Innovative Metabolomics Insights for Better Health

Amino Acids Metabolomics Assay Final Report

Metware Biotechnology Inc.

www.metwarebio.com

Contents

MWXS-23-xxx Amino Acids Targeted Metabolomics Assay Final Report

1 Abstract

Amino acids serve as the building blocks of life, which are directly or indirectly linked to all human diseases and health conditions. Within the body, amino acid metabolism maintains a dynamic equilibrium, with blood amino acids acting as the central pivot and the liver playing a pivotal role in regulating their levels. The development and progression of various ailments, spanning cardiovascular, renal, diabetic, oncologic, geriatric, and neurologic disorders, can result in disturbances to amino acid metabolism and serum amino acid concentrations. Meanwhile, there are over 400 recognized diseases stem from compromised amino acid metabolism. Amino acid testing has evolved into an indispensable diagnostic and disease screening tool, concurrently serving as a reference standard for nutritional supplementation, overall nutritional well-being enhancement, and early disease prevention in all populations. At MetwareBio, we've established an LC-MS/MS-based platform designed for the comprehensive analysis of 94 amino acids and their derivatives, allowing precise targeting and quantitation.

2 The experimental process

Liquid chromatography-tandem mass spectrometry (LC-MS/MS) can detect and quantify compounds with high polarity and poor thermal stability, and accurately quantify them. The overall process is as follows:

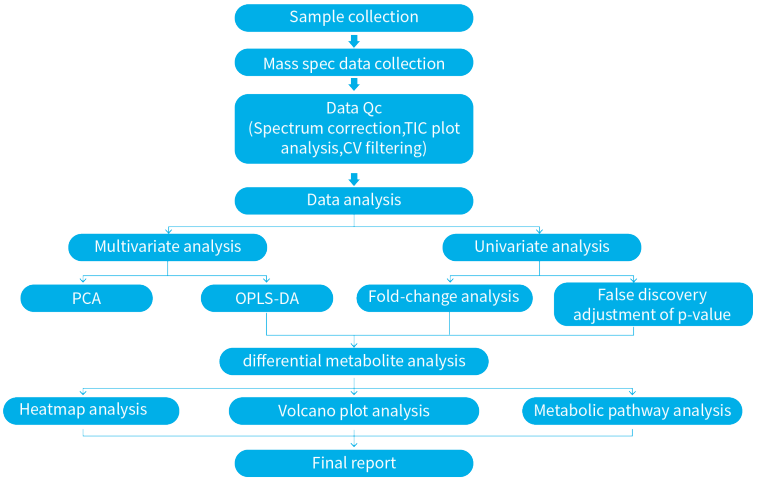

Fig 1: Flow chart of metabolomics analysis

Compounds to be detected:

Table 1: List of compounds in the panel

Number	Compounds	Index
1	2-Aminoethanesulfonic Acid	2-Aminoethanesulfonic-Acid
2	L-Cystine	L-Cystine
3	1,3-Dimethyluric Acid	1,3-Dimethyluric-Acid
4	N-Propionylglycine	N-Propionylglycine
5	N-Isovaleroylglycine	N-Isovaleroylglycine
6	Succinic Acid	Succinic-Acid
7	5-Hydroxy-tryptophan	5-Hydroxy-tryptophan
8	3,7-Dimethyluric Acid	3,7-Dimethyluric-Acid
9	Glycine	Gly
10	L-Alanine	Ala
11	L-Valine	Val
12	L-Leucine	Leu
13	L-Methionine	Met
14	L-Isoleucine	Ile
15	L-Proline	Pro
16	L-Serine	Ser
17	L-Tryptophan	Trp
18	L-Phenylalanine	Phe
19	L-Tyrosine	Tyr
20	L-Cysteine	Cys
21	L-Glutamic acid	Glu
22	L-Aspartate	Asp
23	L-Asparagine Anhydrous	Asn
24	L-Glutamine	Gln
25	L-Lysine	Lys
26	L-Histidine	His
27	L-Arginine	Arg
28	L-Threonine	Thr
29	L-Citrulline	L-Citrulline
30	5-Hydroxy-Tryptamine	5-Hydroxy-Tryptamine
31	L-Homocitrulline	L-Homocitrulline
32	Beta-Alanine	Beta-Alanine
33	Sarcosine	Sarcosine
34	L-Pipecolic Acid	L-Pipecolic-Acid
35	L-Theanine	L-Theanine
36	Ethanolamine	Ethanolamine
37	3-N-Methyl-L-Histidine	3-N-Methyl-L-Histidine
38	Homoserine	Homoserine
39	Creatine	Creatine
40	Kinurenine	Kinurenine
41	L-Cystathionine	L-Cystathionine
42	5-Aminovaleric Acid	5-Aminovaleric-Acid
43	N6-Acetyl-L-Lysine	N6-Acetyl-L-Lysine
44	Phosphorylethanolamine	Phosphorylethanolamine
45	Anserine	Anserine
46	Trans-4-Hydroxy-L-Proline	Trans-4-Hydroxy-L-Proline
47	D-Homocysteine	D-Homocysteine

Table 1: List of compounds in the panel

Number	Compounds	Index
48	α-Aminoadipic acid	α-Aminoadipic-acid
49	L-Ornithine	L-Ornithine
50	L-tyrosine methyl ester	L-tyrosine-methyl-ester
51	2-Aminobutyric acid	2-Aminobutyric-acid
52	(5-L-Glutamyl)-L-Amino Acid	(5-L-Glutamyl)-L-Amino-Acid
53	3-Iodo-L-Tyrosine	3-Iodo-L-Tyrosine
54	P-Aminohippuric Acid	P-Aminohippuric-Acid
55	Glycyl-L-Proline	Glycyl-L-Proline
56	Trimethylamine N-Oxide	Trimethylamine-N-Oxide
57	1,3,7-Trimethyluric Acid	1,3,7-Trimethyluric-Acid
58	3-Hydroxyhippuric Acid	3-Hydroxyhippuric-Acid
59	N8-Acetylspermidine	N8-Acetylspermidine
60	(S)-β-Aminoisobutyric Acid	(S)-β-Aminoisobutyric-Acid
61	S-Sulfo-L-Cysteine	S-Sulfo-L-Cysteine
62	Methionine Sulfoxide	Methionine-Sulfoxide
63	Nα-Acetyl-L-Arginine	Nα-Acetyl-L-Arginine
64	1-Methylhistidine	1-Methylhistidine
65	γ-Glutamate-Cysteine	γ-Glutamate-Cysteine
66	Nα-Acetyl-L-glutamine	Nα-Acetyl-L-glutamine
67	N-Acetyl-L-Tyrosine	N-Acetyl-L-Tyrosine
68	γ-Aminobutyric Acid	γ-Aminobutyric-Acid
69	D-Alanyl-D-Alanine	D-Alanyl-D-Alanine
70	Guanidinoethyl Sulfonate	Guanidinoethyl-Sulfonate
71	Homo-L-arginine	Homo-Arg
72	L-Tryptophyl-L-glutamic acid	TRP-GLU
73	Nicotinuric Acid	Nicotinuric-Acid
74	N-Acetylneuraminic Acid	N-Acetylneuraminic-Acid
75	N,N-Dimethylglycine	N,N-Dimethylglycine
76	4-Acetamidobutyric Acid	4-Acetamidobutyric-Acid
77	L-Carnosine	L-Carnosine
78	6-Aminocaproic Acid	6-Aminocaproic-Acid
79	3-Chloro-L-Tyrosine	3-Chloro-L-Tyrosine
80	S-(5-Adenosyl)-L-Homocysteine	S-(5-Adenosyl)-L-Homocysteine
81	Kynurenic Acid	Kynurenic-Acid
82	N'-Formylkynurenine	N'-Formylkynurenine
83	Urea	Urea
84	argininosuccinic acid	argininosuccinic-acid
85	5-Hydroxylysine	5-Hydroxylysine
86	O-Phospho-L-Serine	O-Phospho-L-Serine
87	N-Acetylaspartate	N-Acetylaspartate
88	L-Homocystine	L-Homocystine
89	3-Aminoisobutanoic Acid	3-Aminoisobutanoic-Acid
90	Glutathione Oxidized	Glutathione-Oxidized
91	L-α-Aspartyl-L-phenylalanine	Asp-Phe
92	N-Glycyl-L-Leucine	N-Glycyl-L-Leucine
93	Creatine Phosphate	Creatine-Phosphate
94	glycylphenylalanine	glycylphenylalanine

Original file path: Final report/data/component.xlsx

2.1 Sample information

This project has 36 samples divided into 6 groups. Sample information is shown in the following table:

Table 2: Sample information table

Species	Tissues	MW_ID	Sample_ID
-	-	A2	A2
-	-	A3	A3
-	-	A4	A4
-	-	A1	A1
-	-	A6	A6
-	-	A7	A7
-	-	B2	B2
-	-	B3	B3
-	-	B4	B4
-	-	B1	B1
-	-	B6	B6
-	-	B7	B7
-	-	C1	C1
-	-	C2	C2
-	-	C5	C5
-	-	C4	C4
-	-	C3	C3
-	-	C7	C7
-	-	D1	D1
-	-	D2	D2
-	-	D5	D5
-	-	D4	D4
-	-	D3	D3
-	-	D7	D7
-	-	E1	E1
-	-	E2	E2
-	-	E5	E5
-	-	E4	E4
-	-	E3	E3
-	-	E7	E7
-	-	F1	F1
-	-	F2	F2
-	-	F5	F5
-	-	F4	F4
-	-	F3	F3
-	-	F7	F7

Original file path: Final report/0.data/sample_info.xlsx

2.2 Reagents and instruments

Table 3: Instrument information

Instrument	Model	Manufacturer
LC-MS/MS	Triple Quad 6500+	SCIEX
Centrifuge	5424R	Eppendorf
Electronic balance	AS 60/220.R2	RADWAG
Multitube vortex oscillator	MIX-200	ShangHaiJingXin
Ultrasonic cleaning apparatus	CD-F15	Olenyer

Table 4: Information of standards and reagents

Reagent	level	Manufacturer
Methanol	HPLC	Thermo fisher
Acetonitrile	HPLC	Thermo fisher
Formic acid	HPLC	Thermo fisher
Chemical standard	99%	Sigma-Aldrich/Zhenzhun.etc

2.3 Sample extraction process

After the sample was thawed and smashed, 0.05 g sample was used for extraction with 500 μL of 70% methanol/water. The sample was vortexed for 3 min under the condition of 2500 r/min and centrifuged at 12000 r/min for 10 min at 4°C. Transfer 300 μL of supernatant into a new centrifuge tube and place the supernatant in -20°C refrigerator for 30 min, and then the supernatant was centrifuged again at 12000 r/min for 10 min at 4°C. After centrifugation, transfer 200 μL of supernatant through Protein Precipitation Plate for further LC-MS analysis.

