenomeSeq<br>○ MaAsLin2 | ○ EdgeR<br>○ DEseq2 |

# Single-factor differential analysis




**Single-factor Statistical Comparisons**

**Taxonomy level** Feature-level

**Experimental factor** SampleType

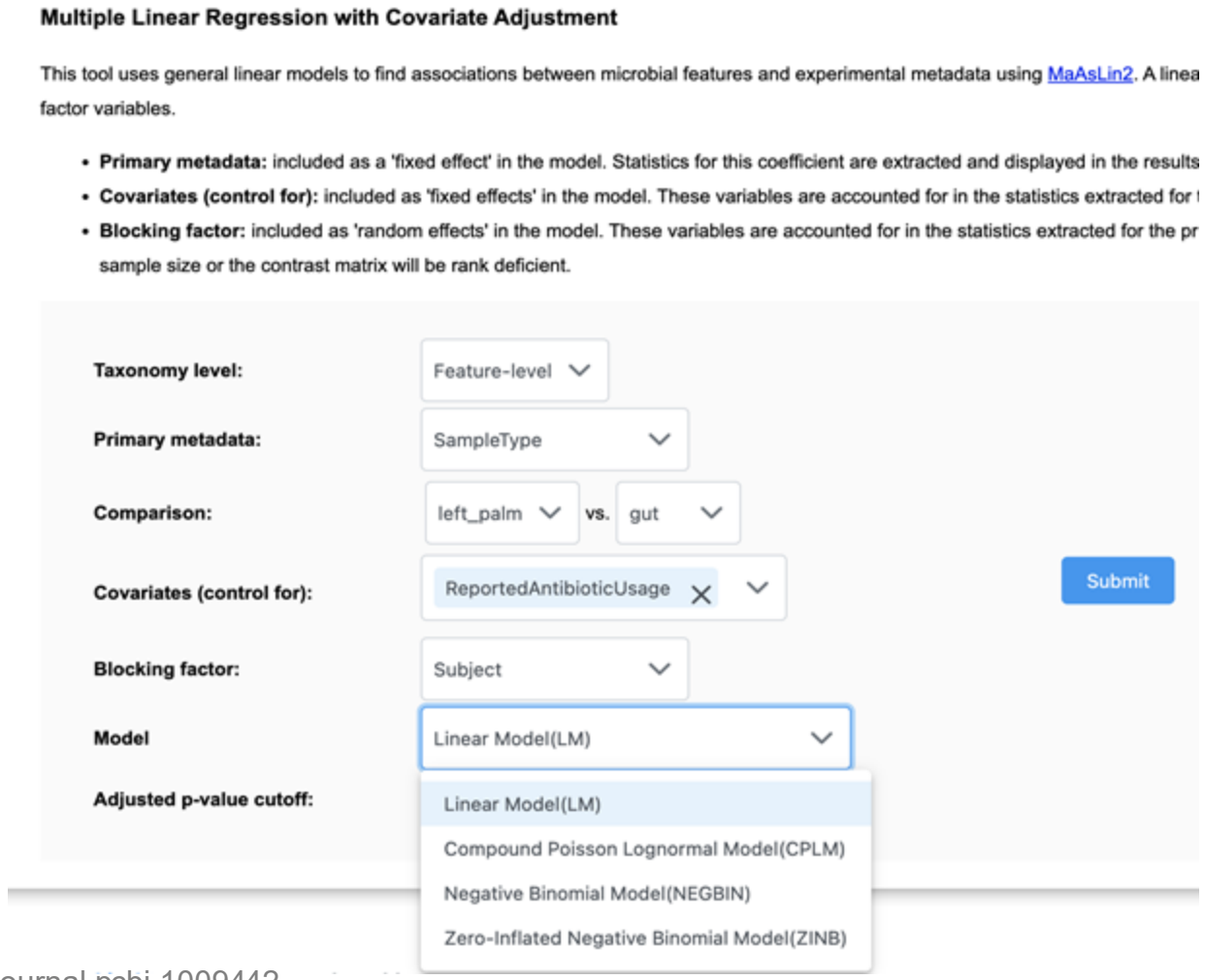
**Statistical method** EdgeR

**Adjusted p-value cutoff**

- Mann-Whitney/Kruskal-Wallis
- T-test/ANOVA
- metagenomeSeq (fitZIG)
- metagenomeSeq (fitFeature)
- EdgeR
- DESeq2

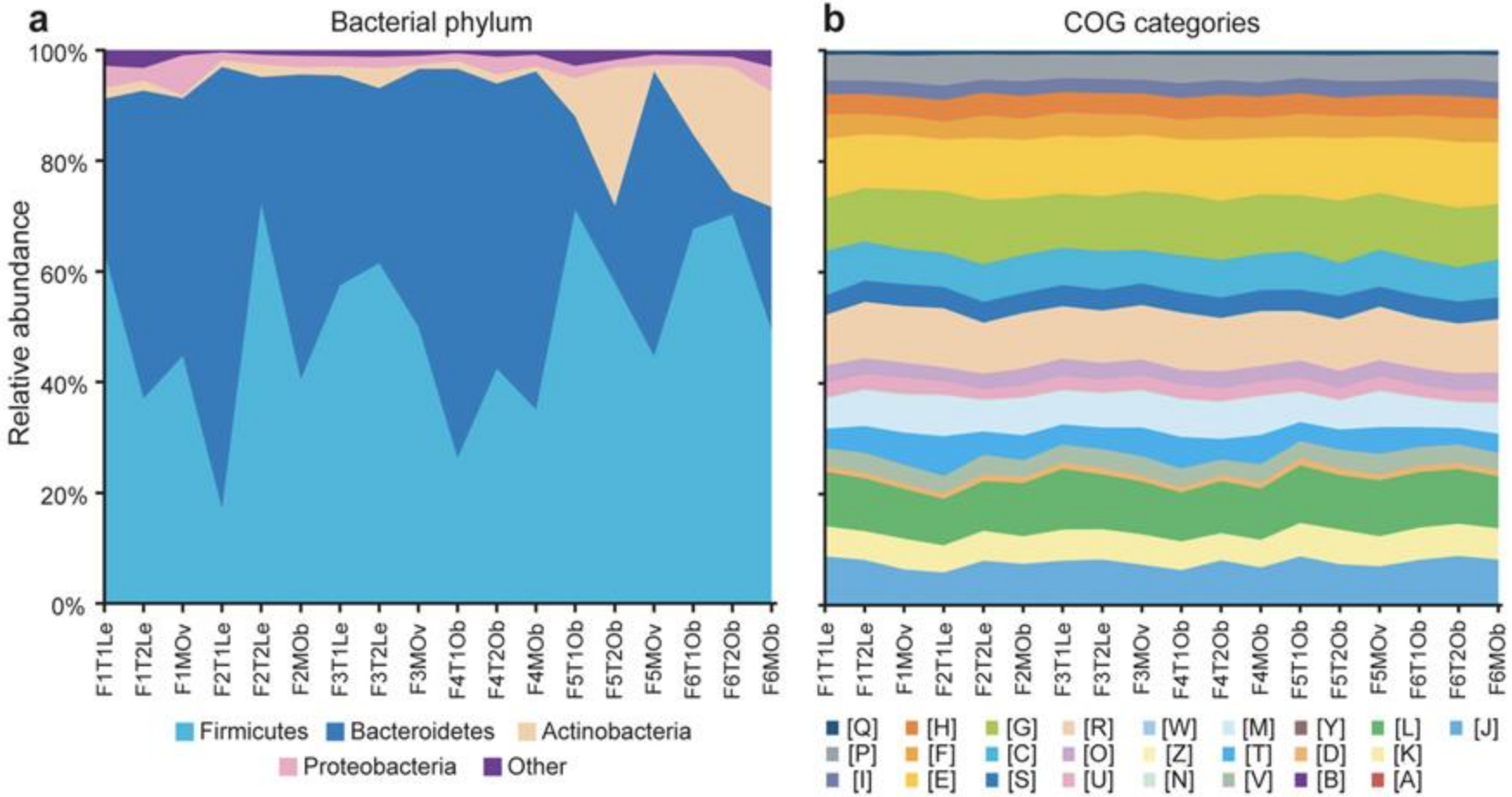
# Multivariable Association with Linear Models (MaAsLin2)

- Estimate multivariable association between phenotypes, environments, exposures, covariates and microbial between phenotype / class labels vs microbial features.
- Using general linear models to accommodate most modern epidemiological study designs, including cross-sectional and longitudinal designs




https://doi.org/10.1371/journal.pcbi.1009442

# Taxonomy Stability vs. Functional Stability




Source: Turnbaugh, P, et al. 2009. *Nature*

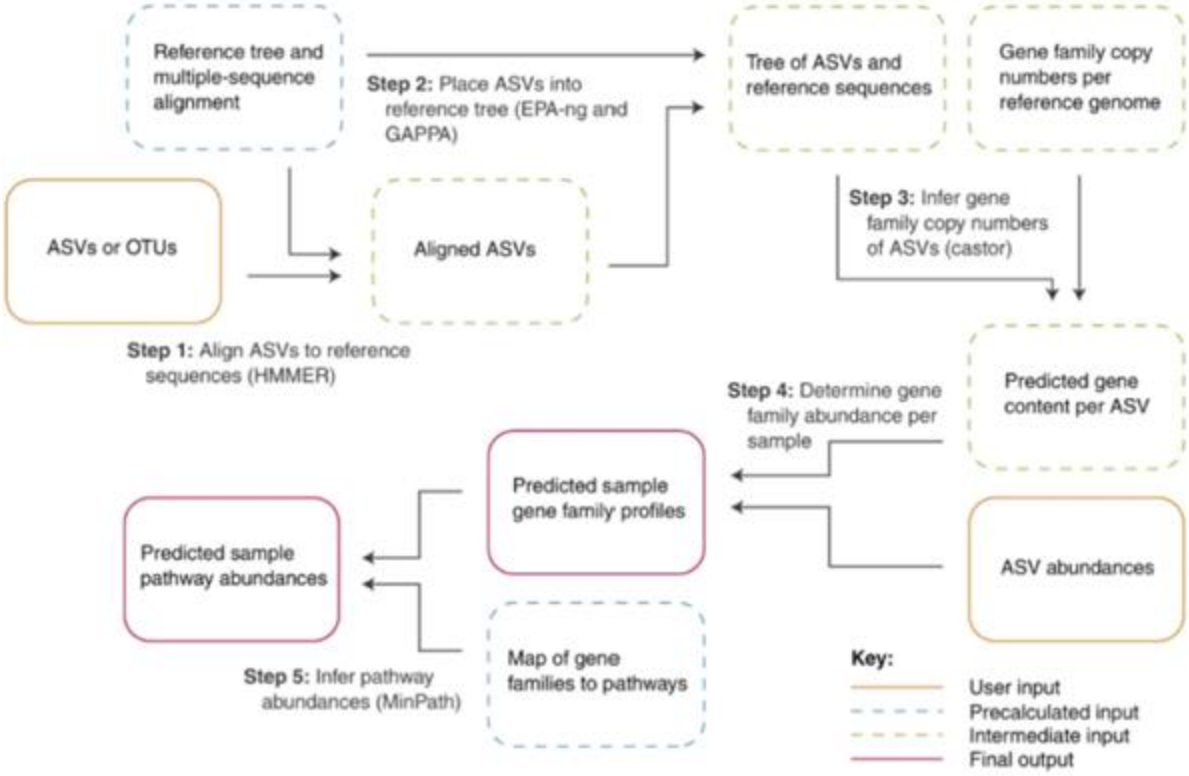
# Predicting Functions

From marker genes to species to potential functions, assuming phylogenetic closeness implies functional similarity.

- Input:
  - Marker gene data (typically 16s rRNA)
  - A set of reference sequence genomes with identified marker genes and predicted protein-coding genes
  - A phylogeny of reference marker genes
- Output:
  - Gene abundance table (mainly metabolic enzymes)
- Tools have been developed and freely available
  - PICRUSt (phylogenetic investigation of communities by reconstruction of unobserved states)
  - Tax4Fun (v1.0 and 2.0)

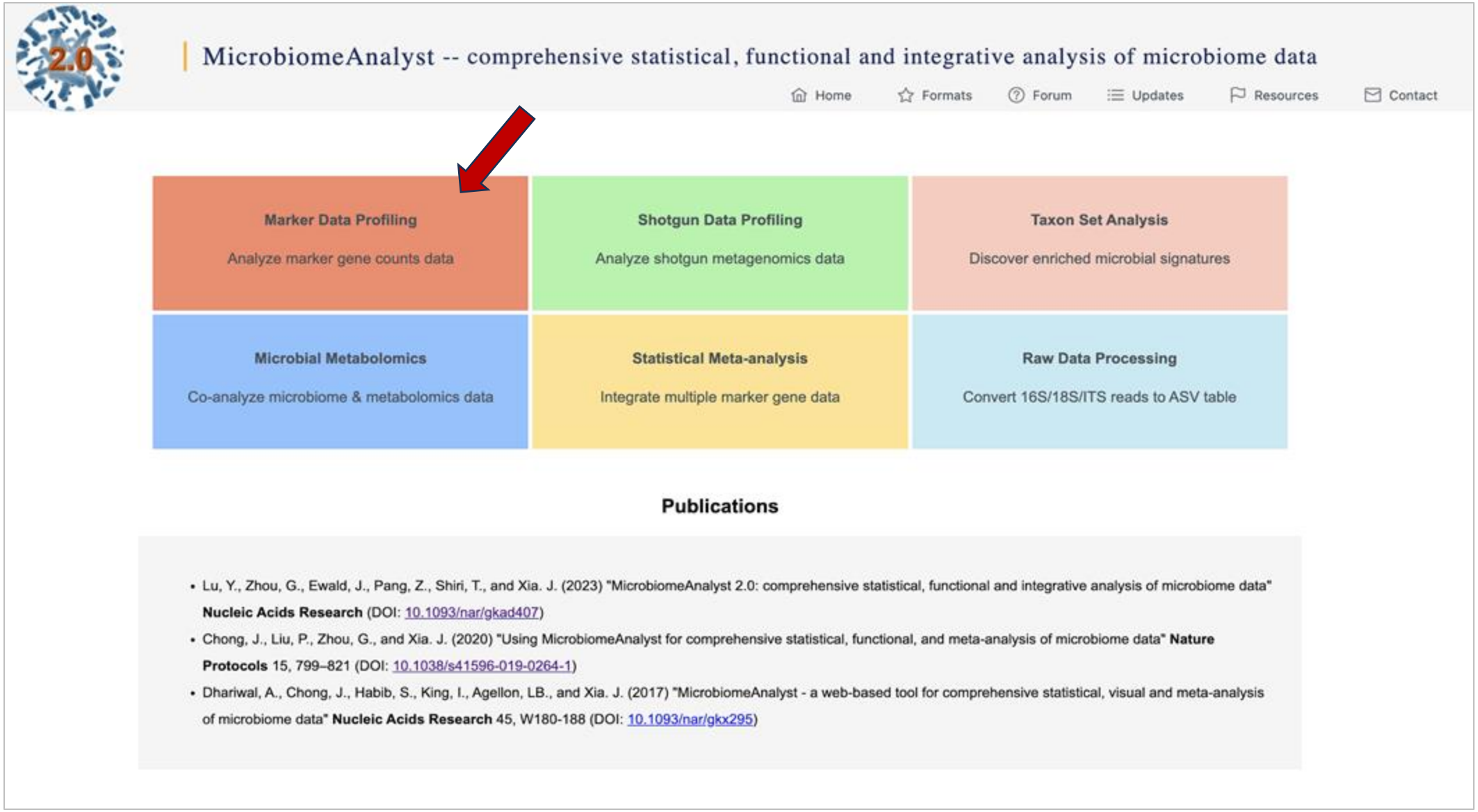
# Predicting functions - PICRUST

- Use KEGG database for functional predictions
- PICRUSt use a machine learning technique that accounts for the evolutionary distance of input sequences from reference genome sequences, to infer gene family abundances from 16S rRNA gene data.
- Can provide coverage for novel sequences not well-represented in reference databases,
  - Time and resources consuming.




https://www.nature.com/articles/s41587-020-0548-6

# Demo 3 – community profiling




**https://www.microbiomeanalyst.ca**

# Basic data integrity check

Basic data filtering are performed by default, as downstream statistics (especially comparative analysis) may not perform properly due to the presence of singletons or constant values.

