Innovative Metabolomics Insights for Better Health

Demo TM Widely Targeted Metabolomics Report

Metware Biotechnology Inc.

www.metwarebio.com

Contents

Demo TM Widely Targeted Metabolomics Report

1 Abstract

Metabolomics is the study of all metabolites and their dynamics in a biological system by performing qualitative and quantitative analyses. The data is often used to study the metabolic basis of observed phenotypes, to understand the response mechanisms under different physical, chemical, or pathological conditions, and to evaluate safety of food and drugs.

For this project, 9 samples were selected and divided into 3 groups for metabolomics study. A total of 1016 metabolites were detected and differential metabolites between sample groups were analyzed. The results of differential metabolite analysis are summarized below.

Table 1: Number of differential metabolites

group name	All sig diff	down regulated	up regulated
A_vs_B	81	61	20
A_vs_C	39	21	18

Number of identified metabolites: Final report/2.Basic_Analysis/Difference_analysis/sigMetabolitesCount.xlsx

2 The experimental process

Ultra-performance liquid chromatography-tandem mass spectrometry (UPLC-MS/MS) is a technique used for accurate qualitative and quantitative analysis for various compounds. The main purpose of metabolomics analysis is to detect and identify metabolites with important biological significance by differentiate statistically significant differential metabolites between sample groups. The overall process is as follows:

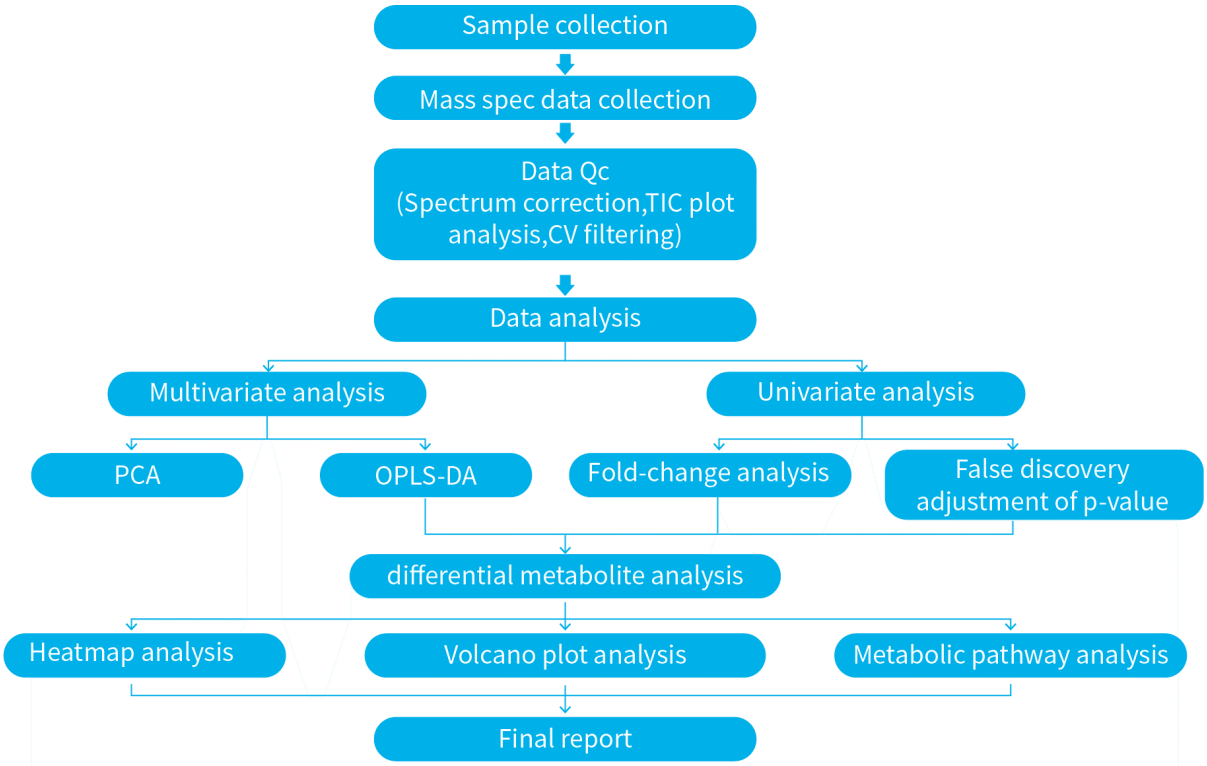


Fig 1: Flow chart of metabolomics analysis

2.1 Sample information and experimental materials and methods

Each sample group and corresponding sample information are as follows:

Table 2: Sample information

Species	Tissue	Sample	Group
Human	Cells	A1	A
Human	Cells	A2	A
Human	Cells	A3	A
Human	Cells	B1	B
Human	Cells	B2	B
Human	Cells	B3	B
Human	Cells	C1	C
Human	Cells	C2	C
Human	Cells	C3	C

Sample information: Final report/1.Data_Assess/all_group/sample_info.xlsx

2.2 Standards and reagents

Table 3: Information of standards and reagents

reagent	level	brand
methanol	HPLC Grade	Thermo Fisher
acetonitrile	HPLC Grade	Thermo Fisher
formic acid	HPLC Grade	Sigma
standard	HPLC Grade	BioBioPha/Sigma-Aldrich

2.3 Sample extraction process

2.3.1 Cell samples class I

Samples stored at -80 °C was thawed on ice. 500 µL solution (Methanol : Water = 4 : 1, V/V) containing internal standard was mixed with the cell sample and vortexed for 3 min. The sample was placed in liquid nitrogen for 5 min, on the dry ice for 5 min, and then thawed on ice and vortexed for 2 min. This freeze-thaw cycle was repeated for three times total. The sample was centrifuged at 12000 rpm for 10 min (4 °C). 300 µL of the supernatant was collected and placed in -20 °C for 30 min. The sample was centrifuged again at 12000 rpm for 3 min (4 °C). A 200 µL aliquot of the supernatant was used for LC-MS analysis.

2.4 Chromatography-mass spectrometry acquisition conditions

2.4.1 Acquisition conditions for untargeted detection

The data acquisition instruments consisted of Ultra Performance Liquid Chromatography (UPLC) (ExionLC 2.0, https://sciex.com/) and Quadrupole-Time of Flight Spectrometry (TripleTOF 6600+, AB SCIEX).

Liquid phase conditions were as follows:

(1) Chromatographic column: ACQUITY HSS T3 (2.1 × 100mm, 1.8 um)

(2) Mobile phase: A phase was ultrapure water (0.1 % formic acid added), B phase was acetonitrile (0.1 % formic acid added);

(3) Column temperature: 40 °C;

(4) Flow rate: 0.4 ml/min;

(5) Injection volume: 5 uL.

Table 4: Elution gradient

Time (min)	Flow rate(mL/min)	A (%)	B (%)
0.0	0.4	95	5
11.0	0.4	10	90
12.0	0.4	10	90
12.1	0.4	95	5
14.0	0.4	95	5

The mass spectrum conditions were as follows:

Table 5: Mass spectrum conditions

Parameter	ESI+	ESI-
Curtain Gas	25	25
IonSpray Voltage	5500	4500
Temperature	500	500
Ion Source Gas1	50	50
Ion Source Gas2	50	50
Declustering Potential	80	-80
Collision Energy	30	-30
Collision Energy Spread	15	15

2.4.2 Acquisition conditions for widely targeted detection

The data acquisition instruments consisted of Ultra Performance Liquid Chromatography (UPLC) (ExionLC 2.0, https://sciex.com/) and tandem mass spectrometry (MS/MS) (QTRAP®6500+, https://sciex.com/).

2.4.2.1 Liquid phase conditions

(1) Chromatographic column: Waters ACQUITY UPLC HSS T3 C18 1.8 µm, 2.1 mm * 100 mm;

(2) Mobile phase: A phase was ultrapure water (0.1 % formic acid added), B phase was acetonitrile (0.1 % formic acid added);

(3) Gradient program: 95:5 V/V at 0 min, 10:90 V/V at 10.0 min, 10:90 V/V at 11.0 min, 95:5 V/V at 11.1 min, 95:5 V/V at 14.0 min;

(4) Flow rate: 0.4 ml/min; Column temperature: 40 °C; Injection volume: 2 µl.

2.4.2.2 Mass spectrum conditions

LIT and triple quadrupole (QQQ) scans were acquired on a triple quadrupole-linear ion trap mass spectrometer (QTRAP), QTRAP® LC-MS/MS System, equipped with an ESI Turbo Ion-Spray interface, operating in positive and negative ion mode and controlled by Analyst 1.6.3 software (Sciex). The ESI source operation parameters were as follows: source temperature 500°C; ion spray voltage (IS) 5500 V (positive), -4500 V (negative); ion source gas I (GSI), gas II (GSII), curtain gas (CUR) were set at 50, 50, and 25.0 psi, respectively; the collision gas (CAD) was high. Instrument tuning and mass calibration were performed with 10 and 100 μmol/L polypropylene glycol solutions in QQQ and LIT modes, respectively. A specific set of MRM transitions were monitored for each period according to the metabolites eluted within this period.

2.5 Qualitative and quantitative principles of metabolites

The mixed samples first underwent untargeted metabolomics detection. Metabolites were analyzed qualitatively with in-house database MWDB, integrated public database (including Metlin, HMDB, and KEGG), AI database, and MetDNA. The identified metabolites were integrated with the in-house database MWDB. Lastly, quantification using MRM mode was performed for all samples based on the newly integrated database.

Metabolites were quantified by triple quadrupole mass spectrometry with multiple reaction monitoring (MRM). In MRM mode, the first quadrupole screens the precursor ions for the target compound and excludes ions of other molecular weights. After ionization induced by the impact chamber, the precursor ion is fragmented, and a characteristic fragment ion is selected through the third quadrupole and excludes the interference of other untargeted ions. By selecting a particular fragment ion, quantification is more accurate and reproducible.

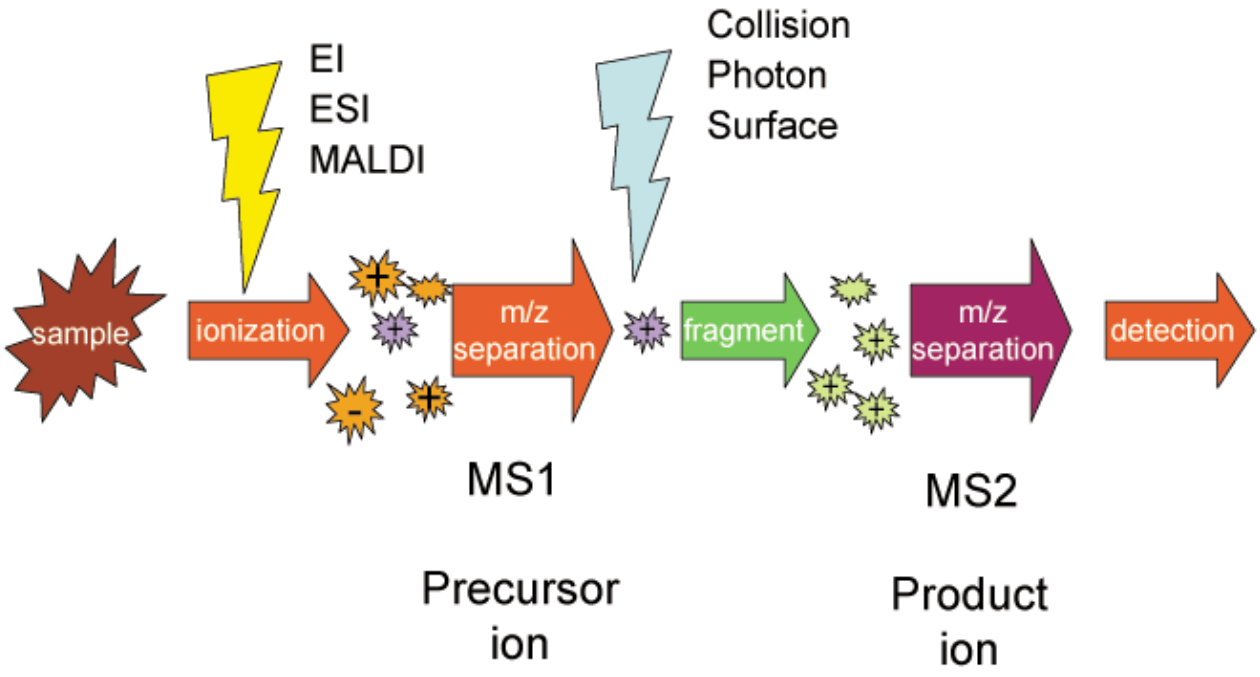


Fig 2:

Schematic diagram of multiple reaction monitoring mode by mass spectrometry

2.6 Data preprocessing

Based on the raw data file ALL_sample_data_raw.xlsx, the missing values were first filled in using 1/5 of the minimum value of each row (metabolite), and then the CV value of the QC sample was calculated, and the metabolites with a CV value less than 0.3 were retained to obtain the final data file ALL_sample_data.xlsx.

ALL_sample_data_raw.xlsx: Final report/1.Data_Assess/all_group/ALL_sample_data_raw.xlsx

ALL_sample_data.xlsx: Final report/1.Data_Assess/all_group/ALL_sample_data.xlsx

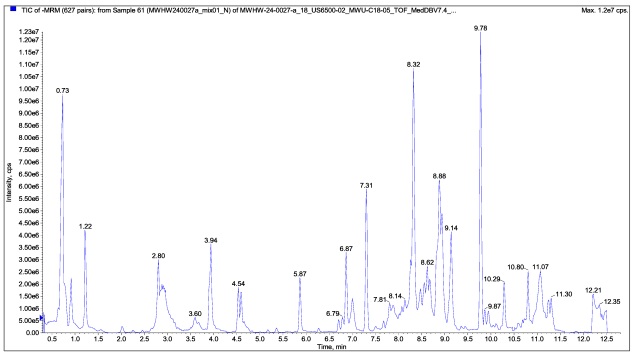
3 Data evaluation

3.1 Results evaluation for Widely Targeted detection

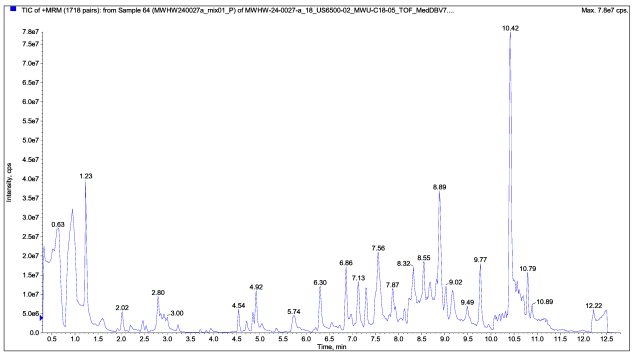
3.1.1 Qualitative and quantitative analysis

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 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).