2.4 Chromatography-mass spectrometry acquisition conditions

The sample extracts were analyzed using an LC-ESI-MS/MS system (UPLC, ExionLC AD, https://sciex.com/; MS, QTRAP® 6500+ System, https://sciex.com/). The analytical conditions were as follows, HPLC: column, ACQUITY BEH Amide (i.d.2.1×100 mm, 1.7 μm); solvent system, water with 2 mM ammonium acetate and 0.04% formic acid (A), acetonitrile with 2 mM ammonium acetate and 0.04% formic acid (B); The gradient was started at 90% B (0-1.2 min), decreased to 60% B (9 min), 40% B (10-11 min), finaly ramped back to 90% B (11.01-15 min); flow rate, 0.4 mL/min; temperature, 40°C; injection volume: 2 μL.

AB 6500+ QTRAP® LC-MS/MS System, equipped with an ESI Turbo Ion-Spray interface, operating in both positive and negative ion modes and controlled by Analyst 1.6 software (AB Sciex). The ESI source operation parameters were as follows: ion source, turbo spray; source temperature 550°C; ion spray voltage (IS) 5500 V (Positive), -4500 V (Negative); curtain gas (CUR) were set at 35.0 psi; DP and CE for individual MRM transitions was done with further DP and CE optimization. A specific set of MRM transitions were monitored for each period according to the amino acid eluted within this period.

2.5 Qualitative and quantitative principles of metabolites

Metabolites were quantified by multiple reaction monitoring (MRM) using triple quadrupole mass spectrometry. In MRM mode, the first quadrupole screened the precursor ions for the target substance and excluded ions of other molecular weights. After ionization induced by the impact chamber, the precursor ions were fragmented, and a characteristic fragment ion was selected through the third quadrupole to exclude the interference of non-target ions. After obtaining the metabolite spectrum data from different samples, the peak area was calculated on the mass spectrum peaks of all substances and analyzed by standard curves.

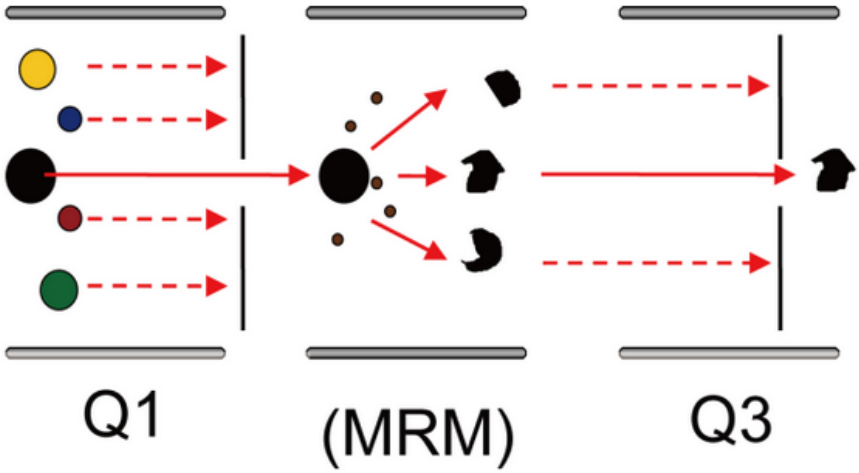

Fig 2:
Schematic diagram of multiple reaction monitoring mode by mass spectrometry

3 Data evaluation

3.1 Data pre-processing

Analyst 1.6.3 was used to process mass spectrum data. The following figure shows the total ions current (TIC) and MRM metabolite detection multi-peak diagram (XIC) of the mixed QC samples. The X-axis shows the Retention time (RT) from metabolite detection, and the Y-axis shows the ion flow intensity from ion detection (intensity unit: CPS, count per second).

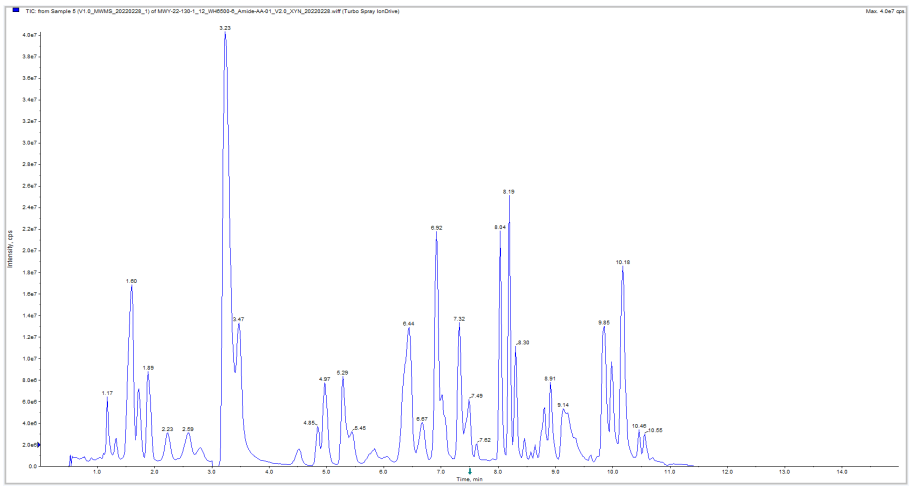

Fig 3: Total ion current diagram of mixed phase mass spectrum analysis

Original file path: Final report/0.data/QC/*QC_MS_TIC.png

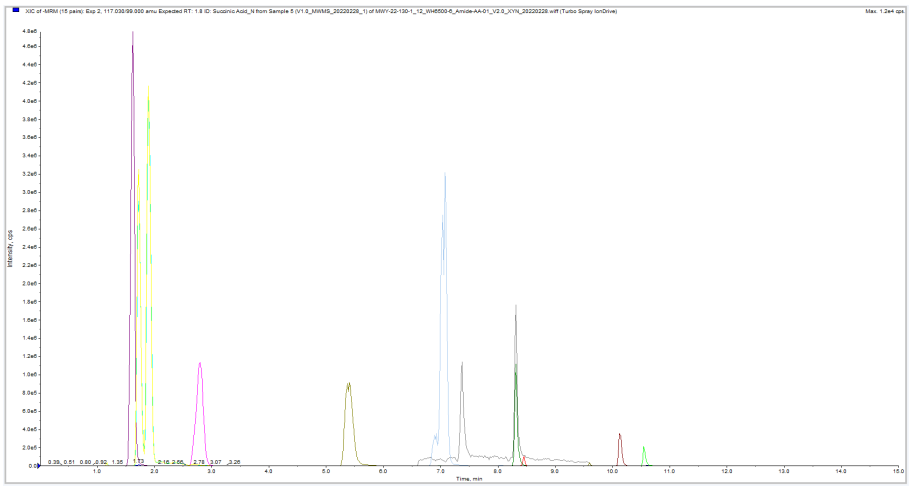

Fig 4: Extraction ion flow chromatogram

Original file path: Final report/0.data/QC/*MRM_detection_of_multimodal_maps*

The mass spectrometry data was analyzed using MultiQuant 3.0.3 software. The mass spectrum peaks detected in different samples were scored and corrected based on retention time and peak shape of the standard. The figure below shows the correction results of quantitative analysis of a substance randomly selected from different samples.

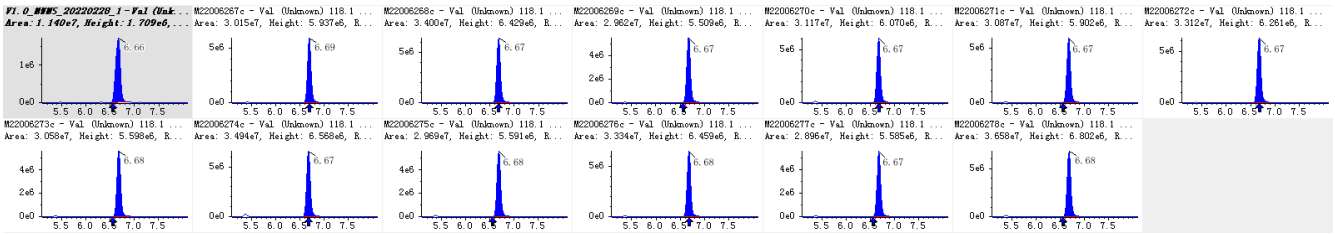

Fig 5: Scoring correction diagram for quantitative analysis of metabolites

Note: The figure shows the quantitative analysis integral correction results of randomly selected metabolites in different samples. The x-axis is the retention time (min) of metabolite detection, the y-axis is the ion flow intensity (CPS) of a certain metabolite ion detection, and the peak area represents the relative content of the substance in the sample.