Default Filtering: Constant features Singleton: None One sample occurrence One total count Update

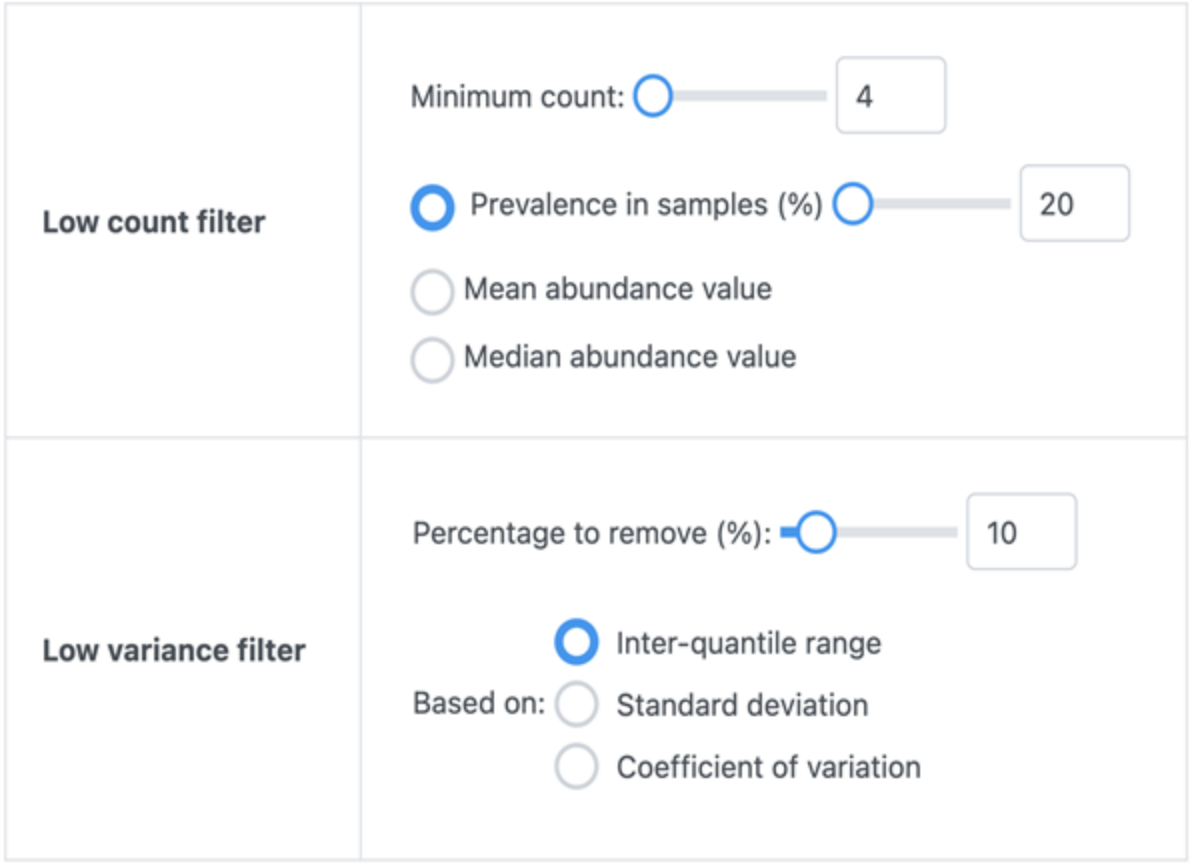
Microbiome data overview | Metadata overview

- Feature abundance table contains raw counts (preferred) or normalized values;
- Features with identical values (i.e. zeros) across all samples will be excluded;
- Features that appear in only one sample will be excluded (considered artifacts);
- For ASV data, which uses actual sequences as IDs, the sequence IDs will be replaced with ASV_1, ASV_2, etc. (refer to the *"ASV_ID_mapping.csv"* from the Downloads page).

| | |
|---|---|
| **Data type:** | OTU abundance table |
| **File format:** | biom |
| **Sample names match (metadata vs. OTU table):** | **Yes** |
| **Normalized counts detected:** | **No** |
| **OTU annotation:** | GreengenesID |
| **OTU number (Post-processing counts/Original counts):** | 2920/3426 |
| **Is any singleton:** | **No** |
| **Singleton removed:** | 0 |
| **Number of experimental factors:** | 7 |
| **Number of experimental factors with replicates:** | 7 [discrete: 7 continuous: 0] |
| **Total read counts:** | 180573 |

# Data Filtering

- **Low count filter** - features with very small counts in few samples are likely due to sequencing errors or low-level contaminations. You need to first specify a minimum count (default 4).
  - A 20% prevalence filter means at least 20% of its values should contain at least 4 counts. You can also filter based on their *mean* or *median* values.
- **Low variance filter** - features that are close to constant throughout the experiment conditions are unlikely to be associated with the conditions under study.



# Data Inputs

- **A count table** containing OTU or ASV counts
- **A taxonomy table**
  - Mapping low-level OTU/ASV assignment to high taxonomy, so that we can do analysis at different taxonomy levels
- **A phylogenetic tree** (optional)
- **A metadata table**
  - Describing the study design and experimental factors

# ASV Table

- The first column name should change to **#NAME** when uploading to MicrobiomeAnalyst
- The ASV sequence will be changed to ASV1, ASV2, ect... in the downstream analysis
- ASV sequence can be used for function prediction by Tax4Fun2.

| ASV | sample1 | sample2 | sample3 | sample4 | sample5 | sample6 | sample7 | sample8 | sample9 | sample10 | s |
|---|---|---|---|---|---|---|---|---|---|---|---|
| TACGGAGGATCCGAGCGTTATCCGGAT | 18 | 53 | 55 | 8638 | 4154 | 11997 | 18174 | 10334 | 10295 | 0 | |
| AACGTAGGTCACAAGCGTTGTCCGGAA | 10 | 0 | 0 | 979 | 1675 | 8346 | 7125 | 8242 | 6844 | 4865 | |
| TACGGAGGGTGCAAGCGTTAATCGGAA | 12315 | 13605 | 30872 | 275 | 366 | 47 | 57 | 15 | 8 | 6 | |
| TACGGAGGATCCGAGCGTTATCCGGAT | 0 | 20 | 40 | 116 | 105 | 341 | 1222 | 81 | 92 | 0 | |
| TACGGAGGATCCGAGCGTTATCCGGAT | 0 | 0 | 0 | 0 | 0 | 7 | 0 | 0 | 0 | 11992 | |
| TACGGAGGATCCGAGCGTTATCCGGAT | 0 | 0 | 0 | 202 | 193 | 0 | 0 | 58 | 48 | 1755 | |
| TACGTATGGTGCAAGCGTTATCCGGATT | 0 | 0 | 14 | 0 | 132 | 1062 | 397 | 1095 | 1460 | 365 | |
| AACGTAGGTCACAAGCGTTGTCCGGAA | 0 | 0 | 0 | 255 | 388 | 171 | 141 | 0 | 0 | 309 | |
| AACGTAGGTCACAAGCGTTGTCCGGAA | 0 | 19 | 14 | 24 | 29 | 73 | 73 | 0 | 0 | 111 | |
| TACGTAGGGGGCAAGCGTTATCCGGAT | 3043 | 13 | 0 | 43 | 52 | 1886 | 466 | 847 | 2212 | 587 | |
| TACGGAGGATCCGAGCGTTATCCGGAT | 0 | 0 | 0 | 201 | 122 | 0 | 0 | 1974 | 2259 | 0 | |
| TACGTATGGTGCAAGCGTTATCCGGATT | 1979 | 0 | 0 | 877 | 2100 | 791 | 589 | 141 | 286 | 562 | |
| TACGGAGGATTCGAGCGTTATCCGGATT | 0 | 0 | 0 | 1469 | 679 | 0 | 0 | 0 | 0 | 0 | |
| TACGTAGGGGGCAAGCGTTATCCGGAA | 0 | 0 | 0 | 282 | 622 | 1392 | 365 | 2624 | 4842 | 518 | |
| TACGTAGGTGGCAAGCGTTGTCCGGAA | 952 | 0 | 0 | 570 | 723 | 7 | 42 | 2052 | 2340 | 5 | |
| TACGTAGGGGGCAAGCGTTATCCGGAT | 0 | 0 | 0 | 125 | 239 | 774 | 445 | 68 | 140 | 187 | |
| TACGTATGGTGCAAGCGTTATCCGGATT | 0 | 0 | 0 | 879 | 1166 | 150 | 50 | 0 | 0 | 472 | |
| TACGTAGGGTGCAAGCGTTAATCGGAAT | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| TACGGAGGATCCGAGCGTTATCCGGAT | 0 | 0 | 0 | 1134 | 1027 | 59 | 169 | 331 | 228 | 26 | |
| TACGGAGGATGCGAGCGTTATCCGGAT | 0 | 0 | 0 | 2982 | 1369 | 171 | 355 | 0 | 0 | 0 | |
| TACGGAGGATCCGAGCGTTATCCGGAT | 0 | 0 | 0 | 127 | 83 | 69 | 94 | 88 | 91 | 295 | |
| TACGGAAGGTCCGGGCGTTATCCGGAT | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| TACGGAGGATCCGAGCGTTATCCGGAT | 13275 | 0 | 0 | 526 | 522 | 0 | 0 | 0 | 0 | 0 | |

# OTU Table

The first column name should change to **#NAME** when uploading to MicrobiomeAnalyst

| OTU_ID | Sample1 | Sample2 | Sample3 | Sample4 | Sample5 | Sample6 | Sample7 | Sample8 | Sample9 | Sample10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 61662 | 0 | 0 | 19 | 14 | 0 | 41 | 0 | 0 | 14 | 0 |
| 67869 | 0 | 0 | 14 | 0 | 0 | 0 | 0 | 19 | 0 | 0 |
| 12985 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 26 | 0 | 0 |
| 39924 | 11 | 0 | 0 | 14 | 29 | 0 | 0 | 16 | 56 | 0 |
| 78292 | 0 | 17 | 14 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 72554 | 0 | 0 | 0 | 0 | 0 | 0 | 177 | 13 | 0 | 0 |
| 55403 | 0 | 17 | 0 | 56 | 0 | 0 | 0 | 0 | 23 | 0 |
| 75160 | 0 | 0 | 56 | 0 | 0 | 25 | 0 | 0 | 0 | 0 |
| 56434 | 0 | 13 | 0 | 0 | 0 | 18 | 962 | 0 | 0 | 0 |
| 19641 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 33 | 0 | 0 |
| 69133 | 0 | 0 | 8 | 0 | 20 | 0 | 0 | 0 | 18 | 0 |
| 62131 | 18 | 8 | 0 | 0 | 0 | 31 | 1838 | 0 | 17 | 5 |
| 24182 | 0 | 0 | 9 | 0 | 0 | 0 | 0 | 0 | 10 | 0 |
| 25179 | 0 | 9 | 0 | 0 | 0 | 0 | 0 | 15 | 41 | 0 |
| 37167 | 0 | 0 | 0 | 0 | 17 | 0 | 0 | 0 | 0 | 12 |
| 99708 | 7 | 0 | 0 | 25 | 0 | 0 | 0 | 0 | 0 | 0 |
| 19096 | 24 | 0 | 6 | 18 | 17 | 0 | 19 | 17 | 0 | 15 |
| 40537 | 0 | 6 | 0 | 0 | 0 | 29 | 0 | 0 | 0 | 0 |
| 66218 | 12 | 0 | 14 | 0 | 13 | 0 | 0 | 10 | 0 | 0 |
| 17988 | 0 | 14 | 0 | 31 | 0 | 0 | 34 | 0 | 0 | 0 |
| 23535 | 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 25 | 0 |
| 16215 | 0 | 0 | 19 | 0 | 8 | 11 | 0 | 31 | 18 | 0 |

OTU IDs annotated based on Greengenes database can be used for function prediction by PICRUSt.

# Taxonomy table

| #TAXONOMY | Kingdom | Phylum | Class | Order | Family | Genus | Species |
|---|---|---|---|---|---|---|---|
| TACGGAGGATCCGAGCGTTA | k__Bacteria | p__Bacteroidetes | c__Bacteroidia | o__Bacteroidale | f__Bacteroidace | g__Bacteroides | s__ |
| AACGTAGGTCACAAGCGTTG | k__Bacteria | p__Firmicutes | c__Clostridia | o__Clostridiales | f__Ruminococca | g__Faecalibacte | s__prausnitzii |
| TACGGAGGGTGCAAGCGTTA | k__Bacteria | p__Proteobacteria | c__Gammapro | o__Enterobacte | f__Enterobacteri | g__Escherichia | s__coli |
| TACGGAGGATCCGAGCGTTA | k__Bacteria | p__Bacteroidetes | c__Bacteroidia | o__Bacteroidale | f__Bacteroidace | g__Bacteroides | s__fragilis |
| TACGGAGGATCCGAGCGTTA | k__Bacteria | p__Bacteroidetes | c__Bacteroidia | o__Bacteroidale | f__Bacteroidace | g__Bacteroides | s__ |
| TACGGAGGATCCGAGCGTTA | k__Bacteria | p__Bacteroidetes | c__Bacteroidia | o__Bacteroidale | f__Bacteroidace | g__Bacteroides | s__caccae |
| TACGTATGGTGCAAGCGTTAT | k__Bacteria | p__Firmicutes | c__Clostridia | o__Clostridiales | f__Lachnospirac | g__[Ruminococ | s__ |
| AACGTAGGTCACAAGCGTTG | k__Bacteria | p__Firmicutes | c__Clostridia | o__Clostridiales | f__Ruminococca | g__Faecalibacte | s__prausnitzii |
| AACGTAGGTCACAAGCGTTG | k__Bacteria | p__Firmicutes | c__Clostridia | o__Clostridiales | f__Ruminococca | g__Faecalibacte | s__prausnitzii |
| TACGTAGGGGGCAAGCGTTA | k__Bacteria | p__Firmicutes | c__Clostridia | o__Clostridiales | f__Lachnospirac | g__[Ruminococ | s__gnavus |
| TACGGAGGATCCGAGCGTTA | k__Bacteria | p__Bacteroidetes | c__Bacteroidia | o__Bacteroidale | f__Bacteroidace | g__Bacteroides | s__ovatus |
| TACGTATGGTGCAAGCGTTAT | k__Bacteria | p__Firmicutes | c__Clostridia | o__Clostridiales | f__Lachnospirac | g__[Ruminococ | s__torques |
| TACGGAGGATTCGAGCGTTA | k__Bacteria | p__Bacteroidetes | c__Bacteroidia | o__Bacteroidale | f__Bacteroidace | g__Bacteroides | s__ |
| TACGTAGGGGGCAAGCGTTA | k__Bacteria | p__Firmicutes | c__Clostridia | o__Clostridiales | f__Lachnospirac | g__Coprococcu | s__ |
| TACGTAGGTGGCAAGCGTTG | k__Bacteria | p__Firmicutes | c__Clostridia | o__Clostridiales | f__Veillonellacea | g__Dialister | s__ |
| TACGTAGGGGGCAAGCGTTA | k__Bacteria | p__Firmicutes | c__Clostridia | o__Clostridiales | f__Lachnospirac | g__[Ruminococ | s__gnavus |
| TACGTATGGTGCAAGCGTTAT | k__Bacteria | p__Firmicutes | c__Clostridia | o__Clostridiales | f__Lachnospirac | g__Roseburia | s__ |
| TACGTAGGGTGCAAGCGTTA | k__Bacteria | p__Proteobacteria | c__Betaproteo | o__Burkholderia | f__Alcaligenacea | g__Sutterella | s__ |
| TACGGAGGATCCGAGCGTTA | k__Bacteria | p__Bacteroidetes | c__Bacteroidia | o__Bacteroidale | f__Bacteroidace | g__Bacteroides | s__caccae |
| TACGGAGGATGCGAGCGTTA | k__Bacteria | p__Bacteroidetes | c__Bacteroidia | o__Bacteroidale | f__Porphyromon | g__Parabactero | s__ |
| TACGGAGGATCCGAGCGTTA | k__Bacteria | p__Bacteroidetes | c__Bacteroidia | o__Bacteroidale | f__Bacteroidace | g__Bacteroides | s__ovatus |
| TACGGAAGGTCCGGGCGTTA | k__Bacteria | p__Bacteroidetes | c__Bacteroidia | o__Bacteroidale | f__Prevotellacea | g__Prevotella | s__ |
| TACGGAGGATCCGAGCGTTA | k__Bacteria | p__Bacteroidetes | c__Bacteroidia | o__Bacteroidale | f__Bacteroidace | g__Bacteroides | s__uniformis |
| TACGTAGGGGGCAAGCGTTA | k__Bacteria | p__Firmicutes | c__Clostridia | o__Clostridiales | f__Lachnospirac | g__Blautia | s__ |
| TACGTAGGGGGCAAGCGTTA | k__Bacteria | p__Firmicutes | c__Clostridia | o__Clostridiales | f__Lachnospirac | g__ | s__ |
| TACGGAGGATCCGAGCGTTA | k__Bacteria | p__Bacteroidetes | c__Bacteroidia | o__Bacteroidale | f__Bacteroidace | g__Bacteroides | s__ |
| TACGTAGGGTGCGAGCGTTA | k__Bacteria | p__Proteobacteria | c__Betaproteo | o__Burkholderia | f__Alcaligenacea | g__Sutterella | s__ |
| TACGGAGGGTGCGAGCGTTA | k__Bacteria | p__Proteobacteria | c__Gammapro | o__Pasteurellale | f__Pasteurellace | g__Haemophilu | s__parainfluenzae |
| TACGGAGGATCCGAGCGTTA | k__Bacteria | p__Bacteroidetes | c__Bacteroidia | o__Bacteroidale | f__Bacteroidace | g__Bacteroides | s__ |

The first column in taxonomy table should be same as the ASV/OTU table.