(a) Demo_QC_MS_TIC-N

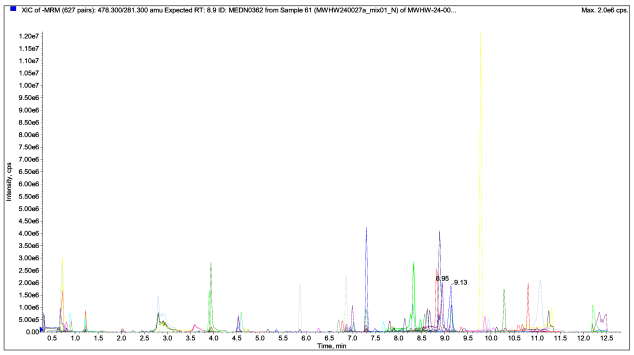


(b) Demo_QC_MS_TIC-P

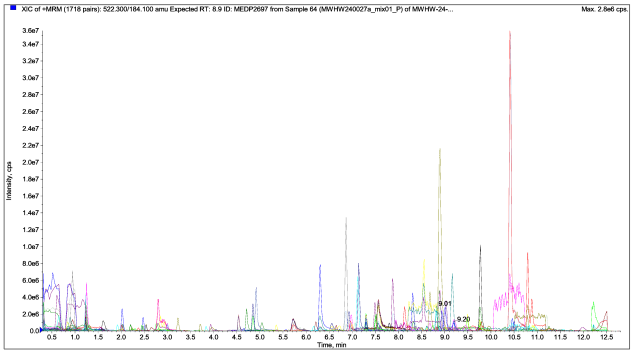
Fig 3: Total ion current diagram of mixed phase mass spectrum analysis

Note: N stands for negative ion mode, P for positive ion mode

Total ion current diagram of mixed phase mass spectrum analysis: Final report/1.Data_Assess/QC/*_QC_MS_TIC*.*



(a) Demo_MRM_detection_of_multimodal_maps-N



(b) Demo_MRM_detection_of_multimodal_maps-P

Fig 4: Multi-peak diagram of MRM metabolite detection

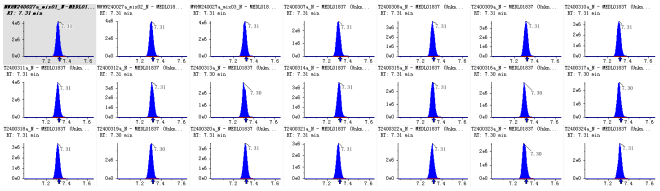
Note: N stands for negative ion mode, P for positive ion mode

Multi-peak diagram of MRM metabolite detection: Final report/1.Data_Assess/QC/*_MRM_detection_of_multimodal_maps*.*

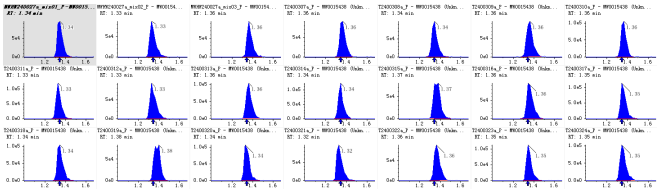
The MRM metabolite detection multi-peak diagram shows the compounds that were detected in the sample, with each mass spectrum peak color representing one detected metabolite. The characteristic ions of each compound were selected by triple quadrupole and measured for their signal intensity (CPS). The mass spectrometry data was analyzed using MultiQuant software and the chromatographic peaks were integrated and corrected. The peak area of each chromatographic peak represents the relative abundance of the corresponding compound.

Mass spectrum peak of each metabolite in different samples was corrected based on retention time and

peak distribution information to ensure the accuracy of qualitative and quantitative analysis. The following figure shows the integral correction results from a randomly selected metabolite in the samples. The X-axis of each sub-plot is the retention time (min), and the Y-axis of each sub-plot is the ion current intensity (CPS) of a certain metabolite ion detection.



(a) Demo_Integral_correction_diagram-N



(b) Demo_Integral_correction_diagram-P

Fig 5: Integral 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 current intensity (CPS) of a certain metabolite ion detection, and the peak area represents the relative content of the substance in the sample.

Integral correction diagram for quantitative analysis of metabolites: Final report/1.Data_Assess/QC/*_Integral_correction_diagram*.*

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

Table 6: Information of metabolite detected in sample

Index	A1	A2
FDATN00665	2009.2715	5146.6752
FDATN01308	95918.2714	122515.2193
FDATN01519	356079.9272	352710.3771
FDATP00838	3242.6491	3242.6491
MADP0518	3137191.1703	1896658.7080
MADP0547	5169438.6535	4023634.6040
MEDL00401	95219.0864	97616.3819
MEDL00977	1610262.1804	8618232.5180
MEDL01793	989.9448	989.9448
MEDL01837	9130892.2215	6813413.2800

Information of metabolite detected in sample: Final report/1.Data_Assess/all_group/ALL_sample_data.xlsx

Compound composition is sample-specific and varies between samples. The analysis of compound composition ratios can help examine the distribution of major compounds in the samples. The proportion of each compound class were analyzed and shown in the ring figure.

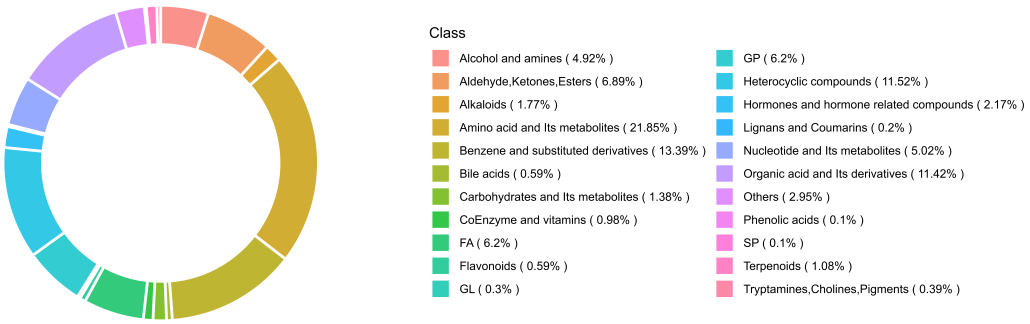


Fig 6: Ring plot of metabolite categories

Note: Each color represents a metabolite class, and the area of the color block indicates the proportion of that class.

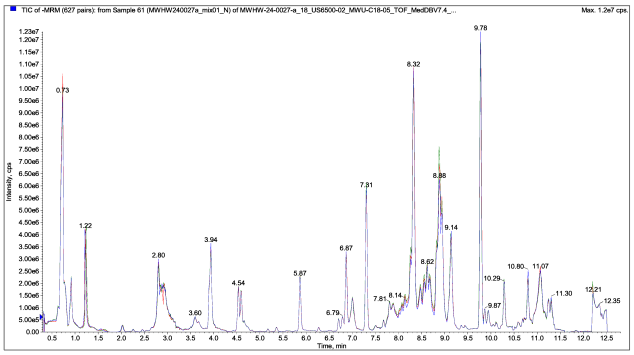
Ring plot of metabolite categories: Final report/1.Data_Assess/Class_Count/Class_Count_Ring.*

3.1.2 Quality control sample analysis

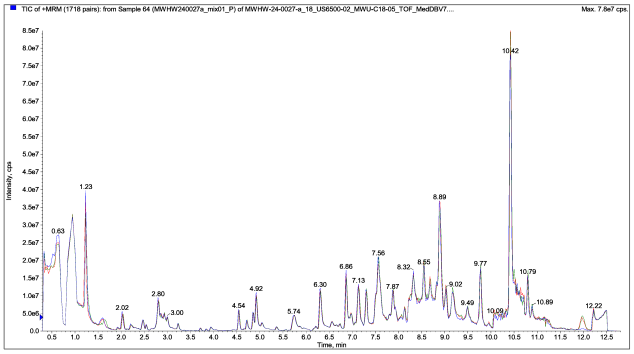
3.1.2.1 Total ion flow chromatogram

A quality control (QC) sample was prepared from a mixture of all sample extracts to examine the reproducibility of the entire metabolomics process. During data collection, one quality control sample was generally inserted for every 10 test samples.

Reproducibility of metabolite extraction and detection process was assessed by analyzing overlapping total ion flow diagram (TIC diagram) from different QC samples. High overlapping rate of TIC diagrams indicates high stability of the instruments throughout the data acquisition process



(a) Demo_QC_MS_tic_overlap-N



(b) Demo_QC_MS_tic_overlap-P

Fig 7: 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. N stands for negative ion mode and P stands for positive ion mode.

TIC overlap diagram detected by QC sample essence spectrum: Final report/1.Data_Assess/QC/*_QC_MS_tic_overlap*.*

3.1.2.2 Peak appearance of internal standards in blank samples

Blank samples were interspersed throughout the experiment, and their peaks can reflect whether there are compound residues from the detection process. The figure below shows that no obvious internal standard peaks were detected in the blank samples, indicating that possibility of cross-contamination between the

samples is minimal.

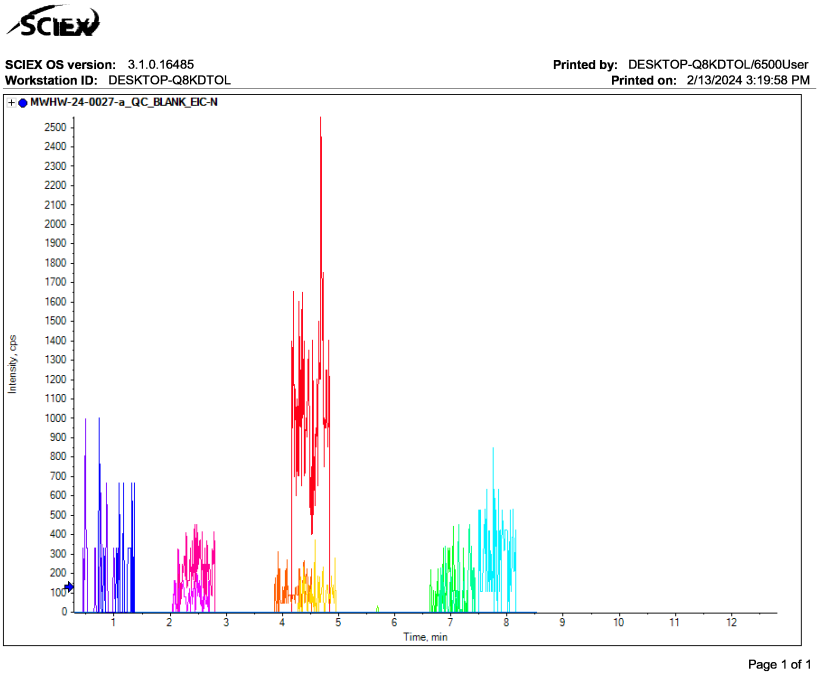


Fig 8: EIC diagram of internal label in blank sample

Note: The signals in the EIC diagram are noise peaks, and the internal standard substance has no obvious signal peak at the corresponding time.

EIC diagram of internal label in blank sample: Final report/1.Data_assess/*/QC/*_*_BLANK_EIC.png

3.1.2.3 Correlation analysis of QC samples

Pearson's correlation analysis was performed on the QC samples. The higher the correlation between QC samples (| r | closer to 1) means that the stability of the entire detection process is optimal.

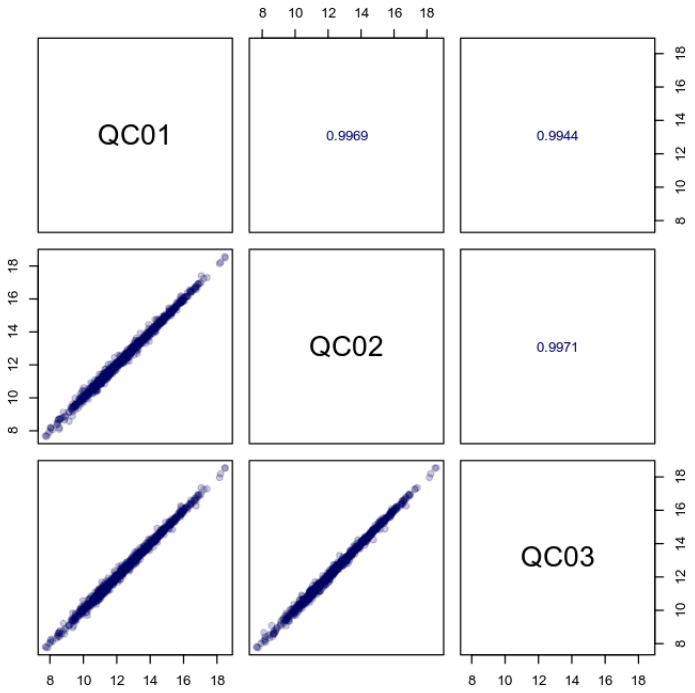


Fig 9: Correlation of the QC sample

Note: The bottom left square of the diagonal line is the correlation scatter plot of the corresponding QC samples. The horizontal and vertical coordinates are the metabolite content (for Log processing), and each point in the plot represents one metabolite. The upper right square of the diagonal line is the Pearson correlation coefficient of the corresponding QC samples.

Correlation of the QC sample: Final report/1.Data_Assess/QC/QC_cor.*

Table of Pearson Correlation Coefficients for all Ssamples: Final report/1.Data_Assess/all_group/ALL_sample_cor.xlsx

3.1.2.4 Stability of internal standards in QC samples

Internal standards with known concentrations were added to the QC samples for assessing variations between samples. The smaller the variation (CV ≤ 15%), the more stable the detection process and the higher the data quality.

Table 7: Stability of internal standard in QC samples

Index	m/z	RT (min)	CV
MWS04187-IS-P	210.1291	2.40	0.0159633
MWS20572-IS-P	198.0982	2.99	0.0226198
MWS3085-IS-P	281.0054	4.34	0.0237477
MWS5078-IS-P	170.0617	1.20	0.0352263
MWS04187-IS-N	208.1135	2.41	0.0048695
MWS2742-IS-N	379.3056	7.78	0.0076959
MWS015201-IS-N	160.1617	6.82	0.0103785
B015202-IS-N	215.1980	7.89	0.0117267

Stability of internal standard in QC samples: Final report/1.Data_assess/*/QC/*_internal_standard_info.xlsx

3.1.2.5 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) was used to analyze the frequency of compound CVs that is smaller than the reference value. The higher the proportion of compounds with low CV value in the QC samples, the more stable the experimental data. As a rule of thumb, the proportion of compounds with CV value less than 0.5 in the QC samples is higher than 85 % indicates that the experimental data is relatively stable. The proportion of compounds with CV value less than 0.3 in the QC samples is higher than 75 % indicates that the experimental data is very stable.

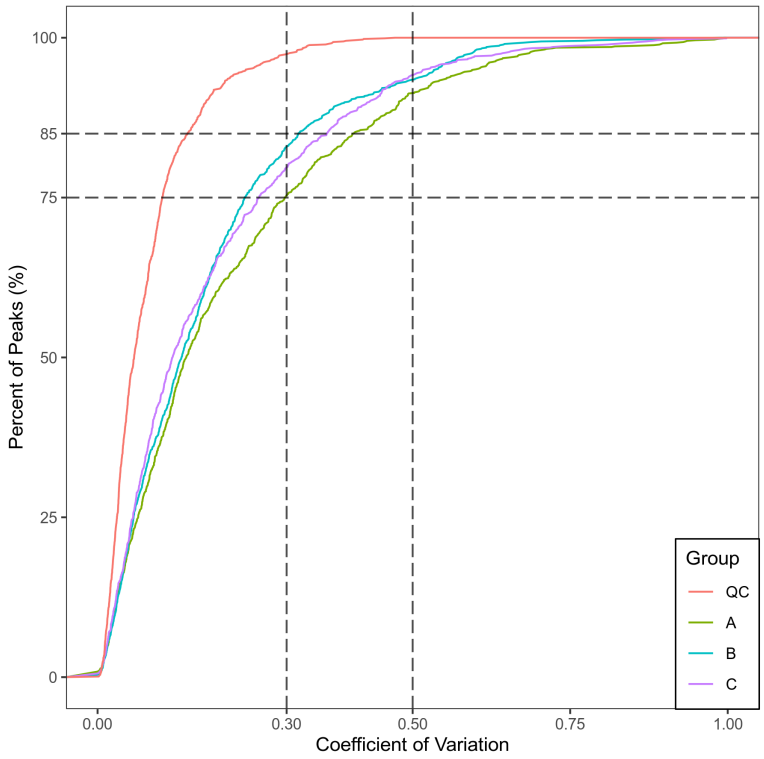


Fig 10: CV distribution of each group

Note: the X-axis represents the CV value, the Y-axis represents the proportion of metabolites. Different colors represent different sample groups. QC indicates quality control samples. The two dash lines on X-axis correspond to 0.3 and 0.5; the two dash line on Y-axis correspond to 75% and 85%.

CV distribution of each group: Final report/1.Data_Assess/QC/*_CV_ECDF.*

3.1.3 Principal Component Analysis (PCA)

3.1.3.1 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 between 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. The 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.