Original file path: Final report/0.data/QC/*Integral_correction.png

3.2 Standard Solution Preparation

Standards were prepared at 10 ng/mL, 20 ng/mL, 50 ng/mL, 100 ng/mL, 200 ng/mL, 500 ng/mL, 1000 ng/mL, 2000 ng/mL, 5000 ng/mL, 10000 ng/mL, and 20000 ng/mL. Mass spectral peak intensity data were collected at each concentration to generate the calibration curve. The standard curves of each substance were plotted with the concentration ratio of external standard to internal standard as the horizontal coordinate and the peak area ratio of external standard to internal standard as the vertical coordinate. The equation of calibration curve are shown in the following table:

Table 5: Equation of calibration curve

Index	Class	RT	Equation
Met	Amino Acid metabolomics	N/A	y = 0.00198 x + 0.00264
Cys	Amino Acid metabolomics	N/A	y = 2.09400e-4 x - 0.01059
Trans-4-Hydroxy-L-Proline	Amino Acid metabolomics	N/A	y = 0.00357 x - 0.01188
Leu	Amino Acid metabolomics	N/A	y = 0.00394 x + 0.00204
Val	Amino Acid metabolomics	N/A	y = 0.00454 x + 0.01186
(5-L-Glutamyl)-L-Amino-Acid	Amino Acid metabolomics	N/A	y = 2.09145e-4 x - 0.00637
Beta-Alanine	Amino Acid metabolomics	N/A	y = 3.77048e-5 x + 0.00112
Asp	Amino Acid metabolomics	N/A	y = 3.80438e-4 x - 1.95362e-4
Ala	Amino Acid metabolomics	N/A	y = 5.57398e-4 x + 0.00332
Thr	Amino Acid metabolomics	N/A	y = 6.43030e-4 x + 0.00849

Final report/0.data/equation.xlsx

3.3 Quantification Results

Concentrations of each compound was obtained by substituting integrated peak area ration of all the detected samples into the equation of calibration curve.

Concentration of solid sample (ng/g) = c*V/1000/m

c: the concentration obtained by substituting the sample peak area ration into the equation of calibration curve (ng/mL);

V: the volume of extraction solution (μL);

m: the mass of the sample (g).

The metabolite ID, concentration and corresponding metabolite names of some metabolites detected in this experiment are shown in the following table:

Table 6: Statistical Table of metabolite quantity

Index	A2	A3
Anserine	53.417	48.2783
Phosphorylethanolamine	94.796	96.9451
Ethanolamine	20.7768	13.1558
(5-L-Glutamyl)-L-Amino-Acid	42.249	47.4256
N6-Acetyl-L-Lysine	70.0808	57.862
N-Propionylglycine	93.526	78.0816
N-Isovaleroylglycine	28.3659	32.1387
N-Glycyl-L-Leucine	106.566	118.223
N-Acetylneuraminic-Acid	93.8616	74.8264
N-Acetylaspartate	90.1164	70.0477

Original file path: Final report/0.data/*level.xlsx

3.4 Sample Quality Control Analysis

3.4.1 Total Ion Chromatogram Analysis

Using the mixed solution as the QC sample, one QC sample was inserted every 10 detection samples for analysis during the detection by the system. The stability of the device during the detection of the project can be assessed by analyzing the overlapped total ion flow chromatograms (TICs) obtained from the mass spectrometry detection and analysis of the same QC samples. The high stability of the testing device is a vital safeguard for the reproducibility and reliability of the data.

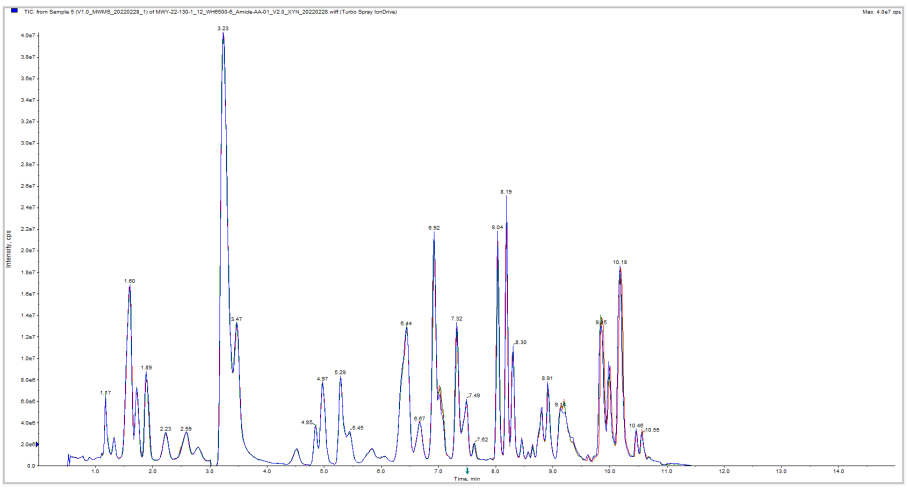

Fig 6: TIC overlap diagram detected by QC sample essence spectrum
Note: Superimposed spectrum from different QC samples. The results showed that the spectrum of total ion flow were highly consistent indicating that the signal stability was good when the same sample was detected at different times by mass spectrometry.

Original file path: Final report/0.data/picture/*QC_MS_tic_overlap*

3.4.2 QC Sample correlation assessment

Pearson correlation analysis was performed on the QC samples. The closer the | r | to 1, the higher the correlation between two samples. The correlation results can be seen in the figure below.

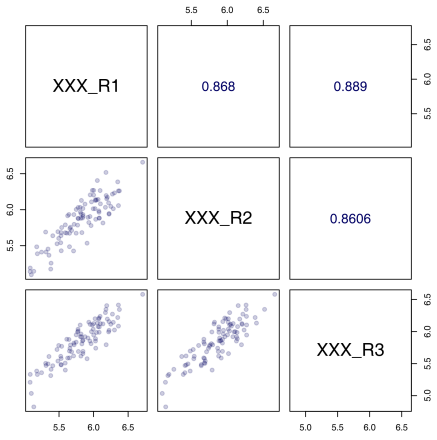

Fig 7: Correlation diagram between QC samples
Note: Diagonal squares represent QC samples name; Left diagonal box represent scatter diagram of QC samples . Both x-axis and y-axis represent metabolite content. Each dot in the diagram represents a metabolite. Right diagonal box represents correlation coefficients of QC samples .

Original file path/1.Data_Assess/pcc/*mix*

3.4.3 CV value distribution of all samples

The Coefficient of Variation (CV) value is the ratio between the standard deviation of the original data and the mean of the original data, which can reflect the degree of data dispersion. The Empirical Cumulative Distribution Function (ECDF) can be used to analyze the frequency of CV of substances that is smaller than the reference value. The higher the proportion of substances with low CV value in QC samples is, the more stable the experimental data is. The proportion of substances with CV value less than 0.3 in QC samples was higher than 80% , indicating that the experimental data were relatively stable. The proportion of substances with CV value less than 0.2 in QC samples was higher than 80%, indicating that the experimental data were very stable.

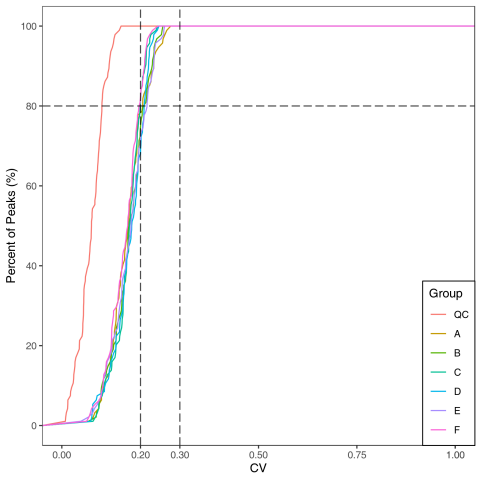

Fig 8: CV distribution of each group

Note: The X-axis represents the CV value, the Y-axis represents the proportion of metabolites with CV value less than a corresponding reference value. Different colors represent different sample groups. QC indicates quality control samples. The two dash lines on X-axis correspond to 0.2 and 0.3; the two dash lines on Y-axis correspond to 80% .

Original file path: Final report/1.Data_Assess/CV/*ECDF*

3.5 Sample quantification histogram

The results of sample content are grouped by statistics, and the statistical results are shown in the following table.

Table 7: Statistical results table

Index	Group	N	Mean
(5-L-Glutamyl)-L-Amino-Acid	F	6	76.559
(5-L-Glutamyl)-L-Amino-Acid	E	6	84.194
(5-L-Glutamyl)-L-Amino-Acid	D	6	12.768
(5-L-Glutamyl)-L-Amino-Acid	C	6	88.298
(5-L-Glutamyl)-L-Amino-Acid	B	6	52.568
(5-L-Glutamyl)-L-Amino-Acid	A	6	46.365
(S)-β-Aminoisobutyric-Acid	F	6	12.714
(S)-β-Aminoisobutyric-Acid	E	6	105.68
(S)-β-Aminoisobutyric-Acid	D	6	14.686
(S)-β-Aminoisobutyric-Acid	C	6	44.93

Original file path: Final report/1.Data_Assess/histogram/groups*.xlsx

The bar chart below shows the content difference of each substance in different groups.