# Integrated table

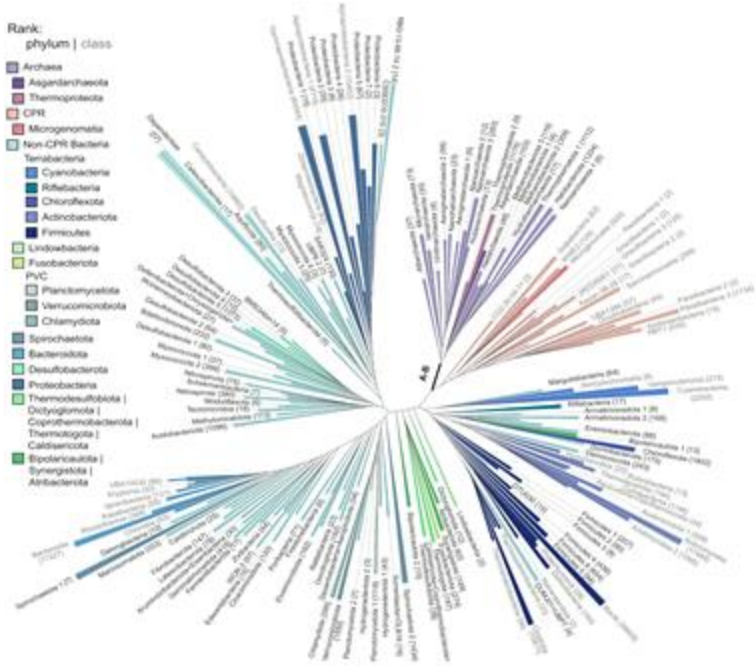
| #NAME | Sample1 | Sample2 | Sample3 | Sample4 | Sample5 | Sample6 | Sample7 | Sample8 | Sample9 | Sample10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Bacteria; Bacteroidota; Bacteroidia; Bacteroidales; Muribaculaceae; uncultur | 579 | 444 | 289 | 228 | 422 | 645 | 405 | 3488 | 988 | 327 |
| Bacteria; Bacteroidota; Bacteroidia; Bacteroidales; Muribaculaceae; uncultur | 345 | 362 | 304 | 176 | 276 | 489 | 353 | 1587 | 602 | 268 |
| Bacteria; Bacteroidota; Bacteroidia; Bacteroidales; Muribaculaceae; uncultur | 449 | 341 | 158 | 204 | 302 | 522 | 231 | 1176 | 465 | 284 |
| Bacteria; Bacteroidota; Bacteroidia; Bacteroidales; Muribaculaceae; uncultur | 430 | 502 | 164 | 231 | 357 | 583 | 69 | 472 | 200 | 158 |
| Bacteria; Bacteroidota; Bacteroidia; Bacteroidales; Rikenellaceae; Alistipes; u | 184 | 321 | 89 | 83 | 41 | 125 | 190 | 1211 | 381 | 207 |
| Bacteria; Bacteroidota; Bacteroidia; Bacteroidales; Muribaculaceae; uncultur | 470 | 331 | 181 | 244 | 353 | 476 | 41 | 115 | 25 | 23 |
| Bacteria; Bacteroidota; Bacteroidia; Bacteroidales; Muribaculaceae; uncultur | 282 | 243 | 163 | 152 | 240 | 396 | 96 | 325 | 167 | 123 |
| Bacteria; Bacteroidota; Bacteroidia; Bacteroidales; Bacteroidaceae; Bacteroi | 172 | 189 | 180 | 130 | 104 | 307 | 140 | 338 | 402 | 151 |
| Bacteria; Bacteroidota; Bacteroidia; Bacteroidales; Muribaculaceae; uncultur | 158 | 130 | 78 | 67 | 155 | 229 | 106 | 609 | 298 | 207 |
| Bacteria; Bacteroidota; Bacteroidia; Bacteroidales; Muribaculaceae; uncultur | 45 | 167 | 89 | 109 | 158 | 202 | 69 | 434 | 307 | 178 |
| Bacteria; Firmicutes; Bacilli; Lactobacillales; Lactobacillaceae; Lactobacillus; | 17 | 168 | 42 | 78 | 269 | 317 | 102 | 55 | 171 | 61 |
| Bacteria; Bacteroidota; Bacteroidia; Bacteroidales; Muribaculaceae; uncultur | 217 | 147 | 98 | 111 | 146 | 258 | 40 | 41 | 0 | 0 |
| Bacteria; Firmicutes; Clostridia; Lachnospirales; Lachnospiraceae; Lachnosp | 93 | 33 | 12 | 9 | 11 | 15 | 325 | 368 | 46 | 87 |
| Bacteria; Firmicutes; Bacilli; RF39; NA; uncultured; uncultured bacterium | 67 | 6 | 7 | 0 | 6 | 5 | 109 | 402 | 342 | 17 |
| Bacteria; Firmicutes; Bacilli; Lactobacillales; Lactobacillaceae; Ligilactobacillu | 52 | 12 | 103 | 43 | 16 | 22 | 129 | 330 | 94 | 48 |
| Bacteria; Bacteroidota; Bacteroidia; Bacteroidales; Muribaculaceae; uncultur | 104 | 65 | 64 | 61 | 81 | 125 | 28 | 107 | 61 | 38 |
| Bacteria; Bacteroidota; Bacteroidia; Bacteroidales; Muribaculaceae; uncultur | 100 | 149 | 116 | 0 | 0 | 195 | 0 | 70 | 0 | 57 |
| Bacteria; Bacteroidota; Bacteroidia; Bacteroidales; Muribaculaceae; uncultur | 69 | 43 | 30 | 20 | 45 | 105 | 31 | 190 | 103 | 37 |
| Bacteria; Firmicutes; Clostridia; Lachnospirales; Lachnospiraceae; Lachnosp | 280 | 51 | 13 | 0 | 15 | 11 | 57 | 147 | 32 | 38 |
| Bacteria; Firmicutes; Bacilli; Erysipelotrichales; Erysipelotrichaceae; Turicibac | 80 | 103 | 52 | 40 | 113 | 126 | 0 | 17 | 24 | 36 |
| Bacteria; Firmicutes; Clostridia; Lachnospirales; Lachnospiraceae; unculturec | 70 | 0 | 0 | 0 | 0 | 0 | 69 | 337 | 43 | 57 |
| Bacteria; Firmicutes; Clostridia; Lachnospirales; Lachnospiraceae; Lachnosp | 54 | 15 | 6 | 0 | 0 | 8 | 136 | 296 | 0 | 55 |
| Bacteria; Firmicutes; Clostridia; Lachnospirales; Lachnospiraceae; Lachnosp | 57 | 0 | 0 | 0 | 0 | 0 | 141 | 284 | 20 | 35 |
| Bacteria; Firmicutes; Clostridia; Oscillospirales; Oscillospiraceae; Oscillibacte | 75 | 0 | 0 | 0 | 0 | 6 | 103 | 200 | 15 | 52 |
| Bacteria; Firmicutes; Bacilli; Lactobacillales; Lactobacillaceae; HT002; uncult | 42 | 45 | 0 | 12 | 45 | 39 | 53 | 30 | 123 | 10 |
| Bacteria; Firmicutes; Clostridia; Lachnospirales; Lachnospiraceae; unculturec | 43 | 36 | 0 | 4 | 10 | 11 | 120 | 109 | 8 | 45 |

# Metadata table

- For categorical metadata, at least two groups and three replicates per groups are required
- No missing values

| #NAME | SampleType | Year | Month | Day | Subject | ReportedAntibioticUsage | DaysSinceExperimentStart | Description |
|---|---|---|---|---|---|---|---|---|
| L1S8 | gut | 2008 | 10 | 28 | 1 | Yes | 0 | 1_Fece_10_28_2008 |
| L1S140 | gut | 2008 | 10 | 28 | 2 | Yes | 0 | 2_Fece_10_28_2008 |
| L1S57 | gut | 2009 | 1 | 20 | 1 | No | 84 | 1_Fece_1_20_2009 |
| L1S208 | gut | 2009 | 1 | 20 | 2 | No | 84 | 2_Fece_1_20_2009 |
| L1S76 | gut | 2009 | 2 | 17 | 1 | No | 112 | 1_Fece_2_17_2009 |
| L1S105 | gut | 2009 | 3 | 17 | 1 | No | 140 | 1_Fece_3_17_2009 |
| L1S257 | gut | 2009 | 3 | 17 | 2 | No | 140 | 2_Fece_3_17_2009 |
| L1S281 | gut | 2009 | 4 | 14 | 2 | No | 168 | 2_Fece_4_14_2009 |
| L2S240 | left palm | 2008 | 10 | 28 | 2 | Yes | 0 | 2_L_Palm_10_28_2008 |
| L2S155 | left palm | 2009 | 1 | 20 | 1 | No | 84 | 1_L_Palm_1_20_2009 |
| L2S309 | left palm | 2009 | 1 | 20 | 2 | No | 84 | 2_L_Palm_1_20_2009 |
| L2S175 | left palm | 2009 | 2 | 17 | 1 | No | 112 | 1_L_Palm_2_17_2009 |
| L2S204 | left palm | 2009 | 3 | 17 | 1 | No | 140 | 1_L_Palm_3_17_2009 |
| L2S357 | left palm | 2009 | 3 | 17 | 2 | No | 140 | 2_L_Palm_3_17_2009 |
| L2S222 | left palm | 2009 | 4 | 14 | 1 | No | 168 | 1_L_Palm_4_14_2009 |

# Phylogenetic Tree (optional)



**Generation steps**:

1. Sequence alignment
2. Model selection: to estimate the evolutionary distances between sequences.
3. Tree construction: Neighbor-Joining, Maximum Likelihood, or Bayesian Inference

**Common tools**:

- MEGA
- RAxML
- Clustal Omega
- MrBayes
- iTOL (Interactive Tree Of Life)
- q2-phylogeny plugin in QIIME

**File Formats**:

- Newick : text-based format (support by MicrobiomeAnlyst)
  ` (A:0.1,B:0.2,(C:0.3,D:0.4):0.5);`
- Nexus: more complex and flexible that can include both trees and data (like DNA sequences, metadata)

# Hands on Practices (15 min)

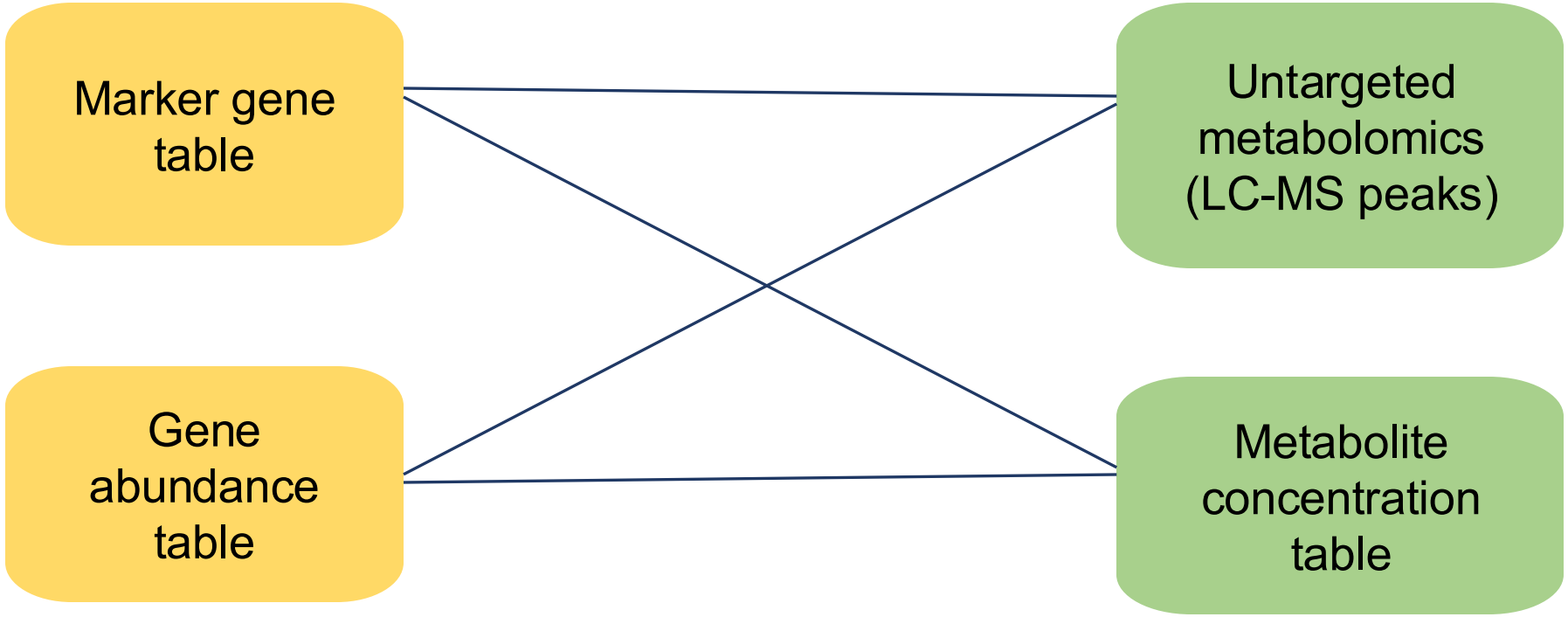
You can either directly use the result table from raw data processing section or download from our github:

https://github.com/xia-lab/Metabolomics_2025/tree/main/docs/mdp_input

# Schedule

| Time | Topics | Lab practices |
|---|---|---|
| 12:00 – 12:15 | General introduction | |
| 12:15 – 1:00 | Metabolomics data processing | Live demo 1 & hands on |
| 1:00 – 1:45 | Microbiome data processing | Live demo 2 & hands on |
| 1:45 – 3:00<br>(15 min break) | Microbial community profiling | Live demo 3 & hands on |
| 3:00 – 3:50 | Microbiome-metabolomics integration | Live demo 4 & hands on |
| 3:50 – 4:15 | Summary and discussion | |

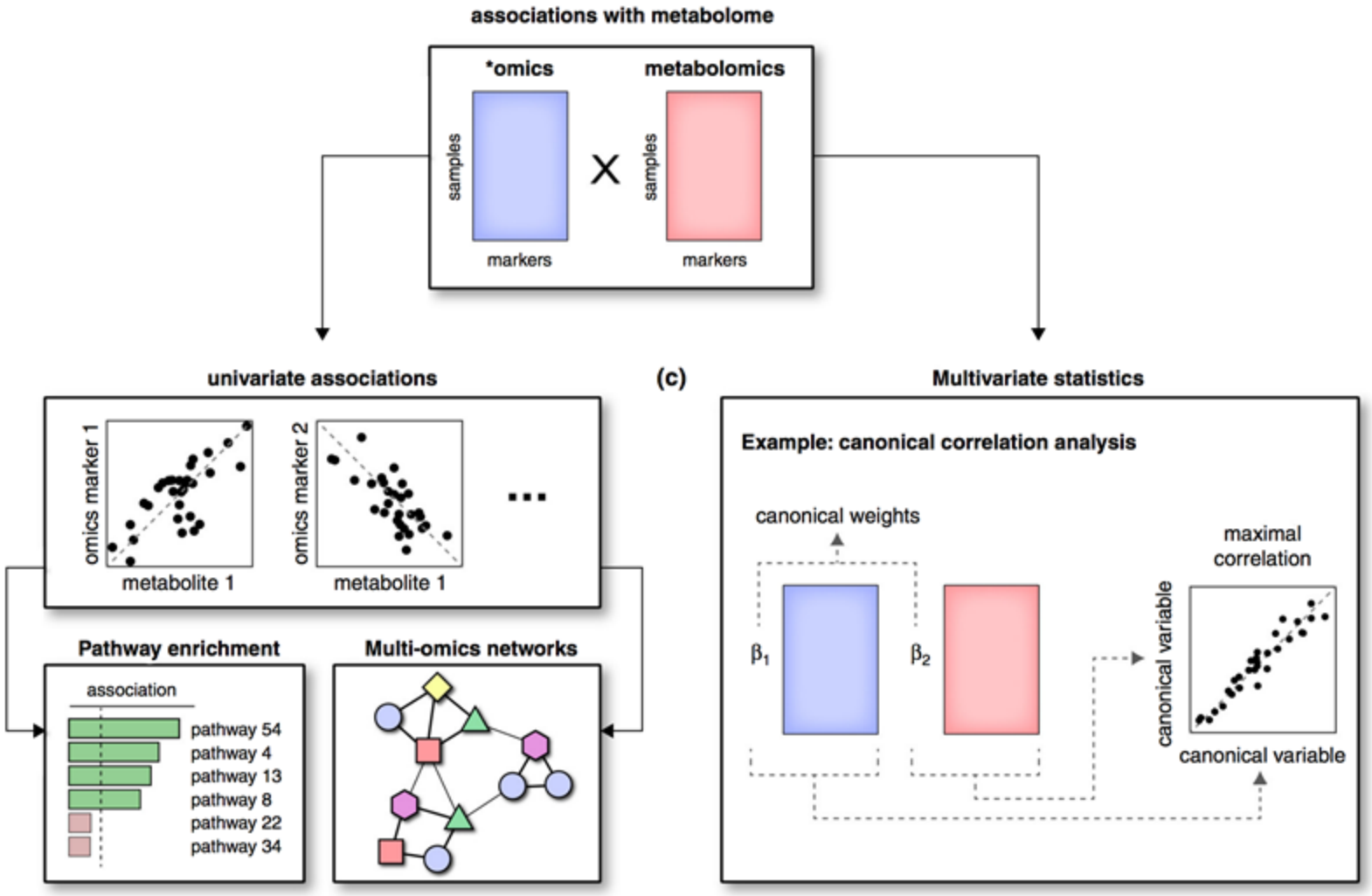
# Paired microbiome – metabolomics data



# Overall strategies for data integration

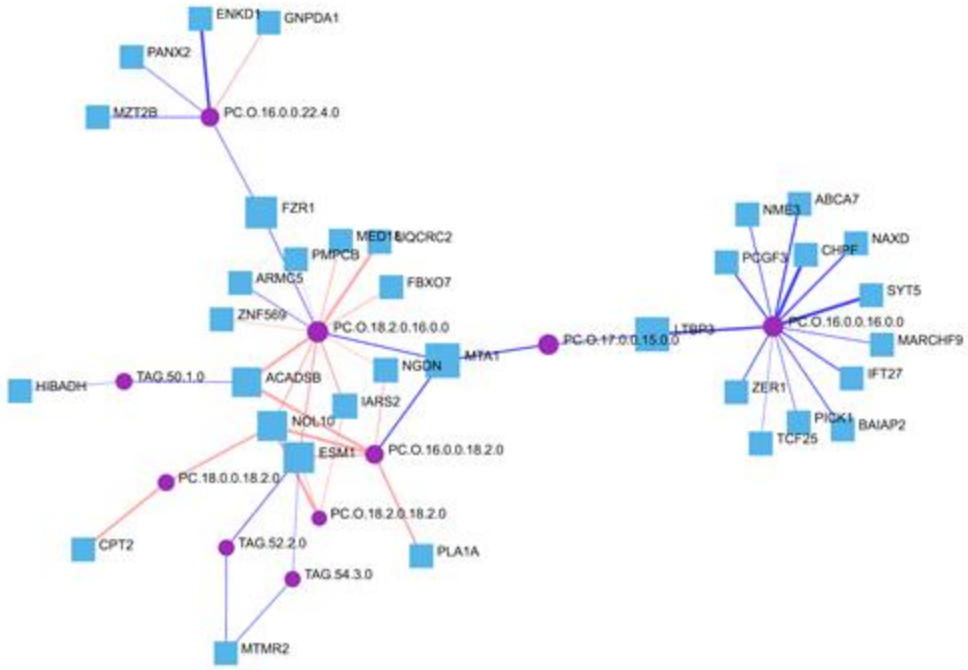
1. General statistical integration
   - Univariate Analysis
     - Feature correlation network
   - Multivariate – cluster analysis
     - SNF, Spectrum clustering
   - Multivariate – dimensionality reduction
     - Procrustean Analysis, DIABLO
2. Functional integration
   - Based on metabolic networks (KEGG)
   - Leveraging genome scale metabolic models (GEMs)

