METWARE BIO

Innovative Metabolomics Insights for Better Health

Demo Untargeted Metabolomics Report

Metware Biotechnology Inc.

www.metwarebio.com

Contents

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.

Untargeted metabolomics is a common approach to metabolomics research. The main idea is to perform qualitative and quantitative analysis, and identify statistically significant differential metabolites between different groups.

（1）For this project, 20 samples were selected and divided into 4 groups for metabolomics study.

Table 1: Number of identified metabolites

-	All	T3_positive	T3_negative
Number of metabolites identified	4860	3166	1694
Number of secondary metabolites identified	3876	2353	1523

Number of identified metabolites:Final_report/1.Data_Assess/metabolitesCount.xlsx

（2）The composition of metabolites is dependent on the sample that changes with experimental conditions. The metabolite composition ratio analysis examines the distribution of major metabolites in the samples. The following ring diagram shows the proportion of each metabolite class:

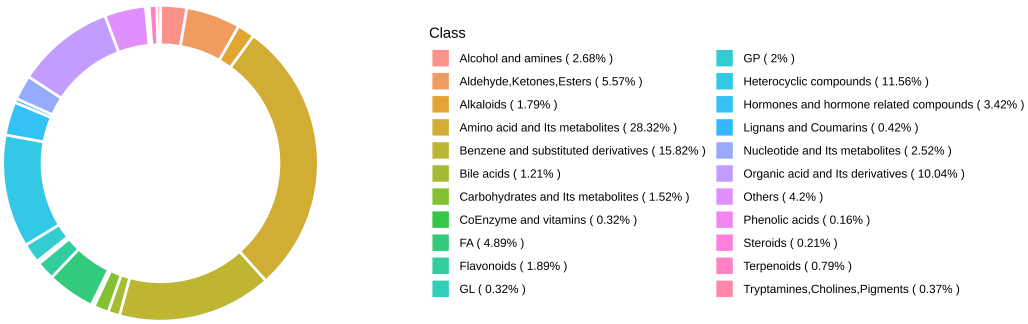


Fig 1: Ring diagram of metabolite categories

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

Ring diagram of metabolite categories:Final_report/1.Data_Assess/*/Class_Count/*_Class_Count_Ring.*

（3）Results of differential metabolite analysis:

Table 2: Number of differential metabolites

group	total	down	up
C_vs_A	1068	681	387
D_vs_B	1455	905	550

Number of differential 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:

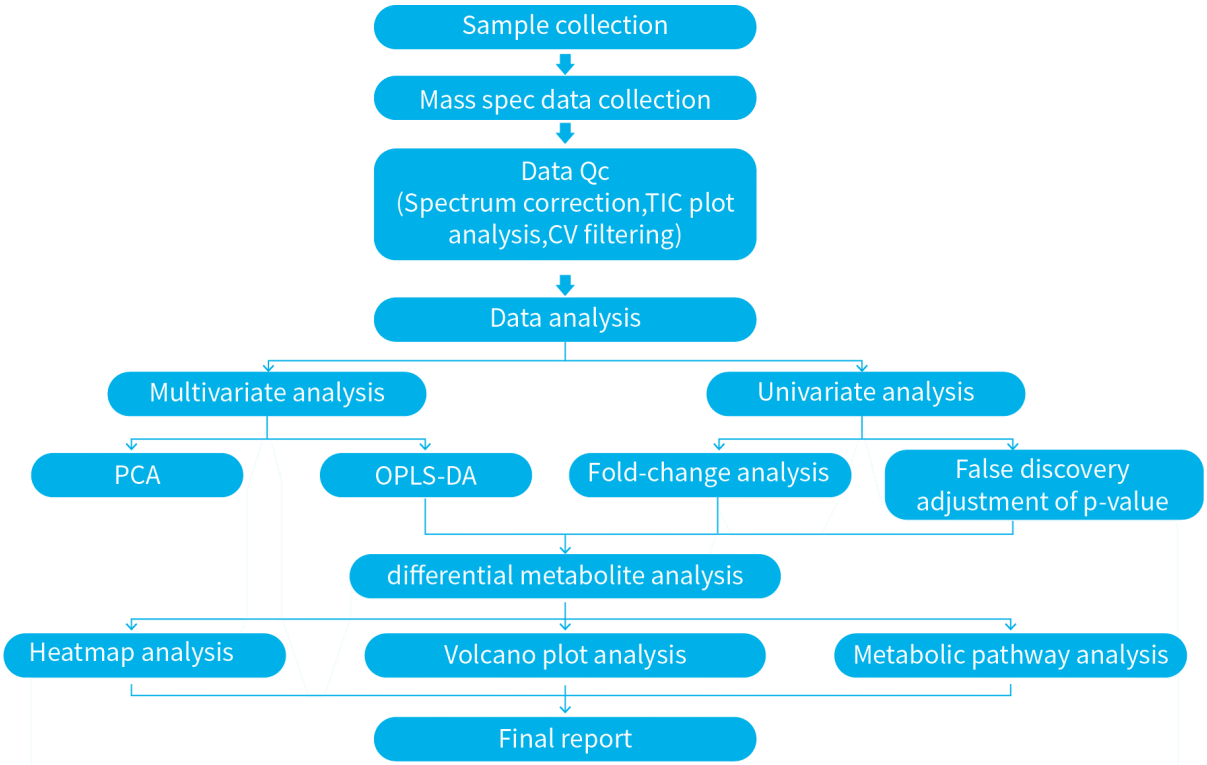


Fig 2: Flow chart of metabolomics analysis

2.1 Sample information

Each sample group and corresponding sample information are as follows:

Table 3: Table of sample information

Species	Tissue	Sample	Group
-	-	A1	A
-	-	A2	A
-	-	A3	A
-	-	A4	A
-	-	B1	B
-	-	B2	B
-	-	B3	B
-	-	B4	B
-	-	C1	C
-	-	C2	C
-	-	C3	C
-	-	C4	C
-	-	C5	C
-	-	C6	C
-	-	D1	D

Table of sample information:Final report/1.Data_Assess/sample_info.xlsx

2.2 Reagents and instruments

Table 4: Information of reagents

Compound	CAS	Purity	Brand	Item No.
Methanol	67-56-1	more than 99.9%	Thermo Fisher	A452-4
Acetonitrile	75-05-8	99.95%	Thermo Fisher	T001014000
Acetic Acid	64-19-7	more than 99.7%	Thermo Fisher	A35500
Ammonium formate	540-69-2	more than 99.0%	Sigma	70221
Ammonium hydroxide solution(25%)	1336-21-6	-	Sigma	543830
Formic Acid	64-18-6	97.5-98.5%	Sigma	00940
Standard	-	more than 98%	isoreag/TRC/TCI/Sigma	-

Table 5: Information of instruments

Name	Instrument	Brand	Address
Mass Spectrometer	TripleTOF 6600+	SCIEX	California, USA
UHPLC	ExionLC AD	SCIEX	California, USA
Centrifuge	5430R	Eppendorf	Hamburg, Germany
Vortex mixer	MI0101002	HFour E's	Guangzhou, China
Electronic balance(1/100000)	AS60/220.R2 Plus	Radwag	Poland
Centrifugal concentrator	CentriVap	LABCONCO	Missouri Kansas, USA
Ultrasonic cleaner	CD-F15	Olenyer	China
Pipette	Research plus	Eppendorf	Hamburg,Germany
Automatic workstation	Biomek i5	Beckman Coulter	California, USA
Uniform-Seal Heat Sealer	FM 5200	foodsaver	California, USA

2.3 Sample extraction process

2.3.1 Solid samples

Samples stored at -80 °C was thawed on ice and homogenized in a ball-mill grinder at 30 Hz for 20 s. 150 µL solution (Methanol : Water = 7:3, V/V) containing internal standard was mixed with the ground sample and mixed in a shaker at 2500 rpm for 5 min. The mixture was placed on ice for 15 min and centrifuged at 12000 rpm for 10 min (4 °C). 150 µL of the supernatant was collected and placed in -20 °C for 30 min. The sample was then centrifuged at 12000 rpm for 3 min (4 °C). A 120 µL aliquot of the supernatant was used for LC-MS analysis.

2.4 Chromatography-mass spectrometry acquisition conditions

2.4.1 Liquid phase conditions (T3)

Chromatographic column: Waters ACQUITY Premier HSS T3 Column 1.8 µm, 2.1 mm * 100 mm

Mobile phase A: ultrapure water (0.1 % formic acid added)

Mobile phase B: acetonitrile (0.1 % formic acid added)

Column temperature: 40 °C; Flow rate: 0.40 mL/min: Injection volume: 2 uL

Table 6: Elution gradient for T3

Time (min)	A (%)	B (%)
0.0	95	5
11.0	10	90
12.0	10	90
12.1	95	5
14.0	95	5

2.4.2 Mass spectrum conditions

Table 7: Mass spectrum conditions for AB Sciex TripleTOF 6600

Parameter	ESI+	ESI-
Duration (min)	14	14
IonSpray Voltage (V)	5500	-4500
Temperature (°C)	550	450
Ion Source Gas1 (psi)	50	50
Ion Source Gas2 (psi)	60	60
Curtain Gas (psi)	35	35
Declustering Potential (V)	60	-60
MS1 Collision Energy (V)	10	-10
MS2 Collision Energy (V)	30	-30
Collision Energy Spread (V)	15	15
MS1 TOF Masses (Da)	50~1000	50~1000
MS2 TOF Masses (Da)	25~1000	25~1000
MS1 Accumulation time (s)	0.2	0.2
MS2 Accumulation time (s)	0.05	0.05
Candidate ions	12	12

2.5 Data preprocessing

The original data file acquired by LC-MS was converted to mzXML format by ProteoWizard. Peak extraction, peak alignment and retention time correction were performed by XCMS program. The peaks with missing rate >50% in each group of samples were filtered. The blank values were filled with KNN, and the peak area was corrected by SVR method. The metabolites were annotated by searching the MetwareBio's in-house database, integrated public database, prediction database and metDNA. Finally, substances with a comprehensive identification score above 0.7 and a CV value of QC samples less than 0.3 were extracted, and then positive and negative mode were combined (substances with the highest qualitative grade and the lowest CV value were retained) to obtain the ALL_sample_data file.

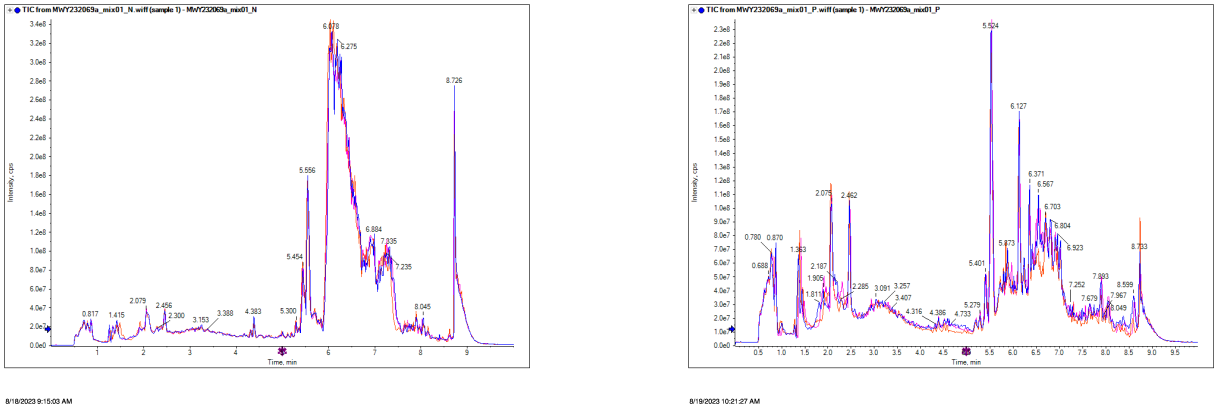
3 Data evaluation

3.1 Quality control sample analysis

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 inserted for every 10 test samples.