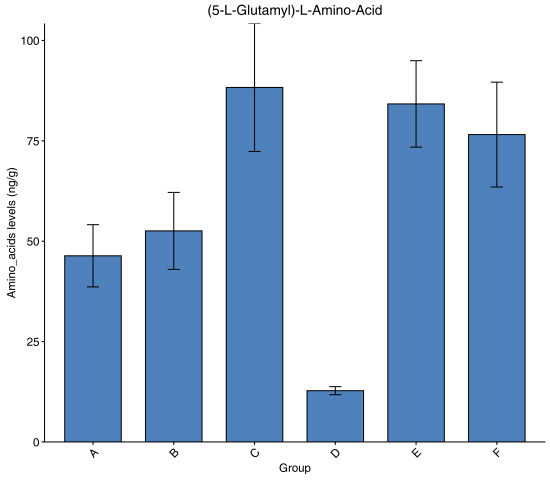

Fig 9: Sample content histogram

Note:The x-axis is the groups, the y-axis is the content, error bars are standard deviations.

Original file path: Final report/1.Data_Assess/histogram/histogram_compounds/*.png

3.6 Principal Component Analysis (PCA)

3.6.1 Principles of principal component analysis

Multivariate statistical analysis can simplify complex high-dimensional data while preserving the original information to the maximum extent by establishing a reliable mathematical model to summarize the characteristics of the metabolic spectrum. Among them, Principal Component Analysis (PCA) is an unsupervised pattern recognition method for statistical analysis of multidimensional data. Through orthogonal transformation, a group of variables that may be correlated are converted into a group of linear unrelated variables that are called principal components. This method is used to study how a few principal components may

reveal the internal structure of multiple variables, while keeping the original variable information (Eriksson et al., 2006). The first principal component (PC1) represents the most variable features in the multidimensional data matrix, PC2 represents the second most variable feature in the data, and so on. prcomp function of R software (www.r-project.org/) was used with parameter scale=True indicating unit variance Scaling (UV) for normalizing the data. See appendix for details of PCA calculation.

3.6.2 Principal component analysis of the sample population

Principal component analysis (PCA) was performed on all the samples (including QC samples) to examine the overall differences between each group and the variation between samples within a group. QC is the Quality control sample mentioned above. PCA plot for the first two principal components is as follows:

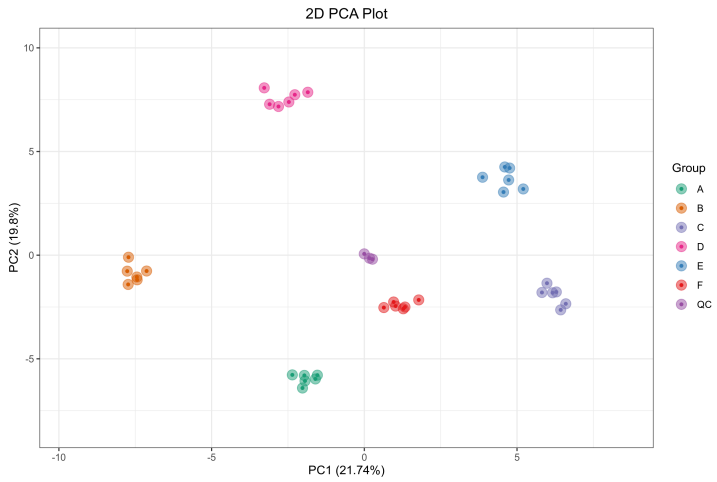

Fig 10: PCA score

diagram of quality spectrum data of each group of samples and quality control sample
Note: PC1 represents the first principal component and PC2 represents the second principal component. Percentage represents the interpretation rate of the principal component to the data set. Each dot in the figure represents a sample, and samples in the same group are indicated in the same color.

Original file path : Final report /1.Data_Assess/*all_pca*

3.6.3 Principal component univariate statistical process control

We plotted the sample control diagram based on principle component analysis results. Each point in the control chart represents a sample, and the X-axis is the injection order of the sample. Due to changes in the instrument, the points on the chart may fluctuate up and down. Generally, PC1 of the QC sample should be within 3 standard deviations (SD) from the normal range.

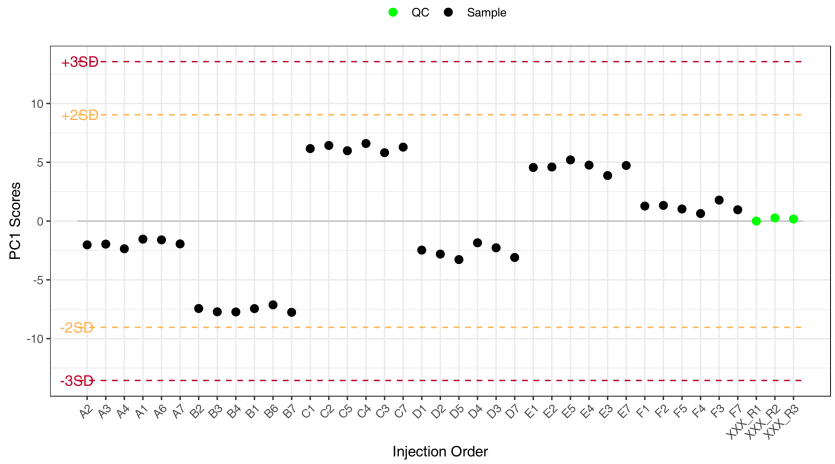

Fig 11: PC1 control diagram of population sample

Note: In the figure, the X-axis is the injection order of the sample, and the Y-axis reflects the PC1 value. The yellow and red lines define plus or minus 2 and 3 standard deviations respectively. The green dots represent QC samples and the black dots represent test samples.

Original file path: Final report/1.Data_Assess/pca/*PC1_QCC*

3.7 Hierarchical Cluster Analysis

3.7.1 Principles of cluster analysis

Hierarchical Cluster Analysis (HCA) is a type of multivariate statistical analysis method. The samples are classified according to their features such that highest homogeneity is achieved between sample from the same group and highest heterogeneity is achieved between samples from different groups. In this report, the compound quantification data was normalized (Unit Variance Scaling, UV Scaling) and heatmaps were drawn by R software Pheatmap package. Hierarchical Cluster Analysis (HCA) was used to cluster the samples.

Fig 12: Sample clustering diagram

Note: X-axis indicates the sample name and the Y-axis are the metabolites. Group indicates sample groups. Z-Score indicates the relative quantification of each metabolite with red representing higher content and green representing lower content. Cluster analysis was performed on both metabolites (vertical cluster tree) and samples (horizontal cluster tree). “all_heatmap_class”: Heat map based on metabolite classification; “all_heatmap_no_cluster”: Showing only heatmap.

Original file path: Final report /1.Data_Assess/*all_heatmap*

4 Analysis results

4.1 Principal component analysis of sample groups

4.1.1 Principal component analysis between sample groups

Principal component analysis was first performed on each pair of sample groups to examine the degree of variation between different groups and between samples within the group.

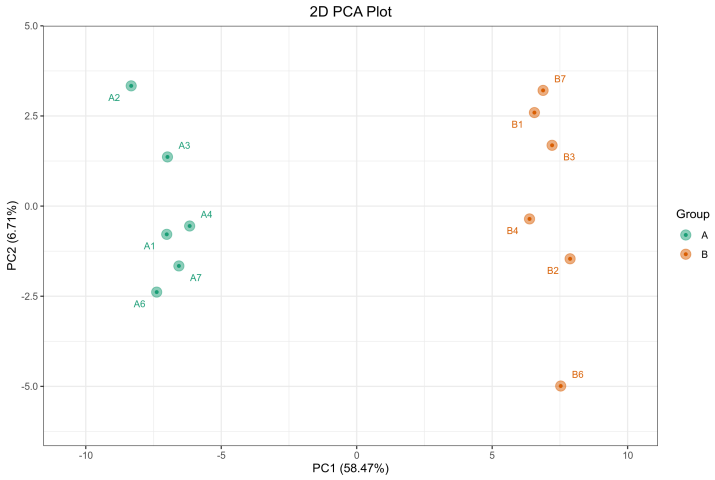

Fig 13: Principal component analysis of different groups

Note: Each group has a PCA plot, PC1 represents the first principal component, PC2 represents the second principal component, and the percentages on the axis represents the interpretation rate of the principal component to the data set. Each dot in the figure represents a sample, samples in the same Group are represented by the same color, and Group is a grouping.

The three-dimensional PCA result is shown in the figure below:

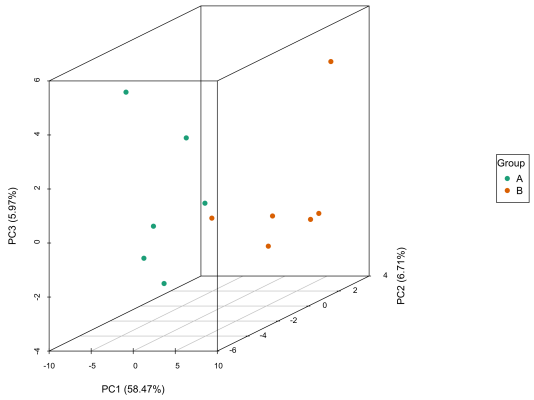

Fig 14: Three-dimensional PCA plot of different groups

Note: PC1 represents the first principal component, PC2 represents the second principal component, and PC3 represents the third principal component.

The explainable variation of the first five principal components is shown in the figure.