# Statistical correlation and co-variations

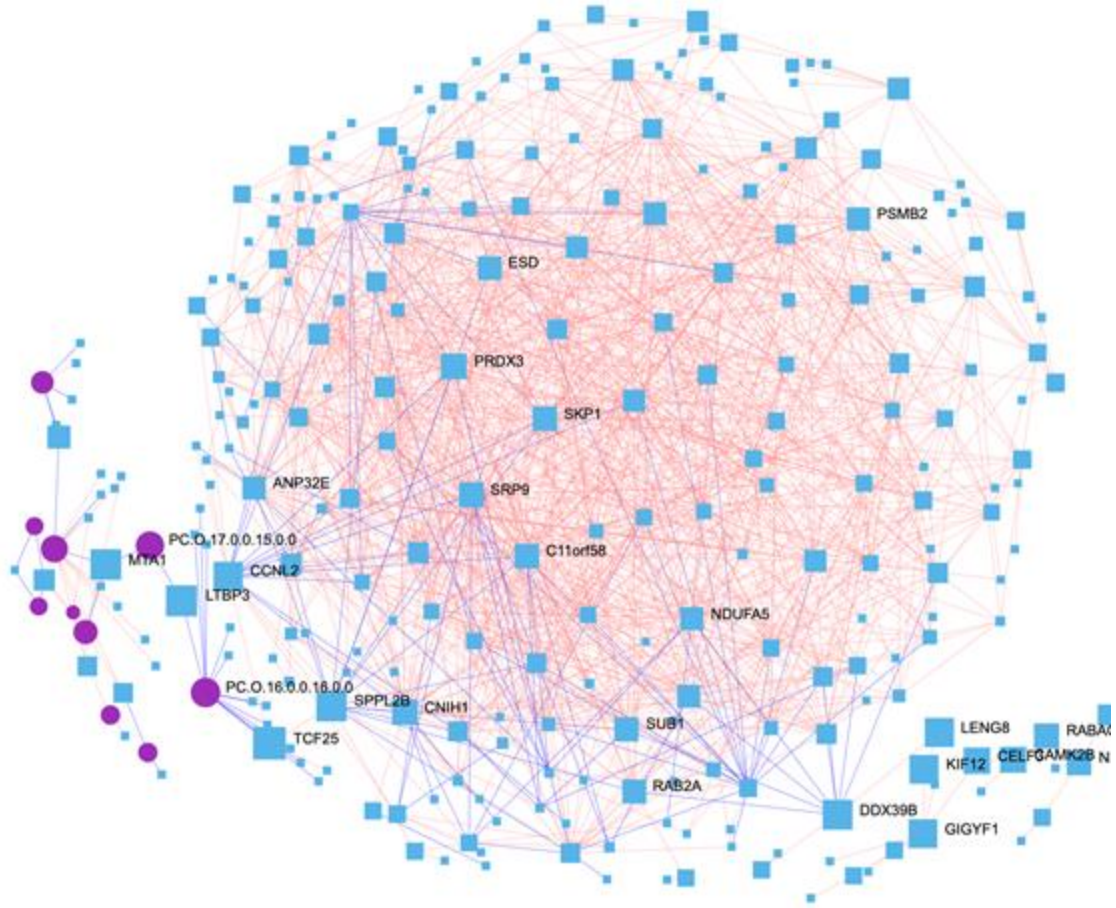



Current Opinion in Biotechnology 2016, 39:198–206

# Multi-omics correlation network




Between omics only




Within and between omics

OmicsAnalyst.ca

# Multivariate co-variance analysis

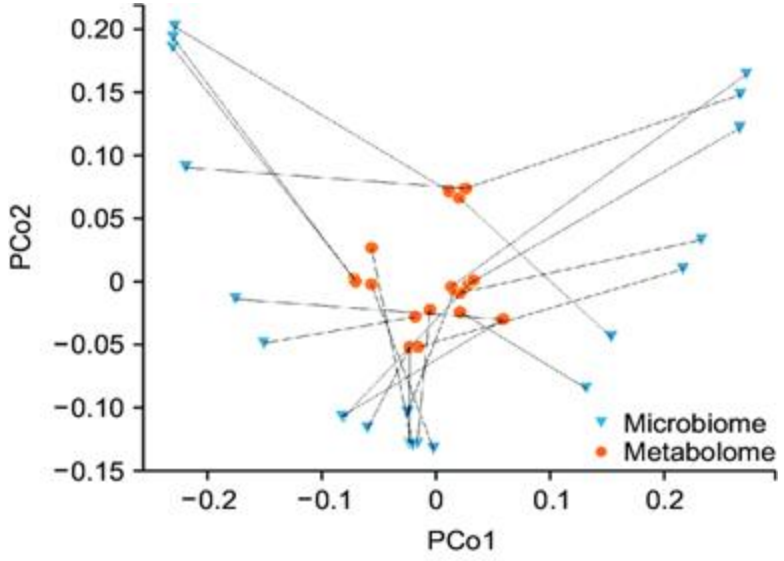
- Visualize if their overall trends are similar
  - PCA or PCoA on individual omics data
  - Procrustean analysis (shape alignment) to see if their main trends agree with each other
- Explicitly compute the shared trend
  - Canonical Correlation Analysis (CCA)
  - DIABLO
  - MOFA
  - Many others

OmicsAnalyst.ca

# Procrustes Analysis

To assess and visualize the similarity of two or more datasets.

- Procrustes essentially computes reduced dimensions for each data set using a method similar to PCA. Then, one of the reduced dimension matrices is rotated until it has maximum similarity with the other.
- Procrustes is asymmetric - the order that the 'omics datasets are considered will impact the results



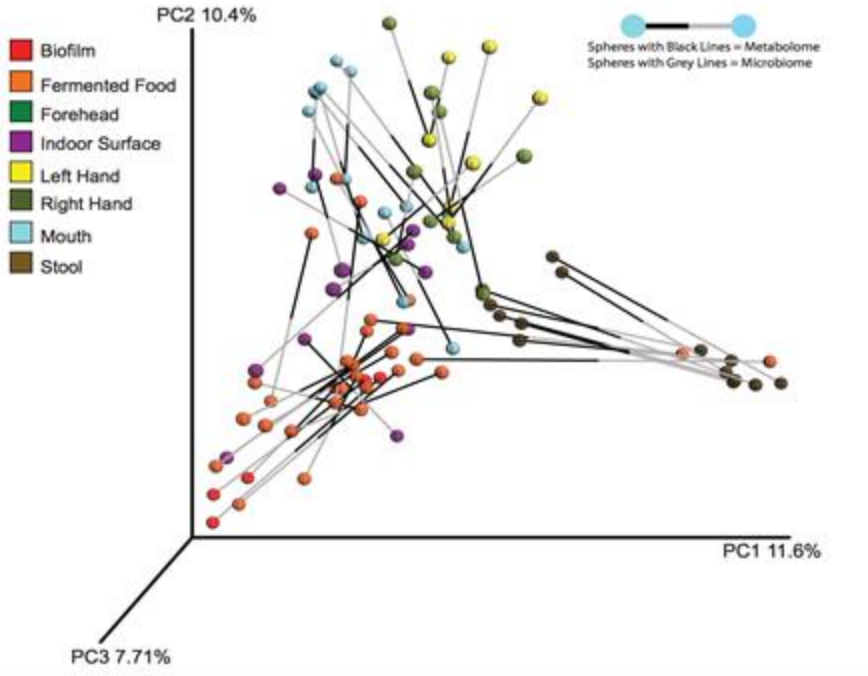

https://pubmed.ncbi.nlm.nih.gov/36950061/

# A case study

AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems

## From Sample to Multi-Omics Conclusions in under 48 Hours

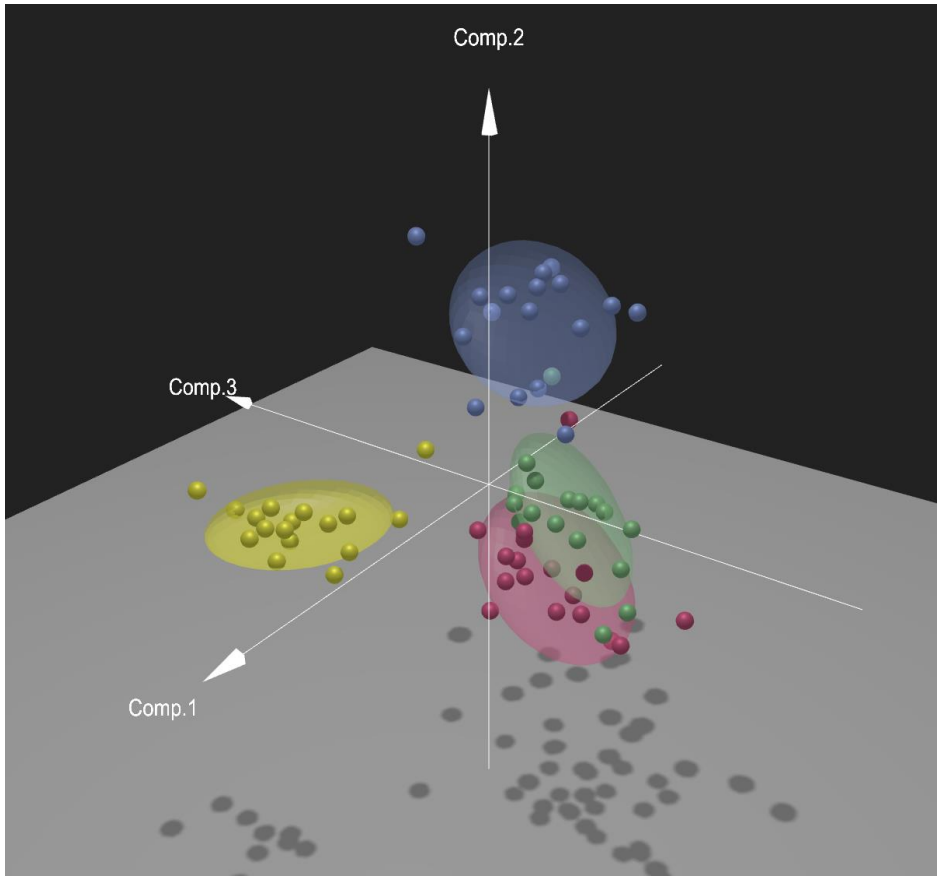
Robert A. Quinn,[a,h] Jose A. Navas-Molina,[b,c] Embriette R. Hyde,[b] Se Jin Song,[b,i] Yoshiki Vázquez-Baeza,[c] Greg Humphrey,[b] James Gaffney,[b] Jeremiah J. Minich,[b] Alexey V. Melnik,[a] Jakob Herschend,[a] Jeff DeReus,[b] Austin Durant,[d] Rachel J. Dutton,[e,h] Mahdieh Khosroheidari,[f] Clifford Green,[f] Ricardo da Silva,[a] Pieter C. Dorrestein,[a,b,g,h] Rob Knight[a,b,c,g]




- Microbial communities and their activities are environment specific
- The metabolite output of the sample type is consistent with the microbial community that produced it.

# DIABLO

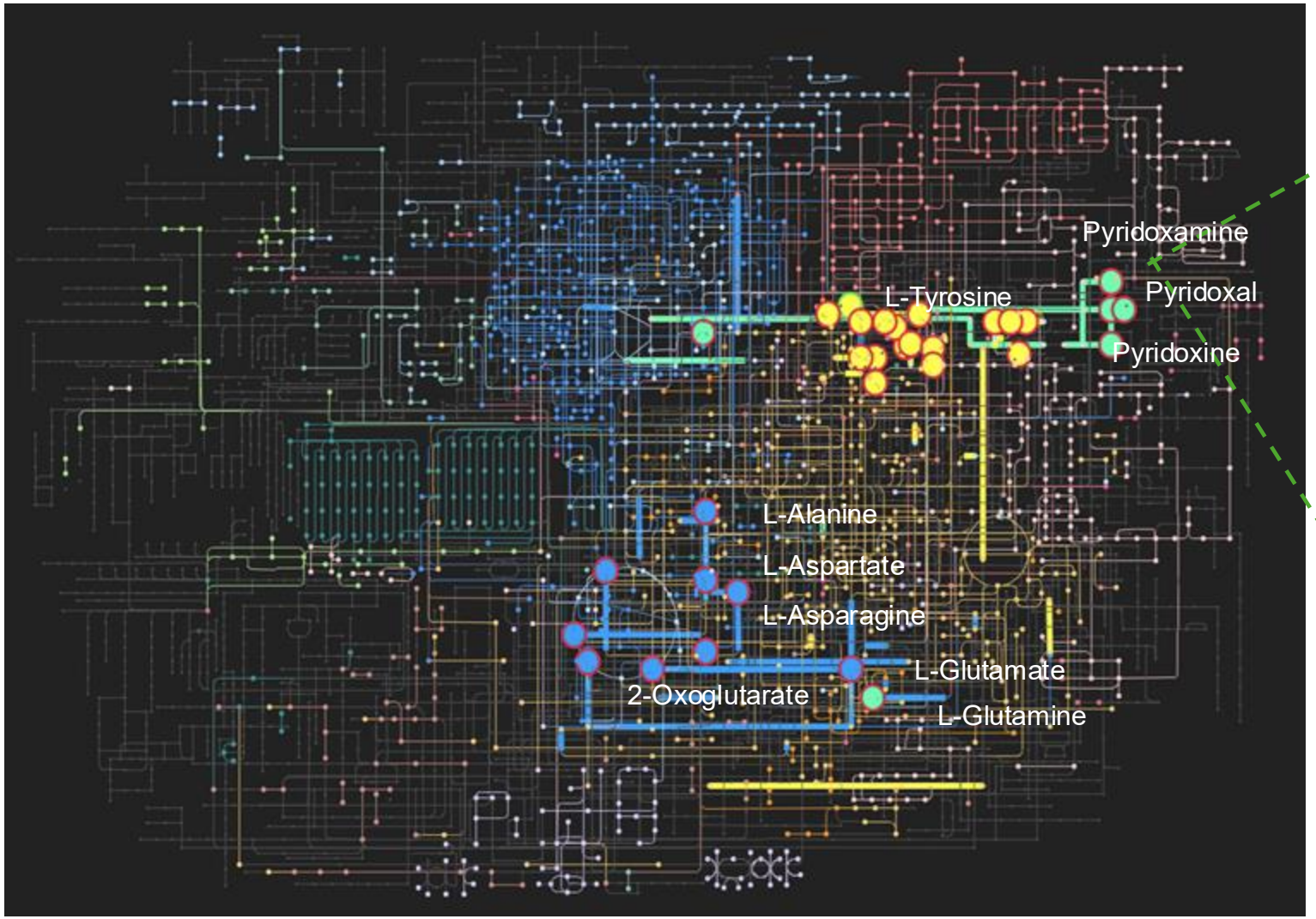
- Data Integration Analysis for Biomarker discovery using Latent cOmponents (DIABLO)
  - DIABLO extends the Partial Least Squares (PLS) regression and Canonical Correlation Analysis (CCA) methods to integrate several datasets, finding latent components that capture covariance across them.
- Can find correlated variables between multiple omics datasets to identify biomarkers or signatures associated with a phenotype of interest

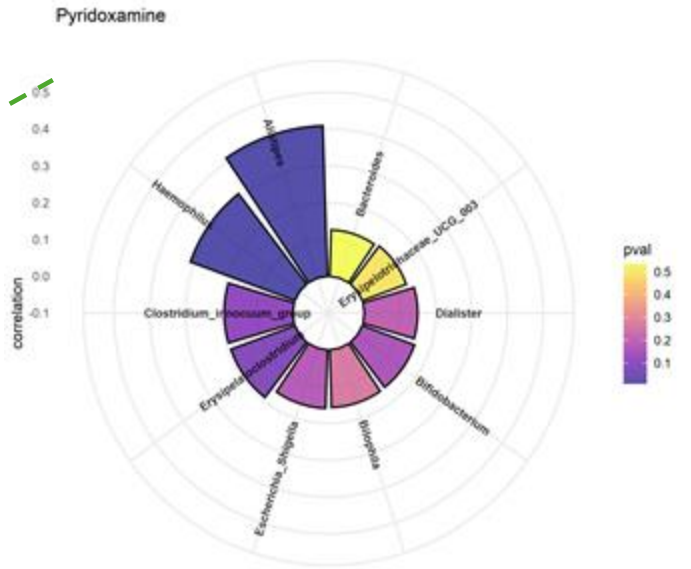


# Functional integration I
# - contextualized metabolic network analysis

- Microbiome data and metabolomics data are projected into metabolic network for visual exploration as well as enrichment analysis.
- The integration strategies are based on microbiome data types:
  - **Marker genes** data will be used to constrain the metabolic network for enrichment analysis of metabolomics data. Users can click a node to view the most correlated microbes of metabolites
  - **Shotgun metagenomics** data, both KOs and metabolomic features will be projected to the selected network for integration analysis.

