Innovative Metabolomics Insights for Better Health

BileAcid Targeted Metabolomics Assay Final Report

Metware Biotechnology Inc.

www.metwarebio.com

Contents

MWY-23-XXX Bile Acid Targeted Metabolomics Assay Final Report

1 Abstract

Changes in physiological activity can change the metabolite profile of an organism. Targeted quantitative detection technology allows sensitive qualitative annotation and highly accurate quantitative analysis of a set of metabolites. MetwareBio has established a LC-MS/MS based analytical method that can quantify 65 bile acid related metabolites.

For this project,24 samples were divided into 4 groups. A total of 44 metabolites were detected based on UPLC-MS/MS system.

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:

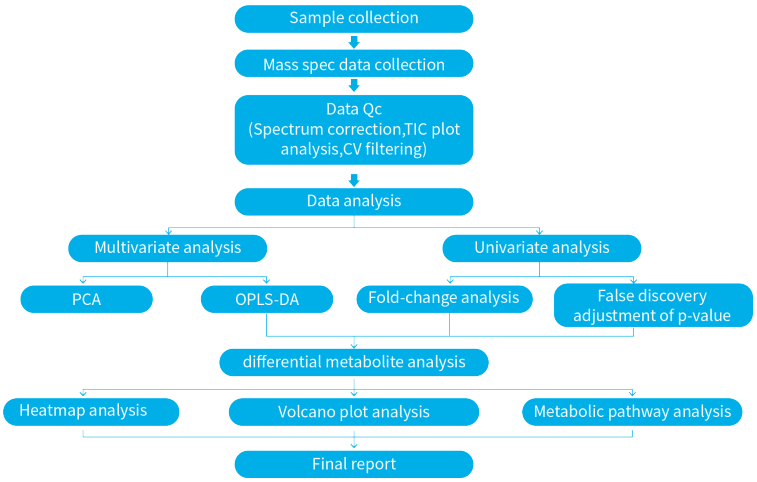

Fig 1: Flow chart of metabolomics analysis

Compounds to be detected:

Table 1: List of compounds in the panel

Number	Compounds	Index
1	taurolithocholic acid-3-sulfate	TLCA-3S
2	Dehydrolithocholic acid	DLCA
3	Isoallolithocholic acid	IALCA
4	isolithocholic acid	ILCA
5	Lithocholic acid	LCA
6	5α-CHOLANIC ACID-3α-OL	alloLCA

Table 1: List of compounds in the panel

Number	Compounds	Index
7	Nor-Deoxycholic Acid	23-DCA
8	5-β-Cholanic Acid-3α-ol-6-one	6-ketoLCA
9	7-ketolithocholic acid	7-KLCA
10	12-ketolithocholic acid	12-KLCA
11	3-oxodeoxycholic acid	3-oxo-DCA
12	murideoxycholic acid	MDCA
13	3β-Ursodeoxycholic Acid	3β-UDCA
14	β-Hyodeoxycholic Acid	3β-HDCA
15	Ursodeoxycholic acid	UDCA
16	Hyodeoxycholic acid	HDCA
17	Isochenodeoxycholic Acid	isoCDCA
18	3β-deoxycholic acid	3β-DCA
19	Chenodeoxycholic acid	CDCA
20	Deoxycholic acid	DCA
21	Isodeoxycholic acid	IDCA
22	norcholic acid	NCA
23	Dehydrocholic acid	DHCA
24	7,12-diketolithocholic acid	7,12-DKLCA
25	6,7-diketolithocholic acid	6,7-DKLCA
26	7-Ketodeoxycholic acid	7-KDCA
27	12-Oxochenodeoxycholic acid	12-oxo-CDCA
28	3-Oxocholic acid	3-oxo-CA
29	Ursocholic acid	UCA
30	ω-muricholic acid	ω-MCA
31	3β-Cholic Acid	3β-CA
32	α-muricholic acid	α-MCA
33	β-muricholic acid	β-MCA
34	hyocholic acid	HCA
35	Cholic acid	CA
36	Glycolithocholic acid	GLCA
37	Glycoursodeoxycholic acid	GUDCA
38	Glycochenodeoxycholic acid	GCDCA
39	Glycodeoxycholic acid	GDCA
40	lithocholic acid-3-sulfate	LCA-3S
41	Glycodehydrocholic acid	GDHCA
42	3β-Glycocholic Acid	βGCA
43	Glycohyocholic acid	GHCA
44	Glycocholic acid	GCA
45	taurolithocholic acid	TLCA
46	Tauroursodeoxycholic acid	TUDCA
47	Taurochenodeoxycholic acid	TCDCA
48	Taurodeoxycholic acid	TDCA
49	Taurodehydrocholic acid	TDHCA
50	glycolithocholic acid-3-sulfate	GLCA-3S
51	Tauro-β-muricholic acid	Tβ-MCA
52	Tauro-ω-muricholic Acid sodium salt	Tω-MCA
53	Tauro-α-muricholicAcid sodium salt	Tα-MCA

Table 1: List of compounds in the panel

Number	Compounds	Index
54	Taurohyocholic acid	THCA
55	Taurocholic acid	TCA
56	Chenodeoxycholic acid-3-β-D-glucuronide	CDCA-3Gln
57	Glycohyodeoxycholic Acid	GHDCA
58	Taurohyodeoxycholic Acid (sodium salt)	THDCA
59	cholic acid 7 sulfate	CA-7S
60	Cholic Acid 3 Sulfate Sodium Salt	CA-3S
61	chenodeoxycholic acid3-sulfate disodium salt	CDCA-3S
62	Deoxycholic Acid 3-O-Sulfate Disodium Salt	DCA-3-O-S
63	Glycoursodeoxycholic Acid 3 Sulfate Sodium	GUDCA-3S
64	Glycochenodeoxycholic Acid 3 Sulfate Disodium Salt	GCDCA-3S
65	Taurocholic Acid 3 sulfate sodium salt	TCA-3S

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

2.1 Sample information

This project has 24 samples divided into 4 groups. Sample information is shown in the following table:

Table 2: Sample information table

Species	**Tissues**	**MW_ID**	**Sample_ID**
_	_	A1	A1
_	_	A2	A2
_	_	A3	A3
_	_	A4	A4
_	_	A5	A5
_	_	A6	A6
_	_	B1	B1
_	_	B2	B2
_	_	B3	B3
_	_	B4	B4
_	_	B5	B5
_	_	B6	B6
_	_	C1	C1
_	_	C2	C2
_	_	C3	C3
_	_	C4	C4
_	_	C5	C5
_	_	C6	C6
_	_	D1	D1
_	_	D2	D2
_	_	D3	D3
_	_	D4	D4
_	_	D5	D5
_	_	D6	D6

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
Ball mill instrument	MM400	Retsch
Centrifugal concentrator	CentriVap	LABCONCO
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
Acetic acid	HPLC	Thermo fisher
Ammonium acetate	LC-MS	Sigma-Aldrich
Chemical standard	99%	Sigma-Aldrich/Zhenzhun.etc

2.3 Sample extraction process

1) Homogenize the sample by adding one steel bead to 20 mg of solid sample, 10 µL of 1 µg/mL internal standard working solution, and 200 µL of 20% methanol in acetonitrile.
2) Shake the homogenized sample at 2500 rpm for 10 min, and then place it in a -20°C for 10 min.
3) Centrifuge at 4°C, 12,000 rpm for 10 min. Then collect the supernatant and concentrate it in a concentrator.
4) After the concentration is completed, reconstitute the sample with 100 µL of 50% methanol-water solution, and set it aside for subsequent LC-MS/MS.