3.1.3.2 Principal component analysis of the sample population

Principal component analysis (PCA) was performed on all the samples (including QC samples) to examine the overall metabolic 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:

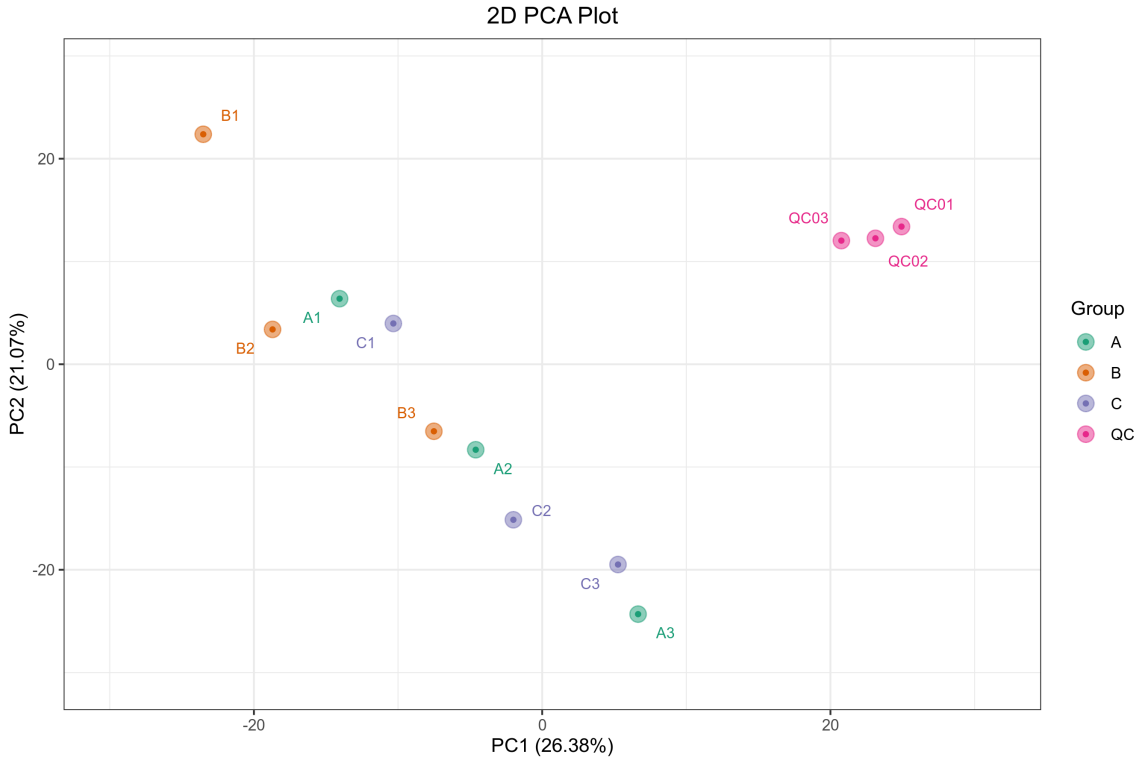


Fig 11: PCA score diagram of quality spectrum data of each group of samples and quality control samples
Note: PC1 represents the first principal component, PC2 represents the second principal component, and PC3 represents the third principal component. Percentage represents the interpretation rate of 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.

Principal component analysis of the sample population: Final report /1.Data_Assess/pca/

3.1.3.3 Principal component univariate statistical process control

We plotted the sample order chart based on principle component analysis results. Each point in the order 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.

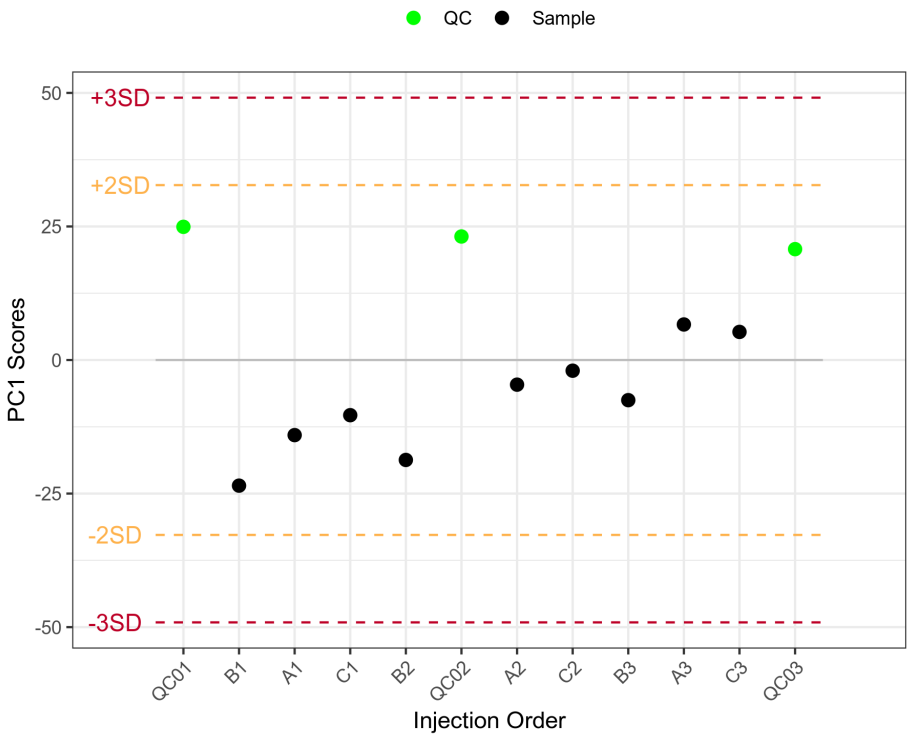


Fig 12: PC1 variation diagram of all the sample

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.

PC1 control diagram of population sample: Final report/1.Data_Assess/pca/*_PC1_QCC.*

3.1.4 Hierarchical Cluster Analysis (HCA)

3.1.4.1 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

3.1.4.2 Hierarchical Cluster Analysis results

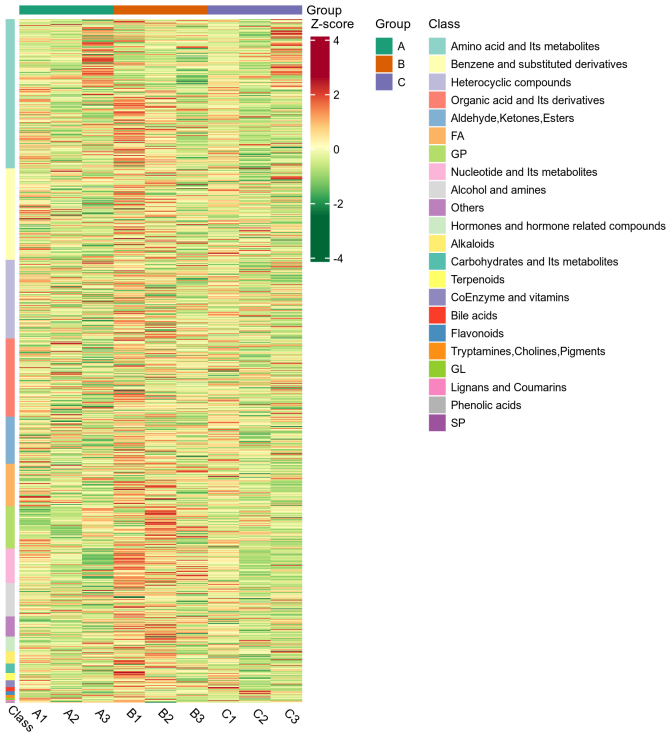


Fig 13: Sample clustering diagram

Note: X-axis indicates the sample name and the Y-axis are the metabolites. Group indicates sample groups. The different colors are the results after standardization of the relative contents (red represents high content, green represents low content). *_all_heatmap_class: Heatmap by metabolites classification, Class represents the first-level classification of metabolites. *_all_heatmap_col-row_cluster: clustering analysis is performed for both metabolites and samples. The clustering tree on the left represents clustering on the metabolites. The clustering tree on the top represent clustering on the samples. *_all_heatmap_row_cluster: clustering analysis is performed for metabolites only.

Hierarchical Cluster Analysis results: Final report/1.Data_Assess/heatmap/

4 Analysis results

4.1 Grouping principal component analysis

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.

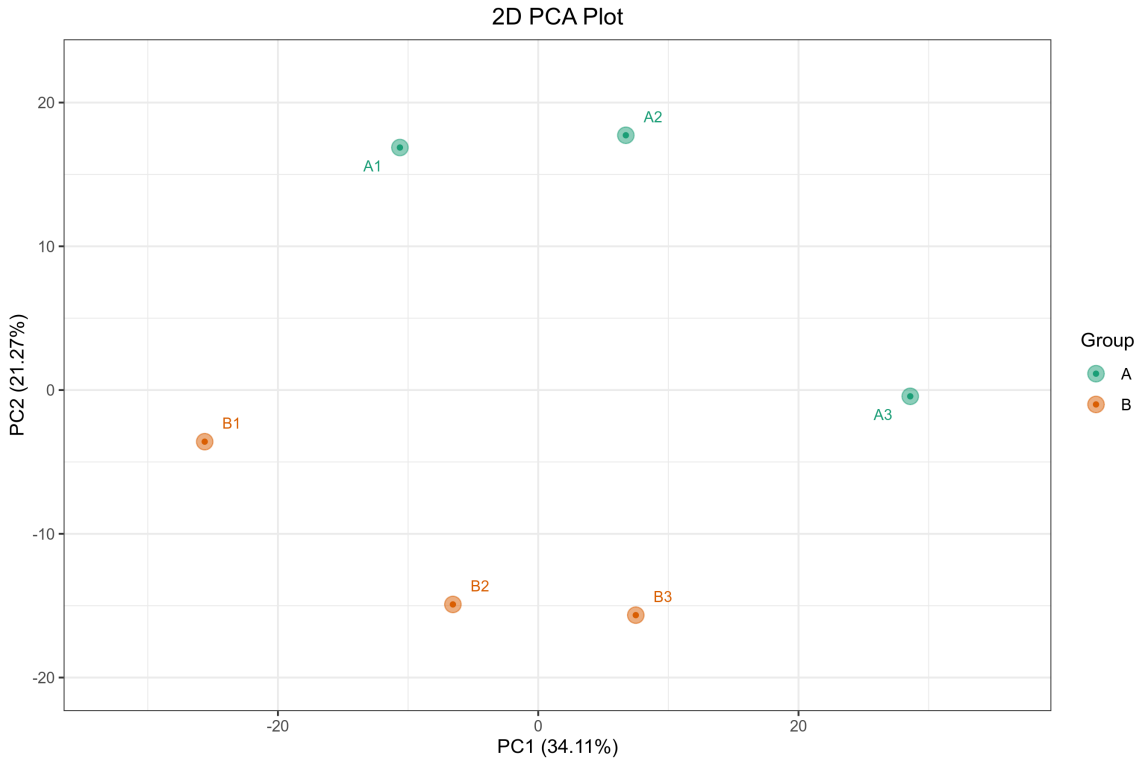


Fig 14: Principal component analysis of different groups

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:

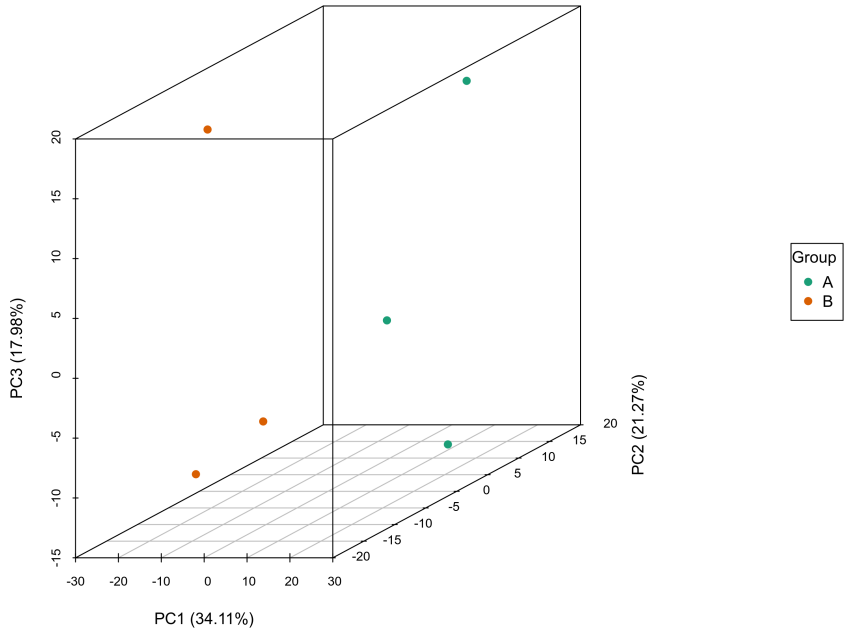


Fig 15: Three-dimensional PCA plot of different groups

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 below:

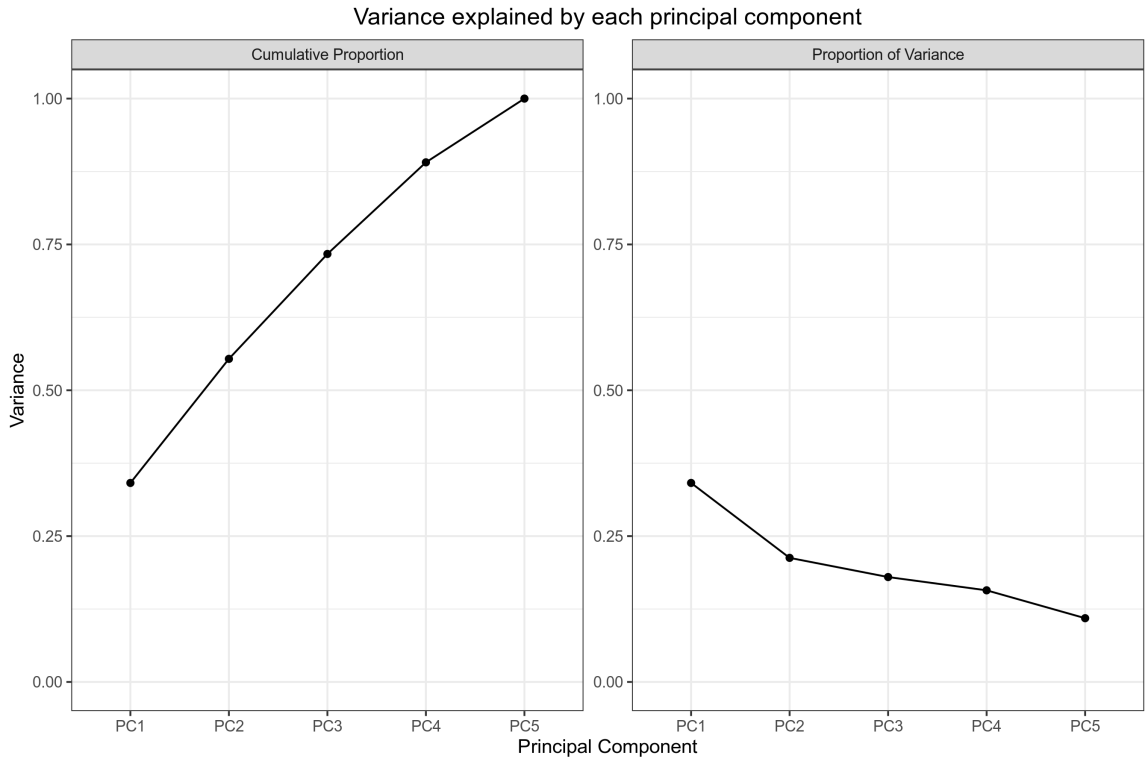


Fig 16: The explainable variation of the first five principal components

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 sample groups: Final report/2.Basic_Analysis/Difference_analysis/*_vs_*/pca/

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 the 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 metabolites. 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 8: Partial results of OPLS-DA

Index	Compounds	VIP
FDATN00665	Diacerein	1.5204626
FDATN01308	Antineoplaston A10	0.8415616
FDATN01519	D-phenylalanine	0.1063203
FDATP00838	Benzethonium chloride	2.0058757
MADP0518	N3-(4-fluorophenyl)-1h-pyrazolo[3,4-d]pyrimidine-3,4-diamine	1.6087124
MADP0547	O-Acetyl-L-homoserine hydrochloride	1.0504053
MEDL00401	Confertifoline	1.0885350
MEDL00977	(E,Z)-2-Amino-3,14-octadecadien-1-ol	0.1040816
MEDL01837	Nordihydrocapsiate	0.0113670
MEDL01870	1-(4-Hydroxy-3-methoxyphenyl)-3-decanone	0.5920043

Partial results of OPLS-DA: Final report/2.Basic_Analysis/Difference_analysis/*_vs_*/*_info.xlsx

OPLS-DA model overview: Final report/2.Basic_Analysis/Difference_analysis/*_vs_*/opls/*_OPLS-DA_model.*

OPLS-DA model summary table: Final report/2.Basic_Analysis/Difference_analysis/*_vs_*/opls/*_OPLS-DA_summary.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).