# Contextualized metabolic network

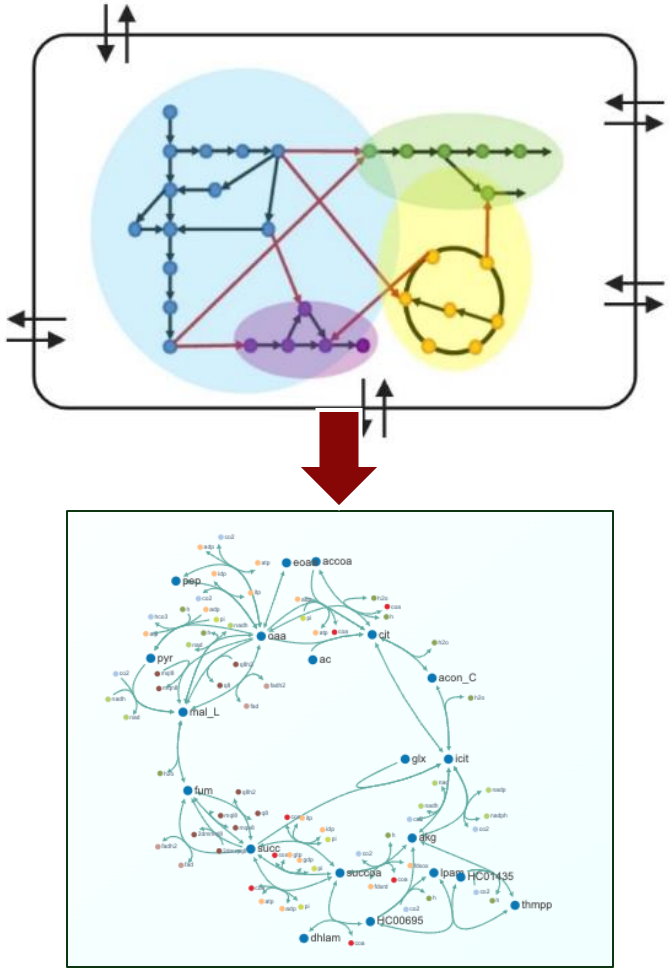



| Microbiome Feature | Metabolomics Feature | Correlation ▼ | P-value ▲ | FDR |
|---|---|---|---|---|
| Alistipes | C00534 | 0.4247146 | 0.00990099 | 0.04950495 |
| Haemophilus | C00534 | 0.3124323 | 0.00990099 | 0.04950495 |
| Erysipelatoclostridium | C00534 | 0.1979814 | 0.04950495 | 0.1485149 |
| Clostridium_innocuum_group | C00534 | 0.1926674 | 0.05940594 | 0.1485149 |
| Escherichia_Shigella | C00534 | 0.1593542 | 0.1188119 | 0.2376238 |
| Bilophila | C00534 | 0.1575959 | 0.1881188 | 0.3094059 |
| Bifidobacterium | C00534 | 0.157558 | 0.2277228 | 0.3094059 |
| Dialister | C00534 | 0.1529811 | 0.2475248 | 0.3094059 |
| Erysipelotrichaceae_UCG_00 | C00534 | 0.1315825 | 0.5544554 | 0.5544554 |
| Bacteroides | C00534 | 0.1291311 | 0.4950495 | 0.550055 |

# Functional integration II
# - genome scale metabolic models (GEMs)

- Computational models that represent the metabolic capabilities of an organism or a community of organisms.
- Reconstructed from annotated genomes
- Organism-specific, capturing the unique metabolic capabilities of an organism as dictated by its genomic content



# Using GEM to predict metabolic potential

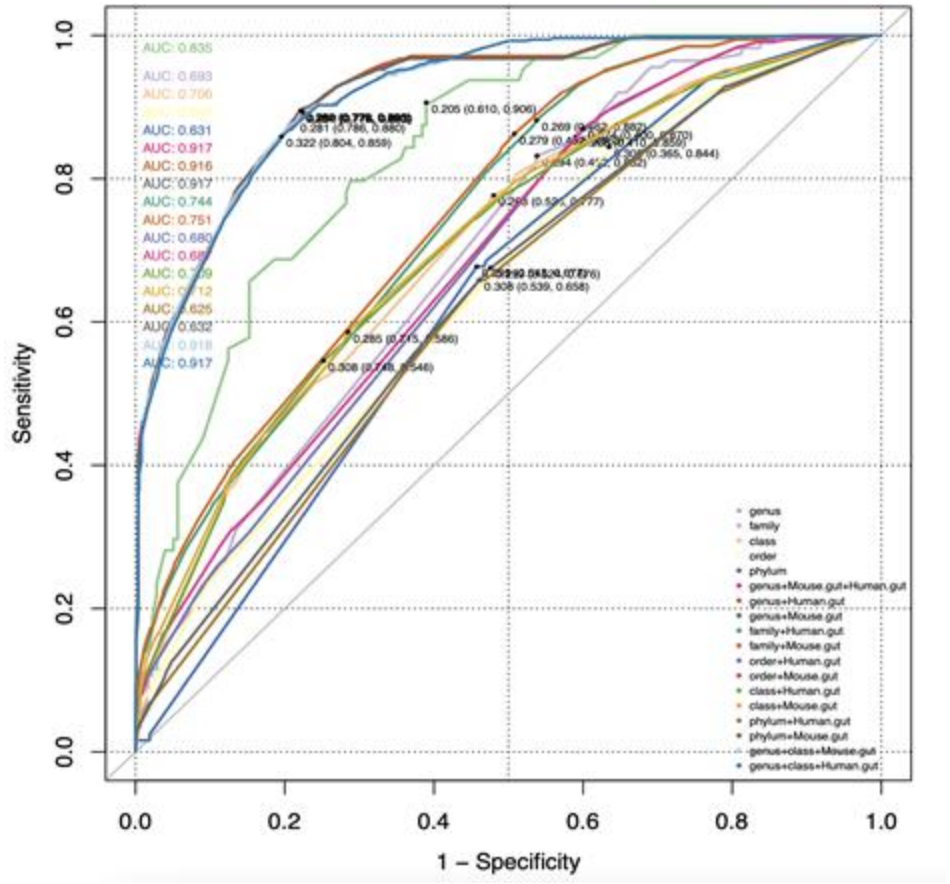



**Table 1.** Odds ratio of dominate genera in mouse gut for bioactive indole generation (red: *p*-value < 0.001; orange: *p*-value < 0.01, yellow: *p*-value < 0.05, blue *p*-value < 0.1 based on Wald test).

| Predictors | IA | Indole | IAAlD | IAM | IAA | ILA | IPA | Tryptamine |
|---|---|---|---|---|---|---|---|---|
| *Bacteroides* | 0.8786 | 310.4118 | 1.5256 | 0.5621 | 2.9424 | 69.0048 | 0.8515 | 0.1484 |
| *Bifidobacterium* | 0.8712 | 0.0421 | 0.4879 | 0.6081 | 1.0597 | 103.1476 | 0.8393 | 8.5582 |
| *Clostridium* | 413.0681 | 2.2526 | 1.6328 | 106.6308 | 0.8401 | 89.0063 | 638.3164 | 3.473 |
| *Desulfovibrio* | 0.9738 | 1.5226 | 1.2451 | 0.8139 | 14.9215 | 0.8931 | 0.9676 | 0.3856 |
| *Enterococcus* | 0.9225 | 2.1667 | 0.8861 | 0.7017 | 0.0392 | 0.6446 | 0.9017 | 1.0872 |
| *Escherichia* | 0.936 | 241.1231 | 0.087 | 0.7997 | 3.9424 | 226.036 | 0.9161 | 9.0402 |
| *Eubacterium* | 0.9783 | 3.7121 | 0.622 | 0.8615 | 16.7253 | 0.8792 | 0.9723 | 0.4817 |
| *Lactobacillus* | 0.8536 | 0.0324 | 1.9794 | 0.5087 | 1.765 | 36.9424 | 0.8227 | 0.4338 |
| *Mouse.gut* | 2.5942 | 2.2931 | 1.1424 | 0.937 | 1.8319 | 6.9211 | 2.9119 | 0.4471 |
| *Parabacteroides* | 0.9668 | 0.1841 | 0.9344 | 0.8384 | 1.2605 | 13.8291 | 0.9569 | 0.4585 |
| *Prevotella* | 0.9145 | 1.6855 | 0.7029 | 0.5947 | 0.473 | 0.7286 | 0.8969 | 0.1589 |
| *Ruminococcus* | 0.9718 | 0.7449 | 0.4029 | 0.8035 | 7.2895 | 0.8861 | 0.9651 | 0.3699 |
| *Streptococcus* | 0.8959 | 0.6288 | 0.7287 | 0.5533 | 0.527 | 0.6756 | 0.8749 | 19.0245 |

**Table 2.** Odds ratio of dominate genera in human gut for bioactive indole generation (red: *p*-value < 0.001; orange: *p*-value < 0.01, yellow: *p*-value < 0.05, blue *p*-value < 0.1 based on Wald test).

| Predictors | IA | Indole | IAAlD | IAM | IAA | ILA | IPA | Tryptamine |
|---|---|---|---|---|---|---|---|---|
| *Bacteroides* | 0.8855 | 1595.5832 | 1.4595 | 0.5265 | 2.6489 | 79.5618 | 0.8575 | 0.1214 |
| *Bifidobacterium* | 0.9025 | 0.0371 | 0.7658 | 0.5674 | 0.7158 | 175.3057 | 0.8781 | 5.5164 |
| *Clostridium* | 414.0254 | 1.8606 | 1.2366 | 91.4966 | 1.6268 | 81.5643 | 606.5603 | 2.8663 |
| *Desulfovibrio* | 0.9683 | 1.55 | 0.6004 | 0.79 | 47.2592 | 0.8512 | 0.9593 | 0.3552 |
| *Enterococcus* | 0.9478 | 2.561 | 1.2673 | 0.7037 | 0.0322 | 0.7858 | 0.9333 | 0.9879 |
| *Escherichia* | 0.9639 | 318.856 | 0.081. | 0.7677 | 4.0763 | 607.7244 | 0.9531 | 4.6206 |
| *Eubacterium* | 0.981 | 2.0208 | 1.2466 | 0.8559 | 10.5365 | 0.9035 | 0.9753 | 0.463 |
| *Human.gut* | 21.3204 | 1.4413 | 0.7778 | 342.7406 | 2.1685 | 20.4853 | 37.332 | 1876.3277 |
| *Lactobacillus* | 0.8757 | 0.1256 | 2.4492 | 0.5055 | 1.5343 | 52.7602 | 0.8462 | 0.412 |
| *Parabacteroides* | 0.976 | 0.1964 | 1.3015 | 0.8271 | 1.681 | 28.2969 | 0.9687 | 0.4121 |
| *Prevotella* | 0.9479 | 3.5273 | 2.0026 | 0.7035 | 1.294 | 0.7879 | 0.9332 | 0.2527 |
| *Ruminococcus* | 0.9663 | 0.6021 | 0.3477 | 0.7784 | 4.9819 | 0.8487 | 0.9562 | 0.3397 |
| *Streptococcus* | 0.8871 | 0.451 | 0.7435 | 0.5304 | 0.3467 | 0.6357 | 0.8596 | 18.3925 |

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8777792/

# Microbe-metabolite correlation heatmap

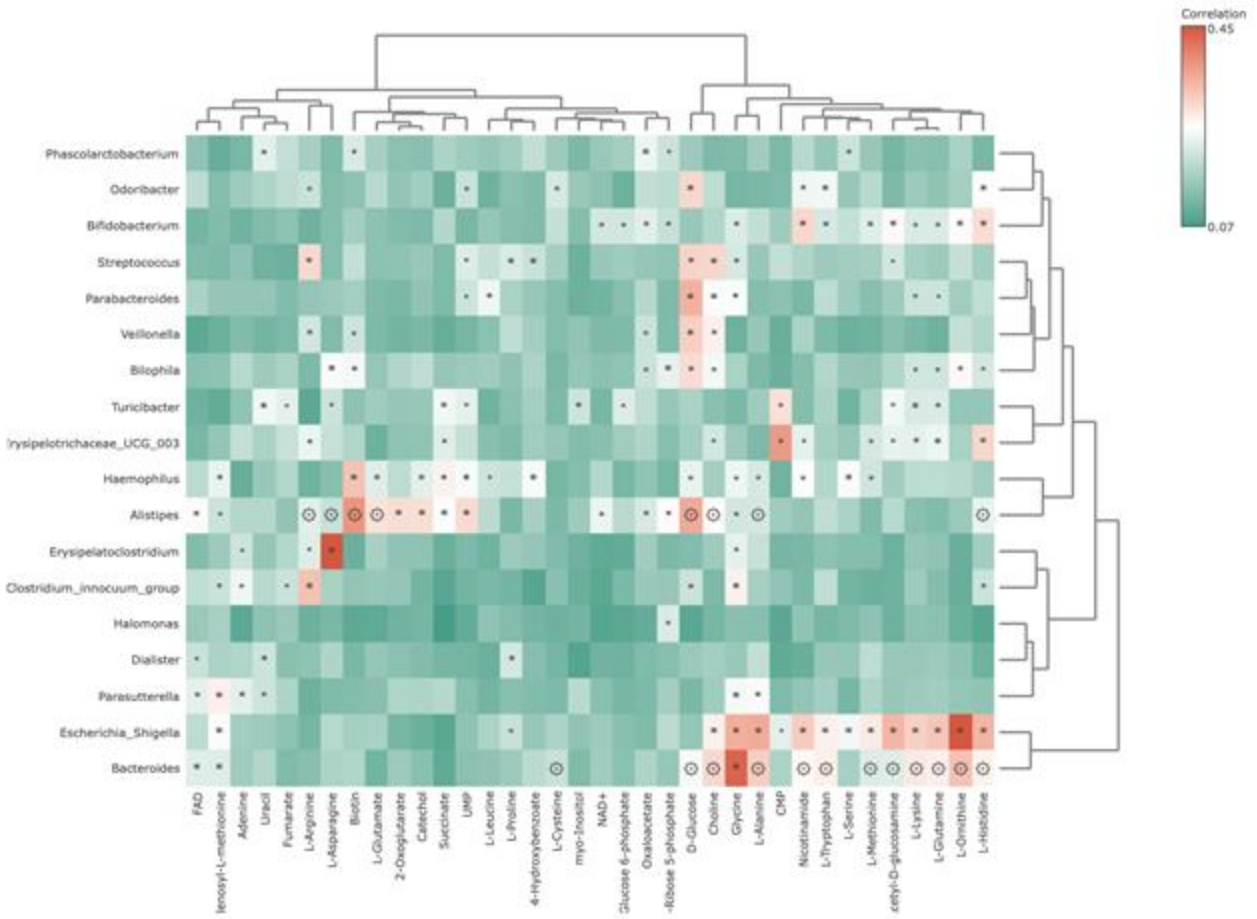
The relationships between the paired microbiome and metabolomics are intuitively presented using an interactive heatmap.

Two types of heatmaps are provided:

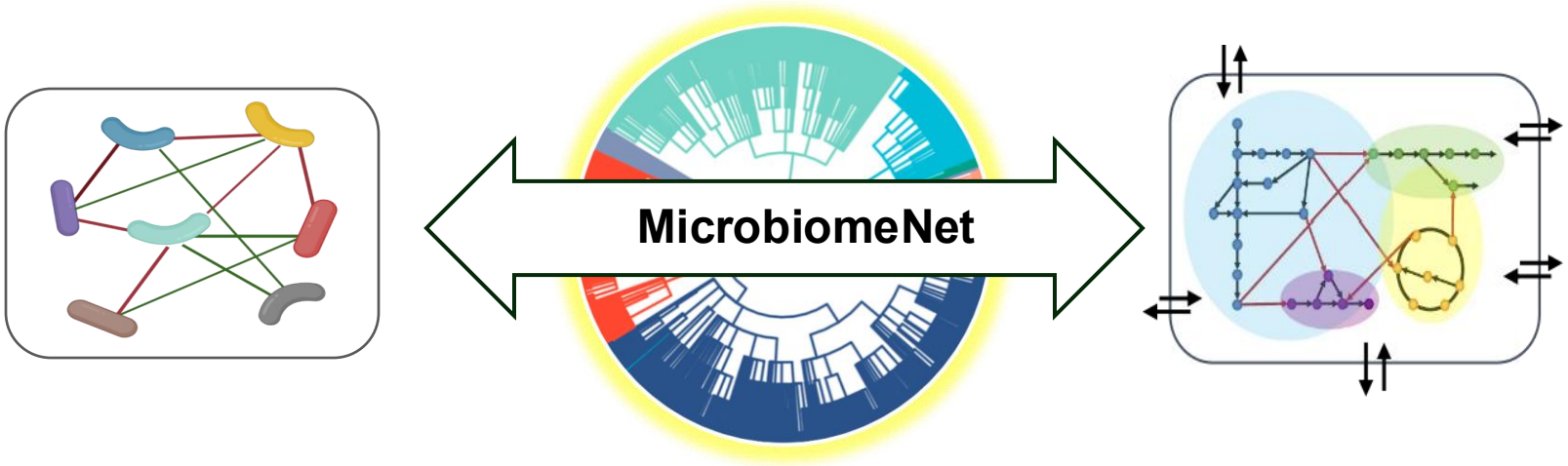
- Correlation heatmap: based on the statistical correlation.
- Probability heatmap:
  - Logistic regression models trained based on high-quality genome-scale metabolic models (GEMs).
  - Reflect the potential of given taxa to produce metabolites

# Overlay statistics and functional results

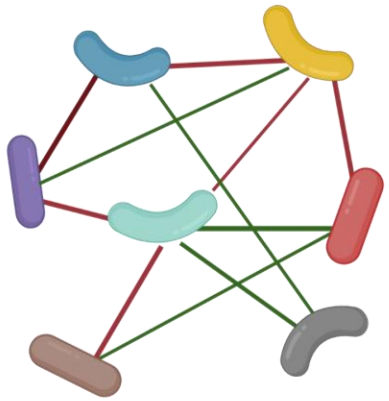
- Statistical pairwise correlation analysis often leads to a high number of false positives, making biological interpretation difficult.
- A model-based correlation based on GEMs to provide a probability heatmap between microbial taxa and their metabolites



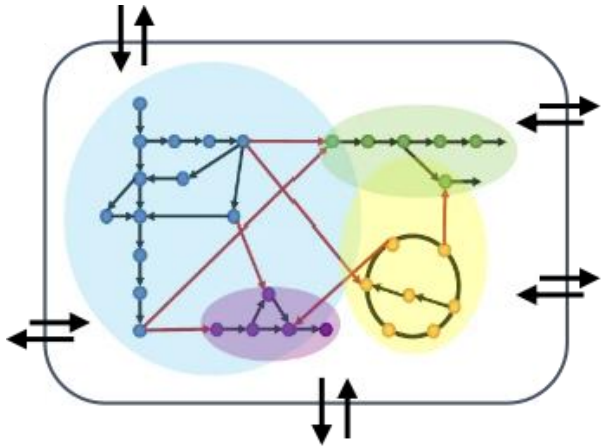
# From statistical associations to potential metabolic interactions



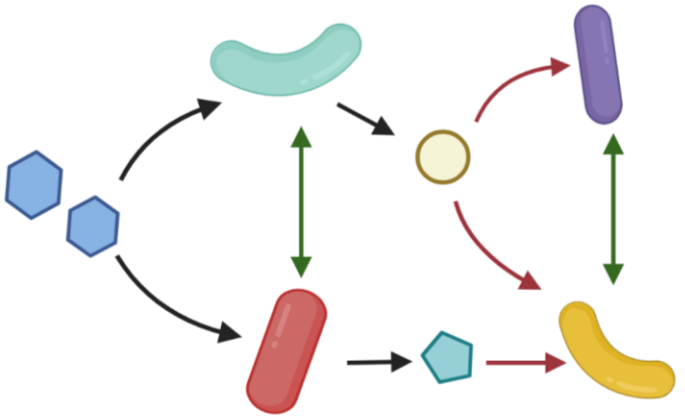
# Understanding co-occurrences through GEMs



Microbial co-occurrence network
> 12,5000



Genome-scale metabolic models (GEMs)
>13,4000



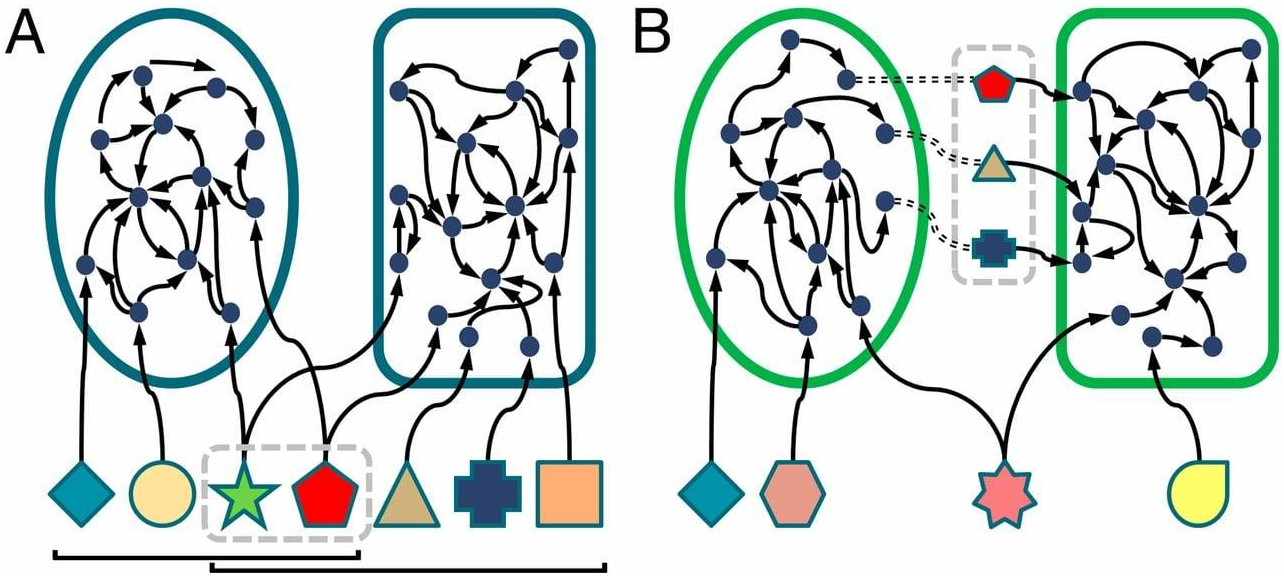
Community interactions & function

# Complement or compete?

**Seed metabolite:**
The minimal group of external compounds required by a specific organism to produce all other compounds in its metabolism.