3.1.1 Total ion current diagram

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_neg_QC_TIC (b) Demo_pos_QC_TIC

Fig 3: 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 was 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_TIC.png

3.1.2 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.

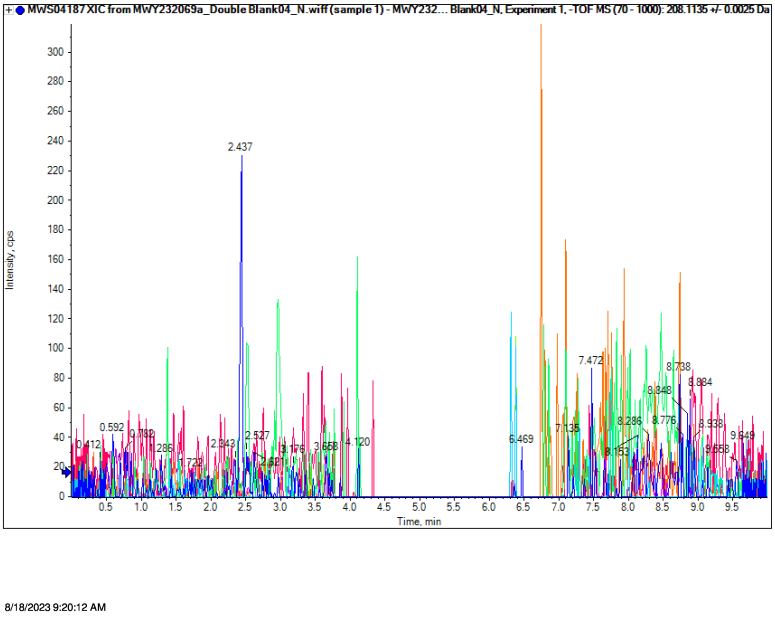


Fig 4: EIC diagram of internal label in blank sample

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

EIC diagram of internal label in blank sample:Final_report/1.Data_Assess/*/QC/*_*_BLANK_EIC.png

3.1.3 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.

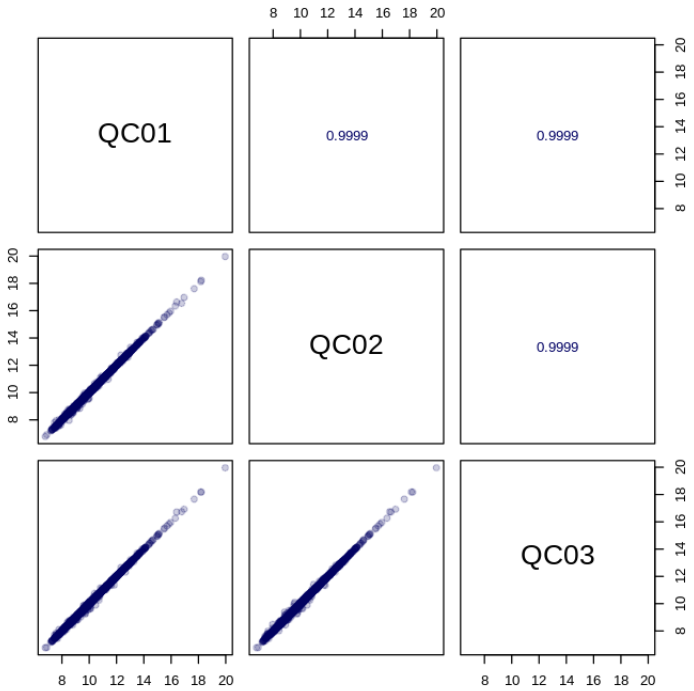


Fig 5: Plot of QC sample correlation

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.

Plot of QC sample correlation:Final report/1.Data_Assess/*/QC/*_QC_correlation.*

3.1.4 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 8: Stability of internal standards in QC samples

Index	Q1 (Da)	RT (min)	CV
MWS04127	195.0449	0.80	0.0296
MWS1055	126.0608	4.46	0.0367
MWS04187	208.1136	2.52	0.0409
MWS4243	198.0324	2.61	0.0442

Stability of internal standards in QC samples:Final_report/1.Data_Assess/*/QC/*_internal_standard.xlsx

3.1.5 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.

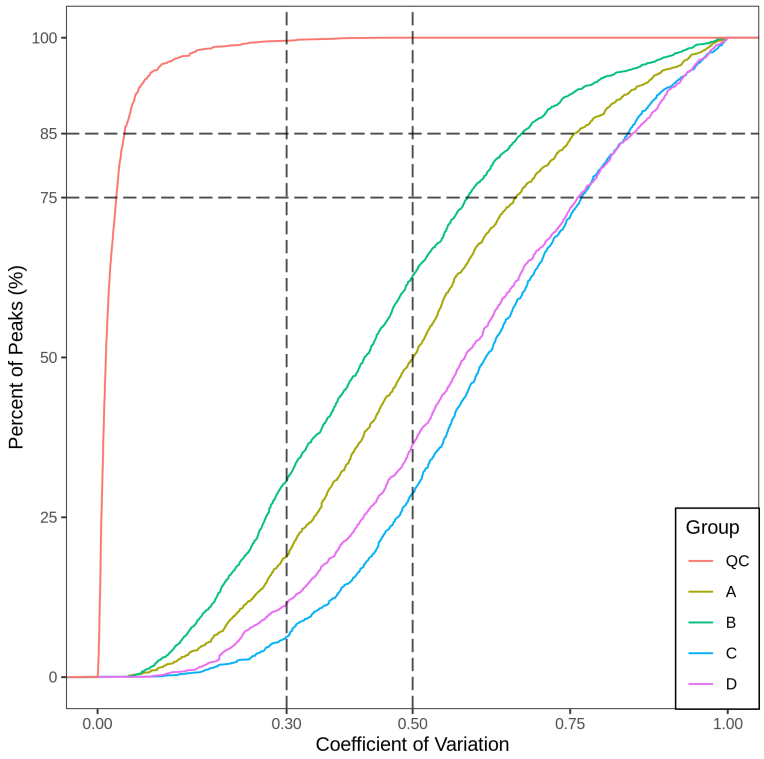


Fig 6: CV distribution of each group

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. Mix indicates QC samples. The two dash lines on X-axis correspond to 0.3 and 0.5; the two dash lines on Y-axis correspond to 75 % and 85 % (If there is only one sample in the group, the CV value cannot be calculated).

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

3.2 Principal Component Analysis (PCA)

3.2.1 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. 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.

3.2.2 Principal component analysis of the sample populations

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:

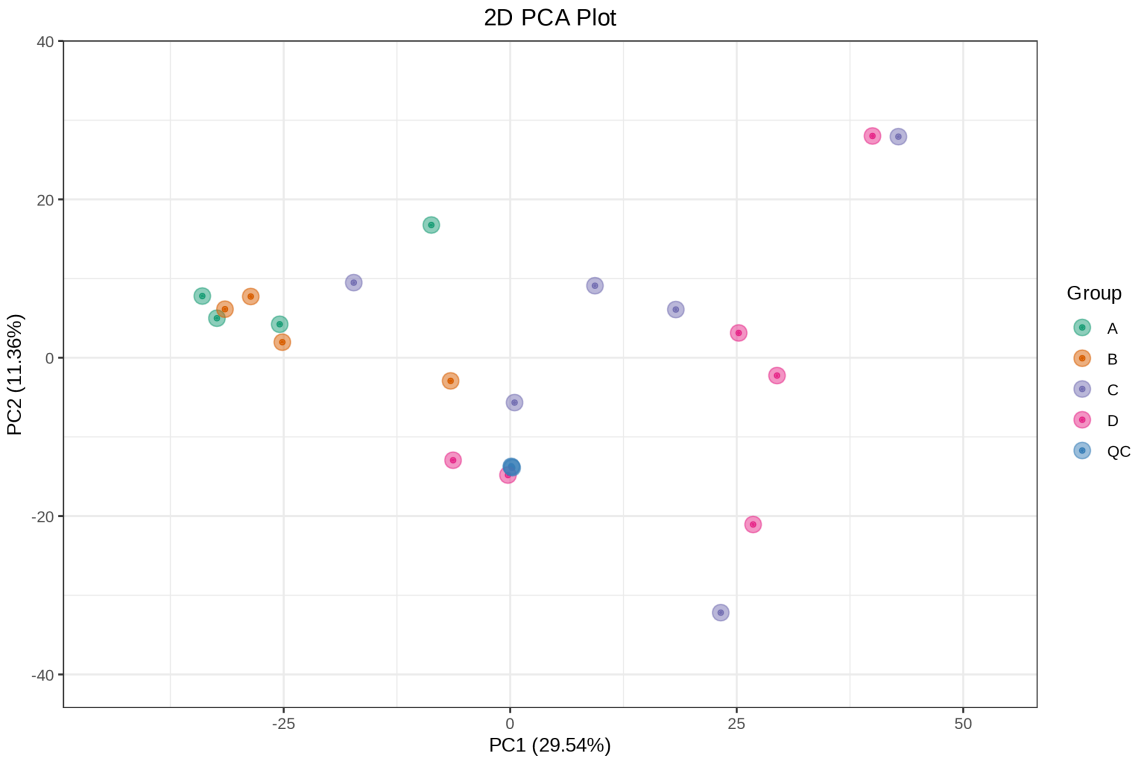


Fig 7: PCA score diagram of quality spectrum data of each group of samples and quality control samples

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.

Principal component analysis of population sample:Final_report/1.Data_Assess/*/pca/

3.2.3 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.