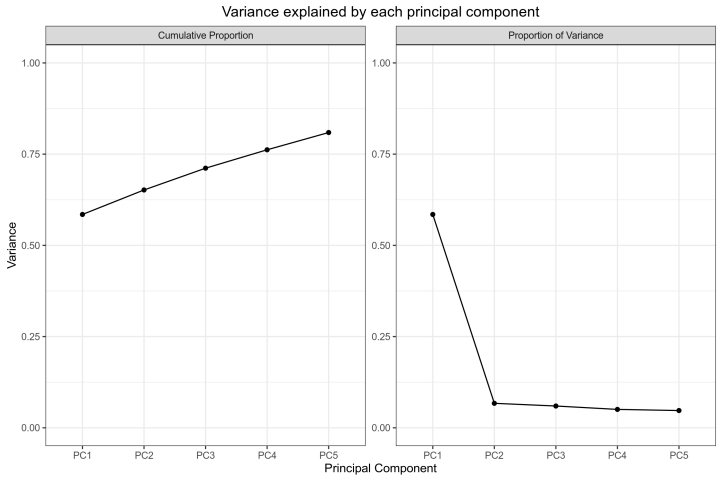

Fig 15: The explainable variation of the first five principal components
Note: The X-axis represents each principal component, the Y-axis represents the explainable variation, the left figure represents the cumulative explainable variation, and the right figure represents the explainable variation of each principal component

Principal component analysis of different groups:Original file path: Final report/2.Basic_analysis/Difference_analysi ID*_vs_group-ID*/pca/group-ID*_vs_group-ID*_pca.*;

Three-dimensional PCA plot of different groups:Original file path: Final report/2.Basic_analysis/Difference_analysis ID*_vs_group-ID*/pca/group-ID*_vs_group-ID*_pca3D.*

The explainable variation of the first five principal components:Original file path: Final report/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/pca/group-ID*_vs_group-ID*_pcaVar.*

4.2 Dynamic distribution of metabolite content differences

To show the overall compound abundance distribution in the samples, compounds were sorted and plotted based on fold-change values from small to large. The distribution of the ranked compounds is shown below with the top 10 up-regulated and top 10 down-regulated compound labelled.

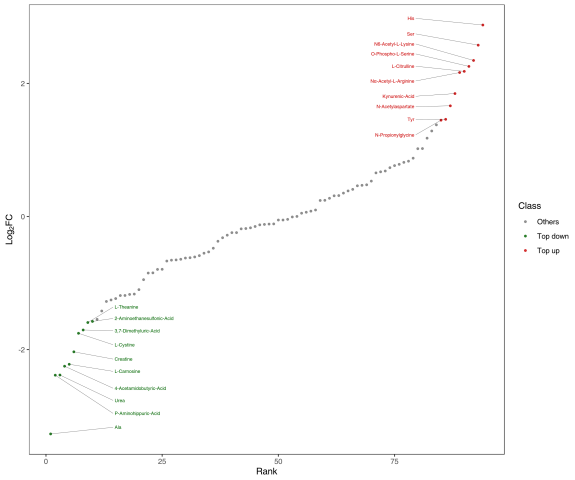

Fig 16: Dynamic distribution of metabolite content differences

Note: In the figure, the X-axis represents the rank number of metabolites based on FC value. The Y-axis represents the log_2FC value. Each point represents a metabolite. The green points represent the top 10 down-regulated metabolites and the red points represent the top 10 up regulated metabolites.

Final report/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/distribution/group-ID*_vs_group-ID*fc_distribution*

4.3 Differential metabolite screening

It is often necessary to combine univariate statistical analysis and multivariate statistical analysis for large high dimensional datasets such as metabolomics datasets to accurately identify differential metabolites. Univariate statistical analysis methods include parametric test and nonparametric test. Multivariate statistical analysis methods include principal component analysis and partial least square discriminant analysis. Based on the results of OPLS-DA (biological repetition ≥ 2), multivariate analysis of Variable Importance in Projection (VIP) from OPLS-DA modeling was used to preliminarily select differential metabolites from different samples. The fold-change and statistical significance (p-value) from univariate analysis can be used in conjunction to further identify differential metabolites. If biological replicates were < 3, differential metabolites are screened based on Fold Change value. If there were ≥ 3 biological replicates, VIP and P-values were used in combination to screen for differential metabolites. The detailed screening criteria is as follows:

For two sets of comparisons:

1.Metabolites with Fold Change ≥ 2 and Fold Change ≤ 0.5 were considered as significant and selected.

A partial result from the screening criteria is seen below:

Table 8: Screening results of differential metabolites

Index	Compounds	Type
Phosphorylethanolamine	Phosphorylethanolamine	up
N6-Acetyl-L-Lysine	N6-Acetyl-L-Lysine	up
N-Propionylglycine	N-Propionylglycine	up
N-Isovaleroylglycine	N-Isovaleroylglycine	down
N-Acetylaspartate	N-Acetylaspartate	up
Methionine-Sulfoxide	Methionine Sulfoxide	up
Lys	L-Lysine	down
Leu	L-Leucine	down
L-Theanine	L-Theanine	down
Nα-Acetyl-L-Arginine	Nα-Acetyl-L-Arginine	up

Final report/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/group-ID*_vs_group-ID*filter.xlsx.

4.3.1 Bar chart of differential metabolites

The following figure shows the result of top differentially expressed metabolites in each comparison with fold-change value shown as $\log_2$ values .

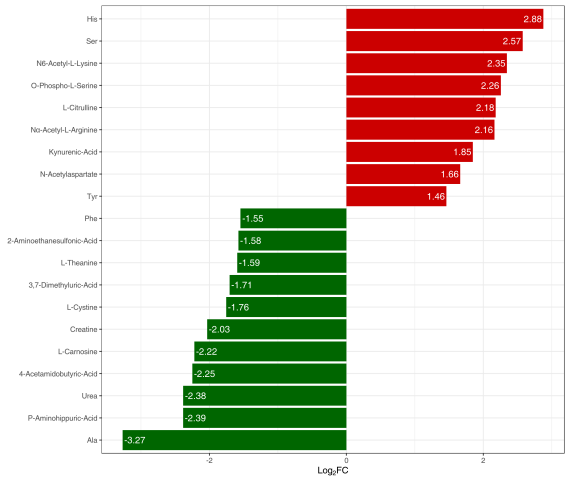

Fig 17: Bar chart of differential metabolites

Note: X-axis refers to log_2FC values of top differential metabolites,the Y-axis refers to metabolites. Red bars represent up-regulated differential metabolites and green bars represent down-regulated differential metabolites.

Final report/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/TopFcMetabolites/group-ID*_vs_group-ID*_TopFcMetabolites.*

4.3.2 Differential metabolite radar map

The top 10 differential metabolites based on Fold-change were selected and plotted on the radar plot.

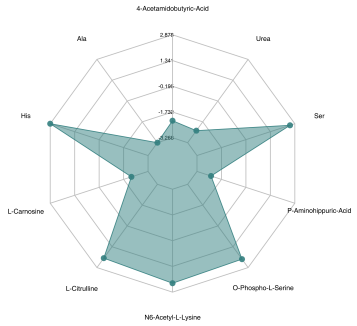

Fig 18: Differential metabolite radar map
Note: The grid lines correspond to the log_2FC. The green colored area is formed from the lines connecting the dots

Final report/2.Basic_Analysis/Difference_analysis/group-ID*_vs_group-ID*/radarchart/*radarchart**

4.3.3 Volcanic plot of differential metabolites

Volcano Plot is mainly used to show the relative differences and the statistical significance of compounds between two groups. We provided the volcano plot of differential compounds using different selection criteria for your consideration. The details of different selection criteria are described in the README document under the volcano plot directory. In addition, the attached results also provided an interactive web version of the volcano plot where you can examine the details of each compound.

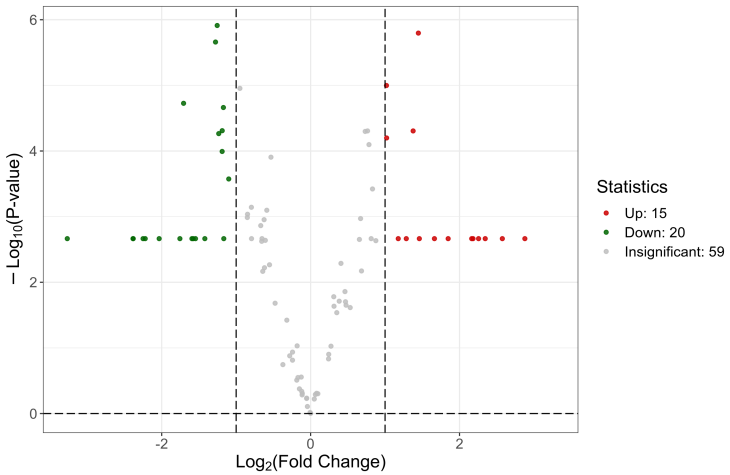

Fig 19: Volcanic plot of differential metabolites

Note: Under the dual screening conditions of FC + Pvalue/FDR, each point in the volcano map represents a metabolite, the horizontal coordinate represents the multiple change of the difference between the metabolites in different groups (log_2FoldChange), and the vertical coordinate represents the significance level of the difference (-log_10p-value). The greater the absolute value of abscissa, the greater the multiple difference of expression between the two samples. The larger the ordinate value is,is, the more significant the differential expression is. In the figure, the green dots represent the down-regulated metabolites,the red dots represent the up-regulated metabolites, and the gray dots represent the metabolites detected but not significantly different.

