Hands-on MetaboAnalyst: Comprehensive Guide

Data Upload

- Peak intensity table
- Sample groups defined
- Metabolite IDs (HMDB, KEGG)

Normalization

- Sample-specific normalization
- Log transformation
- Scaling methods (auto, pareto)

Statistical Analysis

- t-tests, ANOVA
- PCA, PLS-DA
- Volcano plots, heatmaps

Pathway Analysis

- Enrichment analysis
- Topology analysis
- Visual pathway maps

1. Data Upload

Overview

Data upload is the foundational step in MetaboAnalyst analysis. Proper data formatting ensures accurate downstream analysis and interpretation.

Key Components

- **Peak Intensity Table:** A matrix where rows represent metabolites and columns represent samples. Each cell contains the measured intensity or concentration value.
- **Sample Groups:** Classification of samples into experimental conditions (e.g., Control, Treatment, Disease, Healthy).
- **Metabolite IDs:** Standard identifiers linking detected features to known metabolites:
  - HMDB: Human Metabolome Database IDs
  - KEGG: Kyoto Encyclopedia of Genes and Genomes

Best Practices

- Use CSV or TXT format with proper delimiters
- Ensure no missing sample names or group labels
- Remove special characters from metabolite names
- Verify metabolite ID accuracy before upload

Data Upload Structure

Sample ID	Group	Metabolite 1	Metabolite
Sample_01	Control	1250.5	843.2
Sample_02	Control	1189.3	891.7
Sample_03	Treatment	2305.1	1542.8
Sample_04	Treatment	2187.6	1489.4

Group Labels

Peak Intensities

Metabolite Identification

HMDB0000001 → Glucose
KEGG:C00031 → Lactate

2. Normalization

Overview

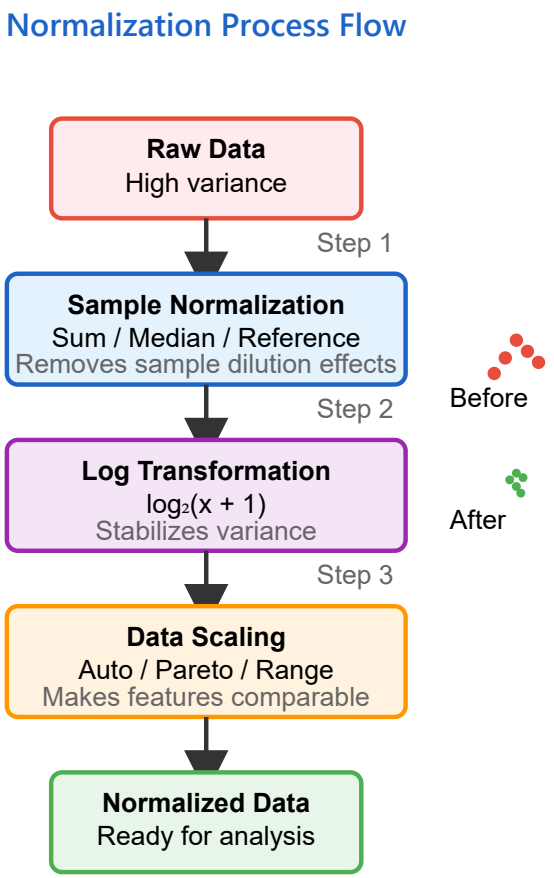
Normalization removes systematic variation and makes samples comparable by addressing technical factors like sample dilution, instrument drift, and analytical variability.

Normalization Methods

- **Sample-Specific Normalization:**
  - By sum: Normalizes to total ion intensity
  - By median: Uses median intensity per sample
  - By reference feature: Uses internal standard
- **Log Transformation:** Reduces heteroscedasticity and makes data more normally distributed. Common choices: log2, log10, or natural log.
- **Scaling Methods:**
  - Auto scaling (unit variance): Mean-centered, divided by SD
  - Pareto scaling: Mean-centered, divided by √SD
  - Range scaling: Scaled to unit range [0,1]

When to Apply

- Always normalize when comparing across samples
- Apply log transformation for wide dynamic ranges



- Use appropriate scaling based on variance structure

3. Statistical Analysis

Overview

Statistical analysis identifies significant metabolic differences between groups and reveals patterns in complex datasets through multivariate and univariate methods.