2.4 Chromatography-mass spectrometry acquisition conditions

Data acquisition was performed on Ultra Performance Liquid Chromatography (UPLC) (ExionLC™ AD, https://sciex.com/) and Tandem Mass Spectrometry (MS/MS) (QTRAP® 6500+, https://sciex.com/).The primary liquid phase conditions consist the following:

1) Chromatography column: Waters ACQUITY UPLC HSS T3 C18 column (1.8 µm, 100 mm × 2.1 mm i.d.)
2) Mobile phase: ultra-pure water (containing 0.01% acetic acid and 5 mmol/L ammonium acetate) for phase A; acetonitrile (containing 0.01% acetic acid) for phase B.

3) Flow rate: 0.35 mL/min; column temperature: 40°C; injection volume: 3 μL.
4) Gradient elution program: A/B 95:5 (V/V) at 0 min, A/B 60:40 (V/V) at 0.5 min, 50:50 (V/V) at 4.5 min, 25:75 (V/V) at 7.5 min, 5:95 (V/V) at 10 min, 95:5 (V/V) at 12.0 min.

The primary mass spectrometry conditions consists the following(ESI-MS/MS Conditions):

Electrospray Ionization (ESI) temperature: 550 °C; mass spectrometry voltage: -4500 V; curtain gas (CUR): 35 psi. In triple quadrupole mass spectrometry, ion pairs were scanned and detected based on optimized declustering voltage (DP) and collision energy (CE).

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.

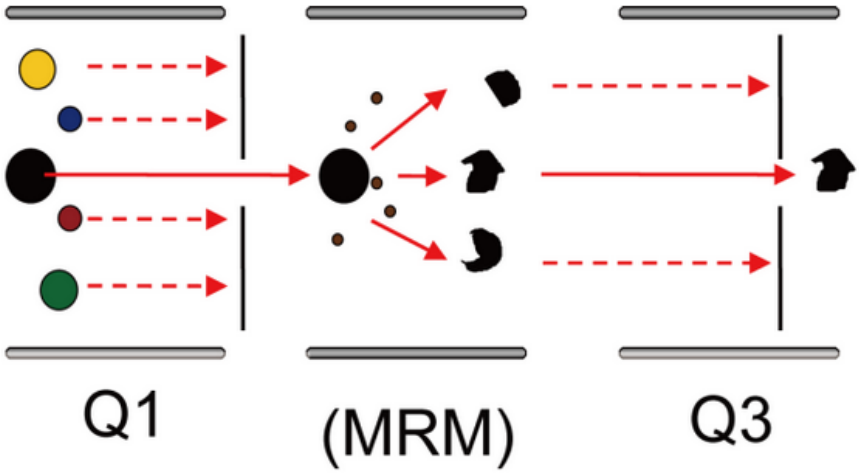

Fig 2:
Schematic diagram of multiple reaction monitoring mode by mass spectrometry

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).

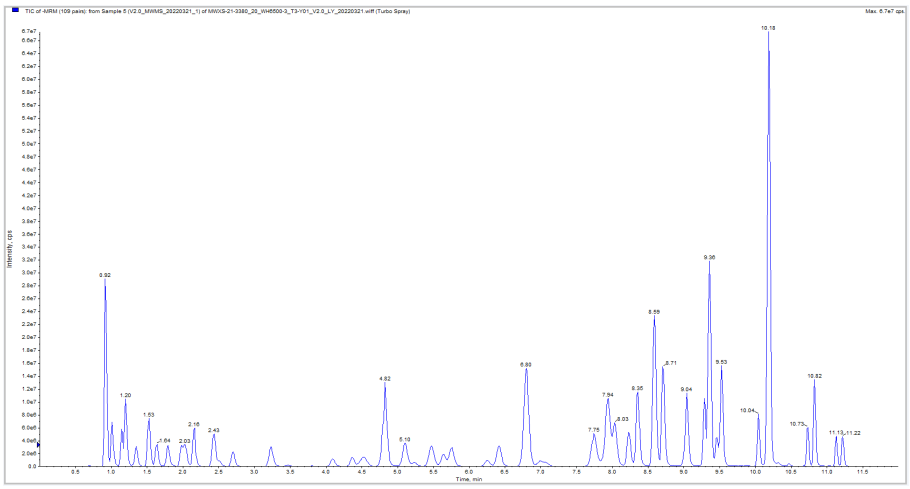

Fig 3: Total ion current diagram of mixed phase mass spectrum analysis

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

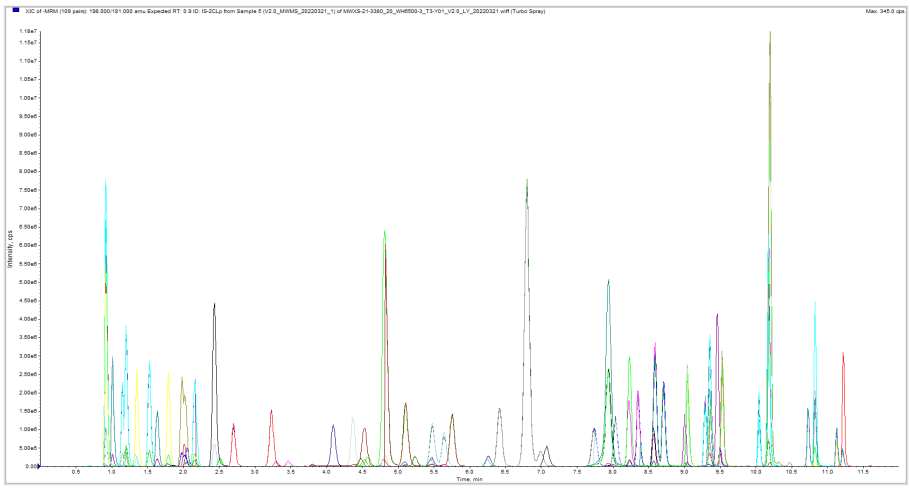

Fig 4: Extraction ion flow chromatogram

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.

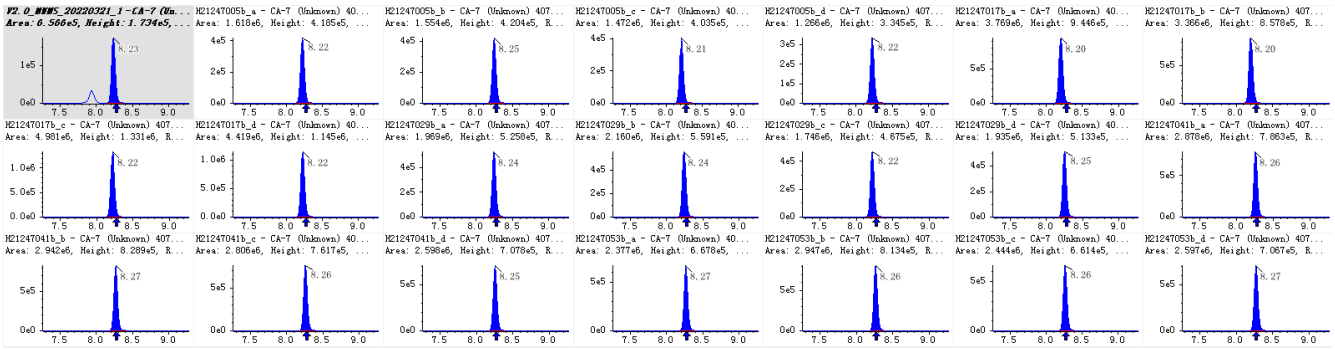

Fig 5: Scoring correction diagram for quantitative analysis of metabolites

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 0.1 ng/mL, 0.2 ng/mL, 0.4 ng/ mL, 1 ng/ mL, 2 ng/ mL, 4 ng/ mL, 10 ng/ mL, 20 ng/ mL, 40 ng/ mL, 100 ng/ mL, 200 ng/ mL, 400 ng/ mL, and 1000 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
CDCA	BAs	10.04	y = 0.00743 x + 0.00310
GUDCA-3S	BAs	1.01	y = 0.02520 x - 7.97879e-4
DCA	BAs	10.17	y = 0.00144 x + 4.10438e-4
3-oxo-DCA	BAs	10.19	y = 0.04094 x + 0.00176
IALCA	BAs	10.73	y = 0.01422 x + 0.00361
IDCA	BAs	10.82	y = 0.00154 x - 3.10450e-5
ILCA	BAs	10.82	y = 0.01764 x + 0.00792
LCA	BAs	11.13	y = 0.01551 x + 0.04294
alloLCA	BAs	11.21	y = 0.00996 x + 0.00518
DLCA	BAs	11.22	y = 0.04407 x + 0.00103