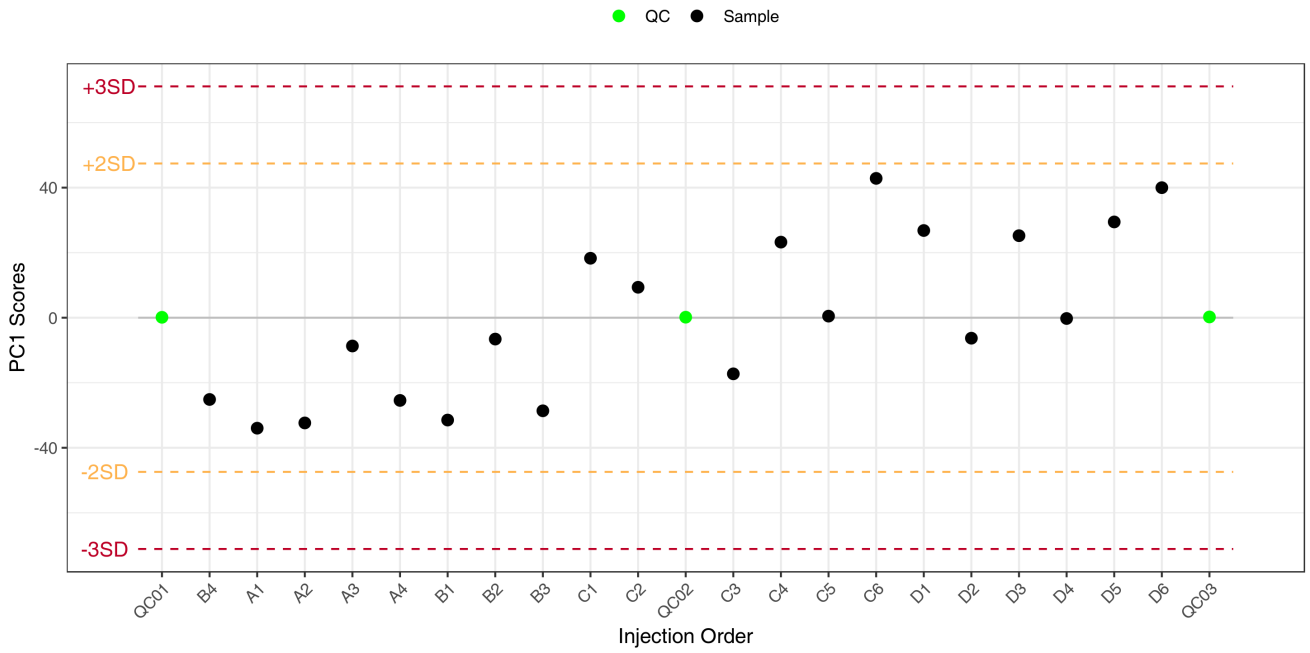


Fig 8: PC1 variation diagram of all the sample

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.*

3.3 Hierarchical Cluster Analysis (HCA)

3.3.1 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.

3.3.2 Hierarchical Cluster Analysis results

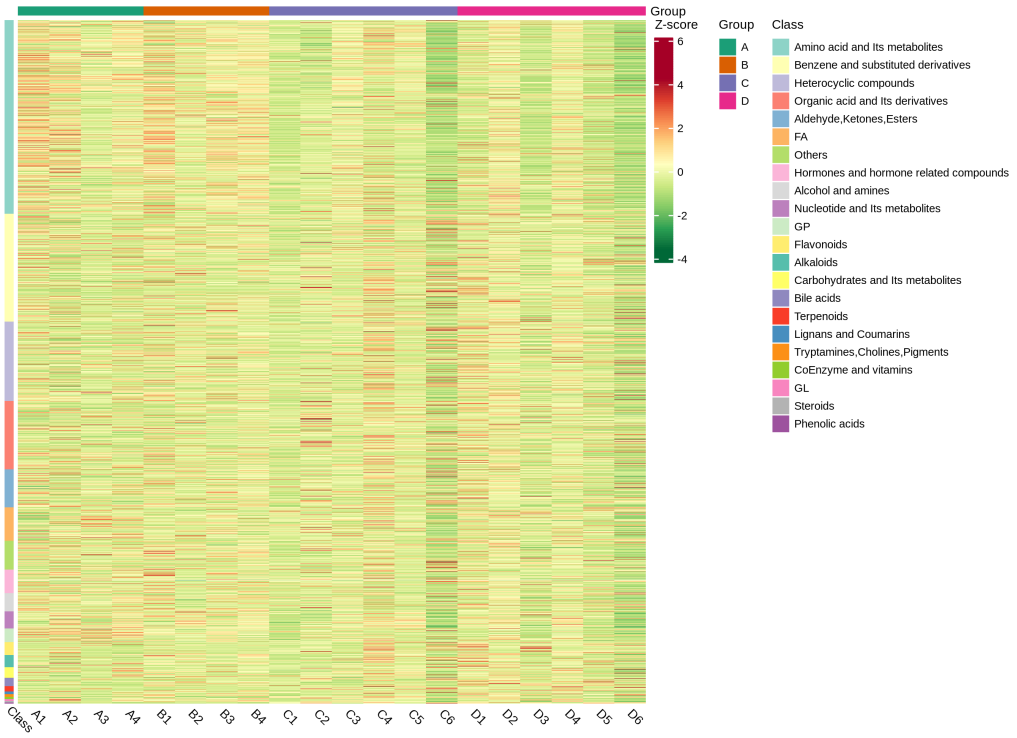


Fig 9: Sample clustering diagram

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 line on the left side of the figure is the metabolite clustering line, and the clustering line on the top of the figure is the sample clustering line. _all_heatmap_row_cluster" clustering analysis is performed for metabolites only, the clustering line on the left side of the figure is the metabolite clustering line.

Clustering analysis result:Final_report/1.Data_Assess/*/heatmap/

4 Analysis results

4.1 Principal component analysis of 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.

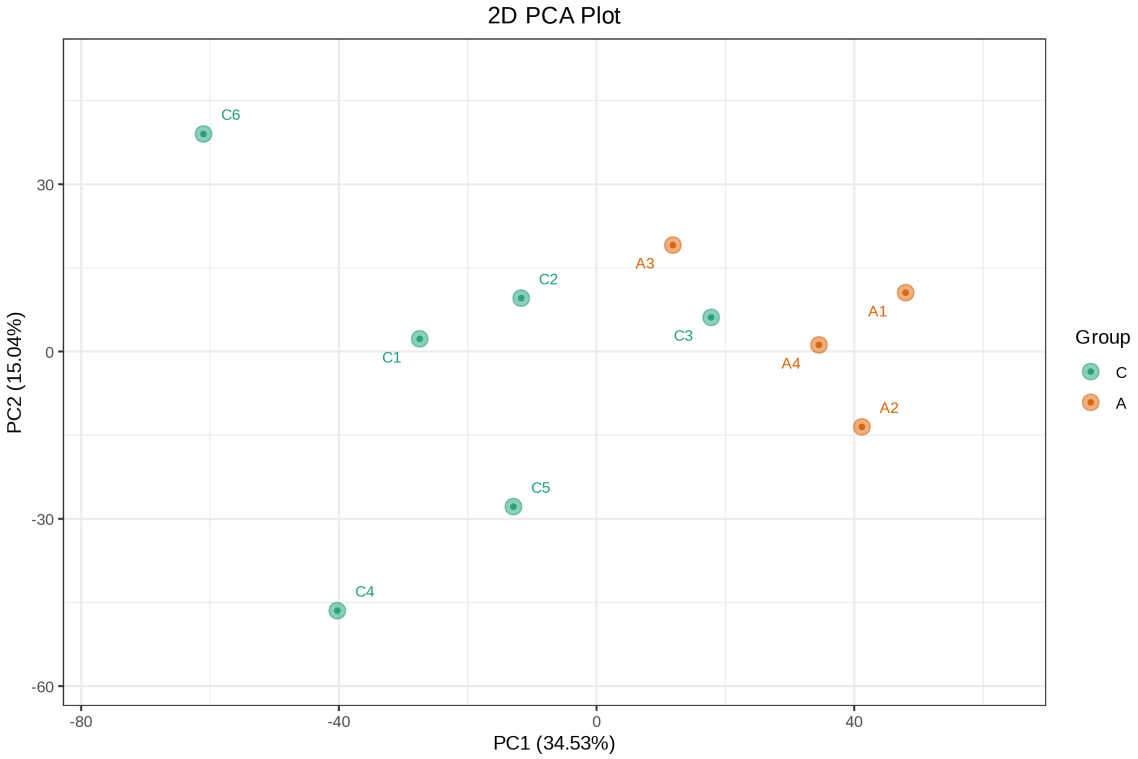


Fig 10: Principal component analysis of different groups

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:

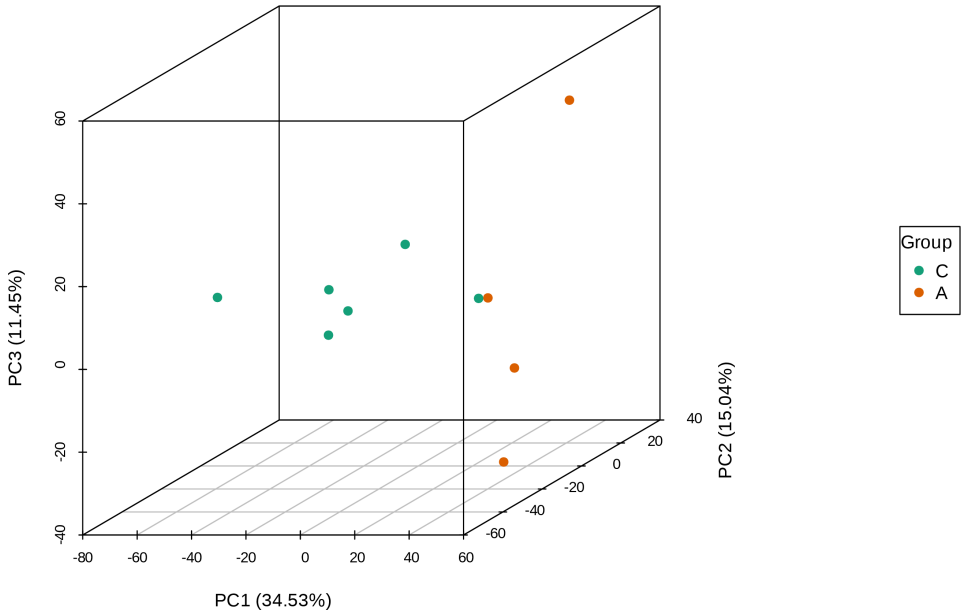


Fig 11: Three-dimensional PCA plot of different groups

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:

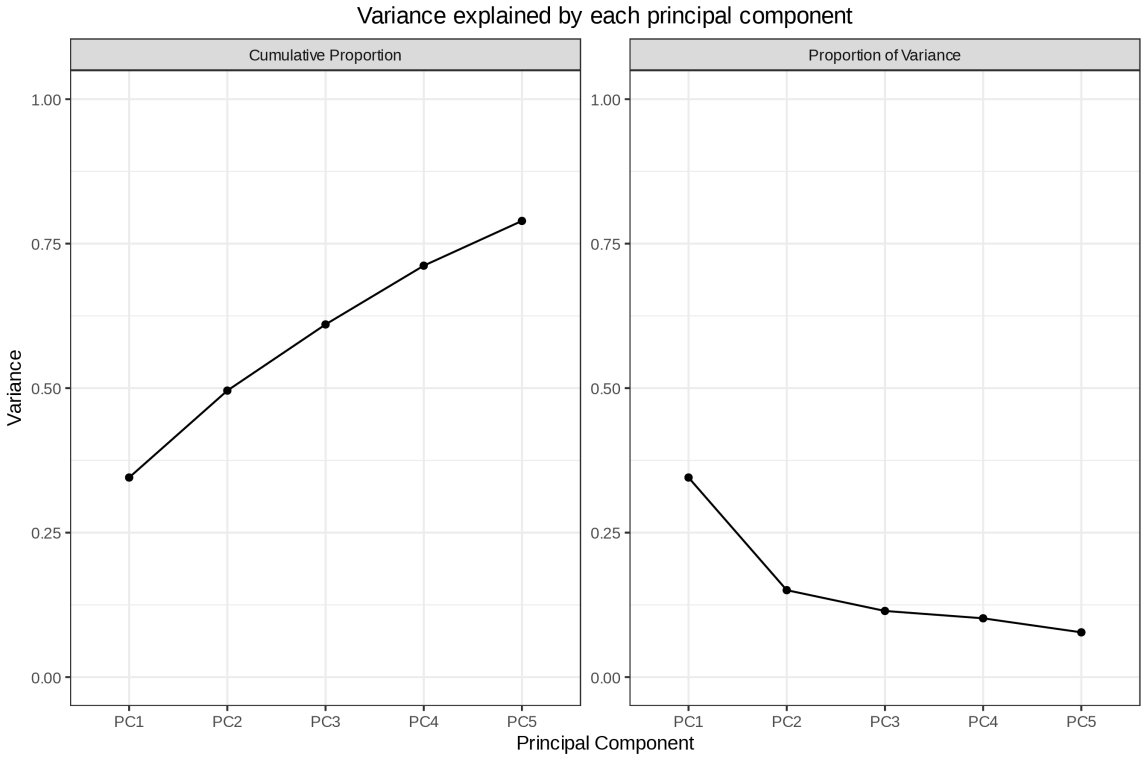


Fig 12: The explainable variation of the first five principal components

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: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 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 lipids. 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 9: Partial results of OPLS-DA