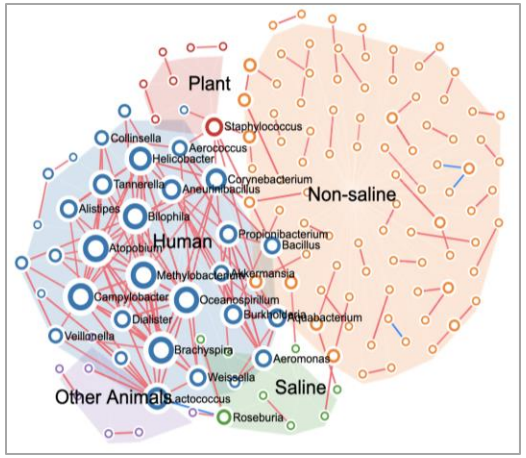
**Competition index:**
The similarity of the nutrition requirement of two microbes, denoted by the proportion of shared seed metabolites.

**Complementarity index:**
The potential for one microbe to produce the seed metabolite required by another microbe.

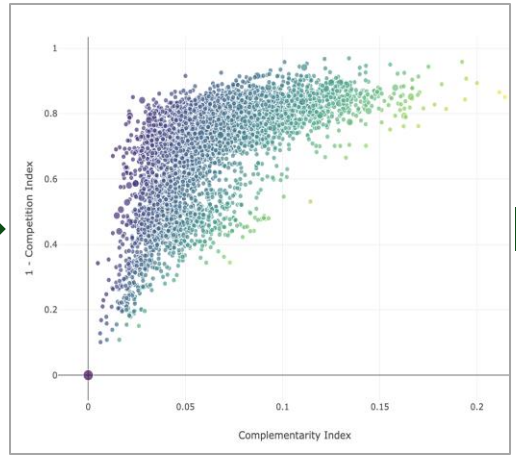



https://www.pnas.org/doi/10.1073/pnas.1300926110

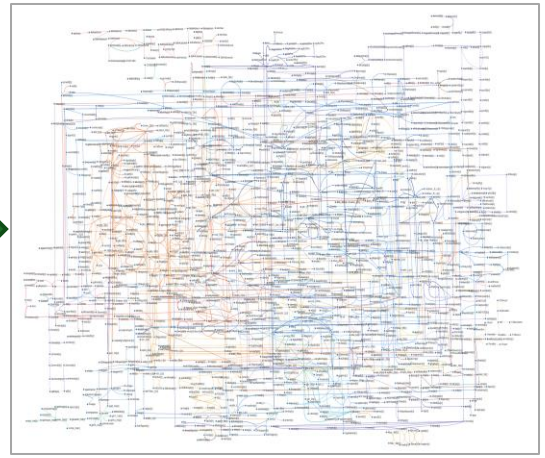
# Conceptual workflow for MicrobiomeNet



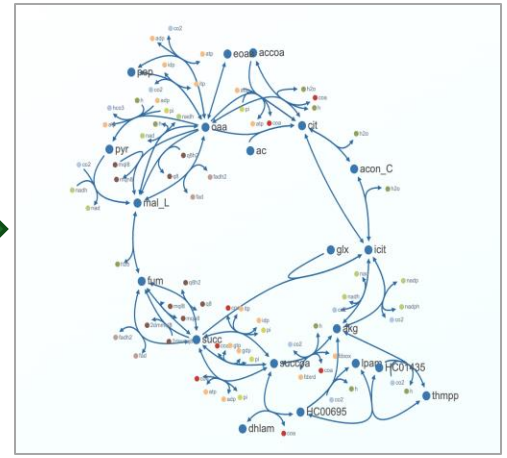

Association Network




Metabolism Compatibility



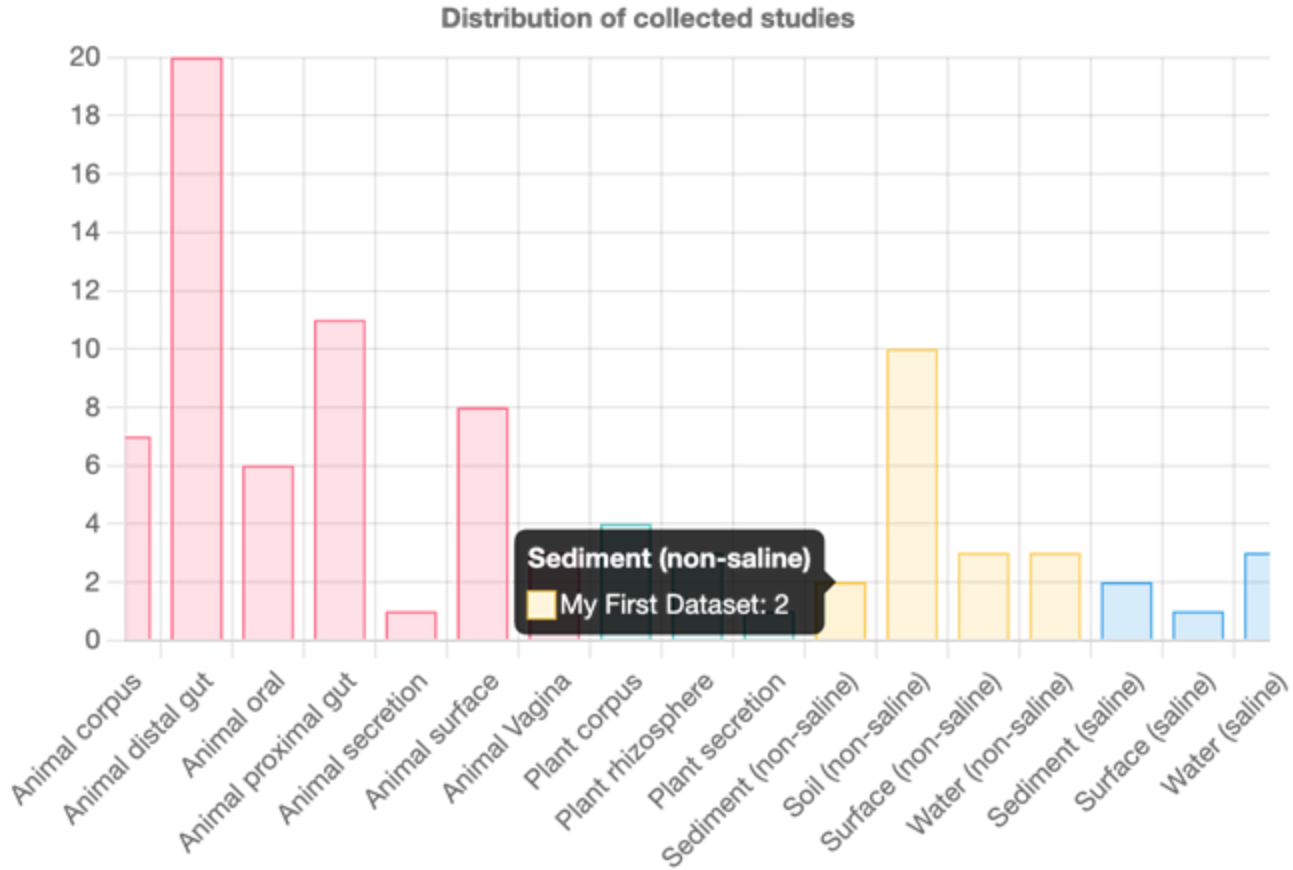
GEM Comparison




Pathway Visualization

# Microbial association networks

- A collection of ~ 5.8 million microbial statistical associations from 76 studies detected using different algorithms
- Including Spearman, Pearson, SparCC, FlashWeave, MINE, Dice Index, and Binding

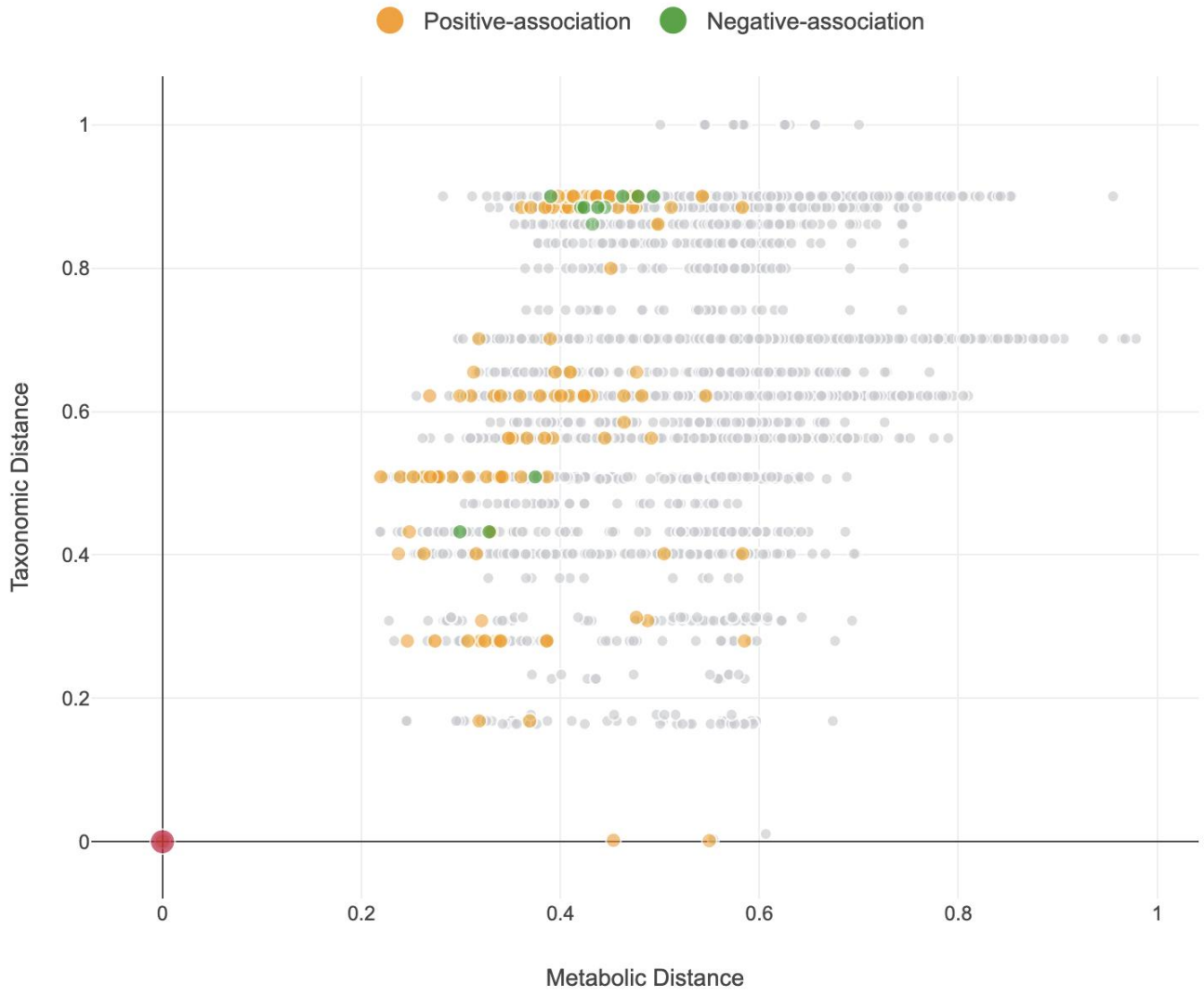


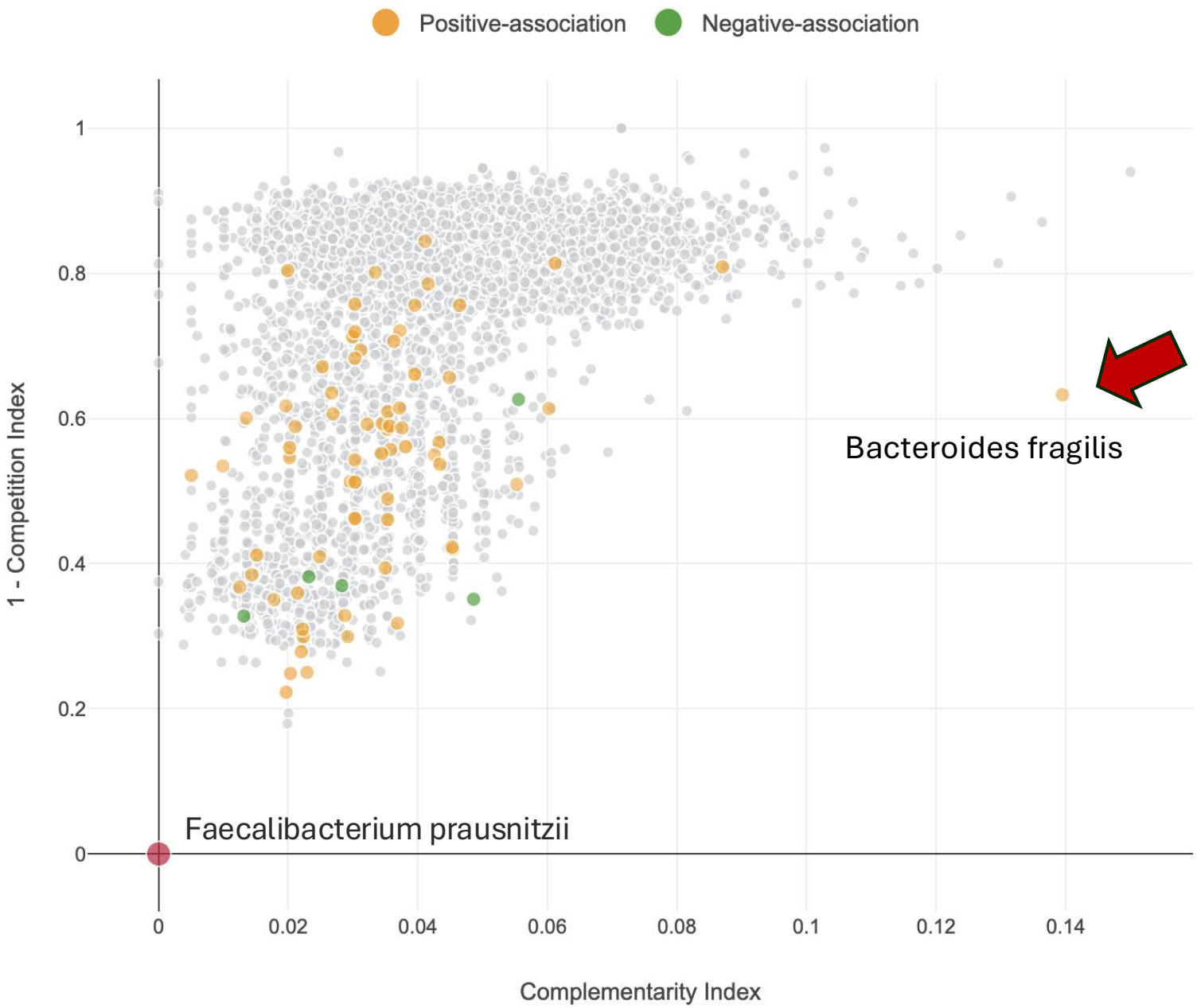
# Neighbourhood Map

A scatter plot integrating metabolic distance (x-axis) and taxonomic distances (y-axis) for all taxa covered in MicrobiomeNet based on GEMs. Both distance are normalized to a scale of 0 to 1 for comparison purposes.

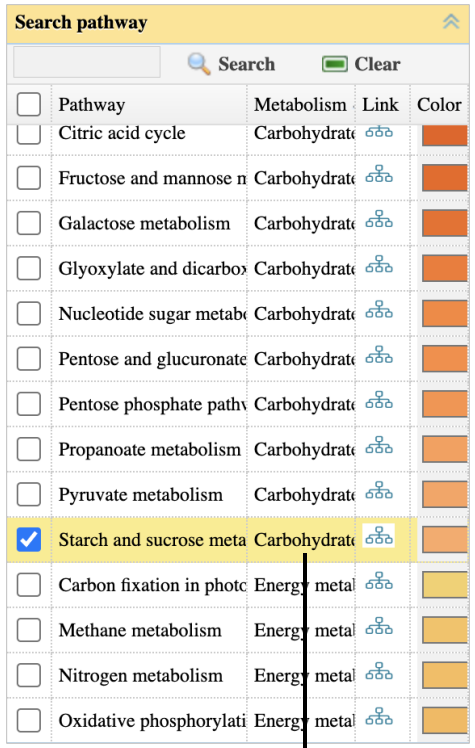
**Metabolic distance:**
Based on the presence or absence of specific reactions in the given taxa.