Final report/2.Basic_Analysis/Difference_analysis/group-ID*_vs_group-ID*/vol/*vol_*

4.3.4 Heatmap of differential metabolites

In order to observe the fold-change of differential compounds more intuitively, we normalized the abundances using unit variance scaling (UV scaling, see appendix for details of calculation formula) and plotted on a heatmap using pheatmap in R.

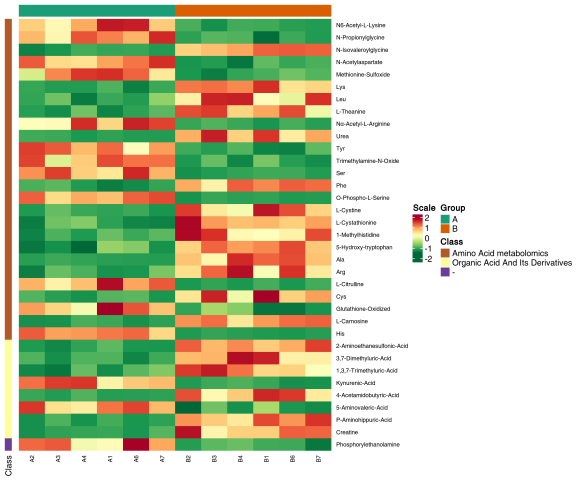

Fig 20: Heatmap of differential metabolites

Note: The X-axis shows the name of the samples and the Y-axis shows the differential metabolites. Different colors in the heatmap represent the values obtained after normalization and reflect the level of relative quantification. The darker the red, the higher the quantification. In contrast, the darker the green, the lower the quantification. The colored bar on top depicts sample groups. If hierarchical clustering is performed, the clustering tree will be shown on the left. If classification was performed on the metabolites, a colored bar will be shown on the left to depict Level 1 classifications.

Heatmap of differential metabolites:Final report/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/heatmap/group-ID*_vs_group-ID*_heatmap.*;

4.3.5 Z-value map of differential metabolites

Z-score plot is to normalize the differential metabolites in different samples by calculating the Z-value. The a-axis represents the z-value, the y-axis represents the differential metabolites, and the dots in different colors represent samples of different groups. The distribution of each differential metabolite among different groups can be seen intuitively. The formula is: $z = (x - \mu) / \sigma$; Where x is a specific score, μ is the mean, and σ is the standard deviation.

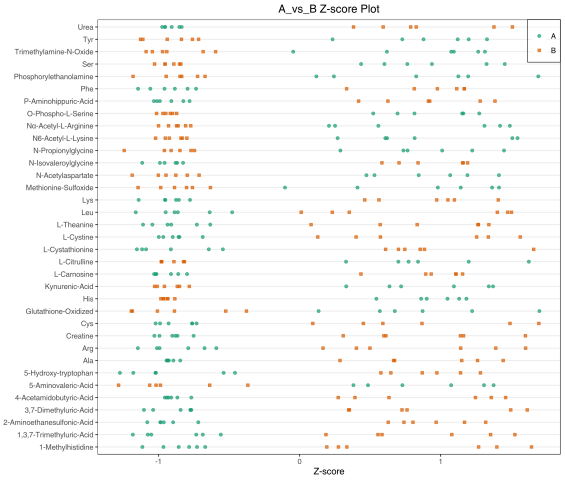

Fig 21: Z-value map of differential metabolites

Note: the X-axis is the value of substance content after normalized treatment, the Y-axis is the number of metabolites, and the points in different colors represent different groups of samples.

/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/zScore/group-ID*_vs_group-ID*_zScore.*.

4.3.6 Correlation analysis of differential metabolites

Compounds may act synergistically or in mutually exclusive relationships amongst each other. Correlation analysis can help measure the compound proximities of significantly different compounds. This analysis will help further understand the mutual regulatory relationship between compounds in the biological process. Pearson correlation was used to perform correlation analysis on the differential compounds identified based on the screening criteria described previously.

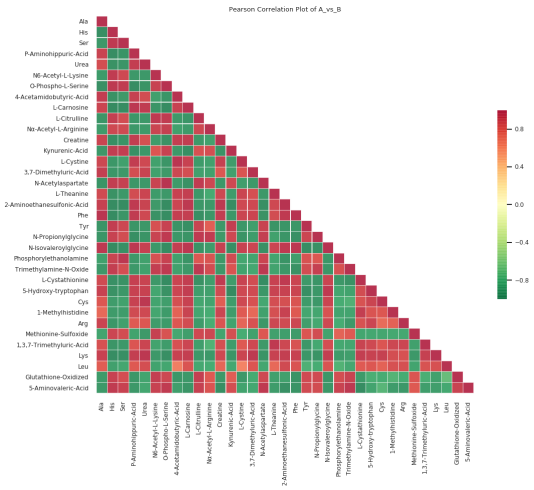

Fig 22: Heat map of correlation of different metabolites

Note: The ID of the metabolites are shown on both horizontal and vertical axis. The colors represent the Pearson correlation coefficient (r) with the scale seen on the right (The darker the red, the stronger the positive correlation; the darker the green the stronger the negative correlation). If there are more than 50 differential metabolites, the figure will only show the top 50 metabolites based on VIP values.

Differential metabolite correlation heat map: Final report/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/cpdCorr/group-ID*_vs_group-ID*_raw_cpdCorr_*.*;

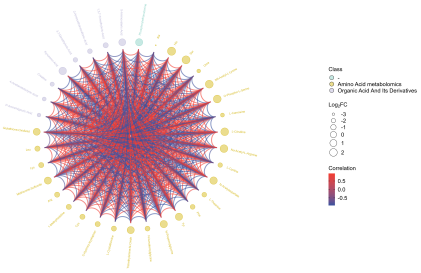

Fig 23: Chord diagram of differential metabolites

Note: The outermost layer shows the metabolite ID. The second layer shows log_2FC value,The larger the dot,the larger the log_2FC value; The color for the first and second layer represent Level 1 metabolite classification. The chords in the inner most layer reflect the Pearson correlation between the connected metabolites. Red chords represent positive correlation, and the blue chords represent negative correlation. Only metabolites with $|r| \geq 0.8$ and $p < 0.05$ are plotted.

Final report//2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/cpdCorr/group-ID*_vs_group-ID*_cpdCorrCir_*.*;

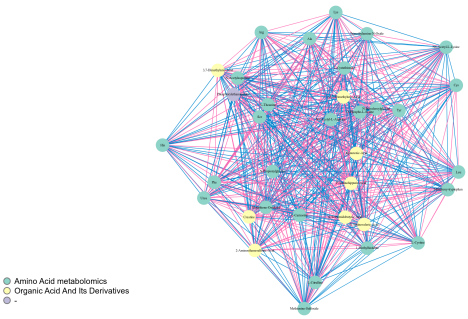

Fig 24: Correlation network diagram of differential metabolites

Note: The points in the figure represent the various differential metabolites, and the size of the points is related to the Degree of connection. The larger the point, the greater the Degree of connection, i.e. the more points (neighbors) connected to it. Red lines represent positive correlations and blue lines represent negative correlations. Line thickness represents the absolute value of Pearson correlation coefficient. The larger the |r|, the thicker the line. Only metabolites with $|r| \geq 0.8$ and $p < 0.05$ are plotted.

Final report/2.Basic_Analysis/Difference_analysis/group-ID*_vs_group-ID*/cpdCorr/*network*

4.3.7 Violin plot of differential metabolites

Violin plot is used to display data distribution and its probability density. The box in the middle represents the interquartile range, and the middle box represents the 95% confidence interval. The black horizontal line is the median, and the outer shape represents the distribution density of the data. The following figure shows the result of top 50 differentially compounds with the largest Log_2FC value.

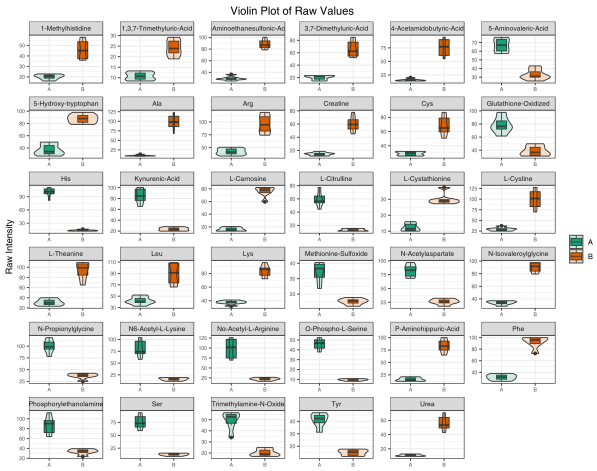

Fig 25: Violin plot of differential metabolites

Note: X-axis refers to sample,the Y-axis refers to content.

Final report/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/fullViolin/group-ID*_vs_group-ID*_fullViolin_Raw.*;

4.3.8 K-means analysis

K-means analysis is a method to examine the trend of relative quantification changes of a metabolite in different sample groups. K-means is performed based on the Z-score normalized relative quantification value.