Final report/0.data/equation.xlsx

3.3 Quantification Results

Concentration of each compound was obtained by substituting integrated peak area ratio 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 ratio 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	A1	A2
12-KLCA	0.08169	0.08658
GCDCA-3S	0.04509	0.04554
GHDCA	0.01226	0.01347
βGCA	0.02943	0.03072
GUDCA-3S	0.07169	0.06080
HCA	0.02331	0.02742
HDCA	0.03462	0.03406
IALCA	0.07184	0.06126
IDCA	0.09259	0.10036
ILCA	0.00515	0.00518

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.

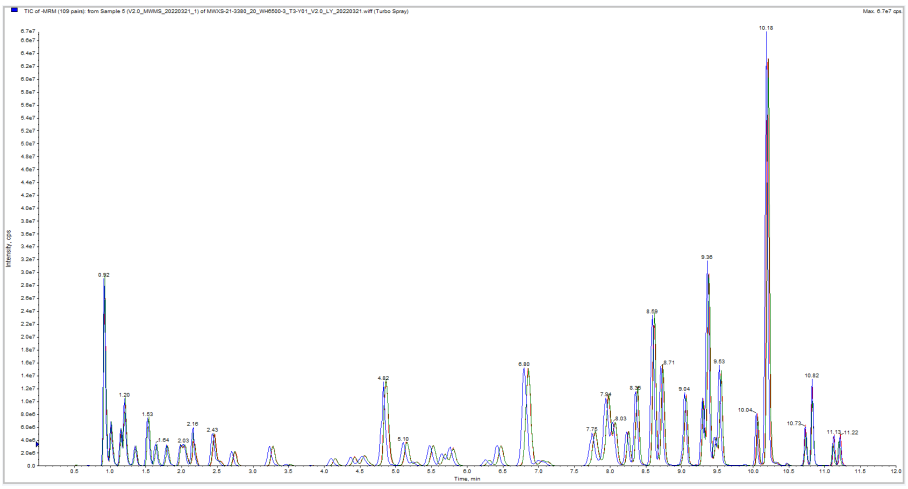

Fig 6: TIC overlap diagram detected by QC sample essence spectrum
Note: Superimposed spectrum from different QC samples. The results showed that the spectrum of total ion flow were highly consistent indicating that the signal stability was good when the same sample was detected at different times by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

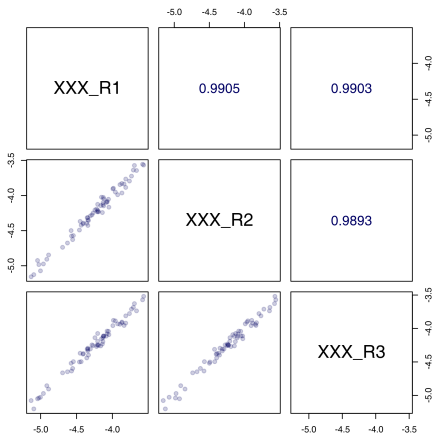

Fig 7: Correlation diagram between QC samples
Note: Diagonal squares represent QC samples name; Left diagonal box represent scatter diagram of QC samples . Both x-axis and y-axis represent metabolite content. Each dot in the diagram represents a metabolite. Right diagonal box represents correlation coefficients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

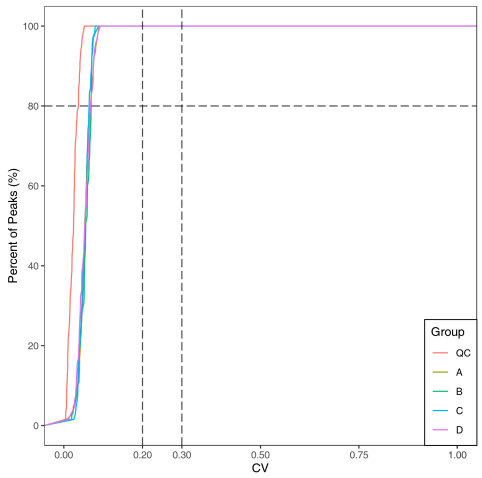

Fig 8: CV distribution of each group

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 Principal Component Analysis (PCA)

3.5.1 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.

3.5.2 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:

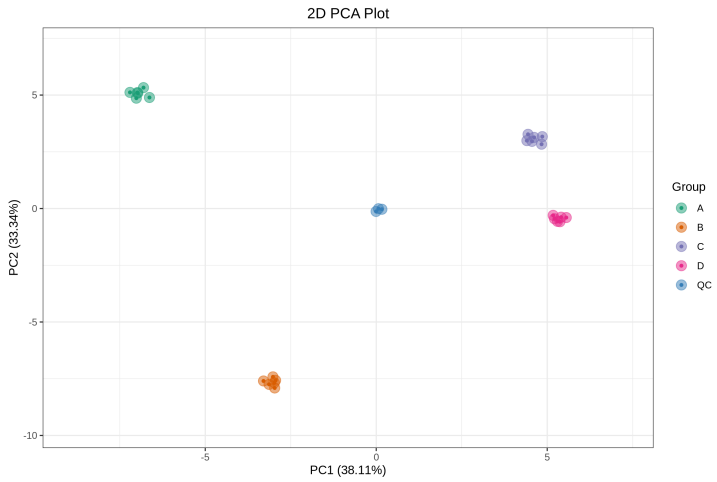

Fig 9: PCA score

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*

3.5.3 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.

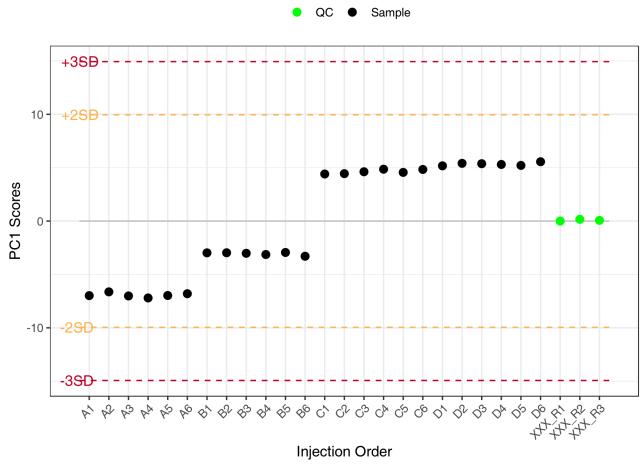

Fig 10: PC1 control diagram of population sample

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*

3.6 Hierarchical Cluster Analysis

3.6.1 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.