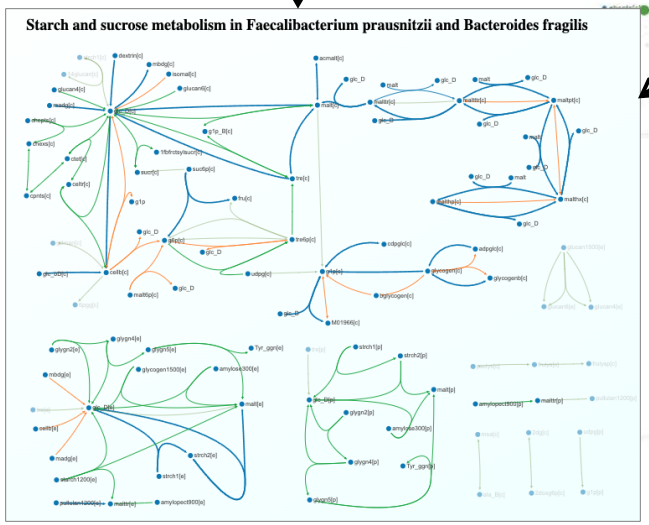
**Taxonomic distances:**
Extract from a phylogenetic tree based on NCBI taxonomy

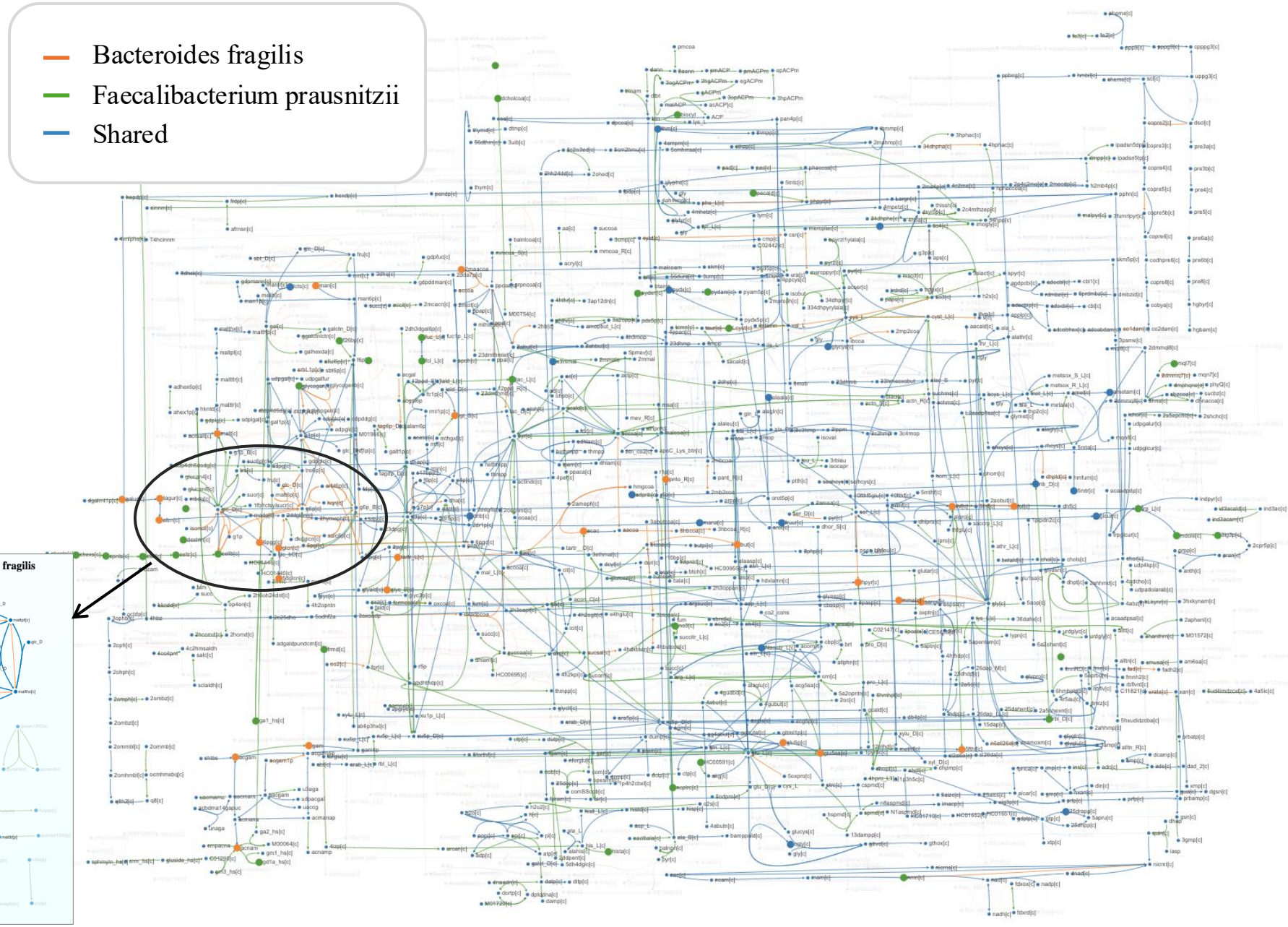



Positive-association
Negative-association
1 - Competition Index
1
0.8
0.6
0.4
0.2
0
0
0.02
0.04
0.06
0.08
0.1
0.12
0.14
Complementarity Index
Bacteroides fragilis
Faecalibacterium prausnitzii

# Metbolism of Faecalibacterium prausnitzii and Bacteroides fragilis



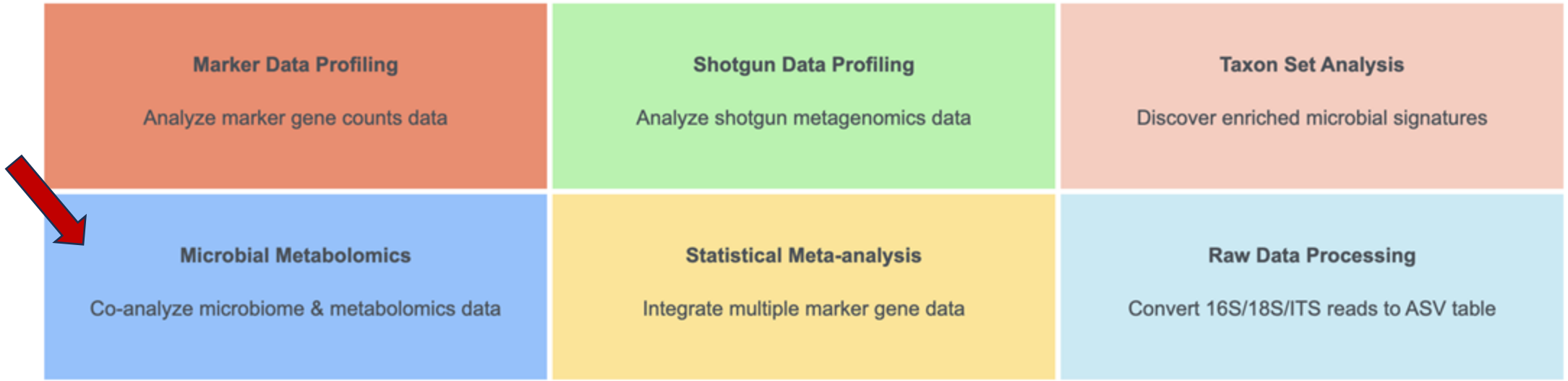



# Demo 4 – microbiome metabolomics data integration

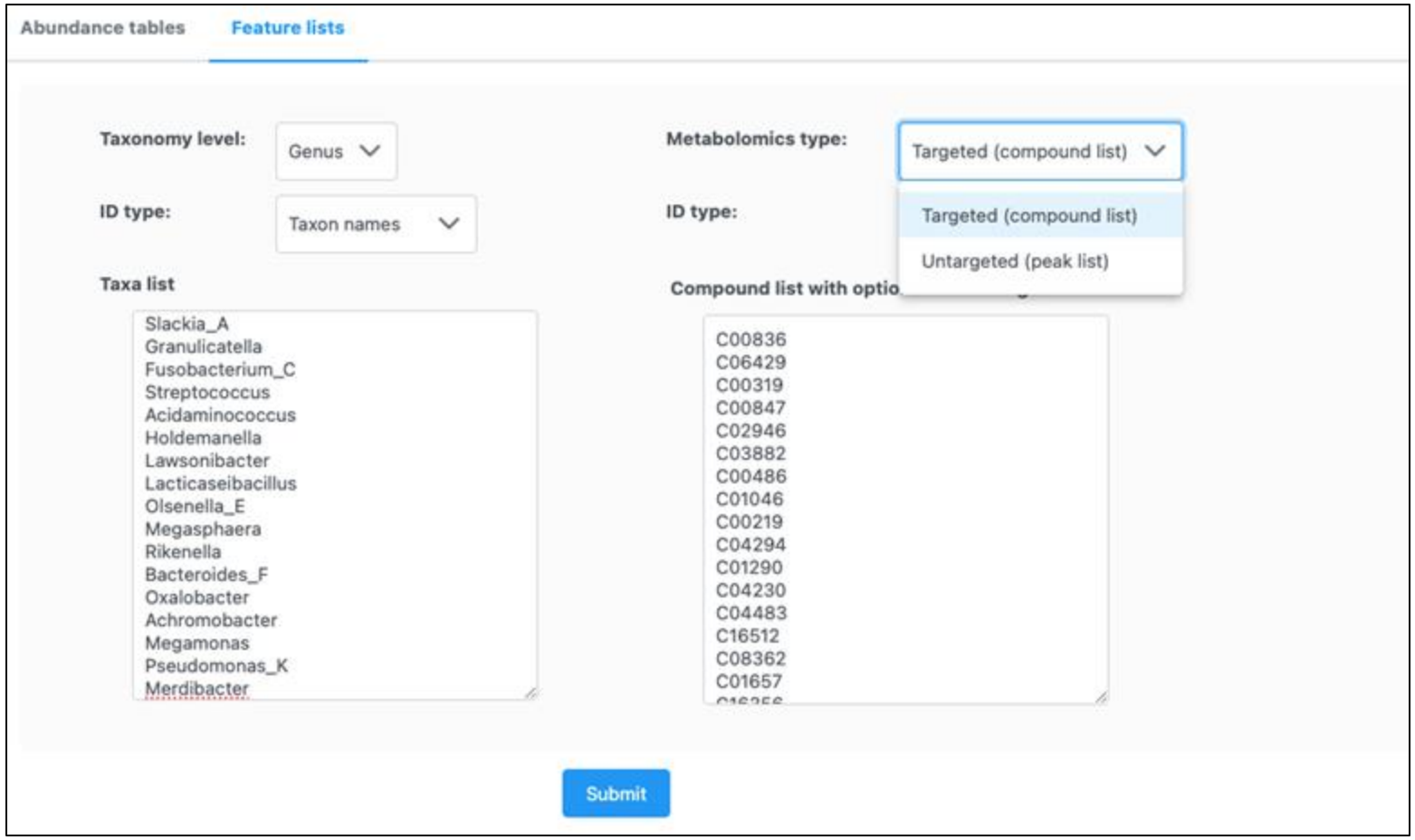
# Shotgun Data Profiling - MicrobiomeAnalyst

- Gene count table from shotgun **metagenomics processing and annotation**



# Input Data Type I

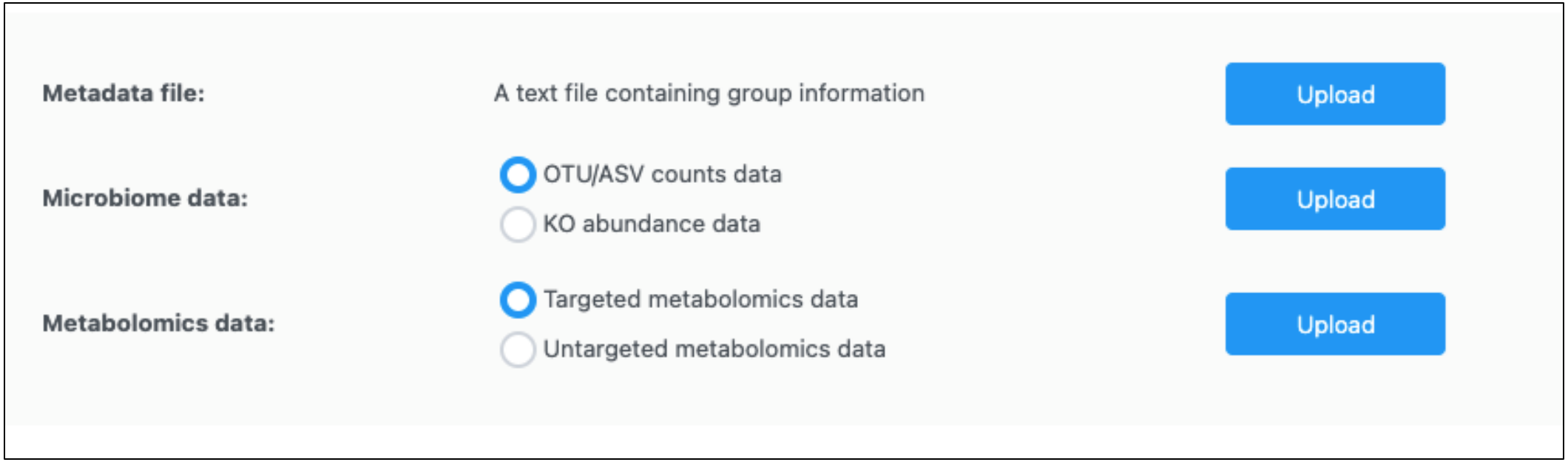
- Taxa list
- Metabolite list/Peak list



# Input Data Type II

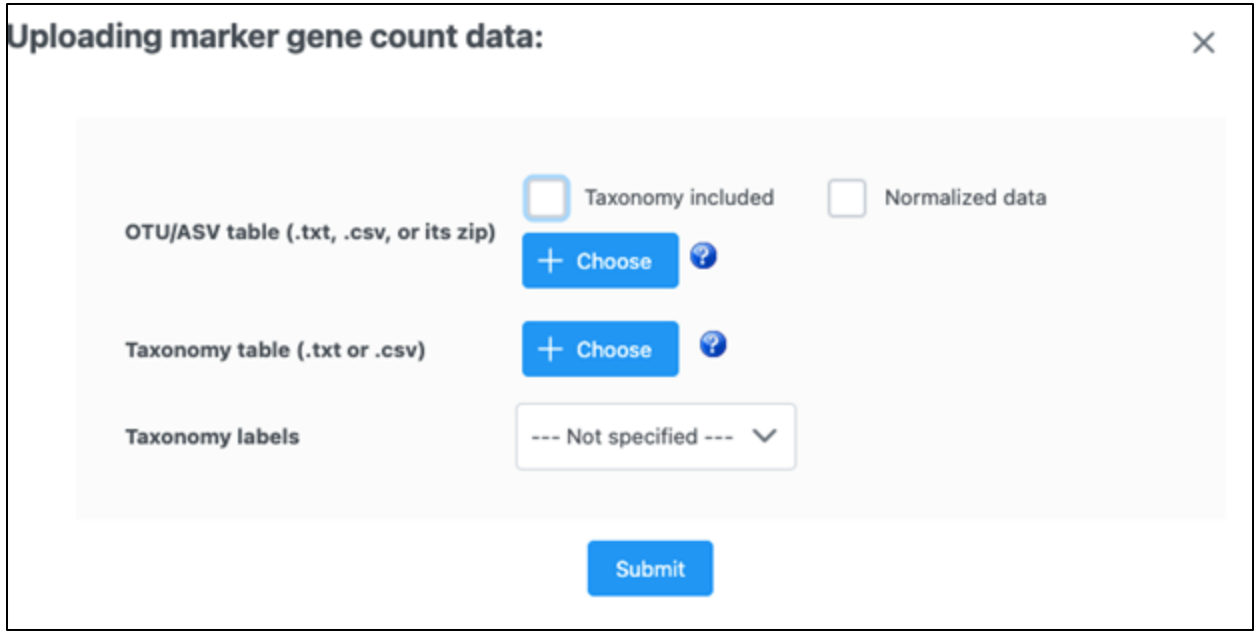
## Paired Abundance Tables:

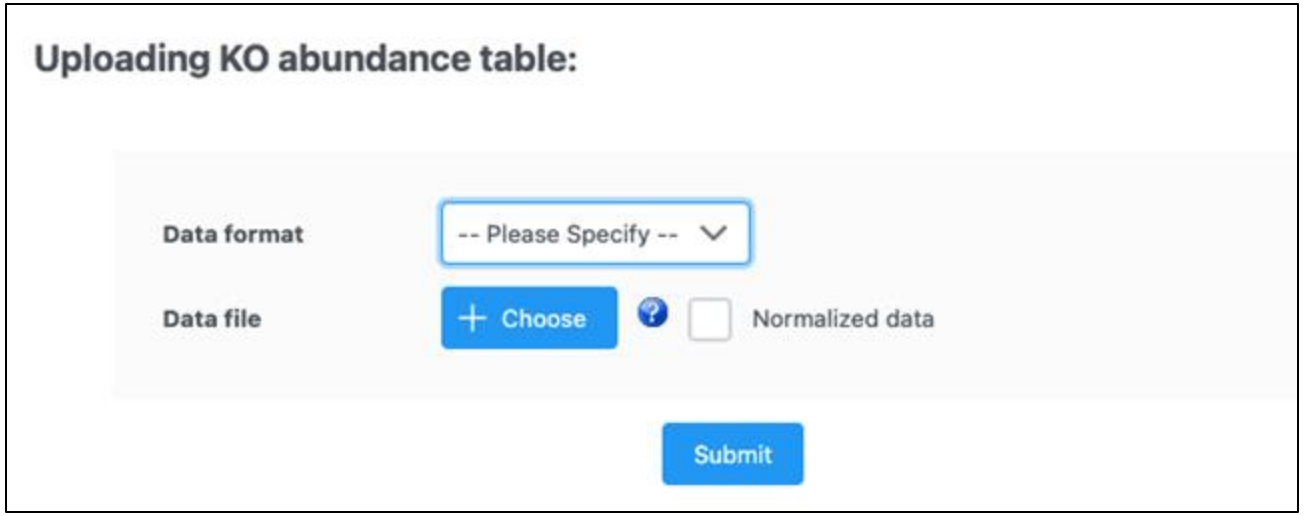
- Only one metadata file is required which is shared by both microbiome and metabolomics data