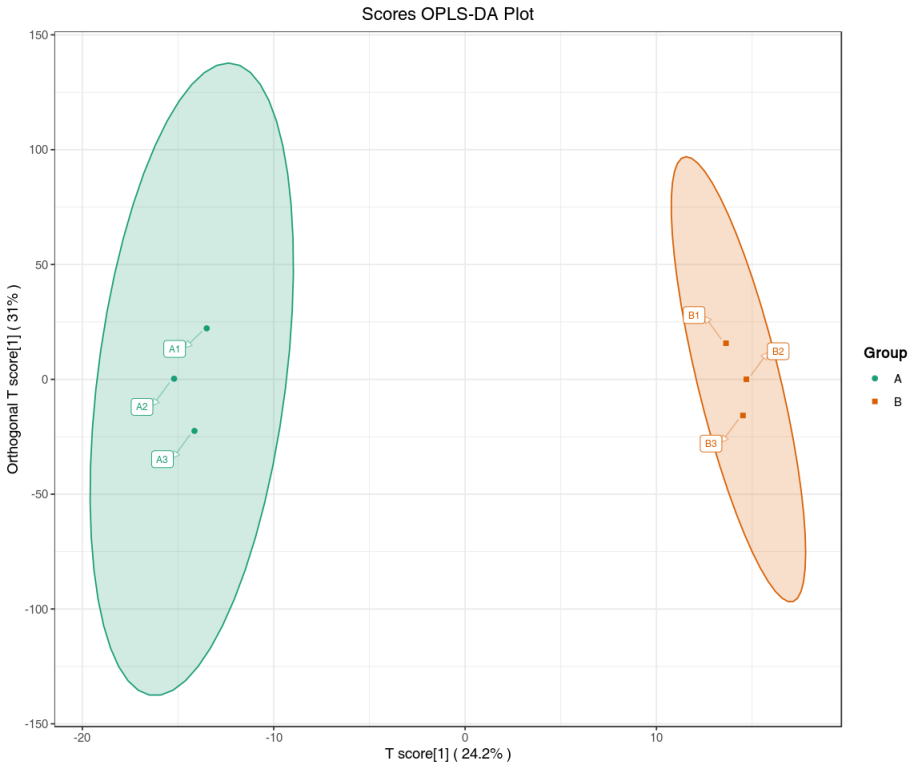


Fig 17: OPLS-DA score diagram

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.

OPLS-DA score diagram: Final report/2.Basic_Analysis/Difference_analysis/*_vs_*/opls/*_OPLS-DA_scorePlot.*

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 coX-axis indicating the model R^2Y, Q^2 values, and the vertical coY-axis is the frequency of the model classification effect. Bootstrapping on the model was

performed for 200 times and if Q^2 P = 0.02, it indicates that the predictability of four random grouping models is better than that of the OPLS-DA model in the Permutation detection. If R^2Y 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.

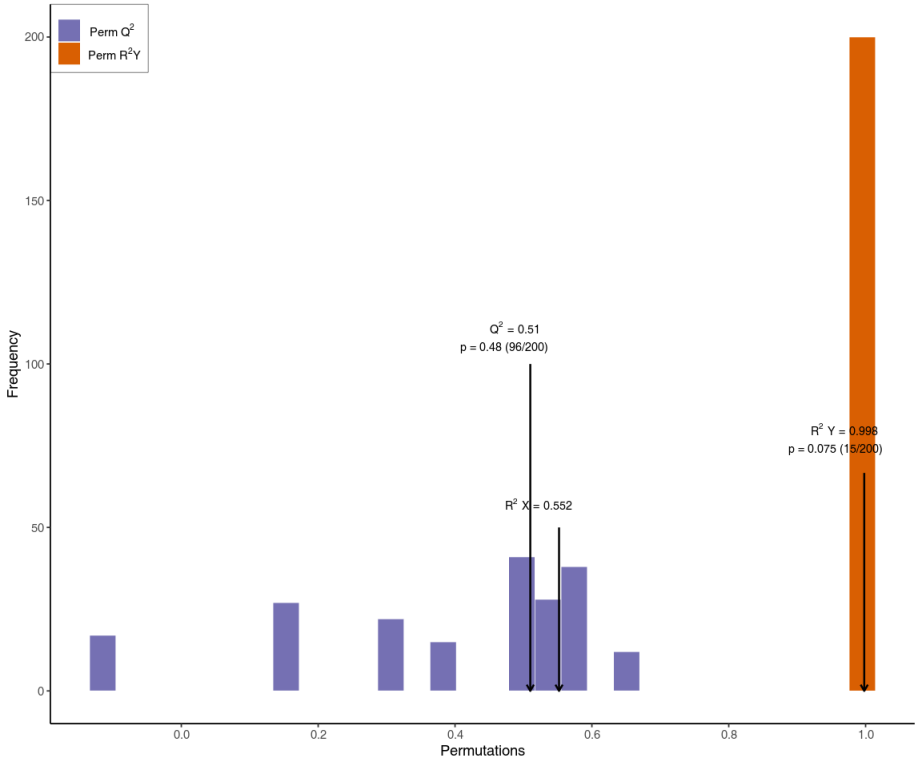


Fig 18: OPLS-DA verification diagram

Note: The X-axis represents the R^2Y and Q^2 values of the model, and the Y-axis is the frequency of the model classification effect in 200 random permutation and combination experiments. The orange in the figure represents the randomization model R^2Y, the purple represents the randomization model Q^2, and the values represented by the black arrows represent the R^2X, R^2Y and Q^2 values of the original model.

OPLS-DA verification diagram: Final report/2.Basic_Analysis/Difference_analysis/*_vs_*/opls/*_OPLS-DA_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.

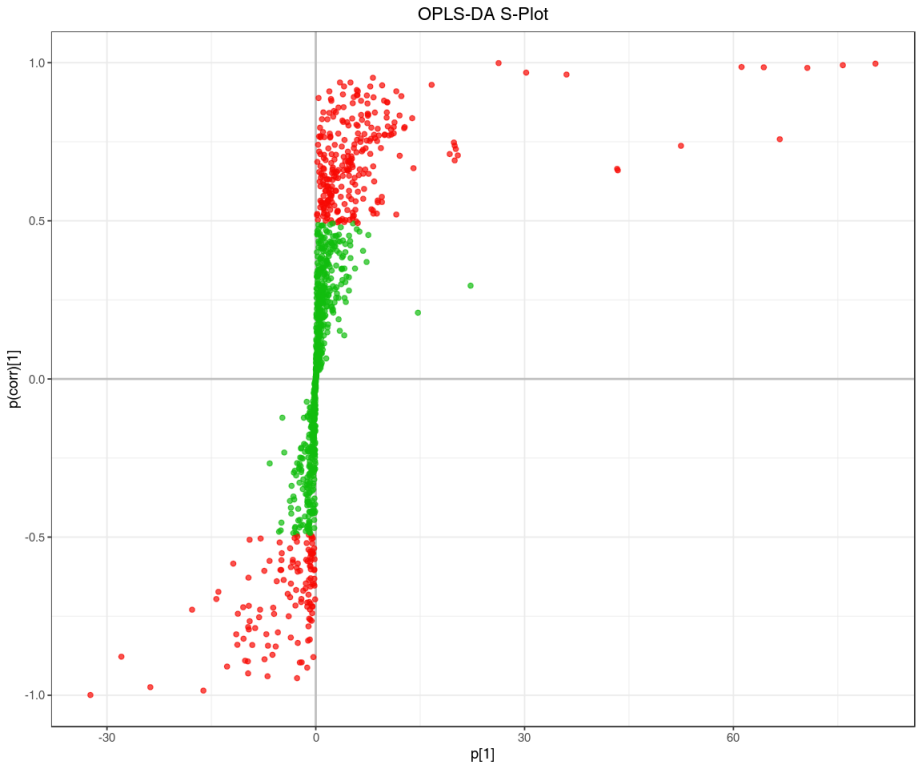


Fig 19: OPLS-DA S-plot

OPLS-DA S-plot: Final report/2.Basic_Analysis/Difference_analysis/*_vs_*/opls/*_OPLS-DA_SPlot.*

4.3 Dynamic distribution of metabolite content differences

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

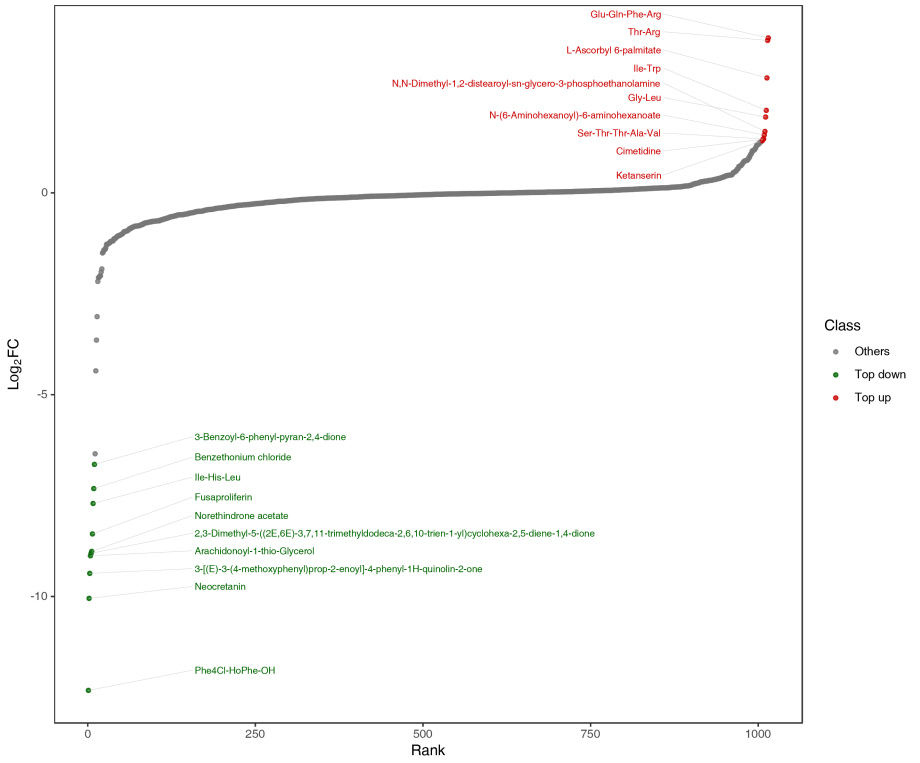


Fig 20: Dynamic distribution of metabolite content difference

Note: In the figure, the X-axis represents the rank number of metabolites based on FC value. The Y-axis represents the $\log_2$FC 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.

Dynamic distribution of metabolite content difference: Final report/2.Basic_Analysis/Difference_analysis/*_vs_*/TopFcMetabolites/*_TopFcDistribution_*.*

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 ≥ 3), multivariate analysis of Variable Importance in Projection (VIP) from OPLS-DA modeling was used to preliminarily select differential metabolites from different samples. Differential metabolites can further be screened by combining the P-value/FDR (when biological replicates ≥ 2) or FC values from univariate analysis. The screening criteria for this project are as follows:

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 are significantly difference.

2. Metabolites with P-value < 0.05 (Student's t test were used when the data follow a normal distribution, otherwise Wilcoxon rank-sum test) were considered as significant differences and selected.

Partial results from the screening criteria is shown below.

Table 9: Screening results of differential metabolites

Index	Compounds	Type
MADP0518	N3-(4-fluorophenyl)-1h-pyrazolo[3,4-d]pyrimidine-3,4-diamine	down
MEDN0105	Taurocholic acid	up
MEDN0159	Flavin Adenine Dinucleotide(FAD)	down
MEDN0280	Taurine	down
MEDN0366	LPE(16:0/0:0)	down
MEDN0368	LPE(14:0/0:0)	down
MEDN0434	B-Pseudouridine	down
MEDN0442	Pantetheine	down
MEDN0555	Hydroxyphenyllactic acid	down
MEDN1056	Iminodiacetic acid	down

Screening results of differential metabolites: Final report/2.Basic_Analysis/Difference_analysis/*_vs_*/*_filter.xlsx

4.4.1 Bar chart of differential metabolites

The following figure shows the result of top 20 differentially expressed metabolites in each comparison with fold-change value shown as $\log_2$ values.

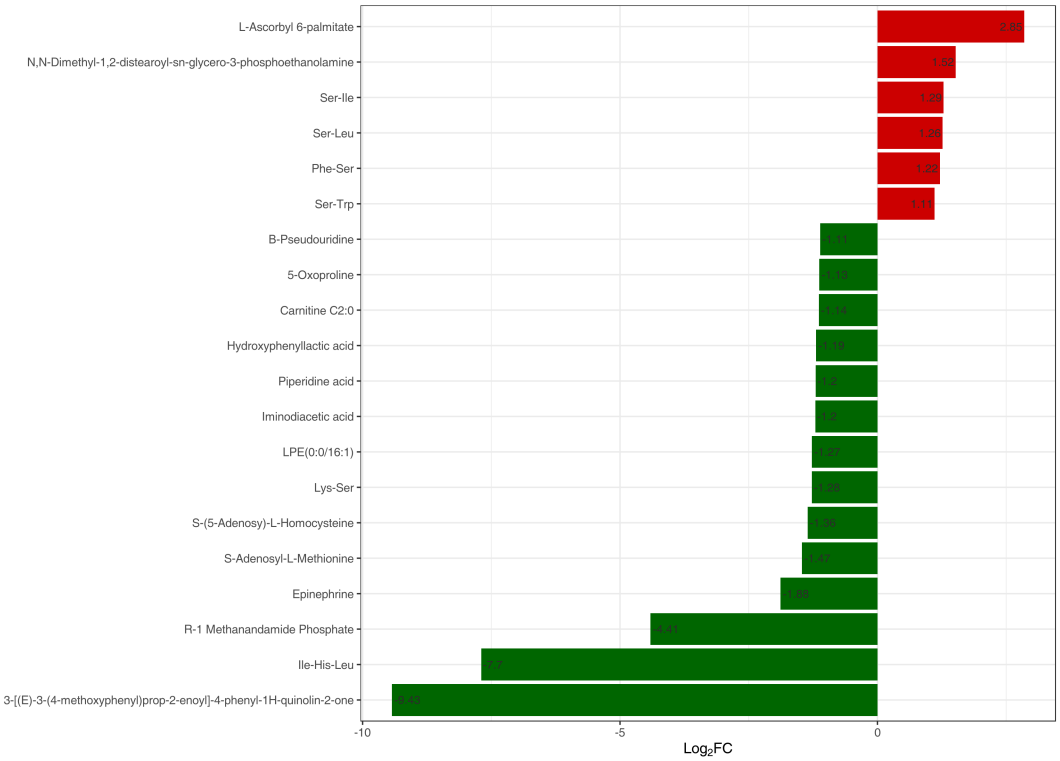


Fig 21: Bar chart of differential metabolites

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.