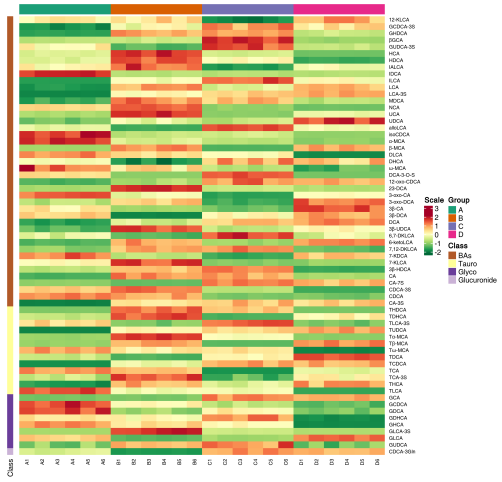

Fig 11: Sample clustering diagram

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.

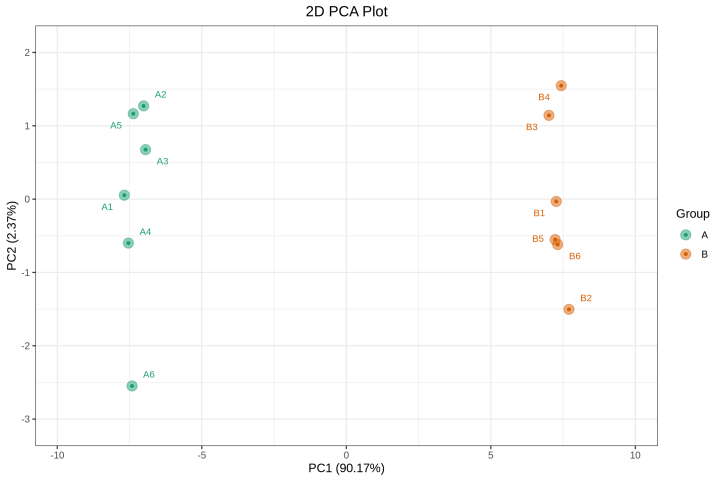

Fig 12: Principal component analysis of different groups

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:

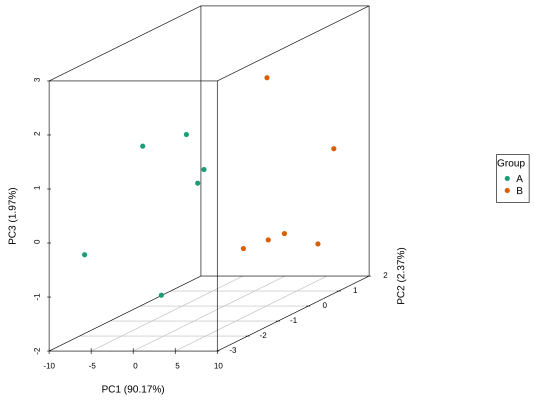

Fig 13: Three-dimensional PCA plot of different groups

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.

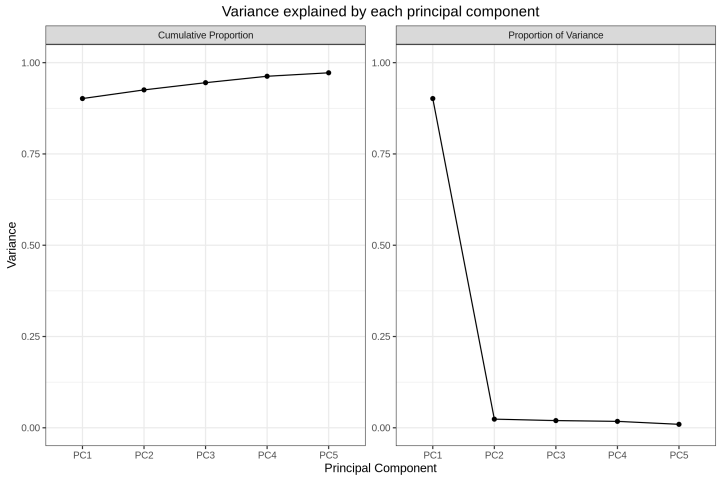

Fig 14: The explainable variation of the first five principal components
Note: The X-axis represents each principal component, the Y-axis represents the explainable variation, the left figure represents the cumulative explainable variation, and the right figure represents the explainable variation of 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 Discriminant Analysis by Orthogonal Partial Least Squares (OPLS-DA)

PCA analysis is often insensitive to variables with small correlation. In contrast, partial least squares-discriminant analysis (PLS-DA) is a multivariate statistical analysis method with supervised pattern recognition, in which components in independent variable X and dependent variable Y are extracted to calculate the correlation between components. Compared with PCA, PLS-DA can maximize the difference between groups and facilitate the search for differential compounds. Orthogonal partial least squares discriminant analysis (OPLS-DA) combines orthogonal signal correction (OSC) and PLS-DA method, which can decompose the x-matrix information into two types (1. information related to Y and 2. irrelevant information) and filter the differential variables by removing the irrelevant differences.

The OPLSR.Anal function in the R package MetaboAnalystR was used for this analysis. The following table shows a partial result from the OPLS-DA model:

Table 7: Partial results of OPLS-DA

Index	Compounds	Type
12-KLCA	12-ketolithocholic acid	insig
GCDCA-3S	Glycochenodeoxycholic Acid 3 Sulfate Disodium Salt	insig
GHDCA	Glycohyodeoxycholic Acid	down
βGCA	3β-Glycocholic Acid	insig
GUDCA-3S	Glycoursodeoxycholic Acid 3 Sulfate Sodium	insig
HCA	hyocholic acid	down
HDCA	Hyodeoxycholic acid	insig
IALCA	Isoallolithocholic acid	insig
IDCA	Isodeoxycholic acid	up
ILCA	isolithocholic acid	down

Original file path: The calculation results of all metabolites of OPLS-DA were compared in groups: /2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/group-ID*_vs_group-ID*_info.xlsx.

4.2.1 Principles of OPLS-DA model

During OPLS-DA modeling, the X matrix information is decomposed into information related to Y and information unrelated to Y. Among them, the variable information related to Y is the predicted principal component, and the information unrelated to Y is the orthogonal principal component (Thevenot et al., 2015).

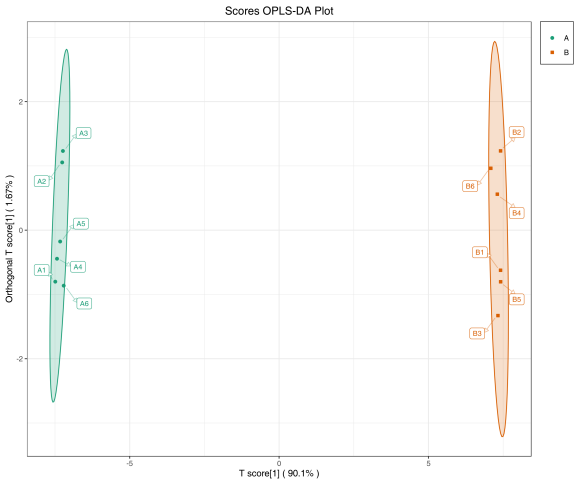

Fig 15: OPLS-DA score diagram

Note: The X-axis represents the predicted principal component, and the difference between groups can be seen in the horizontal direction. The Y-axis represents the orthogonal principal component, and the vertical direction shows the difference within the group. Percentage indicates the degree to which the component explains the data set. Each dot in the figure represents a sample, samples in the same Group are represented by the same color, and Group indicates sample groups.

Original file path:Final report/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/opls/group-ID*_vs_group-ID*_opls_score.*.