## ➢ Microbiome data:

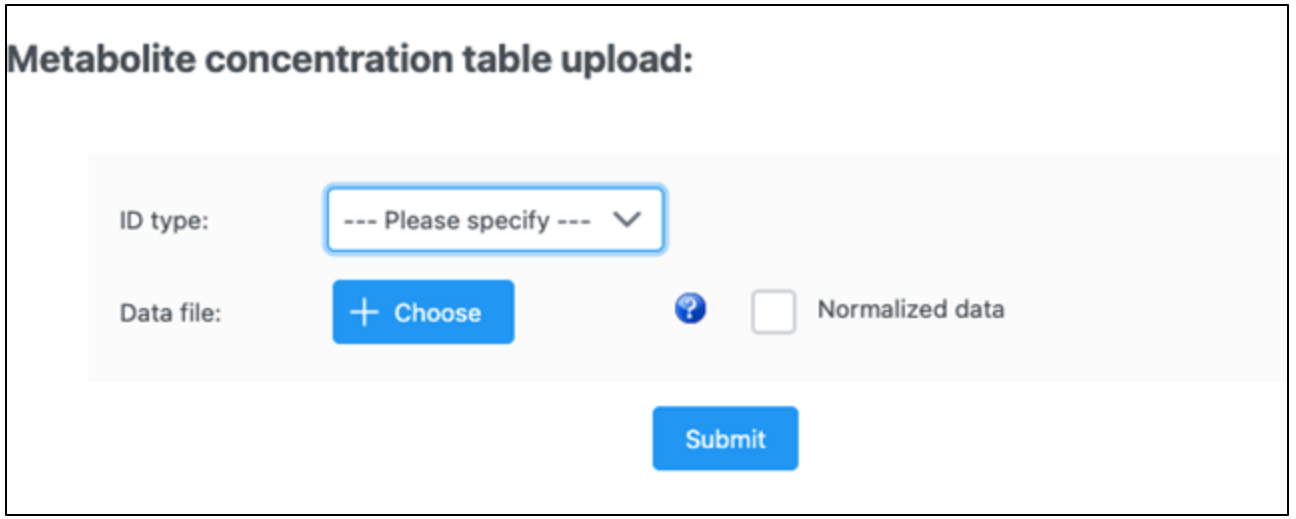
- ASV/OTU abundance table
- KO count table



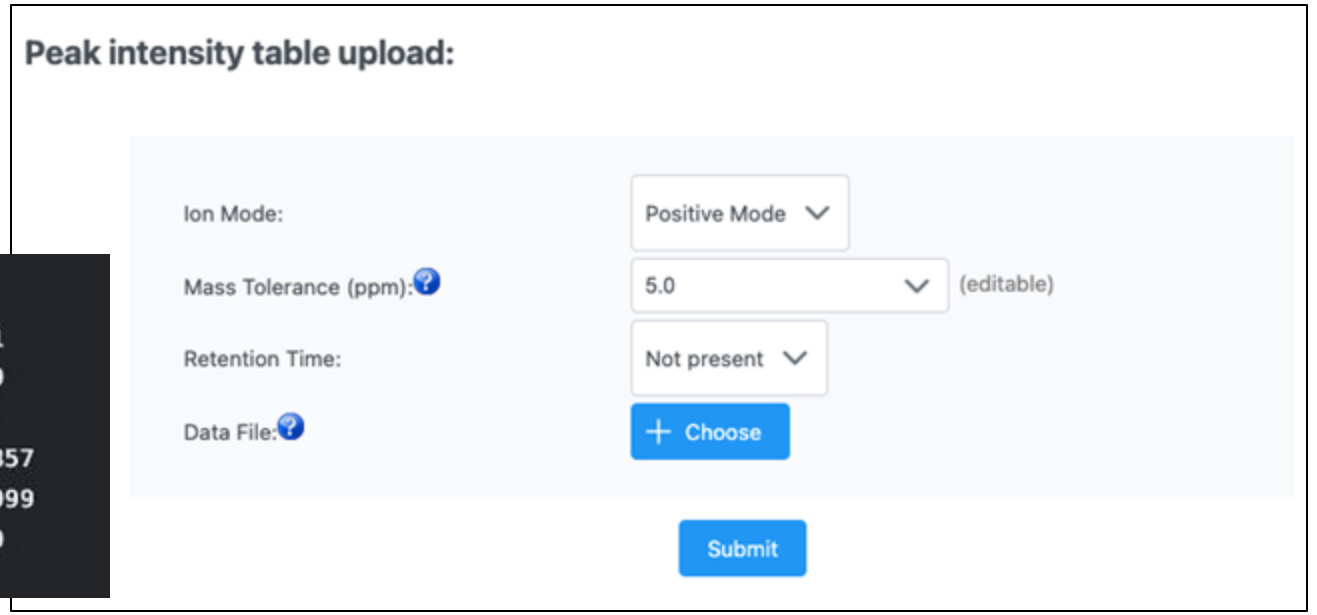



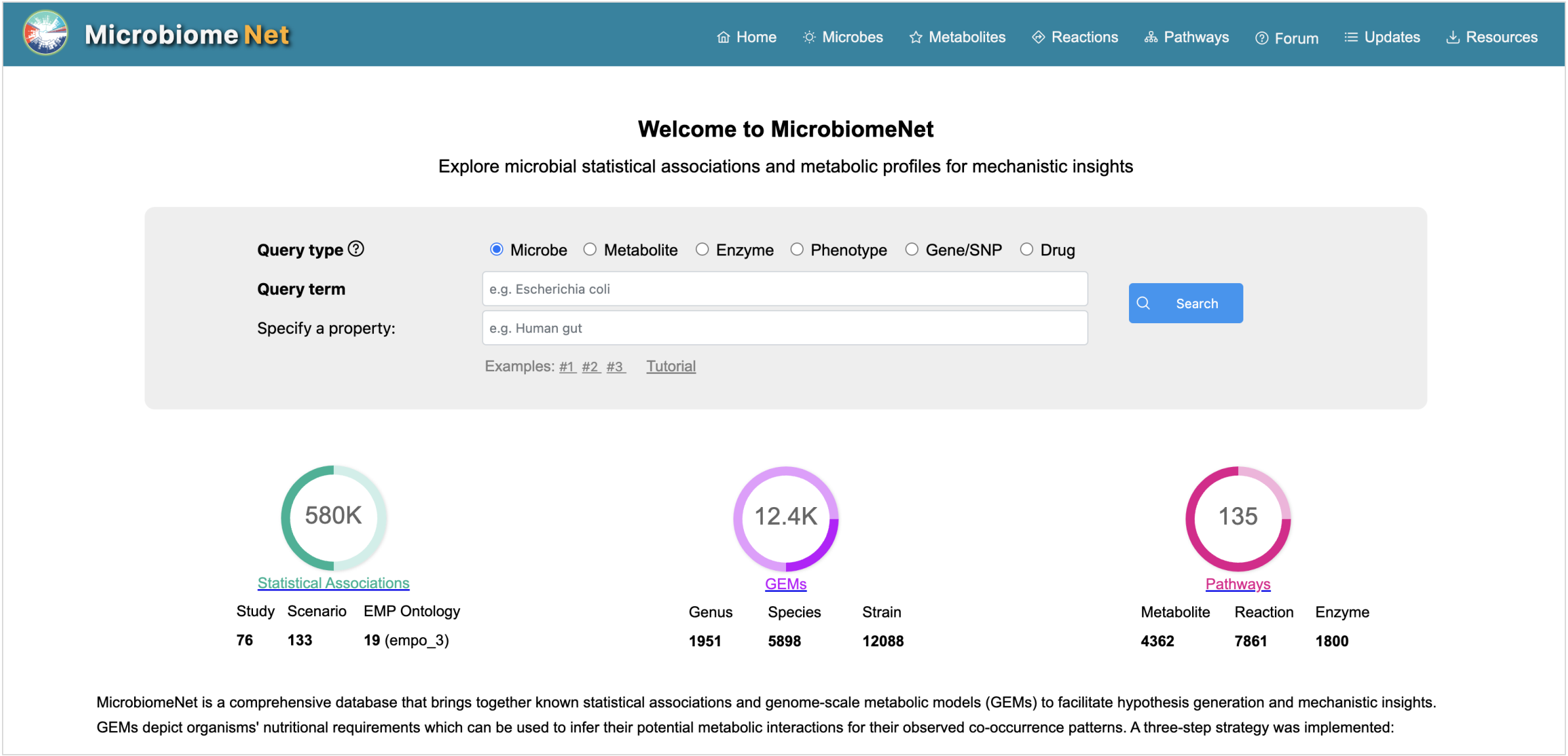
## ➢ Metabolomics data:

- Metabolite concentration table
- Peak intensity table




| #NAME | Sample1 | Sample2 |
|---|---|---|
| 376.3808__0.93 | 119306.8796991 | 181692.0970101 |
| 135.1020__0.98 | 196722.430865503 | 337410.3473859 |
| 374.8815__0.59 | 409570.3544994 | 514658.461539 |
| 86.0603__2.15 | 1087129.323483 | 2038021.41222857 |
| 142.9618__1.35 | 638567.840138999 | 549900.622946999 |
| 129.0663__1.97 | 21840649.9192819 | 4237026.934809 |




Microbiome Net
Home
Microbes
Metabolites
Reactions
Pathways
Forum
Updates
Resources
Welcome to MicrobiomeNet
Explore microbial statistical associations and metabolic profiles for mechanistic insights
Query type
Microbe
Metabolite
Enzyme
Phenotype
Gene/SNP
Drug
Query term
e.g. Escherichia coli
Specify a property:
e.g. Human gut
Search
Examples: #1 #2 #3
Tutorial
580K
Statistical Associations
Study
Scenario
EMP Ontology
76
133
19 (empo_3)
12.4K
GEMs
Genus
Species
Strain
1951
5898
12088
135
Pathways
Metabolite
Reaction
Enzyme
4362
7861
1800
MicrobiomeNet is a comprehensive database that brings together known statistical associations and genome-scale metabolic models (GEMs) to facilitate hypothesis generation and mechanistic insights.
GEMs depict organisms' nutritional requirements which can be used to infer their potential metabolic interactions for their observed co-occurrence patterns. A three-step strategy was implemented:

# Scenario:

1. Whether a microbe can produce a metabolite and how (which pathways);
2. The metabolic capacity of a microbe;
3. Whether two microbes potentially interact with each other, competitively or complementarily;
4. Whether the correlation of interested has been reported in previous studies.

# Hands on Practices (15 min)

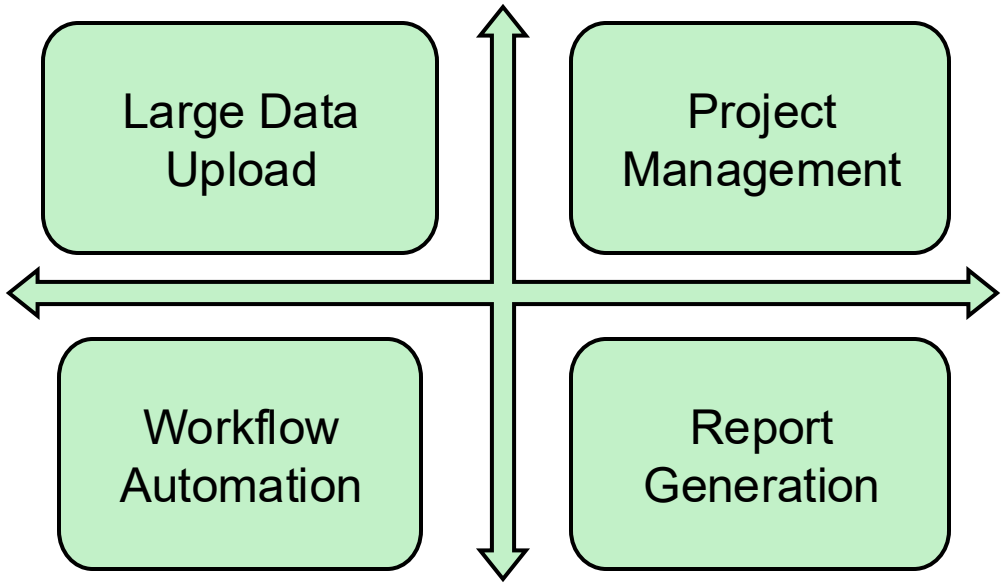
- For intergration analysis in MicrobiomeAnalyst, you can either download the demo data from our github ( https://github.com/xia-lab/Metabolomics_2025/tree/main/docs/mmp_input) or use the files generated from previous sections by yourself.
- Try the example datasets on our website to learn the integration analysis across different data types.
- Search for any microbe or metabolite of interest on MicrobiomeNet for exploration.

# Schedule

| Time | Topics | Lab practices |
|---|---|---|
| 12:00 – 12:15 | General introduction | |
| 12:15 – 1:00 | Metabolomics data processing | Live demo 1 & hands on |
| 1:00 – 1:45 | Microbiome data processing | Live demo 2 & hands on |
| 1:45 – 3:00<br>(15 min break) | Microbial community profiling | Live demo 3 & hands on |
| 3:00 – 3:50 | Microbiome-metabolomics integration | Live demo 4 & hands on |
| 3:50 – 4:15 | Summary and future perspectives | |

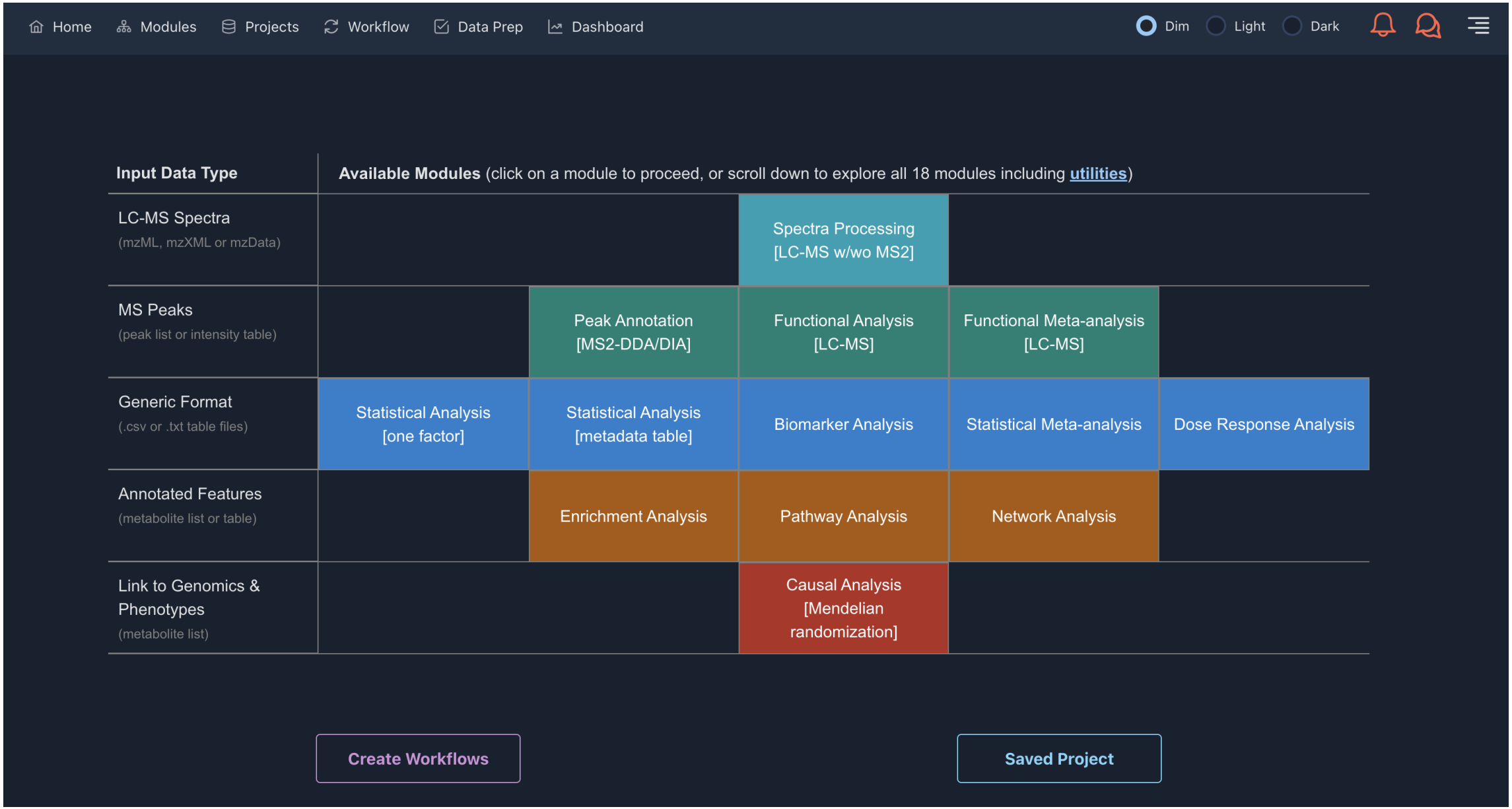
# Help Sustain MetaboAnalyst

- Summer Bootcamp (Tokyo time)
  - Aug. 4 - 8, **9:30 - 16:30**
- Regular Sessions (Montreal time)
  - Saturday morning **9:30 - 12:00**, Sept. - Nov.




Omics Data Science Trainings

"Pro" Tool Suite




# ➔ customize analysis, graphics & report with AI

# Acknowledgements

If you have any questions, please read/post into **OmicsForum** (https://omicsforum.ca)

Contact us:

- yao.lu@xialab.ca
- zhiqiang.pang@xialab.ca
- jeff.xia@xialab.ca