Histogram of multiple difference: Final report/2.Basic_Analysis/Difference_analysis/*_vs_*/TopFcMetabolites/*_TopFcBarChart_*.*

4.4.2 Differential metabolite radar map

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

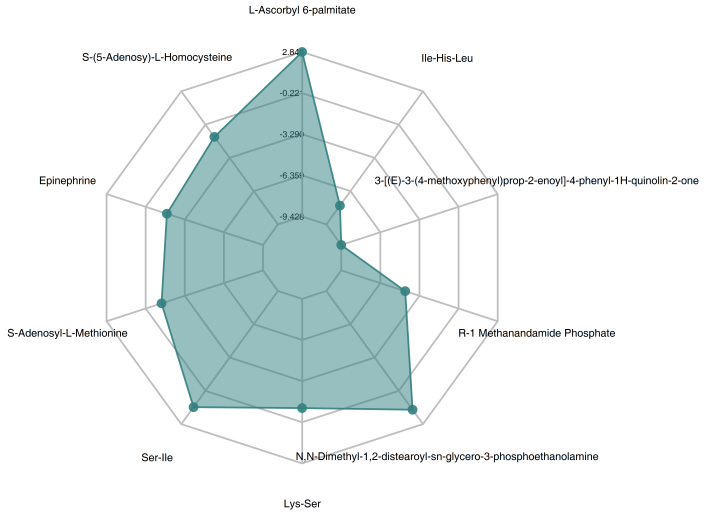


Fig 22: Radar map of differential metabolites

Note: The grid lines correspond to the $\log_2$FC, The green colored area are formed from the lines connecting the dots.

Radar map of differential metabolites: Final report/2.Basic_Analysis/Difference_analysis/*_vs_*/TopFcMetabolites/*_TopFcRadarChart_*.*

4.4.3 VIP value map of differential metabolites

The top 20 metabolites with the largest VIP value from the OPLS-DA model were selected and plotted.

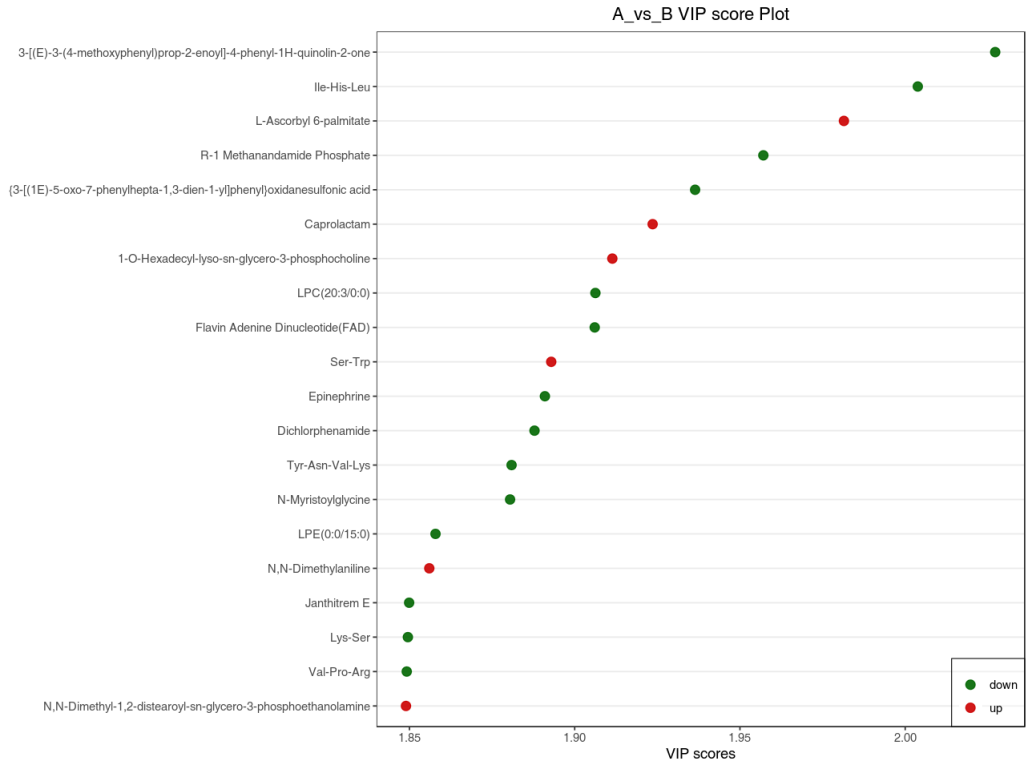


Fig 23: VIP values of differential metabolites

Note: The X-axis represents VIP values, and the Y-axis represents metabolites. Red dots represent up-regulated differential metabolites, and green dots represent down-regulated differential metabolites

VIP values of differential metabolites: Final report/2.Basic_Analysis/Difference_analysis/*_vs_*/ vip-score/*_vipScore.*

4.4.4 Volcano plot of differential metabolites

Volcano Plot is used to show the relative differences and the statistical significance of metabolites between two groups. We provided the volcano plot of differential metabolites 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 metabolite.

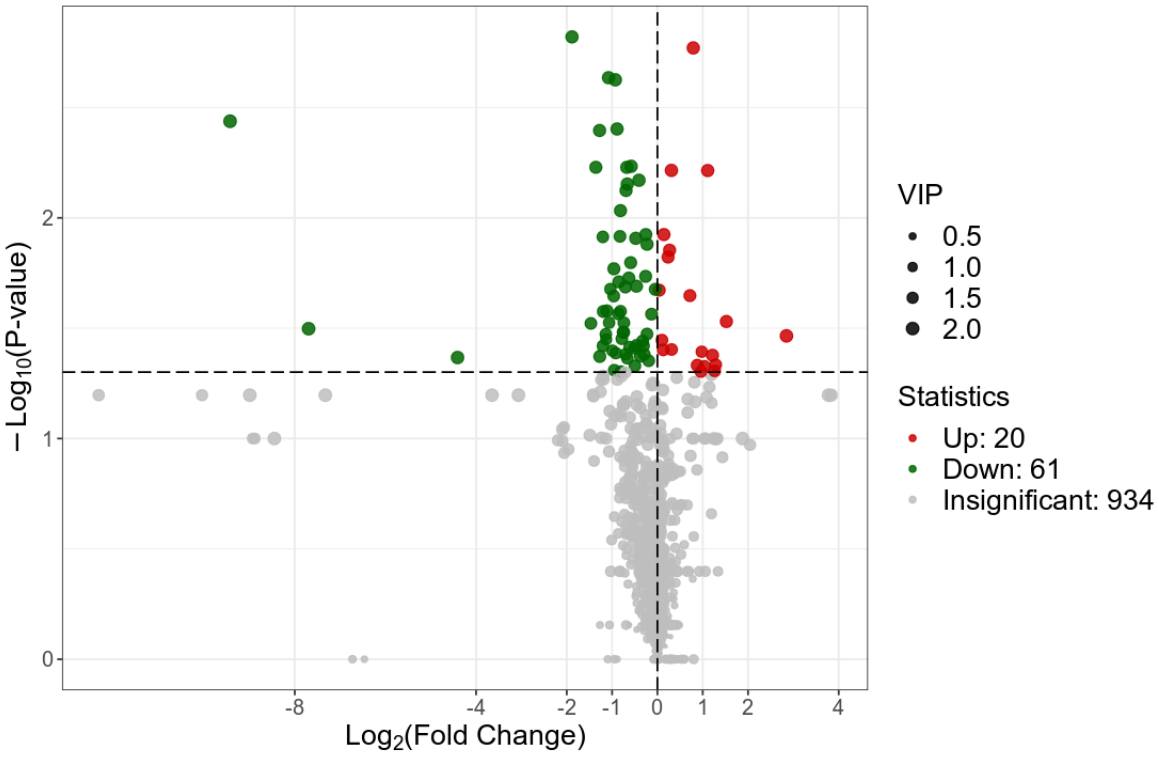


Fig 24: Volcano plot of differential metabolites

Note: Each point in the volcano plot represents a metabolite with green dots represent down-regulated differential metabolite, red dots represent up-regulated differential metabolite, and gray dots represent the detected metabolites but show no significant differences. The X-axis represents the (log2 FC) value of metabolites between two groups. The further away from 0 on the X-axis, the greater the fold-change between two groups. If the metabolites were screened using VIP + FC + P-value, the Y-axis will represent the the level of significant differences ($-\log_{10}$P-value). The size of each dot represents the VIP value.

Volcano maps of differential metabolites: Final report/2.Basic_Analysis/Difference_analysis/*_vs_*/vol/*_volcano_*.*

4.4.5 Scatter plot of differential metabolites

The differential metabolites scatter plot is used to show the abundance differences in compound sub-classes between two groups.

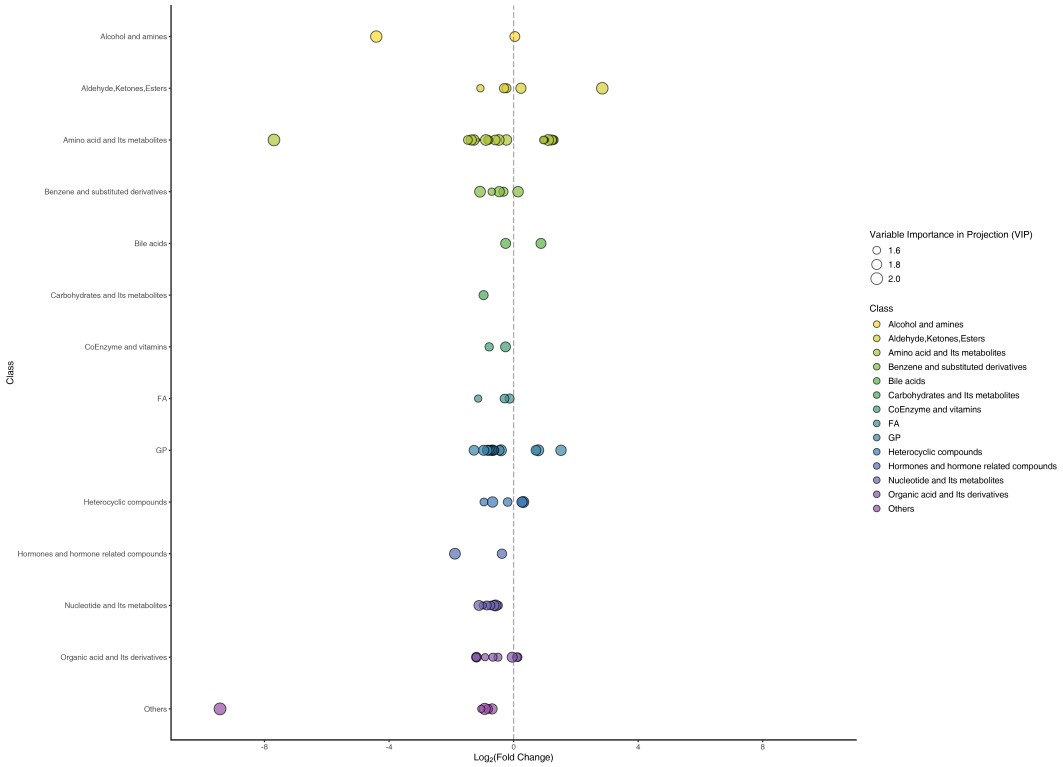


Fig 25: Scatter plot of differential metabolites

Note: Each dot in the graph indicates a metabolite, and different colors indicate different metabolite subclasses; the horizontal coordinate indicates the logarithmic value of the multiplicative difference in the content of a substance in two groups of samples (log_2FC), the larger the absolute value of the horizontal coordinate, the greater the difference in the content of the substance between the two groups of samples, and the size of the dot represents the VIP value.

Scatter plot of differential metabolites: Final report/2.Basic_Analysis/Difference_analysis/*_vs_*/Scatter/

4.4.6 Hierarchical clustering tree

Hierarchical clustering was performed on different sample groups to form a clustering tree showing the similarity between samples. Samples in the same cluster have higher similarity.

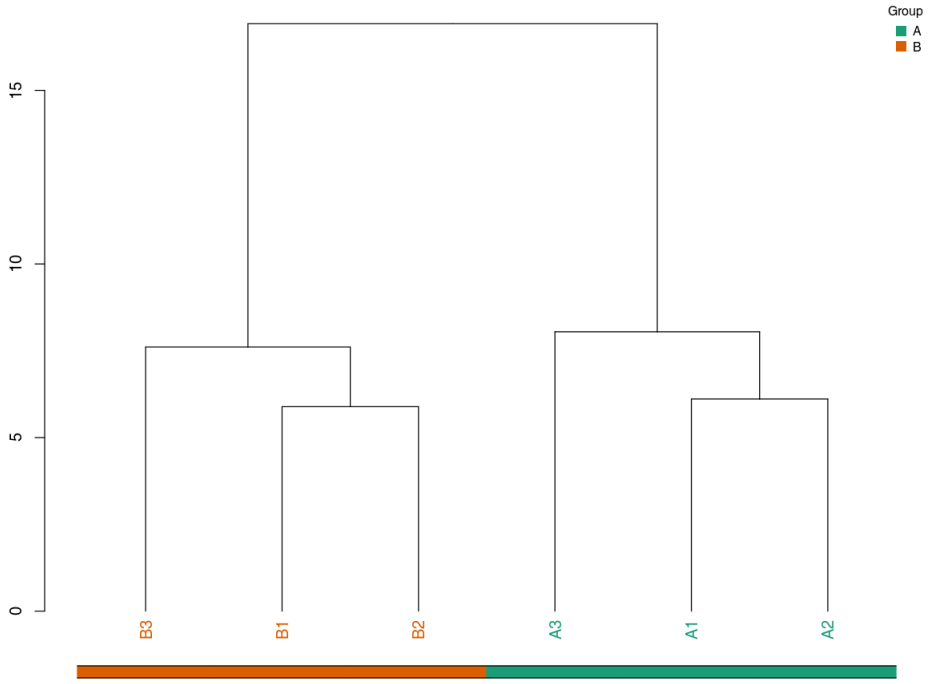


Fig 26: Hierarchical clustering tree

Note: Samples with higher similarity are clustered more closely on the clustering tree.

Hierarchical clustering tree: Final report/2.Basic_Analysis/Difference_analysis/*_vs_*/dendrogram/*_dendrogram.*

4.4.7 Heatmap of differential metabolites

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

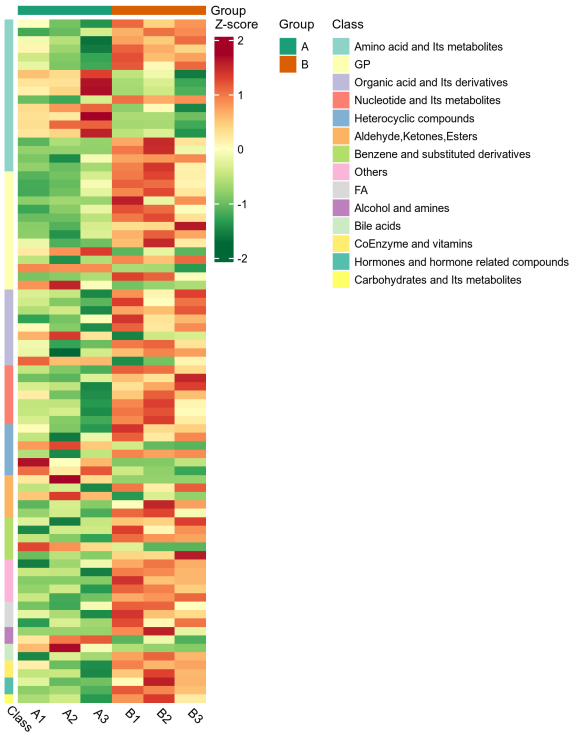


Fig 27: Heatmap of differential metabolites

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 UV scaling 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 or on the top. If classification was performed on the compounds, a colored bar will be shown on the left to depict Level 1 classifications. *_all_heatmap_class: Heatmap by metabolites classification, Class represents the first-level classification of metabolites. *_all_heatmap_colrow_cluster: clustering analysis is performed for both metabolites and samples, the clustering tree on the left side is the metabolite clustering tree, and the clustering tree on the top is the sample clustering tree. *_all_heatmap_row_cluster: clustering analysis is performed for metabolites only, the clustering tree on the left is the metabolite clustering tree.