4.2.2 OPLS-DA model validation

The prediction parameters of the evaluation model are R^2X, R^2Y and Q^2, where R^2X and R^2Y represent the explanatory rate of the model to X and Y matrix respectively, and Q^2 represents the predictability of the model. The closer these three indicators are to 1, the more stable and reliable the model is. $Q^2 > 0.5$ can be considered as an effective model, and $Q^2 > 0.9$ can be considered as an excellent model. The following figure shows the OPLS-DA validation plot with the horizontal coY-axis indicating the model R^2Y, Q^2 values, and the vertical coY-axis is the frequency of the model classification effect. The model performs bootstrapping 200 times and if Q^2's P = 0.02, it indicates that the prediction ability of four random grouping models is better than that of the OPLS-DA model in the Permutation detection. If R^2Y's P = 0.545, it indicated that there were 109 random grouping models in the Permutation detection, whose explanation rate of Y matrix was better than that of the OPLS-DA model. In general, P < 0.05 is the best model.

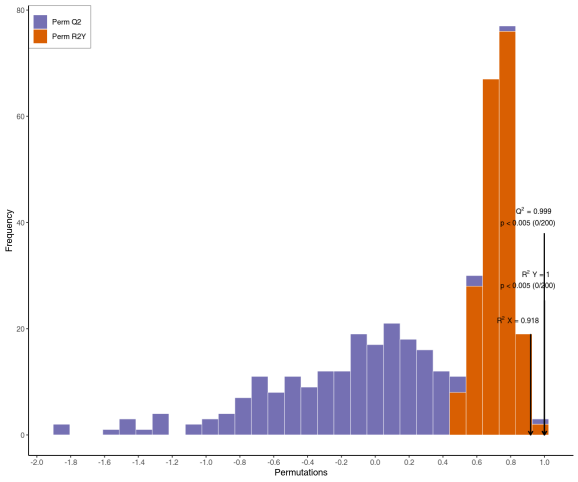

Fig 16: OPLS-DA verification diagram

Original file path:Final report/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/opls/group-ID*_vs_group-ID*_opls_permutation.*.

4.2.3 OPLS-DA S-plot

The figure below shows the OPLS-DA S-plot. The horizontal axis is the covariance between the principal components and metabolites, the vertical axis indicates the correlation coefficient between the principal components and the metabolites. The closer the points are to the top right corner or bottom left corner, the more significant the difference in metabolite abundance. Red dots indicate metabolites with VIP value > 1 and green dots indicate metabolites with VIP value <= 1.

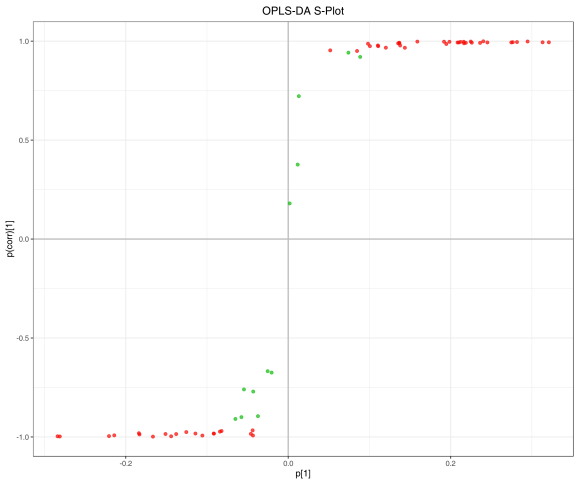

Fig 17: OPLS-DA S-plot

Original file path:Final report/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/opls/group-ID*_vs_group-ID*_opls_splot.*.

4.3 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.

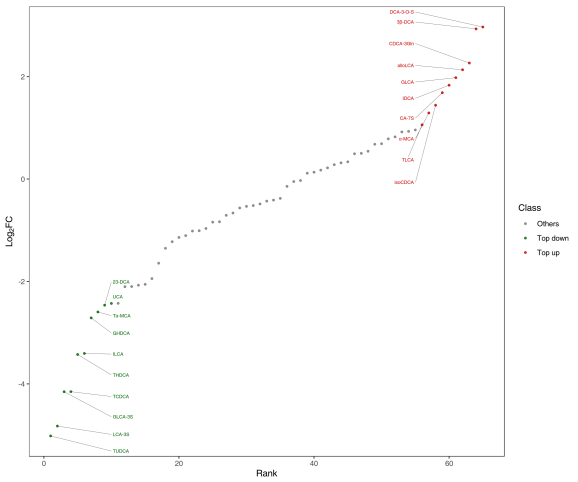

Fig 18: Dynamic distribution of metabolite content differences

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*

4.4 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 VIP > 1 were selected. VIP value represents the effect of the differences between groups for a particular metabolite in various models and sample groups. It is generally considered that the metabolites with VIP > 1 have significant difference.

2.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
GHDCA	Glycohyodeoxycholic Acid	down
HCA	hyocholic acid	down
IDCA	Isodeoxycholic acid	up
ILCA	isolithocholic acid	down
LCA	Lithocholic acid	down
LCA-3S	lithocholic acid-3-sulfate	down
MDCA	murideoxycholic acid	down
UCA	Ursocholic acid	down
alloLCA	5α-CHOLANIC ACID-3α-OL	up
isoCDCA	Isochenodeoxycholic Acid	up

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

4.4.1 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 .

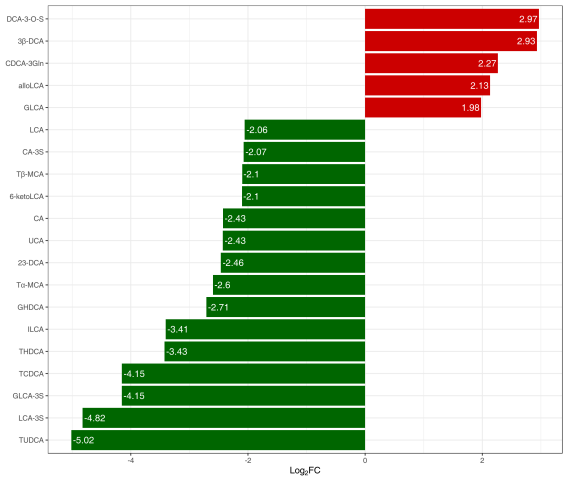

Fig 19: Bar chart of differential metabolites

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.*

4.4.2 Differential metabolite radar map

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

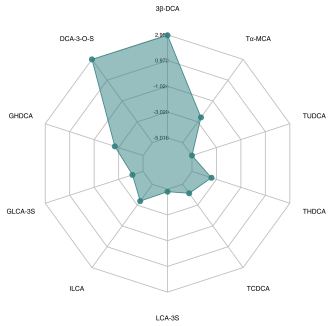

Fig 20: Differential metabolite radar map
Note: The grid lines correspond to the log_2FC. The green colored area is formed from the lines connecting the dots

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

4.4.3 VIP values of differential metabolites

The top 50 metabolites with the largest VIP value in the OPLS-DA model were selected and plotted.

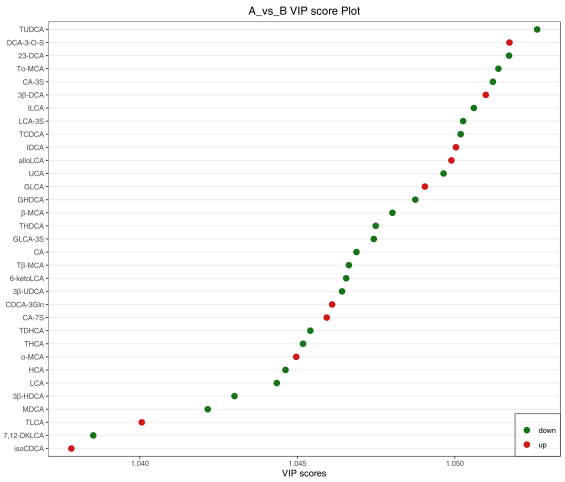

Fig 21: VIP values of differential metabolites
The top 50 metabolites with the largest VIP value in the OPLS-DA model were selected and plotted.