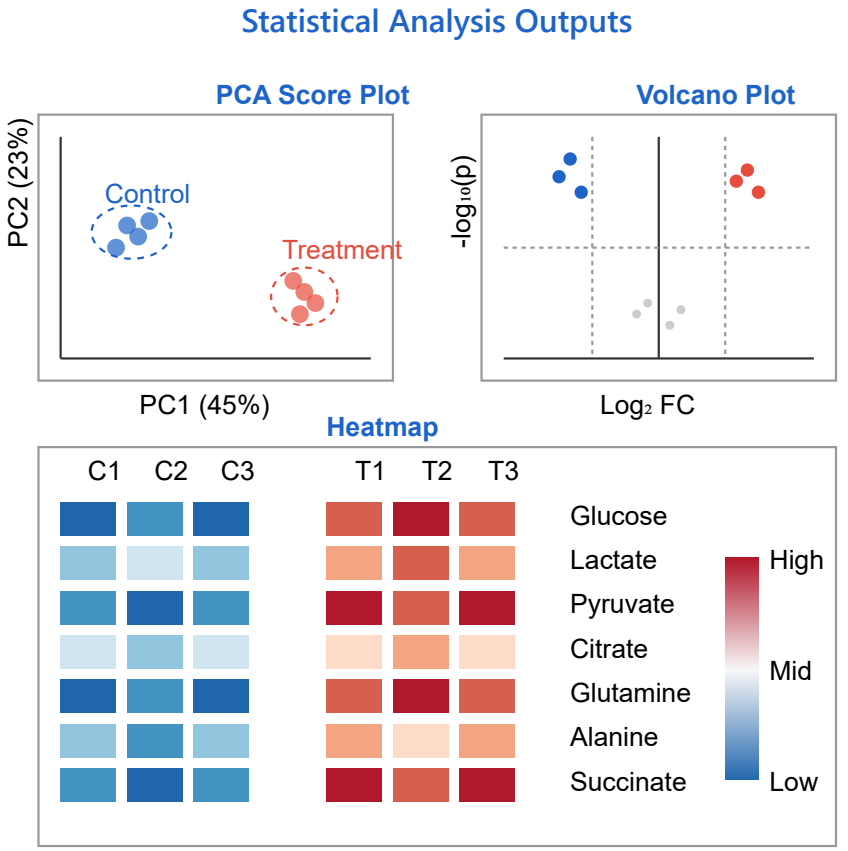
Univariate Methods

- **t-tests:** Compares means between two groups. Options include Student's t-test (equal variance) and Welch's t-test (unequal variance).
- **ANOVA:** Analysis of variance for comparing three or more groups. Identifies which metabolites differ significantly across conditions.
- **Fold Change:** Ratio of mean values between groups, often expressed on log2 scale.

Multivariate Methods

- **PCA (Principal Component Analysis):** Unsupervised method that reduces dimensionality and reveals sample clustering patterns without using group information.
- **PLS-DA (Partial Least Squares Discriminant Analysis):** Supervised method that maximizes separation between predefined groups while identifying discriminating metabolites.

Visualization Tools



- **Volcano Plots:** Display fold change vs. statistical significance, highlighting metabolites that are both large in magnitude and statistically significant.
- **Heatmaps:** Show hierarchical clustering of samples and metabolites, revealing patterns across the entire dataset.

4. Pathway Analysis

Overview

Pathway analysis connects metabolite changes to biological pathways, providing mechanistic insights into metabolic alterations and identifying key regulatory points.

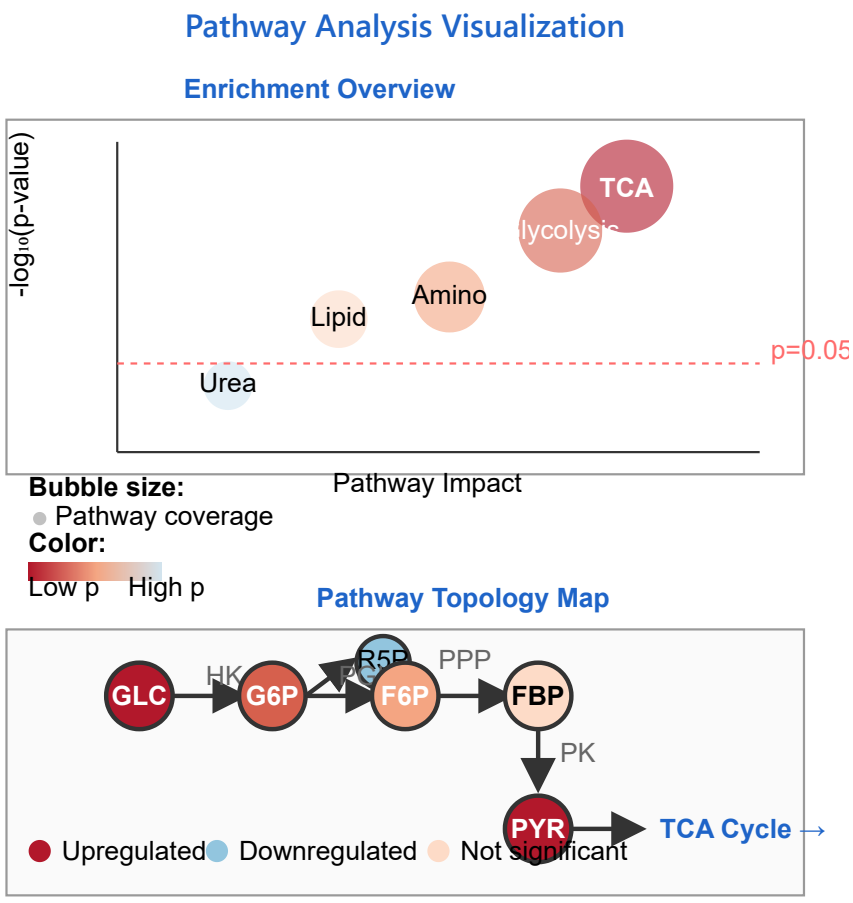
Enrichment Analysis

- **Over-Representation Analysis (ORA):** Tests whether significantly changed metabolites are overrepresented in specific pathways compared to random chance.
- **Hypergeometric Test:** Statistical method to determine pathway significance based on the proportion of pathway metabolites detected.
- **P-value & FDR:** Correction for multiple testing using False Discovery Rate (Benjamini-Hochberg method).

Topology Analysis

- **Impact Score:** Measures the importance of a pathway based on the positions of detected metabolites within the pathway network.
- **Centrality Measures:** Considers betweenness and degree centrality to identify critical pathway nodes.
- **Pathway Impact:** Metabolites at pathway branch points have higher impact than peripheral metabolites.

Visual Pathway Maps



- **KEGG Pathway Integration:** Maps metabolites onto KEGG pathway diagrams with color-coded expression levels.
- **Interactive Views:** Click-through access to metabolite details and related pathways.
- **Pathway Networks:** Shows interconnections between enriched pathways.