Heatmap of differential metabolites: Final report//2.Basic_Analysis/ Difference_analysis/*_vs_*/heatmap/

4.4.8 Correlation analysis of differential metabolites

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

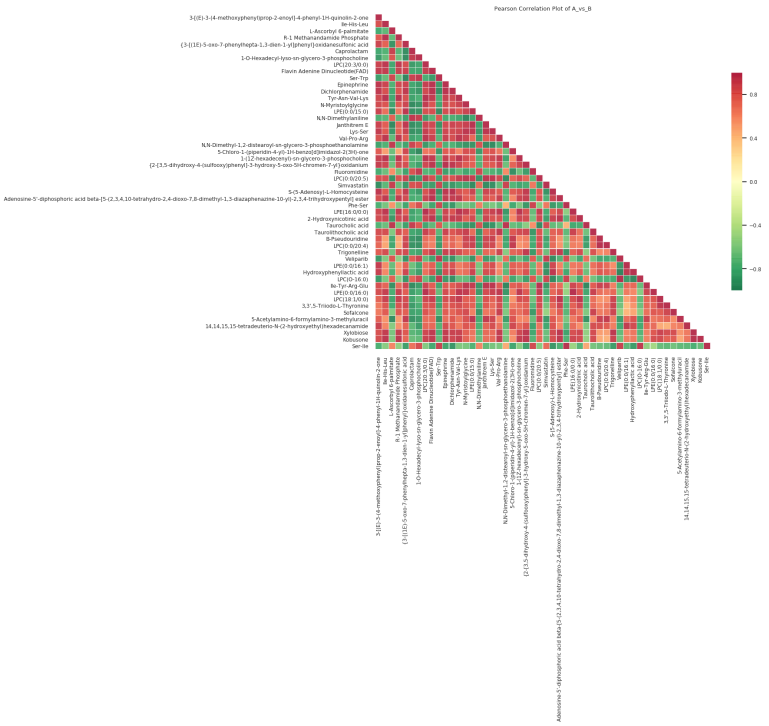


Fig 28: Heatmap of correlation of different metabolites

Note: The ID of the metabolites are shown on both horizontal and verticle axses. 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.

Heatmap of correlation of different metabolites: Final report/2.Basic_Analysis/Difference_analysis/*_vs_*/cpdCorr/*_cpdCorr_*.*

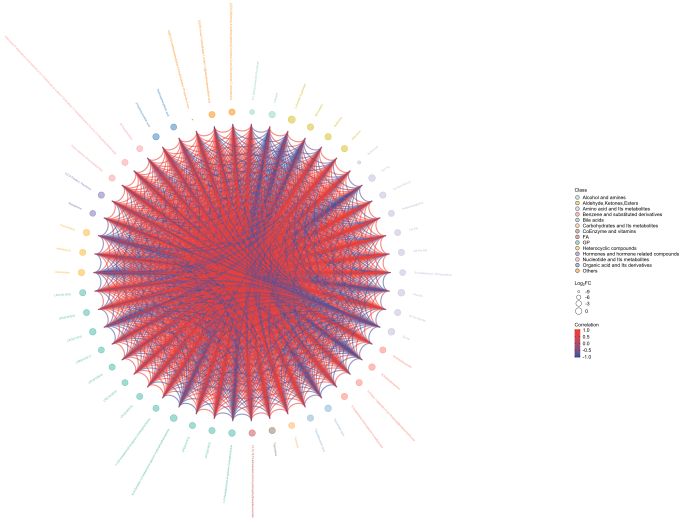


Fig 29: Chord diagram of differential metabolites

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 Class I 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. If there are more than 50 differential metabolites, the figure will only show the top 50 metabolites based on VIP.

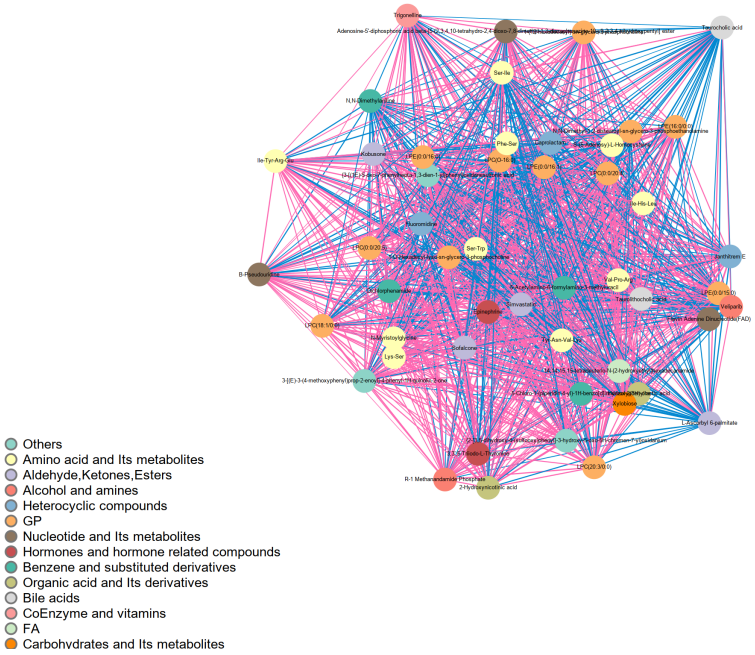


Fig 30: Correlation network diagram of differential metabolites

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 greater the degree of connection, the larger the point, i.e. the more points (neighbors) connected to it. Red lines represent positive correlations and blue lines represent negative correlations. Line thickness represent the absolute value of Pearson correlation coefficient. The larger the |r|, the thicker the line. If there are more than 50 differential metabolites, the figure will only show the top 50 metabolites based on VIP.

Correlation network diagram of differential metabolites: Final report/2.Basic_Analysis/Difference_analysis/*_vs_*/cpdCorr/*_cpdCorrNet_*.*

4.4.9 Z-value map of differential metabolites

Z-score standardization normalizes the relative content of the differential metabolites by calculating Z-scores. The Z-score is calculated by $z = (x - \mu) / \sigma$; Where x is a specific score, μ is the mean, and σ is the standard deviation. The Z-score plot provides a visual representation of the distribution of each differential metabolite across groups. The colored dots in the plot represent samples of different groups.

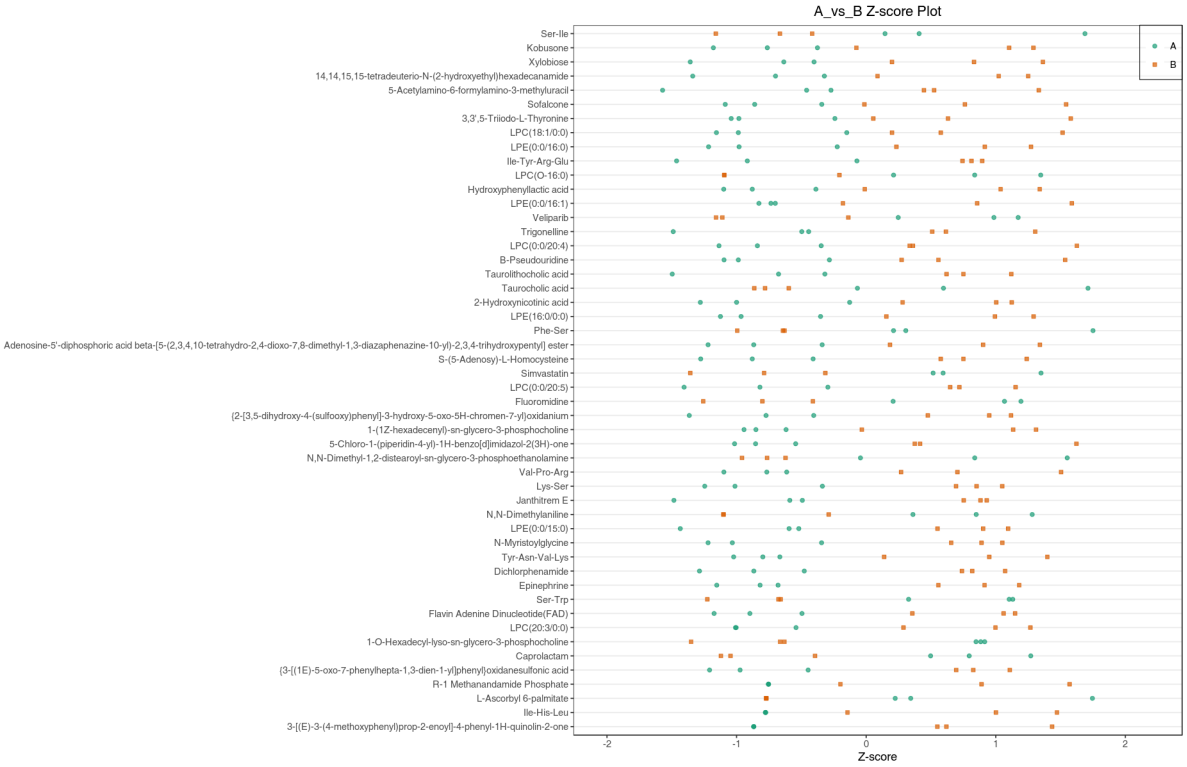


Fig 31: Z-value map of differential metabolites

Note: The X-axis represents the z-score and the Y-axis represents the differential metabolites. The colored dots in the plot represent samples of different groups. If there are more than 50 differential metabolites, the figure will only show the top 50 metabolites based on VIP.

Z-value map of differential metabolites: Final report/2.Basic_Analysis/Difference_analysis/*_vs_*/ zScore/*_zScore*.*

4.4.10 Violin plot of differential metabolites

A violin plot is a combination of a box plot and a density plot, mainly used to show the data distribution and its probability density. The box shape in the middle indicates the interquartile range, the thin black line extending from it represents the 95% confidence interval, the black horizontal line right in the middle is the median, and the outer shape indicates the density of the data distribution.

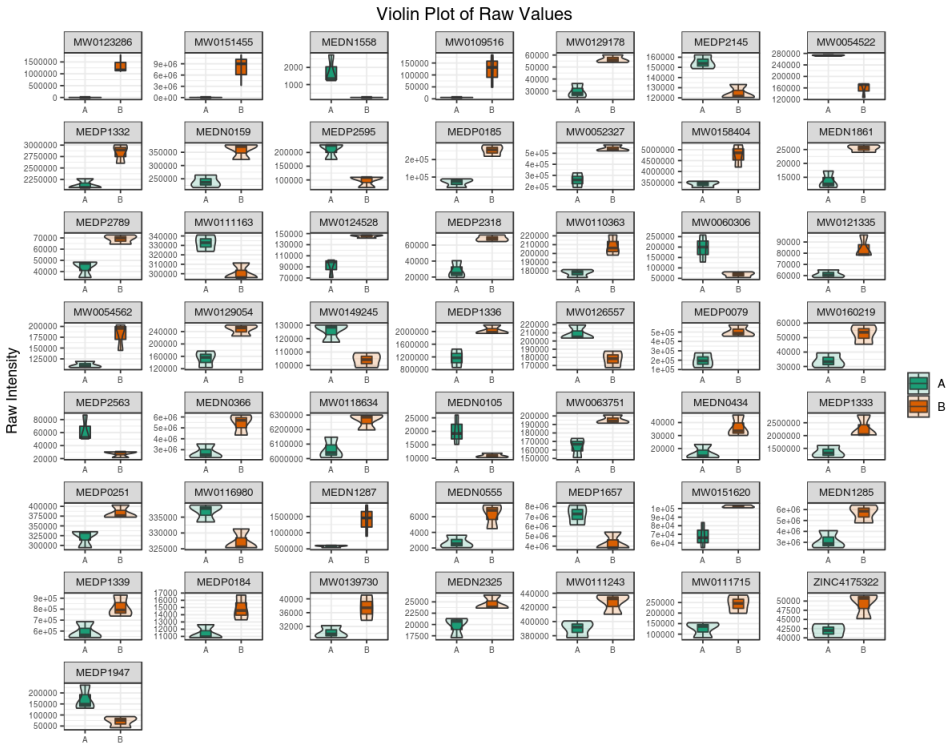


Fig 32: Violin plot of differential metabolites

Note: The horizontal coordinate is the grouping and the vertical coordinate is the relative content of the differential metabolites (raw peak area). If there are more than 50 differential metabolites, the figure will only show the top 50 metabolites based on VIP.

Violin plot of differential metabolites: Final report/2.Basic_Analysis/Difference_analysis/*_vs_*/fullViolin/*_fullViolin*.*

Violin plot of single metabolite: Final report/2.Basic_Analysis/Difference_analysis/*_vs_*/fullViolin/single

4.4.11 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.

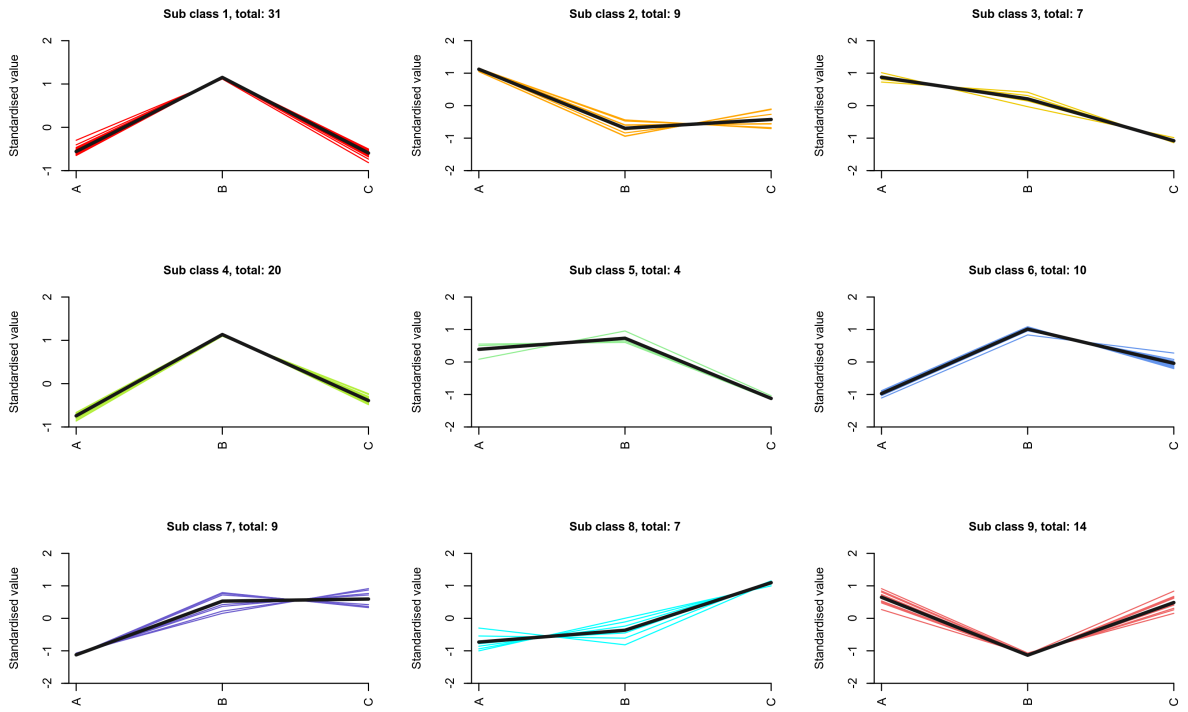


Fig 33: K-Means diagram of differential metabolites

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 ”total” represent the number of metabolites in this cluster.