Index	VIP
MW0057055	1.4383200
MW0107179	0.8042176
MW0107555	1.2124390
MW0009304	1.0059434
MEDN1476	0.1777086
FDATN00717	0.4158045
MW0053418	1.5326597
MW0126293	1.1090842
MW0169549	0.4223619
MW0006917	1.2707853
MW0142582	1.0270233
MW0009611	1.8281976
MW0103343	1.1156874
MW0103332	1.2267477
MW0009652	1.3655840

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).

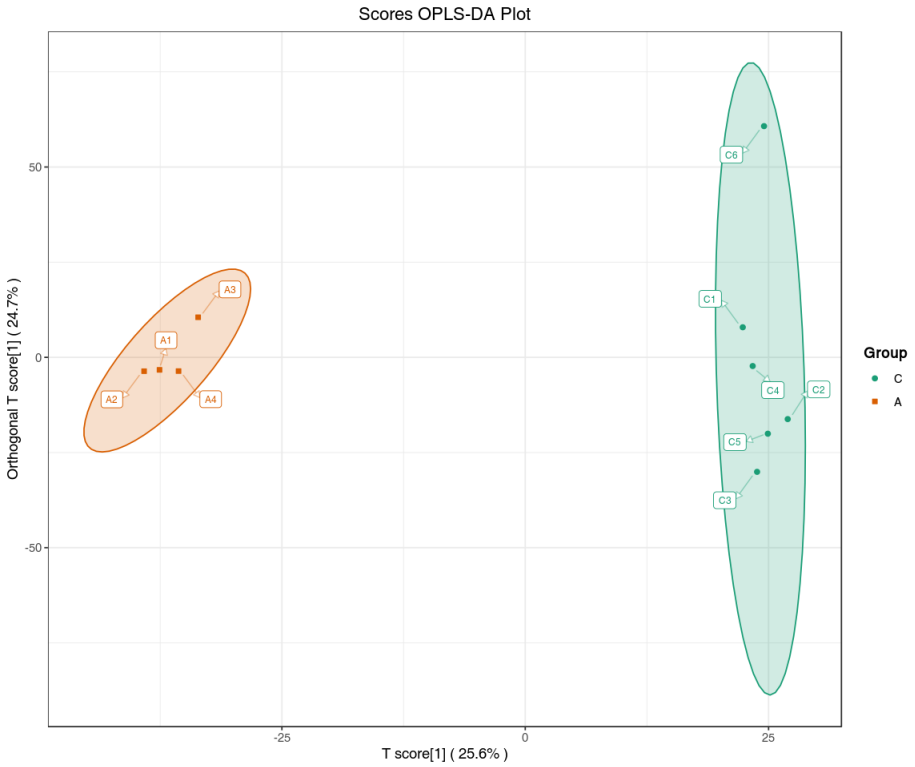


Fig 13: OPLS-DA score diagram

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-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 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 predictability of four random grouping models is better than that of the

OPLS-DA model in the Permutation detection. If R²Y'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.

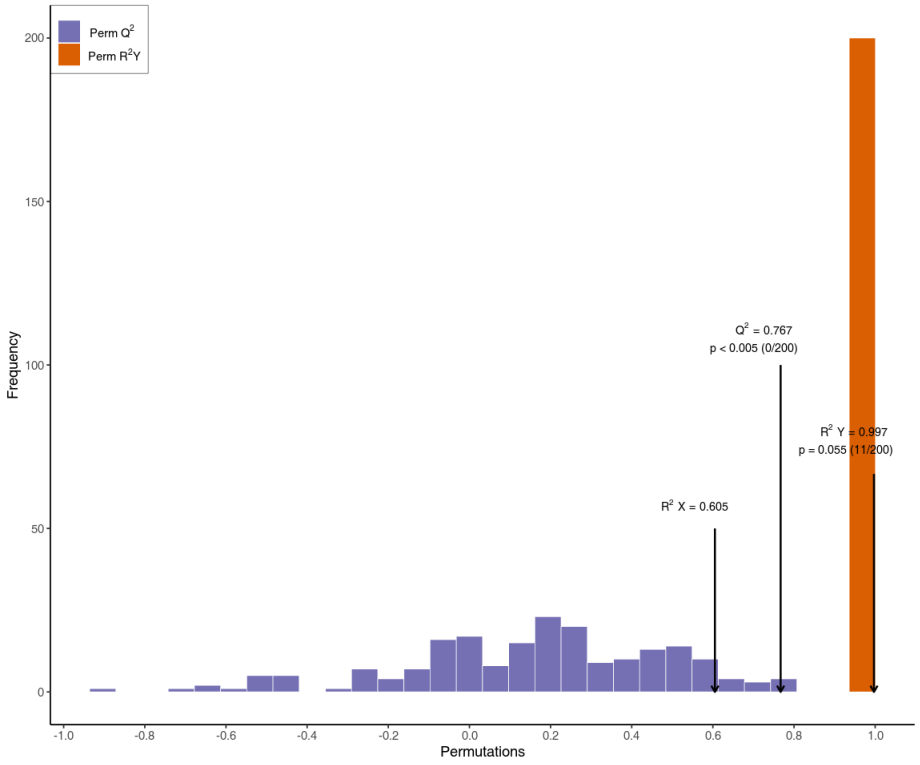


Fig 14: OPLS-DA verification diagram

Note: The orange color represents the R^2Y of the random grouping model, the purple color represents the Q^2 of the random grouping model, and the black arrows represent the values of R^2X, R^2Y, and Q^2 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.

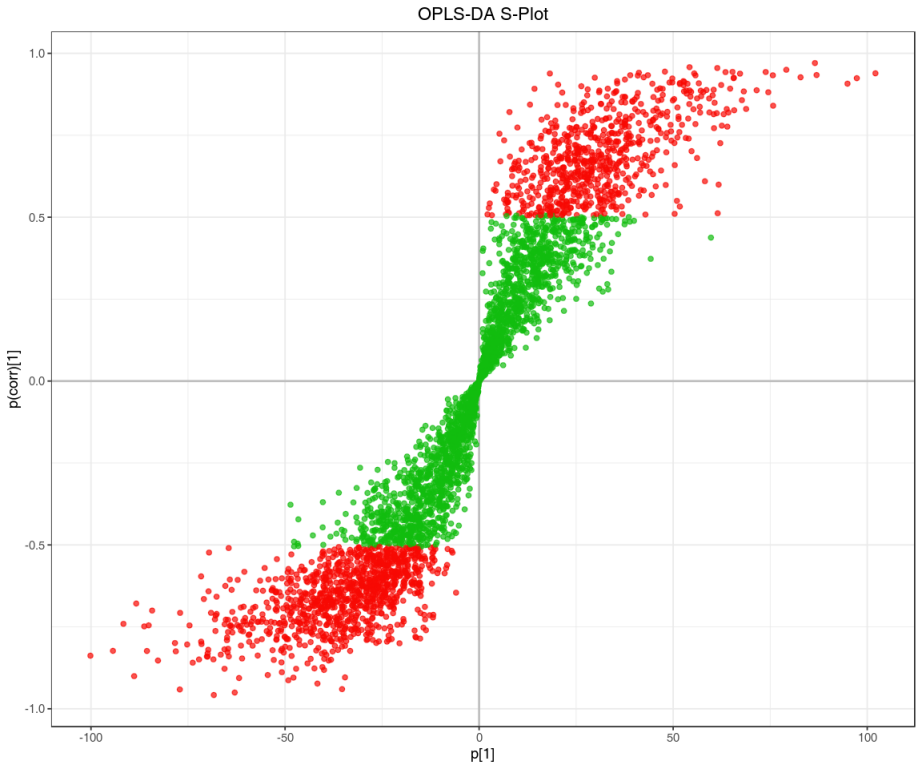


Fig 15: OPLS-DA S-plot

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.

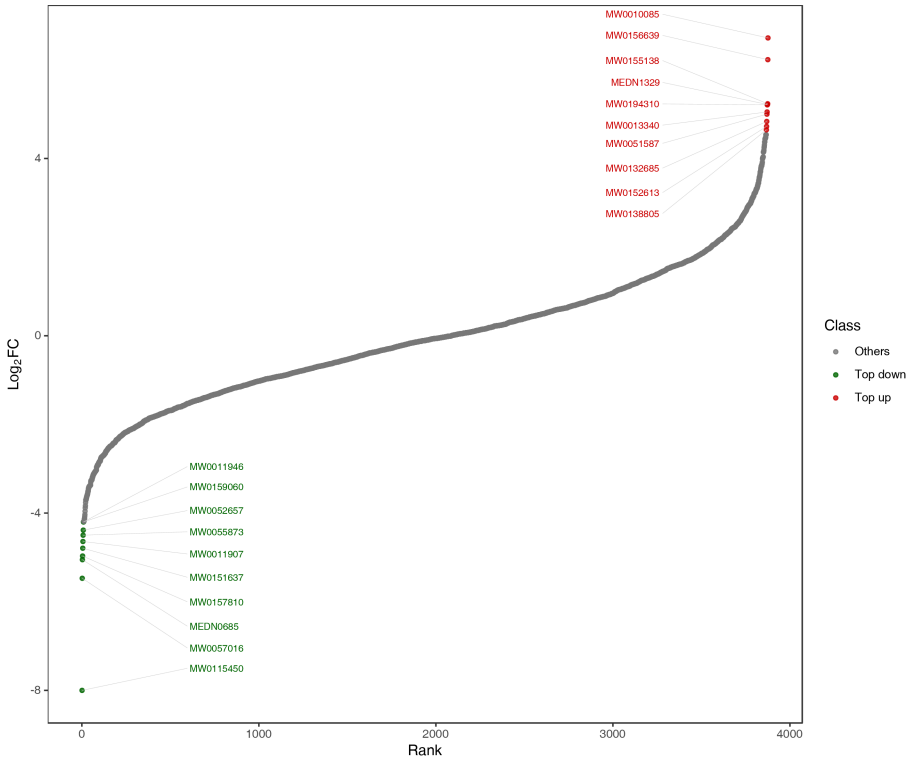


Fig 16: Dynamic distribution of metabolite content difference

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 different..