/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/vipscore/group-ID*_vs_group-ID*_vipScore.*.

4.4.4 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.

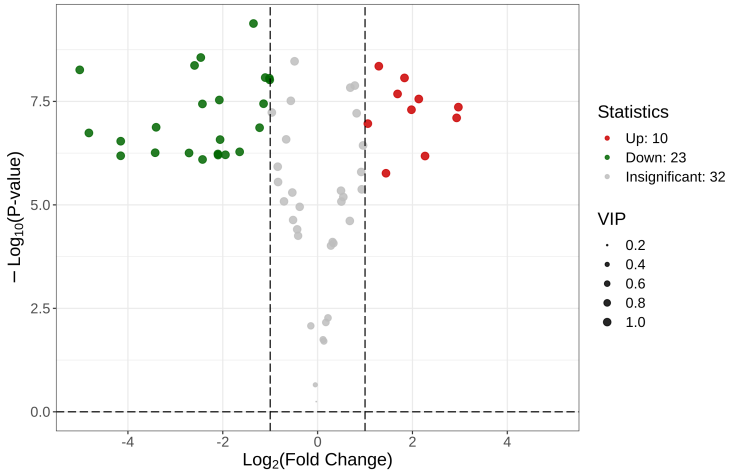

Fig 22: Volcanic plot of differential metabolites

Note: Each point in the volcano plot represents a metabolite with green dots represents the down-regulated differential metabolite, red dots represents the up-regulated differential metabolite, and gray dots represents the detected metabolite but show no insignificant difference. The X-axis represents the (log_2FC) value of metabolite between two groups. The further away from 0 on the X-axis, the greater the fold-change between two groups. If the metabolite were screened using VIP + FC + P-value, the Y-axis will represent the level of significant difference (-log_10p-value).The size of each dot represents the VIP value

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

4.4.5 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.

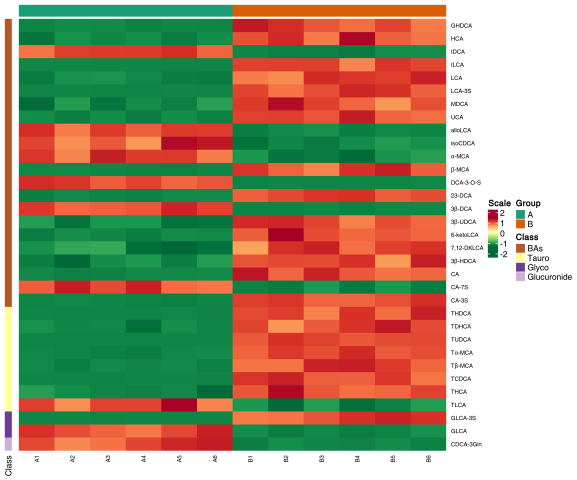

Fig 23: Heatmap of differential metabolites

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.*;

4.4.6 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.

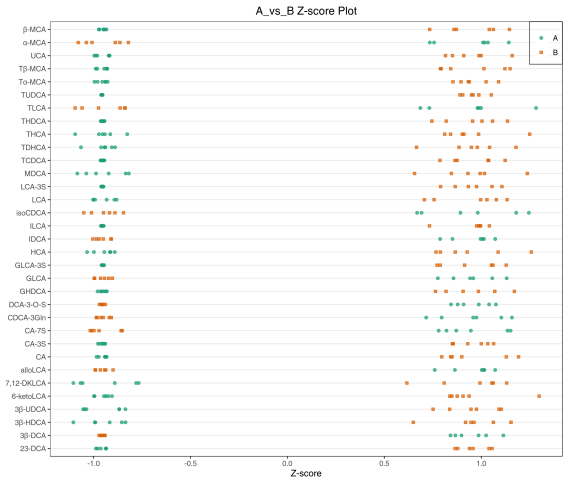

Fig 24: Z-value map of differential metabolites

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.*.

4.4.7 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.

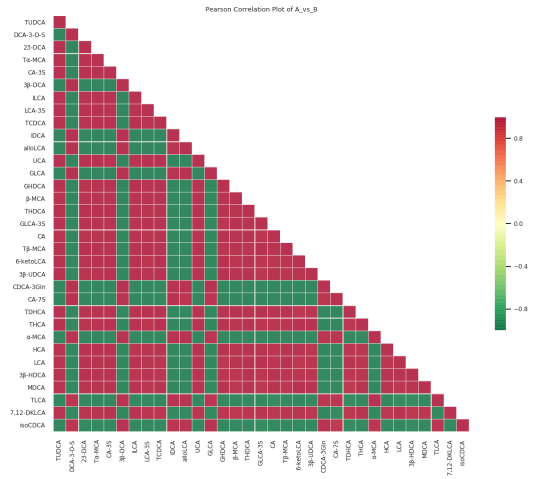

Fig 25: Heat map of correlation of different metabolites

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_*.*;

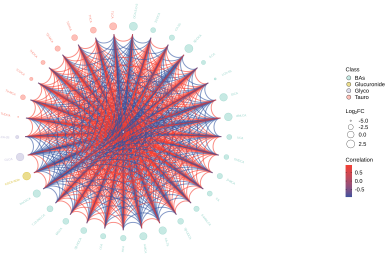

Fig 26: Chord diagram of differential metabolites

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_*.*;

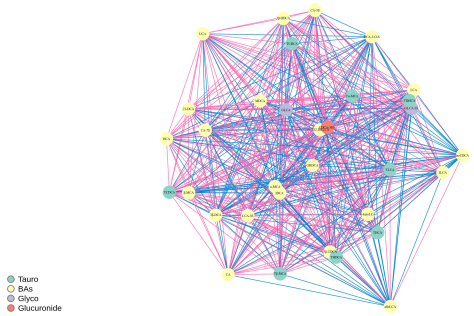

Fig 27: Correlation network diagram of differential metabolites

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*

4.4.8 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.

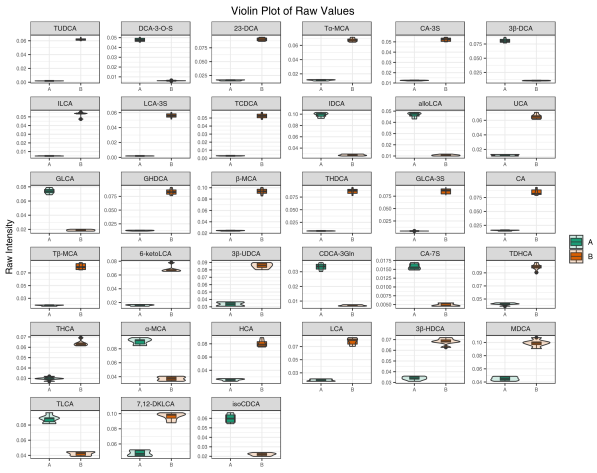

Fig 28: Violin plot of differential metabolites

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.*;

4.4.9 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.