K-Means diagram of differential metabolites: Final report/2.Basic_Analysis/kmeans/kmeans_cluster.*

4.4.12 Venn diagram of differential metabolites

Venn diagram is used to show the number of shared and unique metabolites in different comparison groups. A petal diagram is used for 5 groups or more.

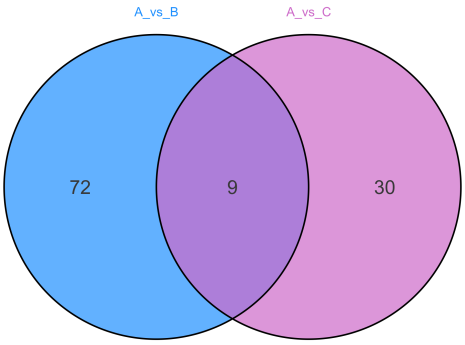


Fig 34: Venn diagram of differences among groups

Note: Each circle represents a comparison group, the number in overlapped parts represents the number of common differential metabolites between comparison groups, and the number in non-overlapped parts represents the number of unique differential metabolites in comparison groups.

Venn diagram of differential metabolites: Final report/2.Basic_Analysis/Venn

4.5 Functional annotation and enrichment analysis of differential metabolites with 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 its abundances as a complete network.

4.5.1 Functional annotation of metabolites

Metabolites are annotated using the KEGG database (Kanehisa et al., 2000), 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.

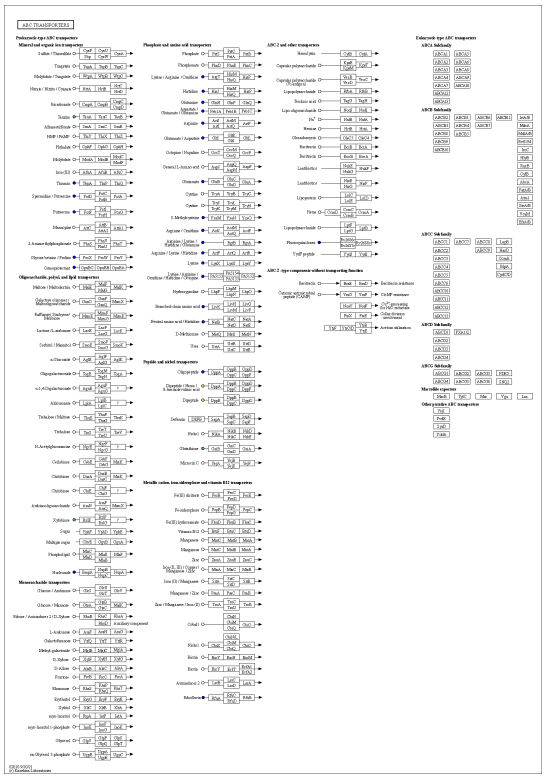


Fig 35: KEGG pathway with detected metabolites

Note: Red circles indicate that the metabolite content was significantly up-regulated in the experimental group; 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; and 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.

KEGG pathway of differential metabolites: Final report/2.Basic_Analysis/Difference_analysis/*_vs_*/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
MADP0518	N3-(4-fluorophenyl)-1h-pyrazolo[3,4-d]pyrimidine-3,4-diamine	down	C75450
MEDN0105	Taurocholic acid	up	C05122
MEDN0159	Flavin Adenine Dinucleotide(FAD)	down	C00016
MEDN0280	Taurine	down	C00245
MEDN0366	LPE(16:0/0:0)	down	C04438
MEDN0368	LPE(14:0/0:0)	down	C04438
MEDN0434	B-Pseudouridine	down	C02067
MEDN0442	Pantetheine	down	C00831
MEDN0555	Hydroxyphenyllactic acid	down	C03672
MEDN1056	Iminodiacetic acid	down	C19911

Table 11: Enrichment statistical of KEGG annotations for differential metabolites

ko_ID	Sig_compound	compound	Sig_compound_all	compound_all
ko00120	2	3	40	309
ko00430	2	4	40	309
ko01100	20	158	40	309
ko04976	3	12	40	309
ko04979	1	7	40	309
ko00740	1	2	40	309
ko01240	4	24	40	309
ko04977	1	16	40	309
ko00920	1	2	40	309
ko02010	11	73	40	309

KEGG annotations for differential metabolites: Final report/2.Basic_Analysis/Difference_analysis/*_vs_*/enrichment/*_filter_anno.xlsx

Enrichment statistical of KEGG annotations for differential metabolites: Final report/2.Basic_Analysis/Difference_analysis/*_vs_*/enrichment/*_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:

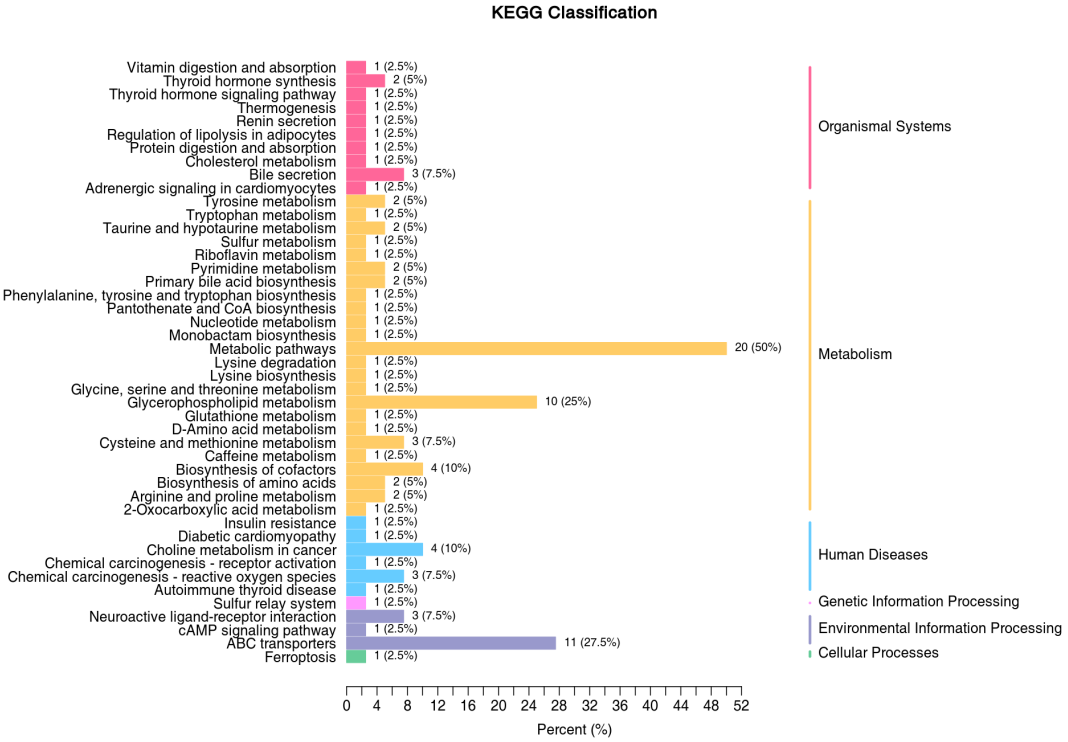


Fig 36: KEGG classification of differential metabolites

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

KEGG classification of differential metabolites: Final report/2.Basic_Analysis/Difference_analysis/*_vs_*/enrichment/*KEGG_barplot.*

4.5.3 Hierarchical Cluster Analysis of differential metabolites in KEGG pathway

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

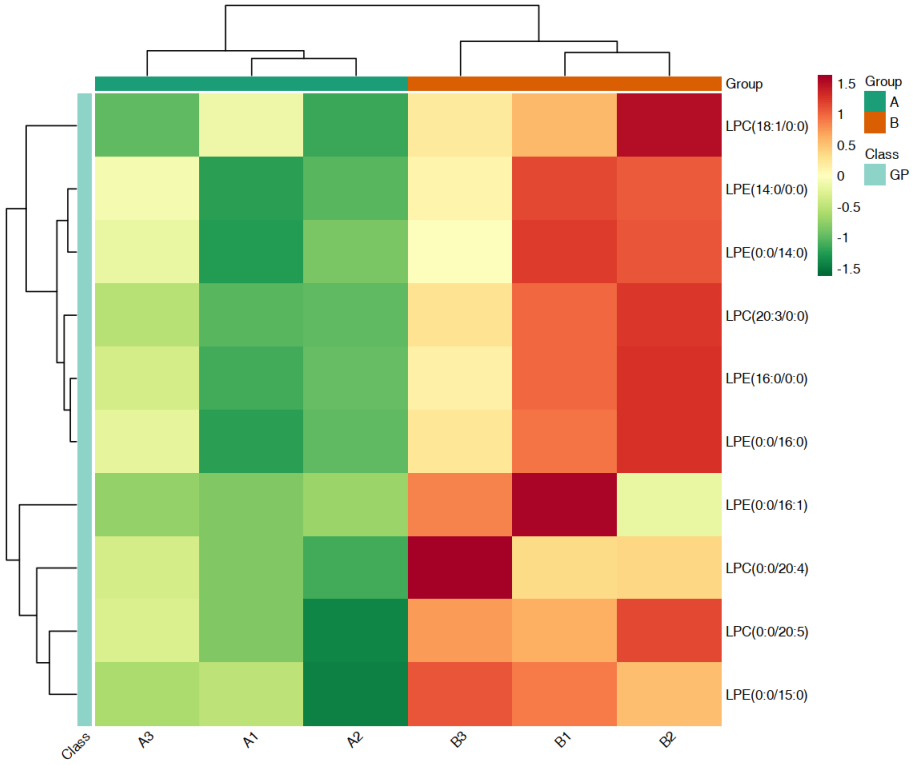


Fig 37: Clustering heat map of differential metabolites in KEGG pathway

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 compound classifications.

Clustering heat map of differential metabolites in KEGG pathway: Final report/2.Basic_Analysis/ Difference_analysis/*_vs_*/enrichment/KEGG_heatmap/*_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 differential metabolites in the corresponding pathway to the total number of metabolites annotated in the same pathway. The greater the value, the greater the degree of enrichment. P-value is 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 is 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 top 20 pathways in terms of P-value are plotted.

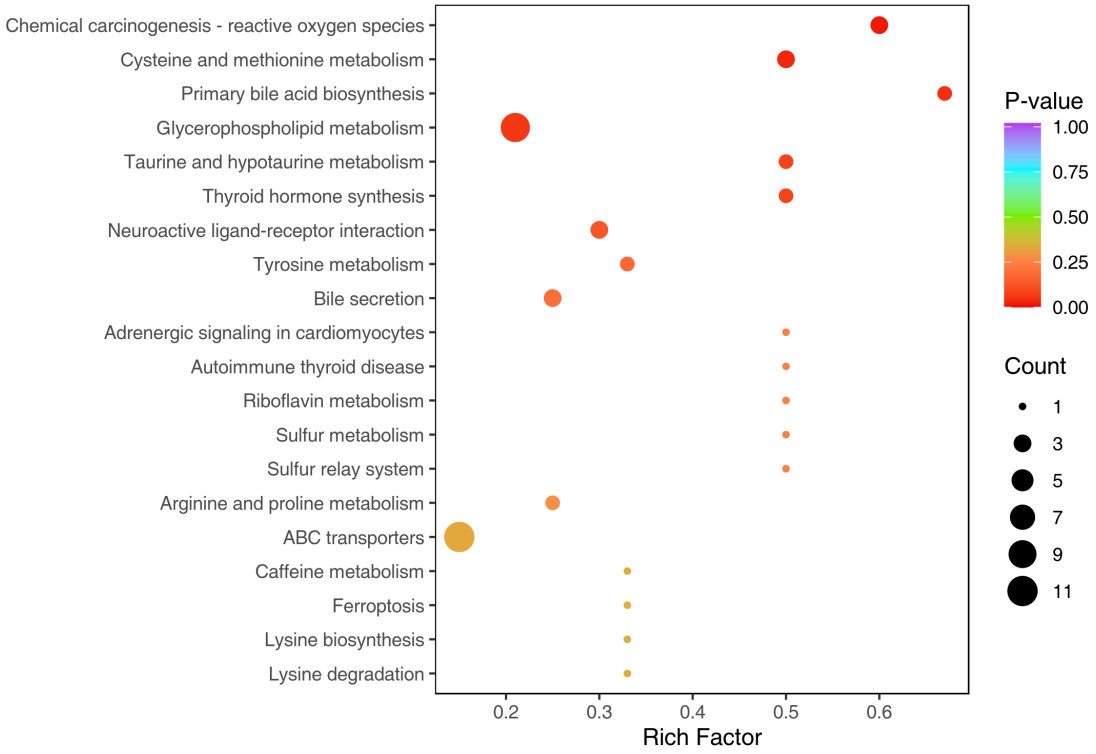


Fig 38: KEGG enrichment diagram of differential metabolites

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.

KEGG enrichment diagram of differential metabolites: Final report/2.Basic_Analysis/Difference_analysis/*_vs_*/enrichment/*_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 formula:

$$\text{DA score} = \frac{\text{up regulated metabolites in a pathway} - \text{down regulated metabolites in a pathway}}{\text{Total number of metabolites annotation in a pathway}}$$

The top 20 pathways in terms of P-value are plotted.

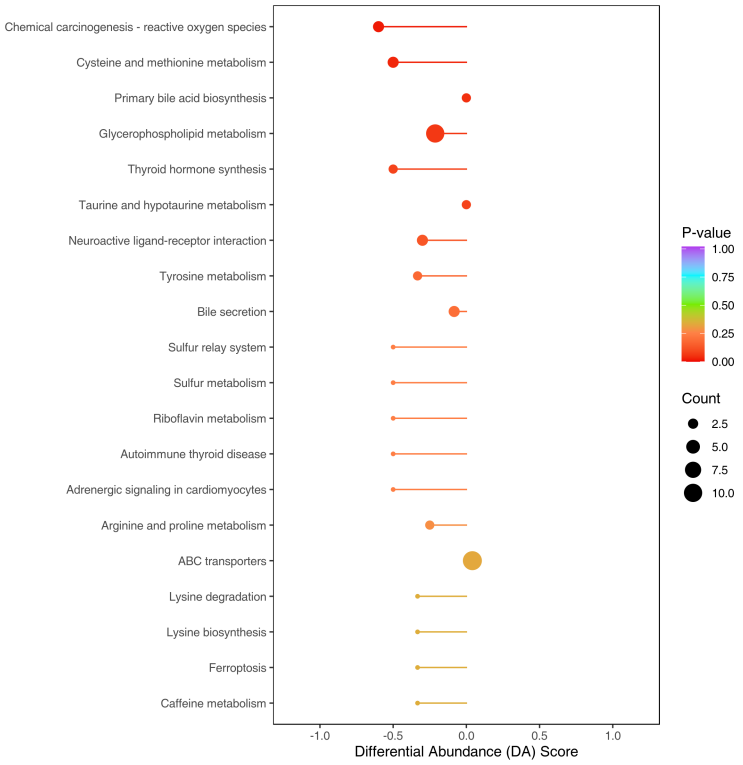


Fig 39: Difference abundance score

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 down-regulated; a dot on the right of the line represents the pathway is up-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.

Difference abundance score: Final report/2.Basic_Analysis/Difference_analysis/*_vs_*/enrichment/*_KEGG_DA_score.*

The table of difference abundance score: Final report/2.Basic_Analysis/Difference_analysis/*_vs_*/enrichment/*_KEGG_DA_score.xlsx

4.5.6 Regulatory network of differential metabolites

Differential metabolites and their corresponding KEGG pathways were used to generate a regulatory interactions network. This analysis is only done for project with samples sourced from human, mouse or rat species.