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 10: Screening results of differential metabolites

Index	Compounds	Type
MW0057055	1,2-Dilinoleoyl-SN-glycero-3-phosphocholine	down
MW0053418	ganoderic acid F	up
MW0006917	Normetanephrine	up
MW0009611	Pyridaben	up
MW0103343	2'-Deoxyguanosine	down
MW0009652	Ranolazine	down
MW0107969	L-Tyrosine ethyl ester	up
MW0152041	Knipholone	up
MW0063564	C24:1 Sphingomyelin	up
MW0107880	L-Homophenylalanine	down
MW0009466	Pimelic Diphenylamide 106	down
MW0003351	2-Phenylpropanal	down
MEDP1884	Prolyl-Histidine	down
MW0148591	Val-Pro-Leu	down
MW0107459	Ile-Pro-Ile	down

Screening results of differential metabolites:Final_report/2.Basic_Analysis/Difference_analysis/*_vs_*/*_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.

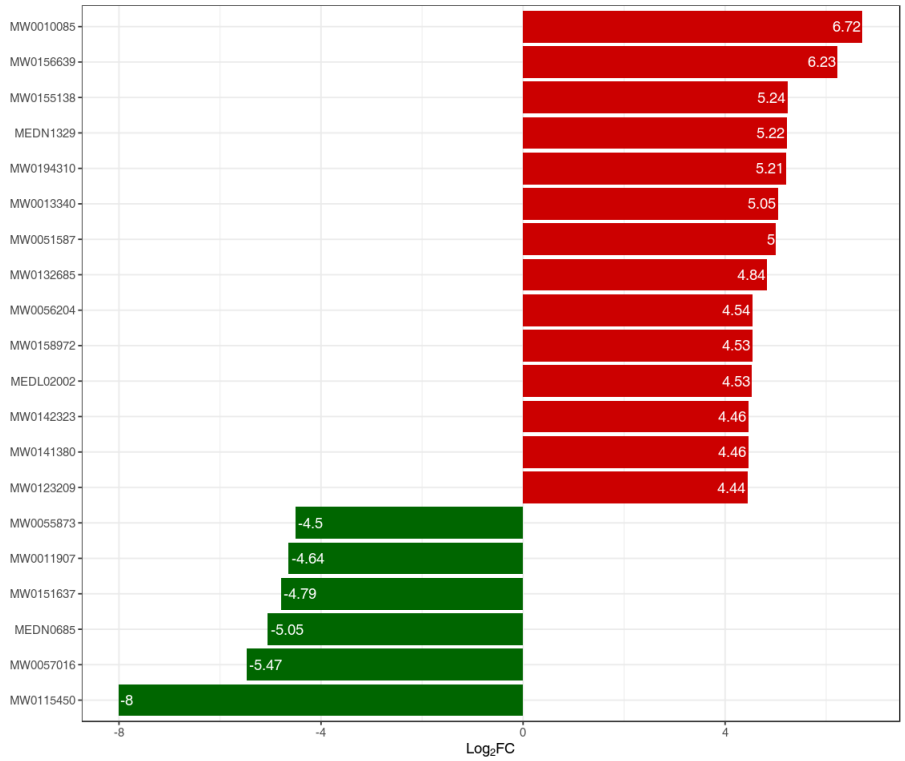


Fig 17: Histogram of multiple difference

Note: X-axis refers to $\log_2$FC 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 Radar map of differential metabolites

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

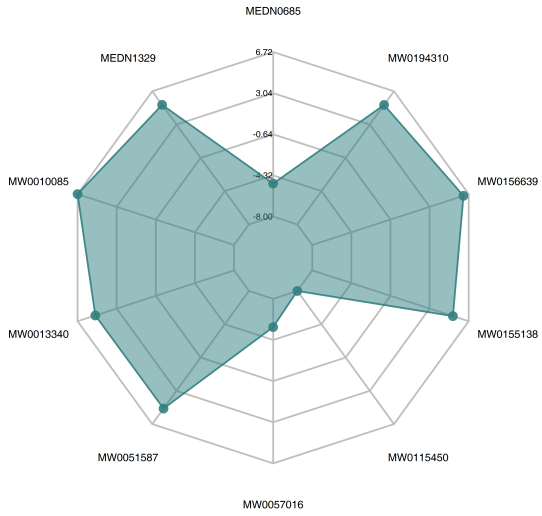


Fig 18: Radar map of differential metabolites

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.

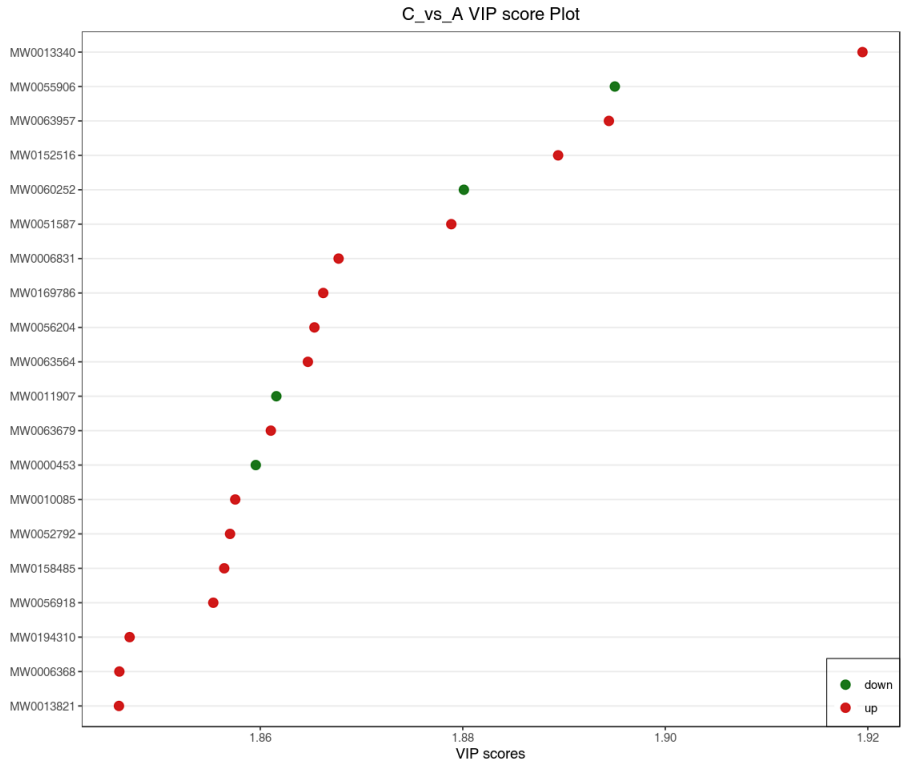


Fig 19: VIP values of differential metabolites

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. Yellow represents metabolites with significant differences in three and more differential comparison groups.

VIP values of differential metabolites:Final_report/2.Basic_Analysis/Difference_analysis/*_vs_*/vipscore/*_vipScore*.*

4.4.4 Volcano plot of differential metabolites

Volcano Plot is mainly 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.

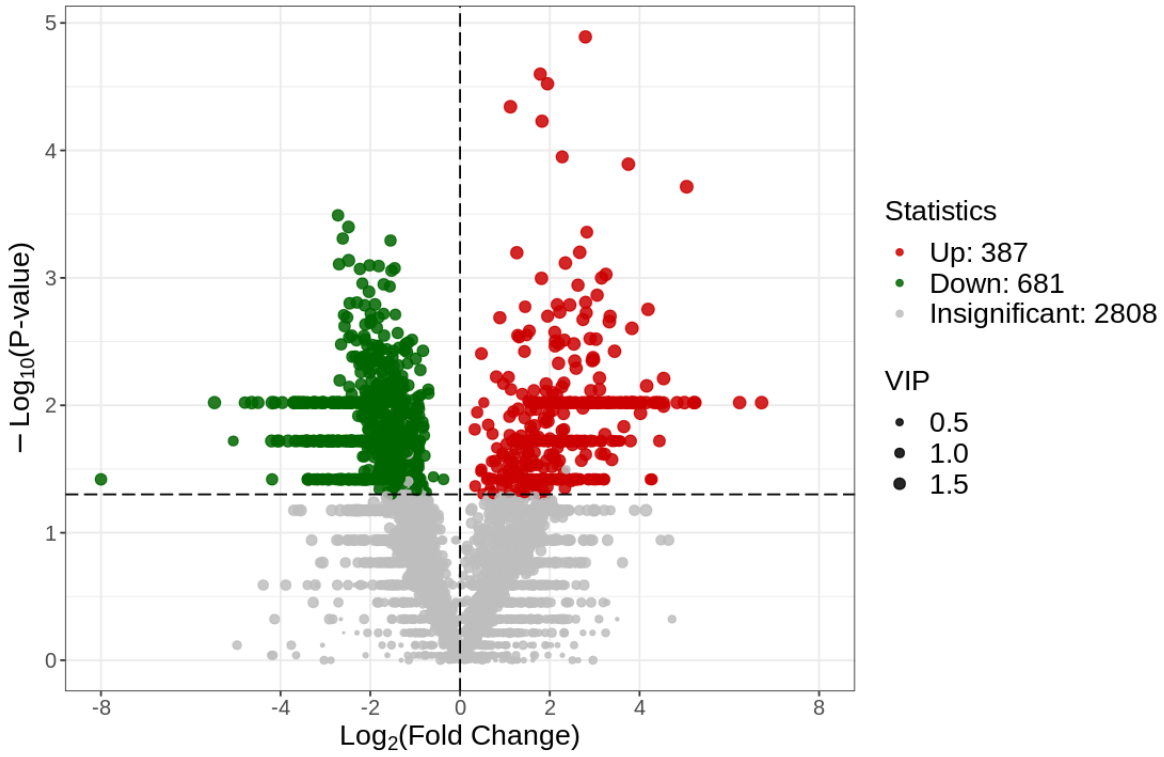


Fig 20: Volcano plot of differential metabolites

Note: Each point in the volcano plot represents a metabolite with green points represent down-regulated differential metabolite, red points represent up-regulated differential metabolite, and gray points represent the detected metabolites but show no significant differences. The X-axis represents the (log_2FC) 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 plot of differential metabolites:Final_report/2.Basic_Analysis/Difference_analysis/*_vs_*/vol/*_vol.*

4.4.5 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.