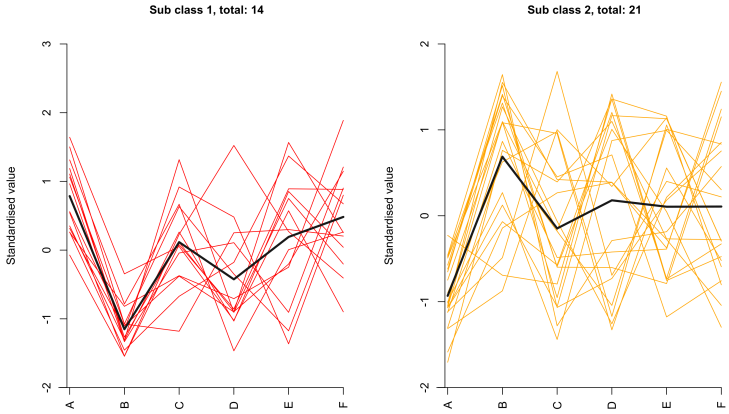

Fig 26: K-Means diagram of differential metabolites

Note: The X-axis represents the sample names and the Y-axis represents the normalized relative quantification. “Sub Class” represents a group of metabolites with the same trend and the number represent the number of metabolites in this cluster.

Figure of K-means clustering:Final report/2.Basic_analysis/kmeans/kmeans_cluster.*

4.3.9 Differential metabolite statistics

The number of different metabolites in each group is shown in the table below:

Table 9: Statistical table of differential metabolites

group name	All sig diff	down regulated	up regulated
A_vs_B	35	20	15

Statistical table of differential metabolites:Final report/2.Basic_analysis/Difference_analysis/sigMetabolitesCount.xl

4.4 Functional annotation and enrichment analysis of differential metabolites in KEGG database

KEGG (Kyoto Encyclopedia of Genes and Genomes) is a database that integrates compounds and genes into metabolic pathways. The KEGG database enabled researchers to study genes with their expression information and compounds with their abundances as a complete network.

4.4.1 Functional annotation of differential metabolites

Metabolites are annotated using the KEGG database, and only metabolic pathways containing differential metabolites are shown. Detailed results are found in the attached results. A portion of the results is shown below:

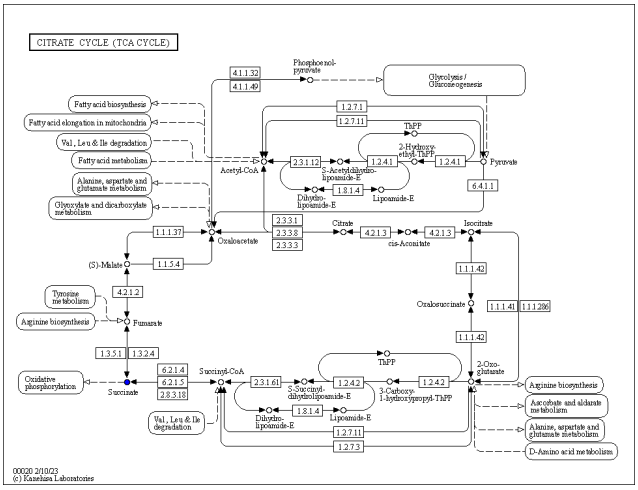

Fig 27: KEGG pathway of metabolites

Note: Red circles indicate that the metabolite content was significantly up-regulated in the experimental group; the blue circles indicate that the metabolite content was detected but did not change significantly; Green circles indicate that the metabolite content was significantly down-regulated in the experimental group. The orange circles indicate a mixture of both up-regulated and down-regulated metabolites.This allows searching for metabolites that may contribute to the phenotypic differences.

Final report/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/enrichment/Graph/ko*.

Statistical analysis of KEGG database annotation of screened metabolites with significant differences. Some of the results are as follows:

Table 10: KEGG annotations for differential metabolites

Index	Compounds	Type	cpd_ID
Phosphorylethanolamine	Phosphorylethanolamine	up	-
N6-Acetyl-L-Lysine	N6-Acetyl-L-Lysine	up	C02727
N-Propionylglycine	N-Propionylglycine	up	-
N-Isovaleroylglycine	N-Isovaleroylglycine	down	-
N-Acetylaspartate	N-Acetylaspartate	up	C01042
Methionine-Sulfoxide	Methionine Sulfoxide	up	-
Lys	L-Lysine	down	C00047
Leu	L-Leucine	down	C00123
L-Theanine	L-Theanine	down	-
Nα-Acetyl-L-Arginine	Nα-Acetyl-L-Arginine	up	-

Table 11: Enrichment Statistics of KEGG annotations for differential metabolites

ko_ID	Sig_compound	compound	Sig_compound_all	compound_all
ko00310	3	6	24	57
ko01100	24	54	24	57
ko00250	2	8	24	57
ko00300	1	2	24	57
ko00470	8	18	24	57
ko00780	2	2	24	57
ko00960	2	4	24	57
ko00970	10	21	24	57
ko01110	12	27	24	57
ko01210	4	11	24	57

Final report/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/enrichment/group-ID*_vs_group-ID*_filter_kegg.xlsx.

Final report/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/enrichment/group-ID*_vs_group-ID*_KEGG.xlsx.

4.4.2 KEGG classification of differential metabolites

The significant differential metabolites were classified based on pathway annotation . The results are as follows:

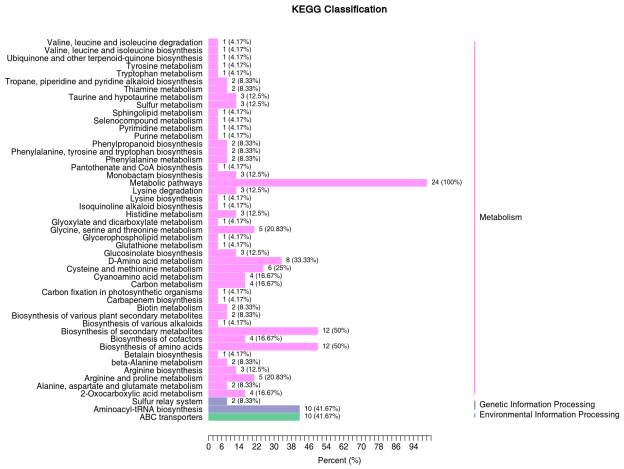

Fig 28: KEGG classification of differential metabolites
Note: the Y-axis shows the name of the KEGG pathway. The number of metabolites and the proportion of the total metabolites are shown next to the bar plot.

Final report/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/enrichment/group-ID*_vs_group-ID_KEGG_barplot.*.

4.4.3 Hierarchical Cluster Analysis of differential metabolites in KEGG signaling pathway

We clustered the metabolites in each pathway base on their relative quantification in order to examine the pattern of metabolite changes in different sample groups. Only pathways with at least 5 differential metabolites were analyzed.

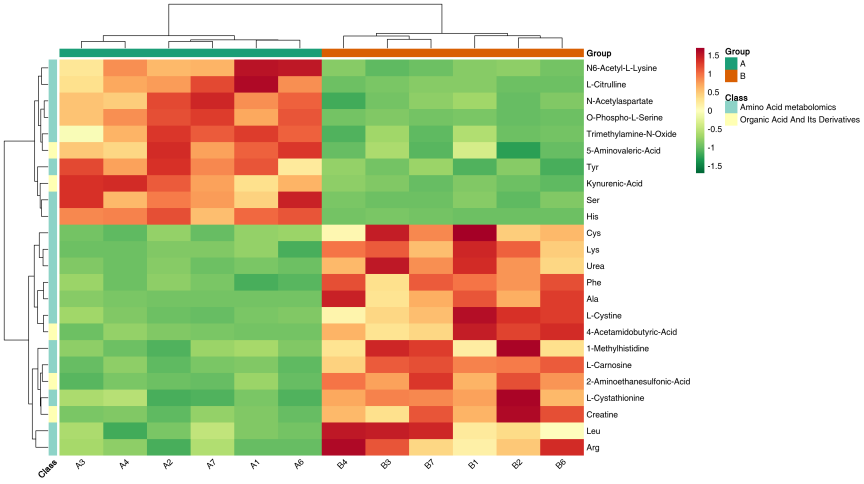

Fig 29: Clustering heat map of differential metabolites in KEGG pathway

Note: The X-axis shows the name of the samples and the Y-axis shows the differential metabolites. Different colors in the heatmap represent the values obtained after normalization and reflects the level of relative quantification. The darker the red, the higher the quantification. In contrast, the darker the green, the lower the quantification. The colored bar on top depicts sample groups. If hierarchical clustering is performed, the clustering tree will be shown on the left. If classification was performed on the metabolites, a colored bar will be shown on the left to depict classifications.

Final report/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/enrichment/group-ID*_vs_group-ID_KEGG_heatmap.*.

4.4.4 KEGG enrichment analysis of differential metabolites

KEGG pathway enrichment analysis was conducted based on the annotation results. We calculated the Rich Factor for each pathway, which is the ratio of the number of differenetial metabolites in the corresponding pathway to the total number of metabolites annotated in the same pathway. The greater the Rich Factor, the greater the degree of enrichment. P-value is the calculated using hypergeometric test as shown below:

$$P = 1 - \sum_{i=0}^{m-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

N represents the total number metabolites with KEGG annotation, n represents the number of differential metabolites in N, M represents the number of metabolites in a KEGG pathway in N, and m represents the number of differential metabolites in a KEGG pathway in M. The closer the p-value to 0, the more significant the enrichment. The size of the dots in the figure represents the number of significantly different metabolites enriched in the corresponding pathway. The results are shown below:

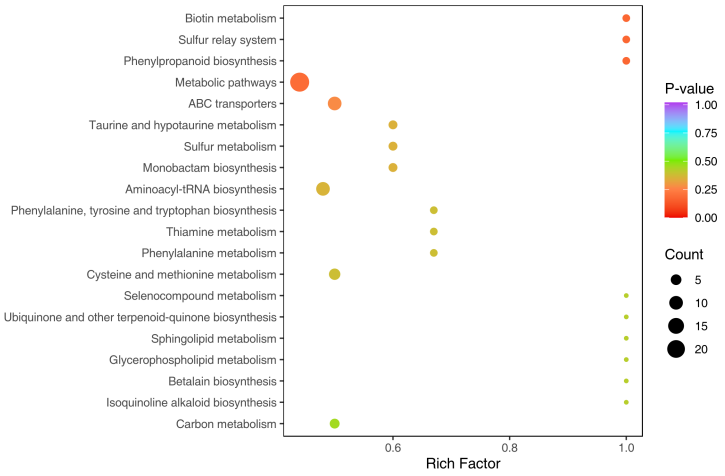

Fig 30: KEGG enrichment diagram of differential metabolites

Note: The X-axis represents the Rich Factor and the Y-axis represents the pathway. The color of points reflects the p-value. The darker the red, the more significant the enrichment. The size of the dot represents the number of enriched differential metabolites.