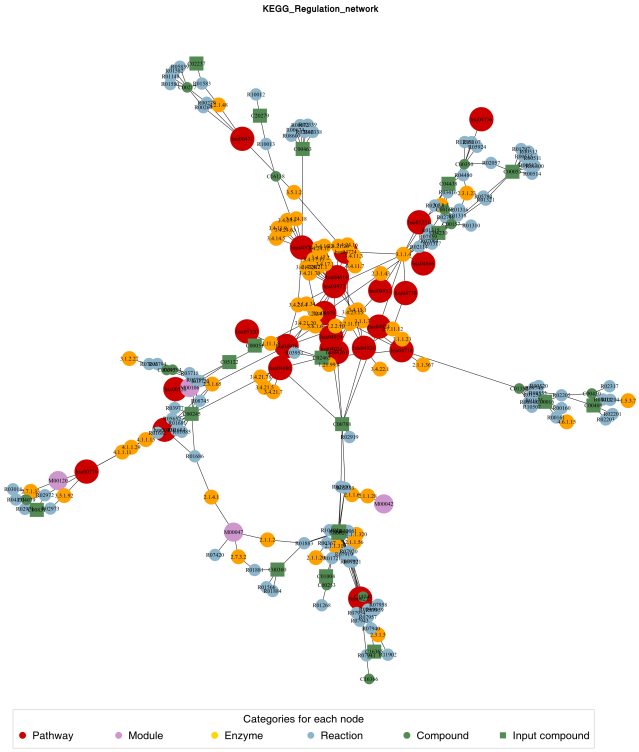


Fig 40: Diagram of the regulatory network of differential metabolites

Note: Red dots represent a metabolic pathway, yellow dots represent a substance-related regulatory enzyme, green dots represent a background substance for a metabolic pathway, purple dots represent a class of substance molecular modules, blue dots represent a substance chemical interaction reaction, and green squares represent the differential metabolites obtained in this comparison.

Diagram of the regulatory network of differential metabolites: Final report/2.Basic_Analysis/Difference_analysis/*_vs_*/enrichment/*_regulation_network.*

4.6 Functional annotation and enrichment analysis with HMDB database

4.6.1 Functional annotation and enrichment analysis of differential metabolites with 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 but 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
SMP0000170	0.00506307679282025
SMP0000497	0.00506307679282025
SMP0000012	0.00506307679282025
SMP0000494	0.00685945992443759
SMP0000218	0.00685945992443759
SMP0000533	0.00685945992443759
SMP0000498	0.00685945992443759
SMP0000006	0.00685945992443759
SMP0000169	0.00685945992443759
SMP0000190	0.00685945992443759

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:

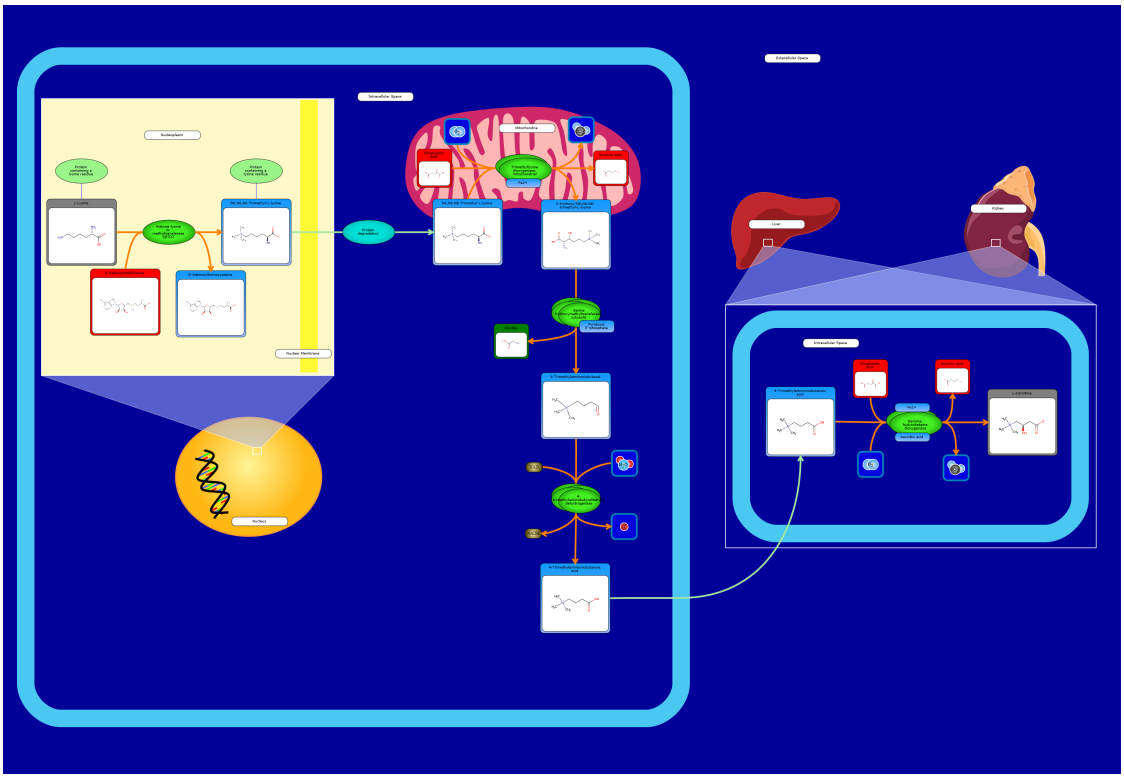


Fig 41: HMDB pathway map of differential metabolites

Note: Boxes with chemical structural formulas represent metabolites, 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:

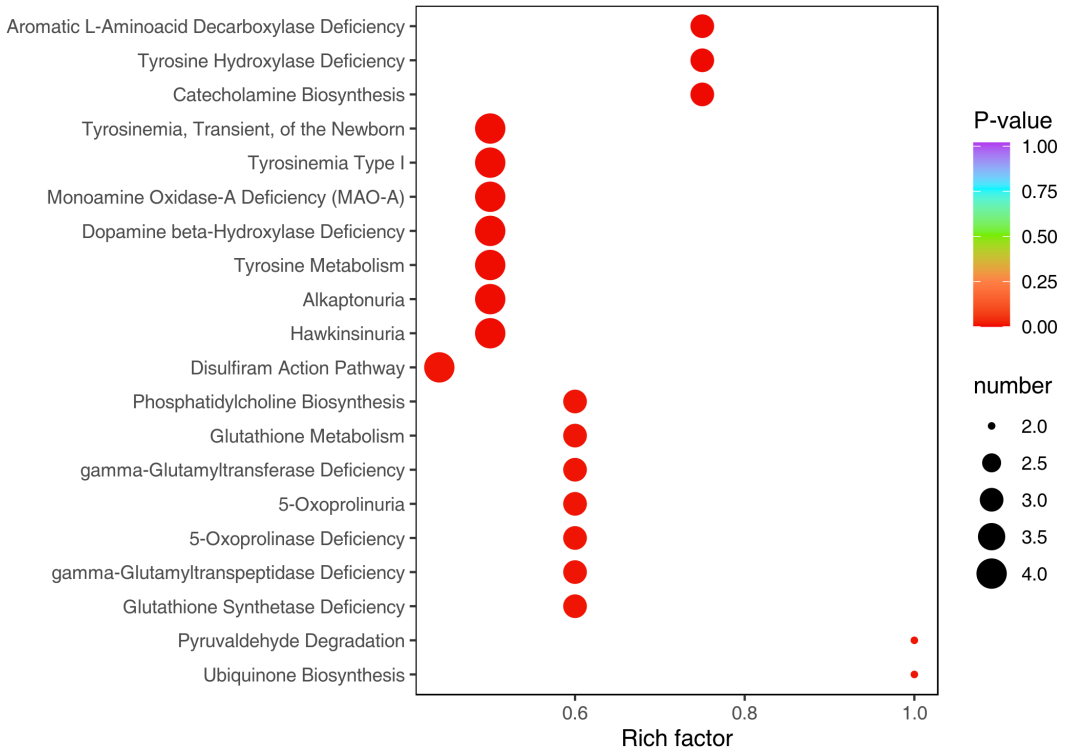


Fig 42: HMDB enrichment diagram of differential metabolites

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.

SMPDB pathway enrichment for differential metabolites: Final report/2.Basic_Analysis/ Difference_analysis/*_vs_*/enrichment/*_SMPDB_primary.xlsx

HMDB pathway map of differential metabolites: Final report/2.Basic_Analysis/Difference_analysis/*_vs_*/enrichment/SMP_primary_pathway

HMDB enrichment diagram of differential metabolites: Final report/2.Basic_Analysis/Difference_ analysis/*_vs_*/enrichment*SMPDB_primary_Enrichment.*

4.7 MSEA enrichment analysis

Conventional enrichment analysis based on hypergeometric distribution relies on up- or down-regulated metabolites and tends to miss metabolites that are not significantly different but are biologically important. Metabolite set enrichment analysis (MSEA) does not require specifying a clear threshold for differential

metabolites. The idea is to establish a series of metabolite sets, each representing a certain biological function, and identify metabolite sets that are significantly different.

Metabolite database from MetaboAnalyst (https://www.metaboanalyst.ca/) includes: (1) human metabolic pathways based on those found in the KEGG database: 84 KEGG pathway metabolic sets (kegg_pathway). (2) biologically significant disease-related metabolic sets for specific biological fluids: 339 blood metabolic sets, 384 urine metabolic sets, and 150 cerebrospinal fluid metabolic sets (csf). The results of the analysis were as follows:

Table 13: Table for MSEA enrichment analysis

name	P-value	foldEnrichment
Primary bile acid biosynthesis	0.037179	3.12545
Taurine and hypotaurine metabolism	0.041038	3.44195
Arginine and proline metabolism	0.048941	3.21885
Glycine, serine and threonine metabolism	0.069600	2.13150
Thiamine metabolism	0.079239	2.89100
Steroid biosynthesis	0.095346	2.67530
Pantothenate and CoA biosynthesis	0.124380	2.38165
Tyrosine metabolism	0.146940	2.23195
Purine metabolism	0.169330	1.59445
Tryptophan metabolism	0.250970	1.55090

The top 50 metabolic sets based on P-value ranking are shown below:

Fig 43: MSEA enrichment analysis graph

Note: The vertical coordinate indicates the name of the metabolic set (sorted by P-value), corresponding to the P-value of the labeled metabolic set; the horizontal coordinate indicates Fold Enrichment, the degree of enrichment; the color indicates P-value, the closer the P-value is to 0, the redder the color is, the more significant the enrichment is.

Table for MSEA enrichment analysis: Final report/2.Basic_Analysis/Difference_analysis/*_vs_*/enrichment/*_msea.xlsx

MSEA enrichment analysis graph: Final report/2.Basic_Analysis/Difference_analysis/*_vs_*/enrichment/*_msea.*

4.8 Diseases association with differential metabolites

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

Table 14: Table of association between differential metabolites and diseases

CompoundName	HmdbDiseases
N3-(4-fluorophenyl)-1h-pyrazolo[3,4-d]pyrimidine-3,4-diamine	-
Taurocholic acid	Hepatocellular carcinoma | Cirrhosis | Colorectal cancer | Crohn's disease | Ulcerative colitis | Metastatic melanoma | Biliary atresia
Flavin Adenine Dinucleotide(FAD)	Anorexia nervosa | Colorectal cancer
Taurine	Heart failure | Sulfite oxidase deficiency, ISOLATED | Epilepsy | Parkinson's disease | Leukemia | Schizophrenia | Irritable bowel syndrome | Ulcerative colitis | Colorectal cancer | Crohn's disease | Gout | Rheumatoid arthritis | Perillyl alcohol administration for cancer treatment | Pancreatic cancer | Periodontal disease | Lung Cancer | Autosomal dominant polycystic kidney disease | Propionic acidemia | Maple syrup urine disease | Eosinophilic esophagitis | Molybdenum cofactor deficiency
LPE(16:0/0:0)	-
LPE(14:0/0:0)	Ulcerative colitis | Iron deficiency
B-Pseudouridine	Canavan disease | Uremia | Colorectal cancer
Pantetheine	-
Hydroxyphenyllactic acid	Colorectal cancer | Supragingival Plaque | Phenylketonuria | Eosinophilic esophagitis
Iminodiacetic acid	-

Table of association between differential metabolites and diseases: Final report/2.Basic_Analysis/ Difference_analysis/*_vs_*/enrichment/*_sigDiseasesTable.xlsx

5 Reference

1. Chen W, Gong L, Guo Z, et al. A Novel Integrated Method for Large-Scale Detection, Identification, and Quantification of Widely Targeted Metabolites: Application in the Study of Rice Metabolomics[J]. Molecular Plant, 2013, 6(6):1769-1780. http://www.cell.com/molecular-plant/fulltext/S1674-2052(14)60263-X
2. Fraga, C.G., et al., Signature-discovery approach for sample matching of a nerve-agent precursor using liquid chromatography-mass spectrometry, XCMS, and chemometrics. Anal Chem, 2010. 82(10): p. 4165-73. http://pubs.acs.org/doi/abs/10.1021/ac1003568
3. Shen, X., Wang, R., Xiong, X. et al. Metabolic reaction network-based recursive metabolite annotation for untargeted metabolomics. Nat Commun, 2019. 10:1516. https://doi.org/10.1038/s41467-019-09550-x
4. L. Eriksson, E.J., N. Kettaneh-Wold, J.Trygg, C. Wikström, and S. Wold, Multi- and Megavariate Data Analysis Part I Basic Principles and Applications, Second edition Umetrics Academy:Sweden, 2006.

https://www.researchgate.net/publication/285755118_Multi-_and_Megavariate_Data_Analysis_Part_I_Basic_Principles_and_Applications_Second_revised_and_enlarged_edition

5. Chen, Y., et al., RRLC-MS/MS-based metabonomics combined with in-depth analysis of metabolic correlation network: finding potential biomarkers for breast cancer. Analyst, 2009.134(10): p. 2003-11. http://dx.doi.org/10.1039/b907243h
6. Thévenot E A, Roux A, Xu Y, et al. 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]. Journal of Proteome Research, 2015, 14(8):3322-35. https://dx.doi.org/10.1021/acs.jproteome.5b00354
7. Kanehisa, M. and S. Goto, KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res, 2000. 28(1): p. 27-30. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC102409/
8. Chong, J. and Xia, J., MetaboAnalystR: an R package for flexible and reproducible analysis of metabolomics data. Bioinformatics, bty528. https://doi.org/10.1093/bioinformatics/bty528
9. Viant M R, Kurland I J, Jones M R, et al. How close are we to complete annotation of metabolomes?[J]. Current opinion in chemical biology, 2017, 36: 64-69. https://www.sciencedirect.com/science/article/pii/S1367593117300054
10. Liang L, Rasmussen M L H, Piening B, et al. Metabolic dynamics and prediction of gestational age and time to delivery in pregnant women[J]. Cell, 2020, 181(7): 1680-1692. e15. https://www.sciencedirect.com/science/article/pii/S009286742030564X

6 Appendix

6.1 Software list and version

Table 15: Software used

Analysis	Software	Version	Method
KNN	R (impute)	1.56.0	default parameters
PCA	R (base package)	4.1.2	UV (unit variance scaling)
Heatmap	R (ComplexHeatmap)	2.9.4	UV (unit variance scaling)
Pearson Correlation	R (base package)	4.1.2	-
Correlation plot	R (corrplot)	0.92	-
OPLS-DA	R (MetaboAnalystR)	1.0.1	log2 + mean centering
Radar plot	R (fmsb)	0.7.1	-
Chord diagram	R (igraph; ggraph)	1.2.11; 2.0.5	-
Network diagram	R (igraph)	1.2.11	-
K-Means	R (base package)	4.1.2	UV (unit variance scaling)

In all the analyses of this project, two main approaches were taken to pre-process the data, which were calculated as follows:

(1) Unit variance scaling (UV)

Unit variance scaling (UV), also known as Z-score normalization / auto scaling, is a method of normalizing data based on the mean and standard deviation of the original data. The processed data conforms to a standard normal distribution with a mean of 0 and a standard deviation of 1.

Calculation method:

Original data centering divided by the standard deviation of the variable.

The formula is as follows:

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

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

(2) Zero-centered (Ctr)

Calculation method:

Original data minus the mean value of the variable.

The formula is as follows:

$$x' = x - \mu$$