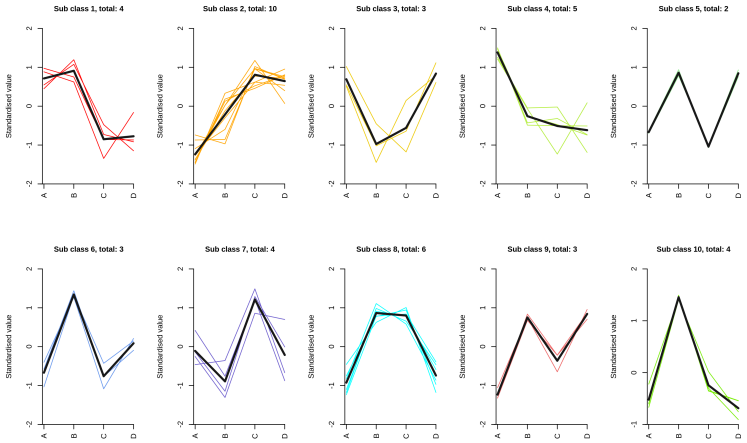

Fig 29: K-Means diagram of differential metabolites

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.*

4.4.10 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	33	23	10
A_vs_C	35	21	14

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

4.4.11 Venn diagram of differences among groups

Venn diagram was used to show the relationship between different metabolites in each group. Show petals in 5 groups or more. The results are shown below:

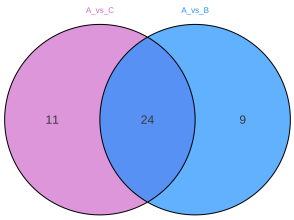

Fig 30: Venn diagram of differences among groups
Note: Each circle in the figure represents a comparison group, the number of circles and overlapped parts represents the number of common differential metabolites between comparison groups, and the number of non-overlapped parts represents the number of unique differential metabolites in comparison groups.

/2.Basic_analysis/Venn

4.5 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.

4.5.1 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:

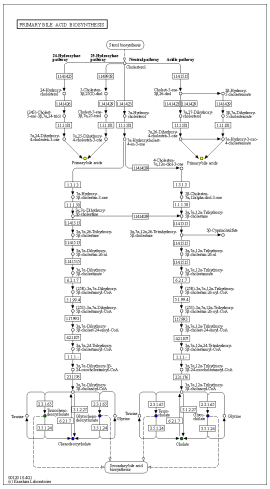

Fig 31: KEGG pathway of metabolites

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
GHDCA	Glycohyodeoxycholic Acid	down	-
HCA	hyocholic acid	down	C17649
IDCA	Isodeoxycholic acid	up	C17661
ILCA	isolithocholic acid	down	C17658
LCA	Lithocholic acid	down	C03990
LCA-3S	lithocholic acid-3-sulfate	down	-
MDCA	murideoxycholic acid	down	C15515
UCA	Ursocholic acid	down	C17644
alloLCA	5α-CHOLANIC ACID-3α-OL	up	-
isoCDCA	Isochenodeoxycholic Acid	up	C17660

Table 11: Enrichment Statistics of KEGG annotations for differential metabolites

ko_ID	Sig_compound	compound	Sig_compound_all	compound_all
ko04976	3	9	3	9
ko00120	2	6	3	9
ko01100	1	4	3	9
ko04979	1	4	3	9

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.

4.5.2 KEGG classification of differential metabolites

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

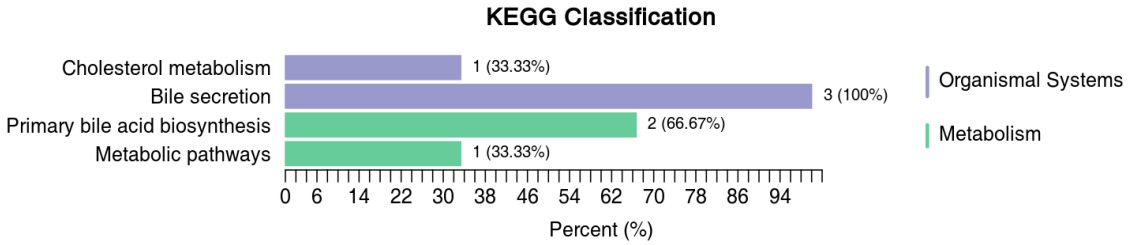

Fig 32: KEGG classification of differential metabolites
Note: the Y-axis shows the name of the KEGG pathway. The number of metabolites and the proportion of the total metabolites are shown next to the bar plot.

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

4.5.3 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.

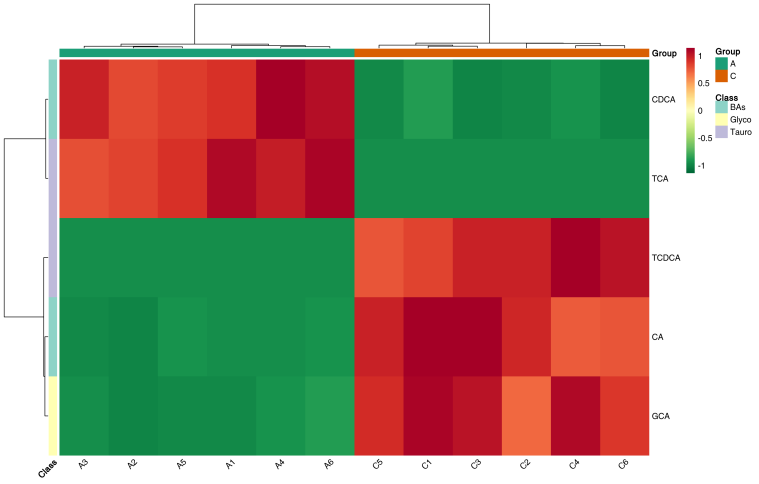

Fig 33: Clustering heat map of differential metabolites in KEGG pathway

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.*.

4.5.4 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:

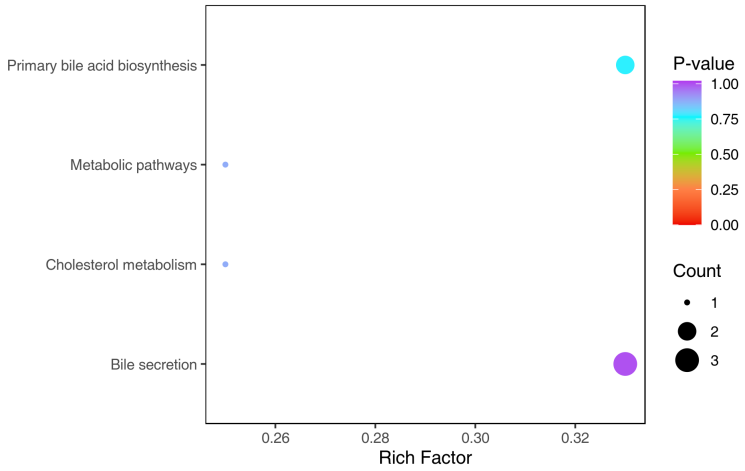

Fig 34: KEGG enrichment diagram of differential metabolites

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.*.