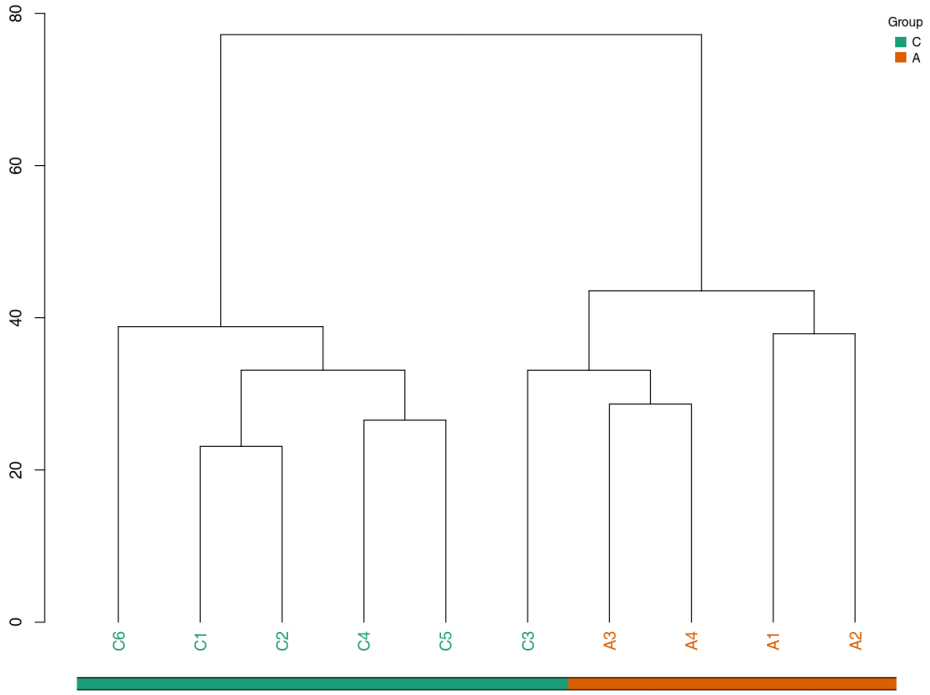


Fig 21: Hierarchical clustering tree

Note: Each row in the figure represents one sample. Samples with high similarity to each other are grouped into the same cluster.

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

4.4.6 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 on a heatmap using pheatmap in R.

Fig 22: Cluster 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 standardization 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. _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 line on the left side of the figure is the metabolite clustering line, and the clustering line on the top of the figure is the sample clustering line. _all_heatmap_row_cluster: clustering analysis is performed for metabolites only, the clustering line on the left side of the figure is the metabolite clustering line.

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

4.4.7 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.

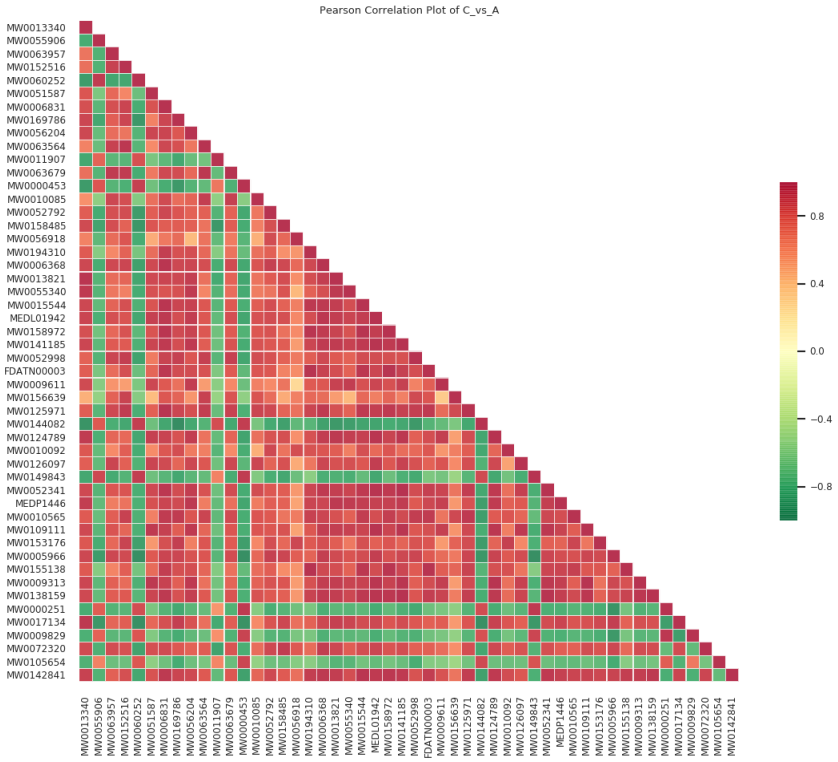


Fig 23: Heatmap of correlation of different metabolites

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.

Original file path (The directory structure in positive and negative ion mode is the same, so only the files in positive ion mode are linked to):

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

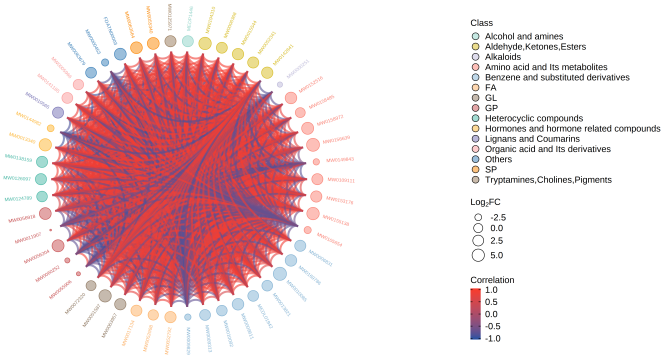


Fig 24: Chord diagram of differential metabolites

Note: The outermost layer shows the metabolite ID. The larger the point, the larger its corresponding 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. If there are more than 50 differential metabolites, the figure will only show the top 50 metabolites based on VIP.

Chord diagram of differential metabolites:Final_report/2.Basic_Analysis/Difference_analysis/*_vs_*/cpdCorr/*_cpdCorrCir_*.*

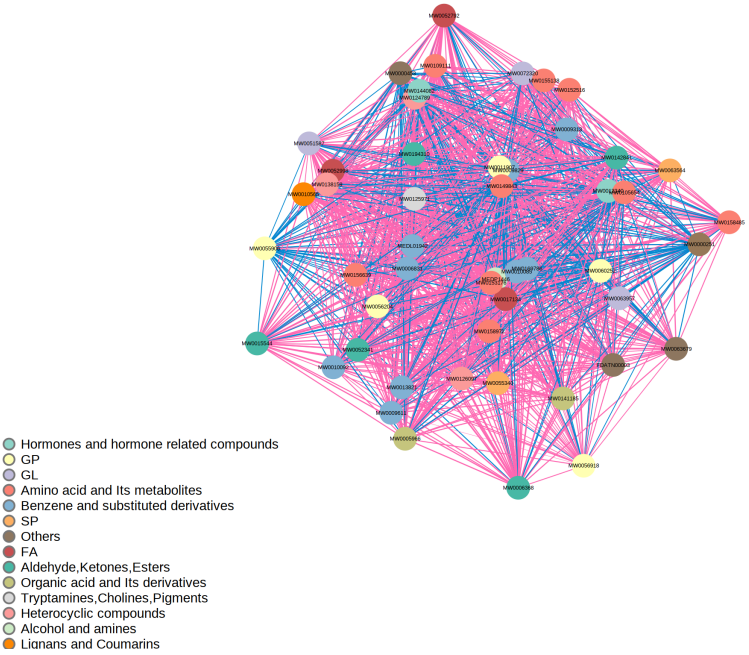


Fig 25: Correlation network diagram of differential metabolites

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 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_*.*

4.4.8 Z-value map of differential metabolites

Z-score standardization normalizes the relative content of the differential metabolites by calculating Z-scores. The Z-score plot provides a very visual representation of the distribution of each differential metabolite across groups.

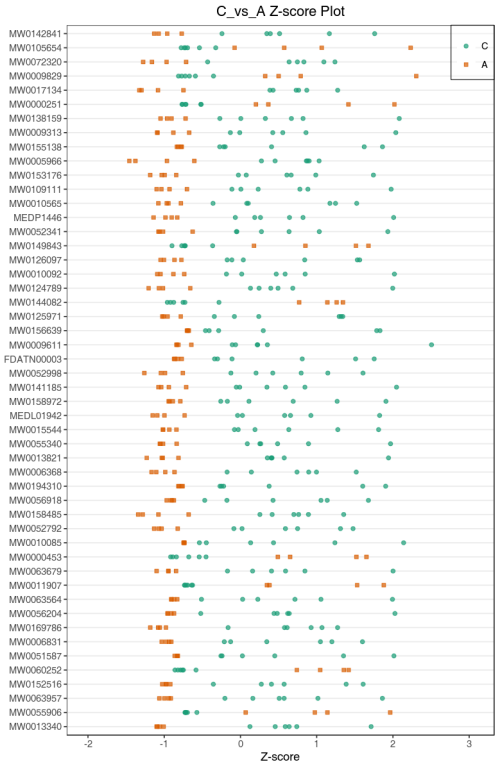


Fig 26: Z-value map of differential metabolites

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 diagram of differential metabolites:Final_report/2.Basic_Analysis/Difference_analysis/*_vs_*/zScore/*_zScore*.*

4.4.9 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.

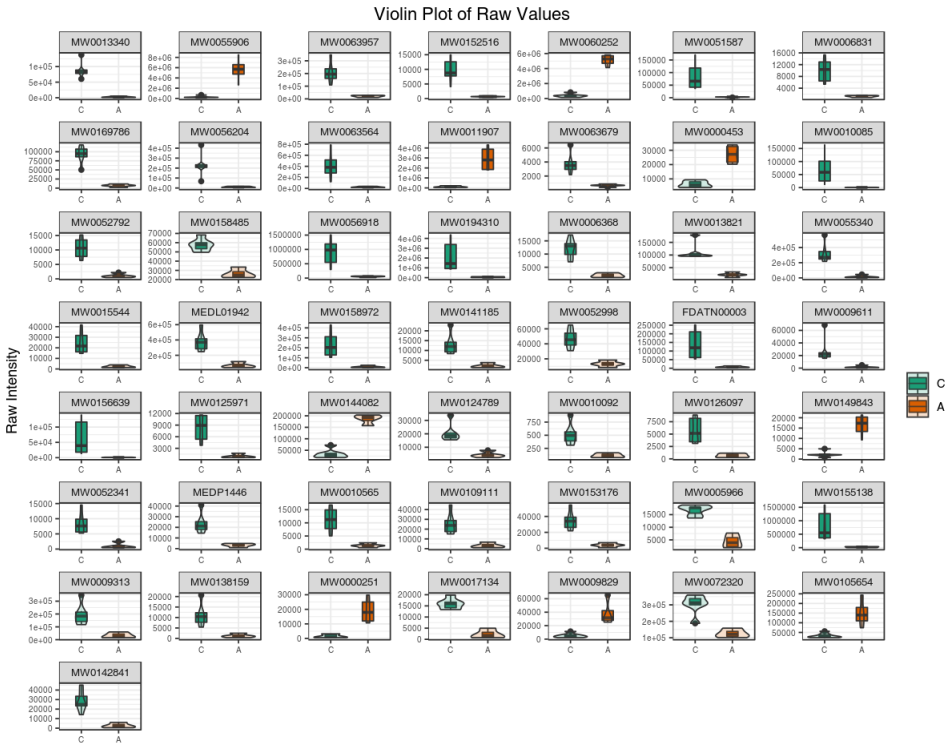


Fig 27: Violin plot of differential metabolites

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/singleViolin

4.4.10 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 UV standardized relative quantification value.