Final report/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/enrichment/group-ID*_vs_group-ID*_KEGG_Enrichment.*.

4.4.5 Overall changes in KEGG metabolic pathway

Differential Abundance Score (DA Score) is a score based on changes in metabolites in a pathway. DA Score can capture the overall changes of all Differential metabolites in a pathway with the following formulat:

DA score=(up regulated metabolites in a pathway-down regulated metabolites in a pathway)/(Total number of metabolites annotation in a pathway)

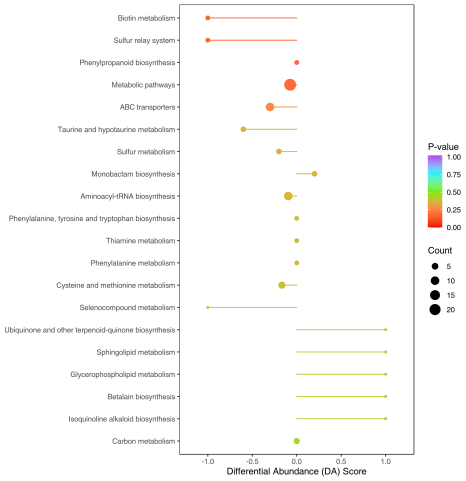

Fig 31: Difference abundance score

Note: The Y-axis represents the name of differential pathway, and the X-axis represents DA Score. DA Score reflects the overall change of all metabolites in the metabolic pathway. A Score of 1 indicates that the expression trend of all identified metabolites in this pathway is up-regulated, and -1 indicates that the expression trend of all identified metabolites in this pathway is down-regulated. The length of the line represent the absolute value of DA-score while the size of the dot at the end of the line represent the number of differential metabolites. A dot on the left of the line represent the pathway is up-regulated; a dot on the right of the line represents the pathway is down-regulated. The color of the line and dot represent the p-value. The darker the red, the smaller the p-value and the darker the purple, the larger the p-value.

Final report/2.Basic_Analysis/Difference_analysis/group-ID*_vs_group-ID*/enrichment/*DA_score*

4.5 ROC curve analysis of differential metabolites

The Receiver Operating Characteristic Curve (ROC curve) is a quantitative tool employed when distinguishing between two conditions or natural states becomes challenging, demanding precision from testers, professional diagnosticians, predictive analysts, or decision-makers. Widely applied in medicine for clinical diagnosis and population screening studies, conducting ROC curve analysis gains significance when dealing with over 30 samples.

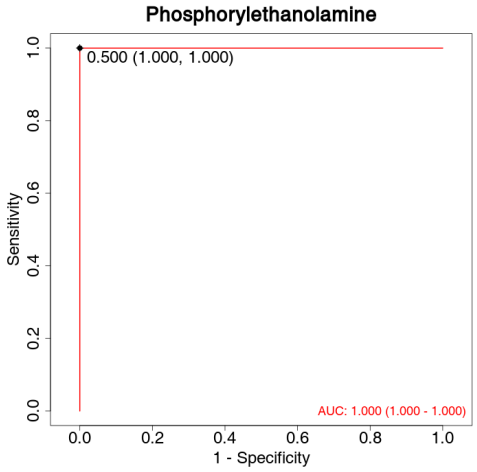

Fig 32: ROC curves for differential metabolites

Note:The horizontal axis signifies 1-specificity, denoting the false positive rate derived from the formula: false positives/(false positives + true negatives). On the other hand, the vertical axis represents sensitivity, indicating the true positive rate computed as true positives/(true positives + false negatives). The region enclosed by the ROC curve and the horizontal axis is termed the Area Under the Curve (AUC), serving as a quantitative assessment metric for the ROC curve. The AUC value ranges between (0.5, 1], with a proximity to 1 indicating superior predictive performance of the model. Within the graph, red text indicates the AUC value and its corresponding 95% confidence interval, while the black text showcases the optimal threshold along with specificity and sensitivity values in parentheses.

Final report/2.Basic_Analysis/Difference_analysis/NC_vs_BT/ROC/*ROC*

5 References

1. Guo S , Duan J A , Qian D , et al. Rapid Determination of Amino Acids in Fruits of Ziziphus jujubaby Hydrophilic Interaction Ultra-High-Performance Liquid Chromatography Coupled with Triple-Quadrupole Mass Spectrometry[J]. Journal of Agricultural & Food Chemistry, 2013, 61(11):2709-2719.
2. Hiraoka N , Toue S , Okamoto C , et al. Tissue amino acid profiles are characteristic of tumor type, malignant phenotype, and tumor progression in pancreatic tumors[J]. Scientific Reports, 2019, 9(1).
3. An Z , Hu T , Lv Y , et al. Targeted amino acid and related amines analysis based on iTRAQ-LC-MS/MS for discovering potential hepatotoxicity biomarkers[J]. Journal of Pharmaceutical and Biomedical Analysis, 2019, 178:112812.
4. Li X , Wong C C , Tang Z , et al. Determination of amino acids in colon cancer cells by using UHPLC-MS/MS and [U-13C5]-glutamine as the isotope tracer[J]. Talanta, 2016, 162:285-292.

5. Tsochatzis E D , Begou O , Gika H G , et al. A hydrophilic interaction chromatography-tandem mass spectrometry method for amino acid profiling in mussels[J]. J Chromatogr B Analyt Technol Biomed Life, 2016, 1047:197.
6. B Z L A , B M J T , B C Z , et al. A reliable LC-MS/MS method for the quantification of natural amino acids in mouse plasma: Method validation and application to a study on amino acid dynamics during hepatocellular carcinoma progression[J]. Journal of Chromatography B, 2019, 1124:72-81.
7. Le A , Ng A , Kwan T , et al. A rapid, sensitive method for quantitative analysis of underivatized amino acids by liquid chromatography-tandem mass spectrometry (LC-MS/MS).[J]. J Chromatogr B Analyt Technol Biomed Life, 2014, 944:166-174.
8. Gray N , Plumb R S , Wilson I D , et al. A validated UPLC-MS/MS assay for the quantification of amino acids and biogenic amines in rat urine[J]. Journal of Chromatography B, 2019, 1106-1107:50-57.

6 Appendix

6.1 Analytical methods

1.PCA

Unsupervised PCA (principal component analysis) was performed by statistics function prcomp within R (www.r-project.org). The data was unit variance scaled before unsupervised PCA.

2.Hierarchical Cluster Analysis and Pearson Correlation Coefficients

The HCA (hierarchical cluster analysis) results of samples and metabolites were presented as heatmaps with dendrograms, while pearson correlation coefficients (PCC) between samples were caculated by the cor function in R and presented as only heatmaps. Both HCA and PCC were carried out by R package pheatmap. For HCA, normalized signal intensities of metabolites (unit variance scaling) are visualized as a color spectrum.

3.Differential metabolites selected

Significantly regulated metabolites between groups were determined by absolute Log_2FC (fold change).

4.KEGG annotation and enrichment analysis

Identified metabolites were annotated using KEGG compound database (http://www.kegg.jp/kegg/compound/), annotated metabolites were then mapped to KEGG Pathway database (http://www.kegg.jp/kegg/pathway.html). Pathways with significantly regulated metabolites mapped to were then fed into MSEA (metabolite sets enrichment analysis), their significance was determined by hypergeometric test's P-Values.

6.2 List of software and versions

Table 12: Software used

Analysis	Software	Version
PCA	R (base package)	3.5.1
Pearson Correlation	R (base package; Hmisc)	3.5.1; 4.4.0
Correlation plot	R (corrplot)	0.84
Heatmap	R (heatmaply; ComplexHeatmap)	1.2.1; 2.7.1.1009
OPLS-DA	R (MetaboAnalystR)	1.0.1
Radar plot	R (fmsb)	0.7.0
Chord diagram	R (igraph; ggraph)	1.2.4.2; 2.0.2
Network diagram	R (igraph)	1.2.4.2
Regulatory network diagram	R (FELLA)	1.10.0

Data processing methods were mainly adopted in the analysis process in two ways:

(1) unit variance scaling (UV)

Unit variance Scaling (UV) is also called Z-Score standardization, i.e., auto scaling. This method standardizes data according to mean and standard deviation of original data. The processed data conform to the standard normal distribution, that is, the mean value is 0 and the standard deviation is 1.

Calculation method: Divide the original data center by standard deviation.

The formula is as follows:

$$x' = \frac{x - \mu}{\sigma}$$

Where μ is the mean value and σ is the standard deviation.

(2) Centralization/zero-mean-centered (Ctr)

Calculation method: subtract the mean of the variables from the original data.

The formula is as follows:

$$x' = x - \mu$$