4.5.5 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)

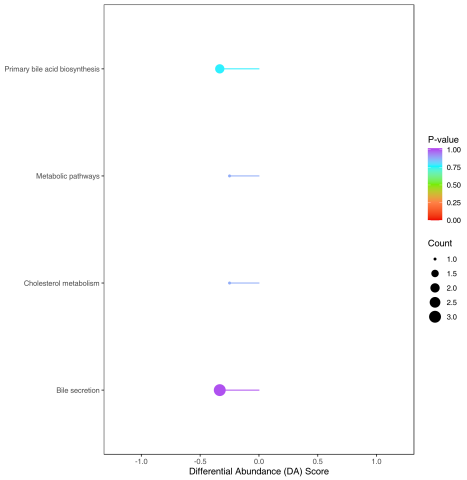

Fig 35: Difference abundance score

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*

4.6 Functional annotation and enrichment analysis in HMDB database

4.6.1 Functional annotation and enrichment analysis of differential metabolites in HMDB database

HMDB is a widely used database that has collected more than 40,000 endogenous metabolites and more than 5000 related protein or gene information. Records in this database links to external databases (such as KEGG, Metlin, Biocyc, etc.) and also contains mass spectra and NMR spectra data. The HMDB sub-database SMPDB also provides a detailed overview of human metabolism, metabolic disease pathways, and metabolite signaling and drug activity pathways.

Pathway enrichment analysis was performed only with the Primary Pathways. The results are as follows:

Table 12: SMPDB pathway enrichment for differential metabolites

primary_SMPDB_ID	p_value
”SMP0000035”	1
”SMP0000314”	1
”SMP0000318”	1
”SMP0000317”	1
”SMP0000316”	1
”SMP0000315”	1
”SMP0000720”	1

The differential metabolites from the top 20 HMDB Primary Pathways pathways with P-value were annotated and visualized using the HMDB database. Detailed information about each group can be found in the corresponding data files. Partial results are shown below:

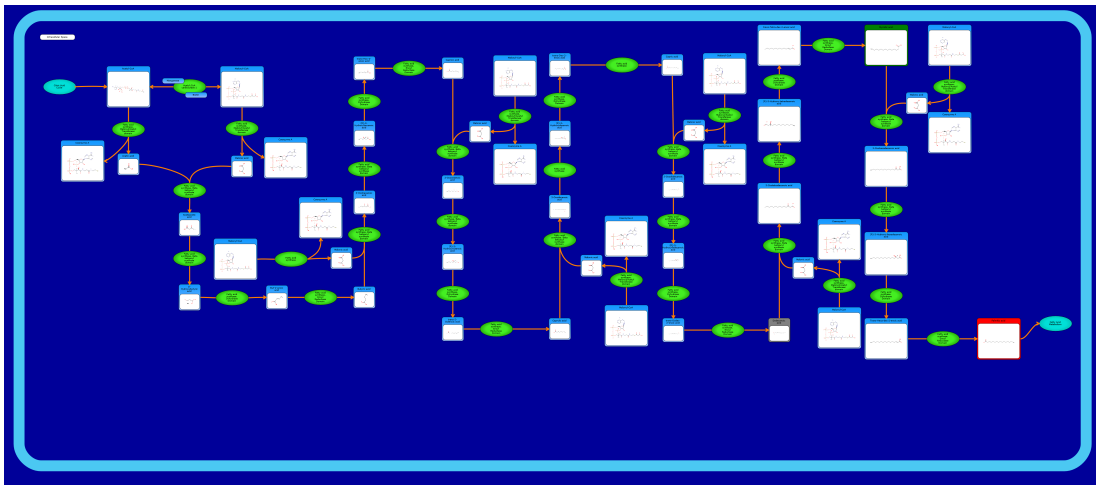

Fig 36: HMDB pathway map of differential metabolites
Note: Red indicated that the metabolite content was significantly up-regulated in the experimental group, Gray indicated that the metabolite content was detected but did not change significantly, Green indicated that the metabolite content was significantly down-regulated in the experimental group. and blue represents metabolites in the pathway that were not detected in this experiment.The causes of phenotypic differences among study subjects were sought through metabolic pathways.

The top 20 HMDB Primary Pathways based on P-value ranking were chosen for Rich Factor plot. The Rich Factor is the ratio of the number of differential metabolites in the corresponding pathways to the total number of metabolites annotated to the same pathway. The higher the value is, the greater the degree of enrichment. The closer P-value is to 0, the more significant the enrichment is. The size of the dots in the figure represents the number of differential metabolites enriched into the corresponding pathway. The results are shown below:

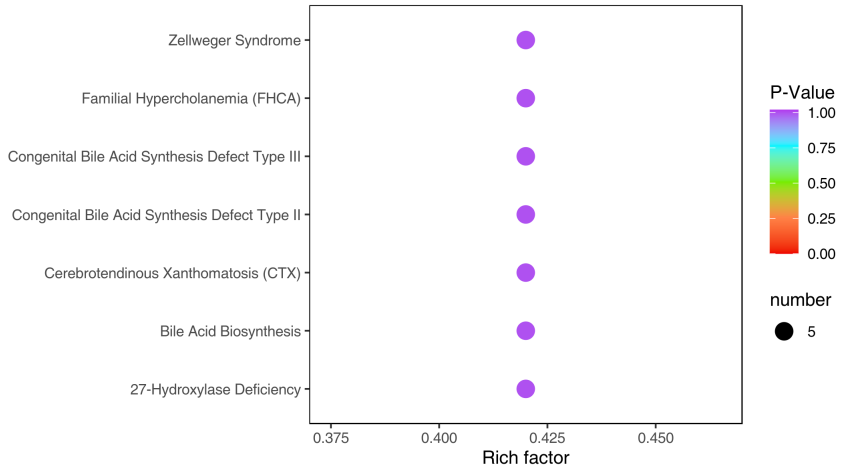

Fig 37: HMDB enrichment diagram of differential metabolites

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.

Statistical table of differential metabolite enrichment in HMDB database:Final report/2.Basic_analysis/Difference_an ID*_vs_group-ID*/enrichment/group-ID*_vs_group-ID*_SMPDB_primary.xlsx;

HMDB pathway map of metabolites:Final report/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/enrichment/SMP_primary_pathway;

HMDB enrichment diagram of differential metabolites:Final report/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/enrichment/group-ID*_vs_group-ID*SMPDB_primary_Enrichment.*.

4.7 Associated diseases

We annotated disease information according to the HMDB database for differential metabolites. Some of the results are shown below :

Table 13: Table of association between differential metabolites and diseases

CompoundName	HmdbDiseases
Glycohyodeoxycholic Acid	-
hyocholic acid	Colorectal cancer | Primary biliary cirrhosis
Isodeoxycholic acid	-
isolithocholic acid	-
Lithocholic acid	Cystic fibrosis | Biliary atresia | Colorectal cancer | Primary biliary cirrhosis
lithocholic acid-3-sulfate	-
murideoxycholic acid	-
Ursocholic acid	Primary biliary cirrhosis
5α-CHOLANIC ACID-3α-OL	Cystic fibrosis | Biliary atresia | Colorectal cancer | Primary biliary cirrhosis
Isochenodeoxycholic Acid	-

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

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. Zheng H , Zhang Q , Quan J , et al. Determination of sugars, organic acids, aroma components, and carotenoids in grapefruit pulps[J]. Food Chemistry, 2016, 205(Aug.15):112-121.
4. Thevenot, E. A., et al. (2015). Analysis of the Human Adult Urinary Metabolome Variations with Age, Body Mass Index, and Gender by Implementing a Comprehensive Workflow for Univariate and OPLS Statistical Analyses. J Proteome Res 14(8): 3322-3335.
5. Chen, W., et al. (2013). A novel integrated method for large-scale detection, identification, and quantification of widely targeted metabolites: application in the study of rice metabolomics. Mol Plant 6(6): 1769-1780.
6. 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.

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 VIP and absolute Log_2FC (fold change). VIP values were extracted from OPLS-DA result, which also contain score plots and permutation plots, was generated using R package MetaboAnalystR. The data was mean centering before OPLS-DA. In order to avoid overfitting, a permutation test (200 permutations) was performed.

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 14: 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:

$$x' = x - \mu$$