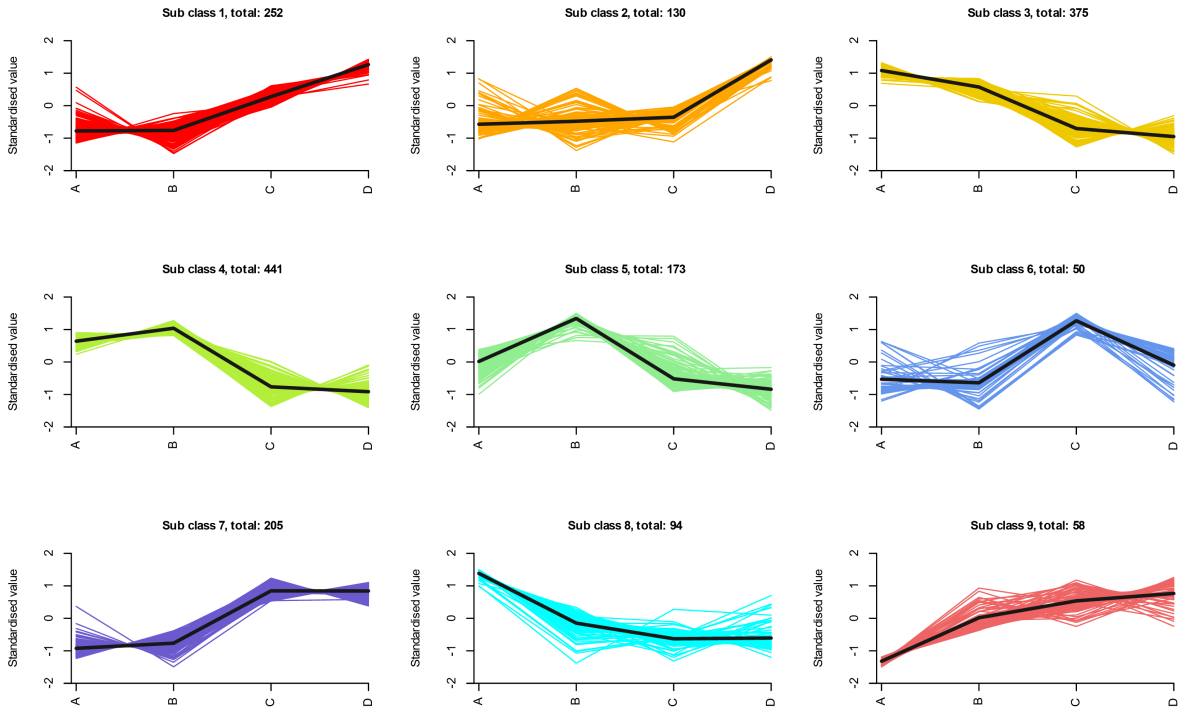


Fig 28: K-Means diagram of differential metabolites

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

K-means diagram of differential metabolites:Final_report/2.Basic_Analysis/kmeans/*_kmeans_cluster.*

4.4.11 Venn diagram of differential metabolites

Venn diagram is used to show the relationship the number of shared and unique metabolites in different comparison groups. A petal diagram is used for groups 5 or more. The results are shown below:

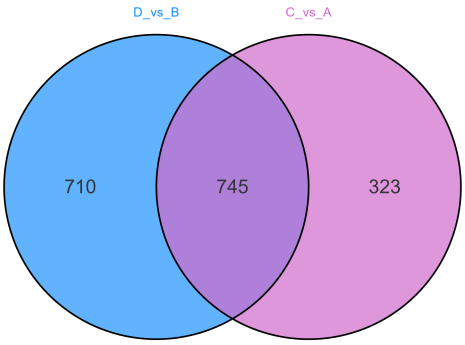


Fig 29: Venn diagram of differences among groups

Note: Each circle in the figure 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 differences among groups: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 differential 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:

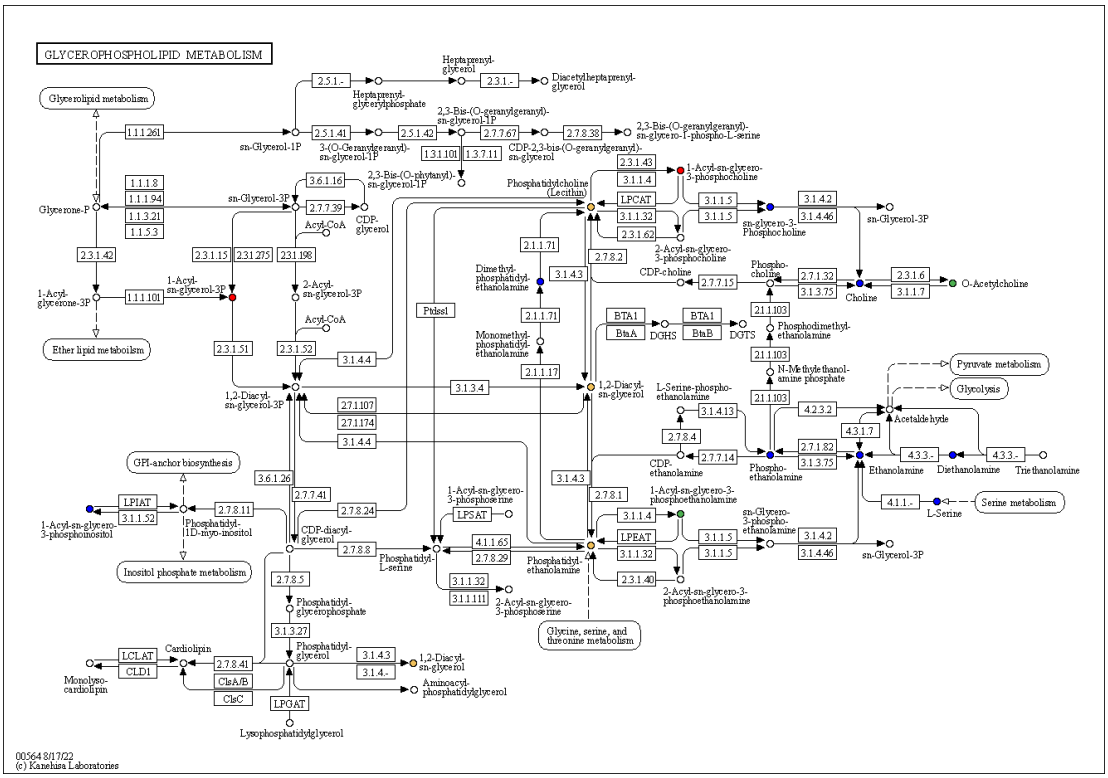


Fig 30: KEGG pathway of differential metabolites

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- 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 11: KEGG annotations for differential metabolites

Index	Compounds	Type	cpd_ID
MW0057055	1,2-Dilinoleoyl-SN-glycero-3-phosphocholine	down	C00157
MW0053418	ganoderic acid F	up	-
MW0006917	Normetanephrine	up	C05589
MW0009611	Pyridaben	up	C18614
MW0103343	2'-Deoxyguanosine	down	C00330
MW0009652	Ranolazine	down	-
MW0107969	L-Tyrosine ethyl ester	up	C01458
MW0152041	Knipholone	up	-
MW0063564	C24:1 Sphingomyelin	up	C00550
MW0107880	L-Homophenylalanine	down	C17235
MW0009466	Pimelic Diphenylamide 106	down	-
MW0003351	2-Phenylpropanal	down	-
MEDP1884	Prolyl-Histidine	down	-
MW0148591	Val-Pro-Leu	down	-
MW0107459	Ile-Pro-Ile	down	-

Table 12: Enrichment Statistics of KEGG annotations for differential metabolites

ko_ID	Sig_compound	compound	Sig_compound_all	compound_all
ko00564	41	126	151	613
ko00590	22	83	151	613
ko00591	22	79	151	613
ko00592	20	72	151	613
ko01100	116	482	151	613
ko04723	25	91	151	613
ko05231	20	73	151	613
ko00350	7	16	151	613
ko00230	3	18	151	613
ko01232	4	24	151	613
ko02010	6	44	151	613
ko00600	3	20	151	613
ko04071	4	19	151	613
ko04217	3	16	151	613
ko00400	2	11	151	613

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

Enrichment Statistics 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:

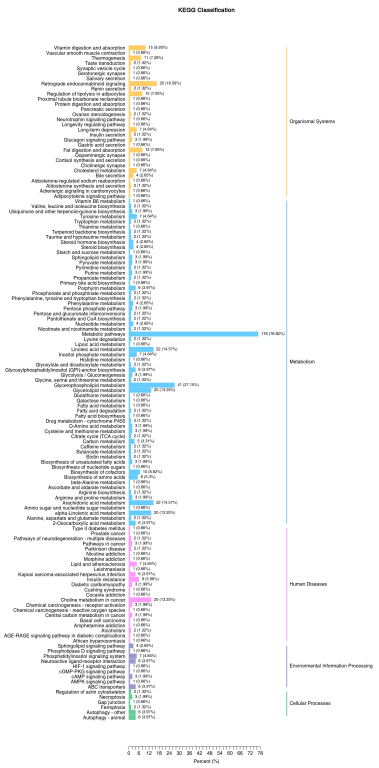


Fig 31: KEGG classification of differential metabolites

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.

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

Five significantly enriched KEGG metabolic pathways were selected for clustering analysis. Only pathways with at least 5 differential metabolites are shown.

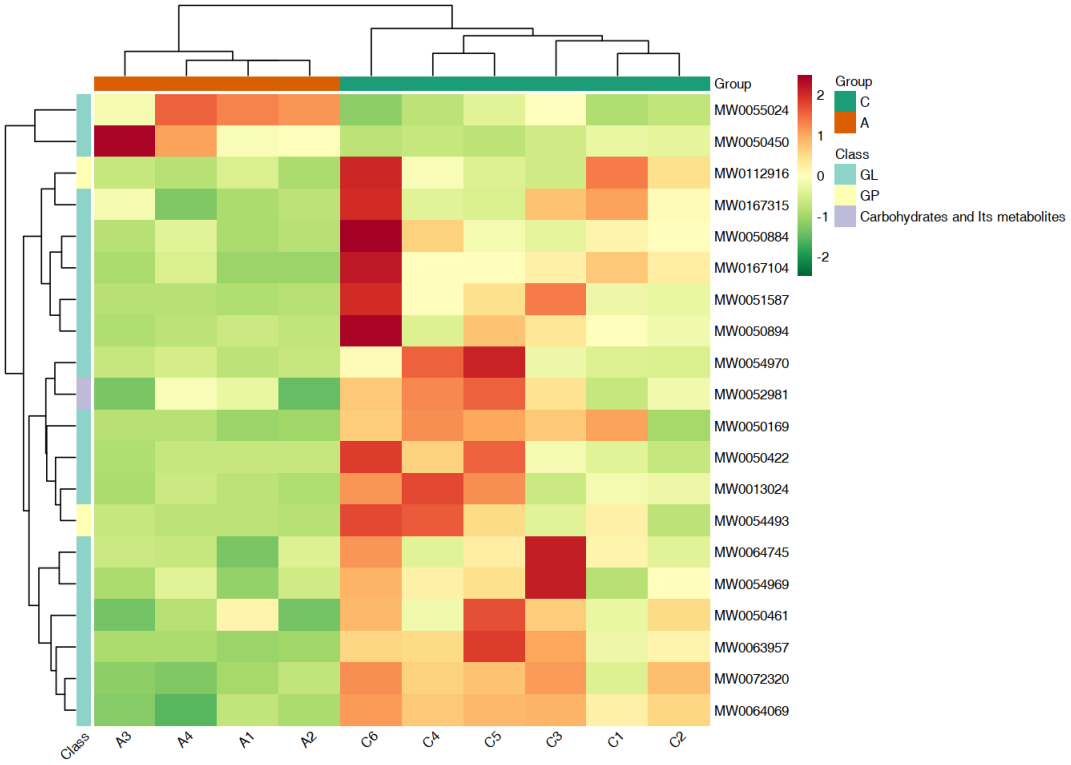


Fig 32: Clustering heat map of differential metabolites in KEGG pathway

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 Level 1 classifications.

Clustering heatmap of differential metabolites in KEGG pathway:Final_report/2.Basic_Analysis/Difference_analysis/

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 was 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 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 top 20 pathways in terms of P-value were selected for presentation from smallest to largest.

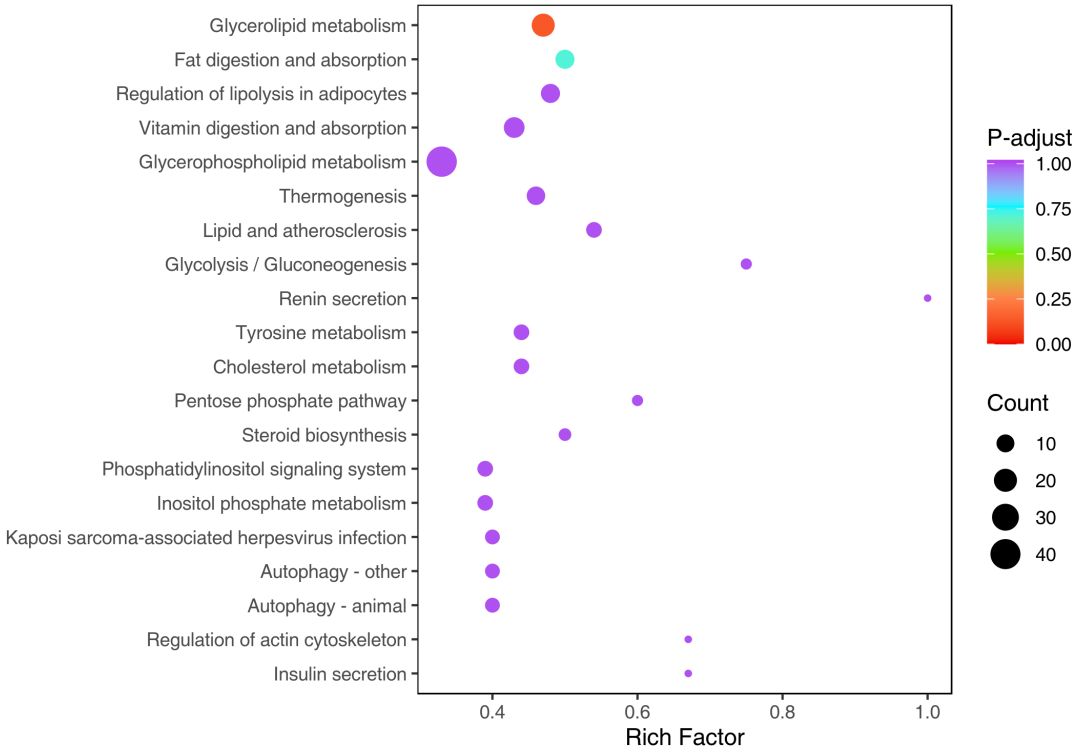


Fig 33: KEGG enrichment diagram of differential metabolites

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.*

4.6 Functional annotation and enrichment analysis with 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 13: SMPDB pathway enrichment for differential metabolites

primary_SMPDB_ID	P-value
SMP0000128	0.100976099704507
SMP0000563	0.100976099704507
SMP0000581	0.100976099704507
SMP0000374	0.100976099704507
SMP0000560	0.100976099704507
SMP0000574	0.100976099704507
SMP0000573	0.100976099704507
SMP0000562	0.100976099704507
SMP0000482	0.100976099704507
SMP0000558	0.109909606456136
SMP0000196	0.109909606456136
SMP0000559	0.109909606456136
SMP0000334	0.109909606456136
SMP0000060	0.109909606456136
SMP0000212	0.109909606456136

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:

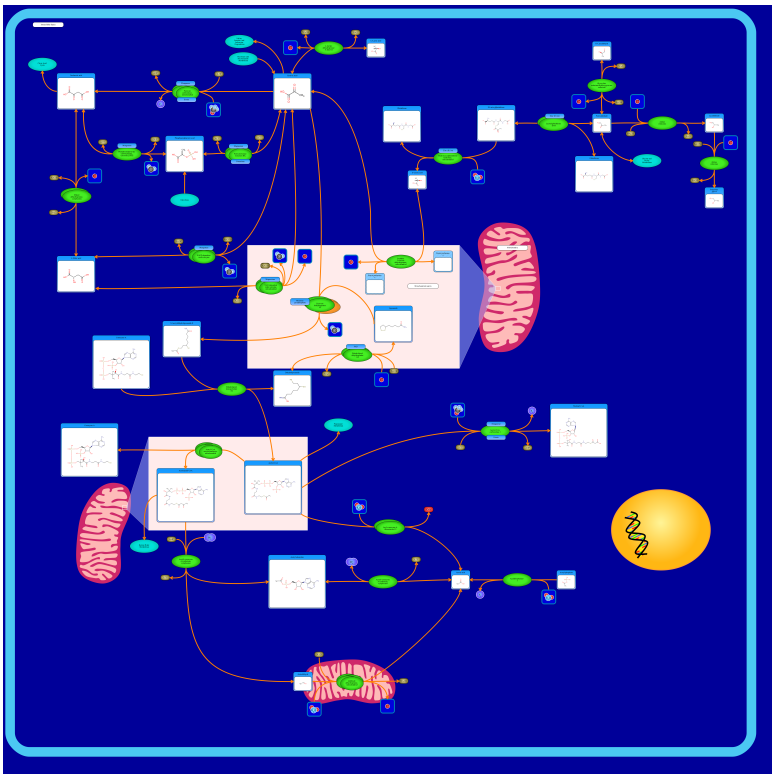


Fig 34: HMDB pathway map of differential metabolites

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:

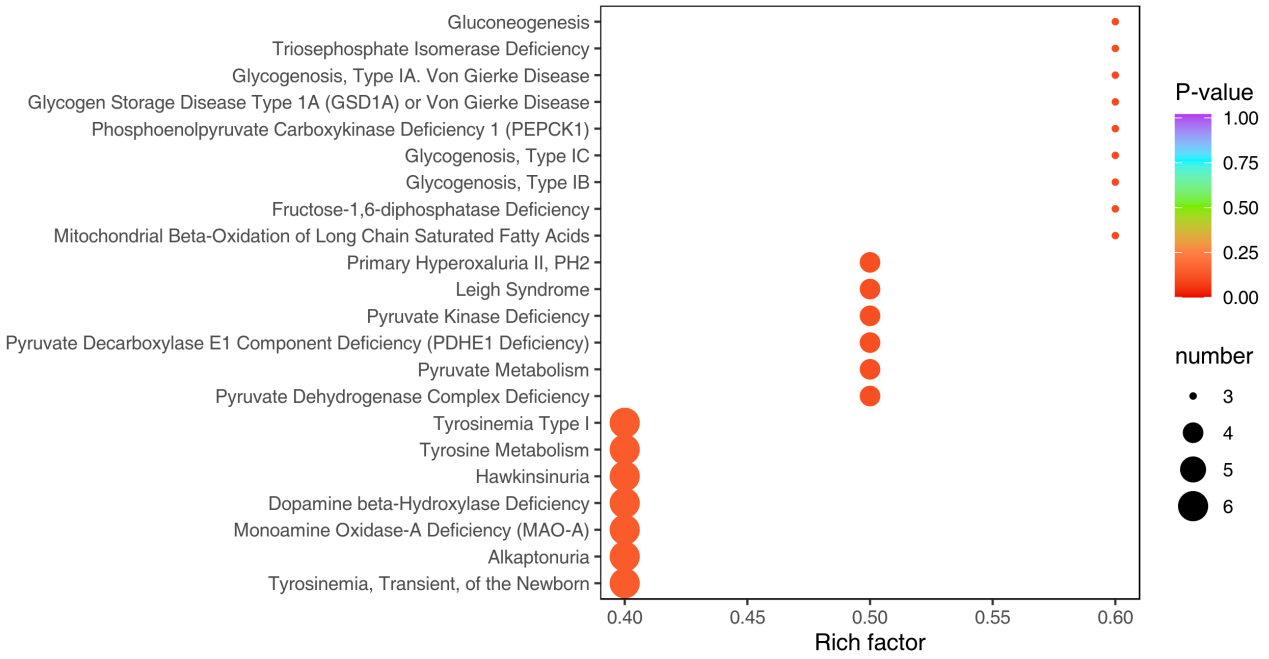


Fig 35: HMDB enrichment diagram of differential metabolites

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 14: Table for MSEA enrichment analysis

name	P-value	foldEnrichment
Steroid biosynthesis	0.030958	4.1452615
Fatty acid degradation	0.042204	3.7940779
Purine metabolism	0.057401	2.3721537
Glycine, serine and threonine metabolism	0.165450	1.9887499
Fatty acid biosynthesis	0.197080	1.7852579
Glycerophospholipid metabolism	0.205030	1.7200072
Pantothenate and CoA biosynthesis	0.215360	1.5581856
D-Glutamine and D-glutamate metabolism	0.306910	1.1662317
Glyoxylate and dicarboxylate metabolism	0.306910	1.1662317
Nitrogen metabolism	0.306910	1.1662317
Selenocompound metabolism	0.337780	1.0350104
Primary bile acid biosynthesis	0.354090	0.9712897
Starch and sucrose metabolism	0.361750	0.9431194
Neomycin, kanamycin and gentamicin biosynthesis	0.368050	0.9191792
Galactose metabolism	0.371000	0.9087391

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

Fig 36: 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 15: Table of association between differential metabolites and diseases

CompoundName	KeggDiseases
1,2-Dilinoleoyl-SN-glycero-3-phosphocholine	-
ganoderic acid F	-
Normetanephrine	-
Pyridaben	-
2'-Deoxyguanosine	-
Ranolazine	-
L-Tyrosine ethyl ester	-
Knipholone	-
C24:1 Sphingomyelin	-
L-Homophenylalanine	-
Pimelic Diphenylamide 106	-
2-Phenylpropanal	-
Prolyl-Histidine	-
Val-Pro-Leu	-
Ile-Pro-Ile	-

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

5 References

1. 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.
2. 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.
3. 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.
4. Kanehisa, M. and S. Goto, KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res, 2000. 28(1): p. 27-30.
5. 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.
6. 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.

6 Appendix

6.1 List of software and versions

Table 16: Software used

Analysis	Software	Version
PCA	R (base package)	4.1.2
Pearson Correlation	R (base package)	4.1.2
Heatmap	R (ComplexHeatmap)	2.9.4
OPLS-DA	R (MetaboAnalystR)	1.0.1
Radar map	R (fmsb)	0.7.1
Chord diagram	R (igraph; ggraph)	1.2.11; 2.0.5
Correlation network diagram	R (igraph)	1.2.11
Modulation network diagram	R (FELLA)	1.2.0

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：

$$x' = x - \mu$$