Innovative Metabolomics Insights for Better Health

Demo Flavonoids Metabolomics Report

Metware Biotechnology Inc.

www.metwarebio.com

Contents

Demo Flavonoids Metabolomics Report

1 Abstract

Flavonoids are a class of compounds that exist in nature. The C6-C3-C6 characteristic structure is two aromatic rings connected by a central three-carbon chain with ketone carbonyl group and a basic oxygen atom in the first position. Most plants contain flavonoids that play important roles in plant growth, development, flowering, fruiting, and antibacterial and disease prevention.

(1) For this project, 15 samples were selected, and divided into 5 groups, with 3 biological replicates for each group. A total of 635 metabolites were detected based on UPLC-MS/MS system and identified using MWDB metabolite database.

(2) Results of differential metabolite analysis:

Table 1: Number of differential metabolites

group name	All sig diff	down regulated	up regulated
A_vs_B	174	53	121
C_vs_D	63	22	41

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 differentiating statistically significant differential metabolites between sample groups. The overall process is as follows:

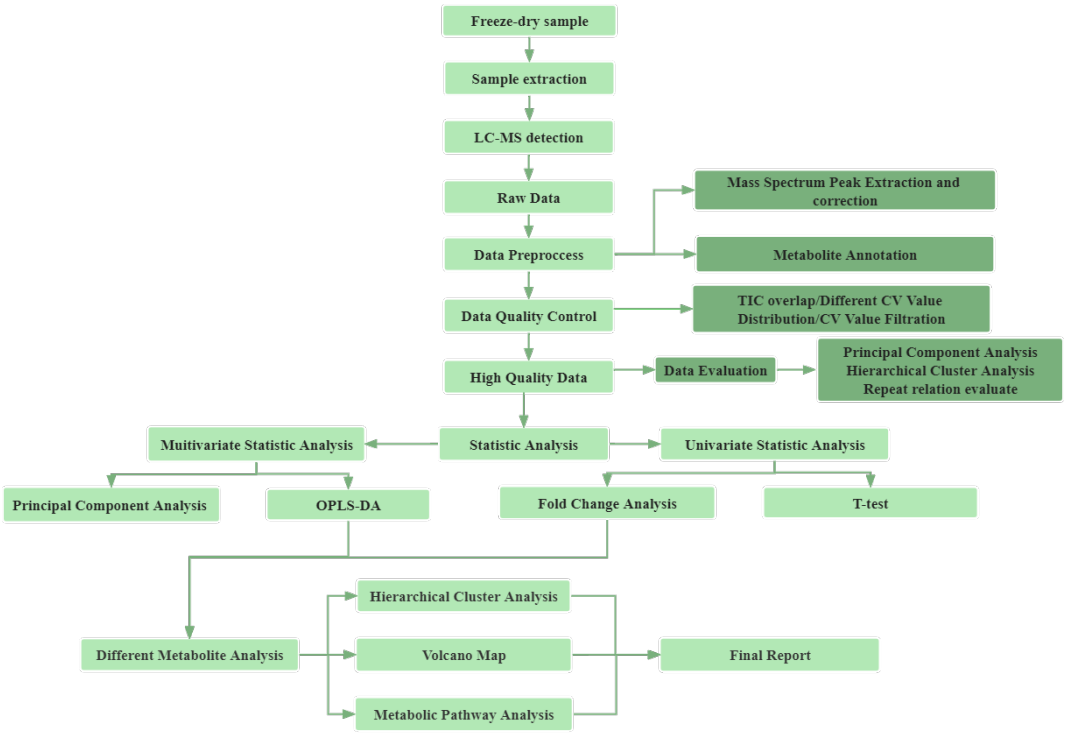


Fig 1: Flow chart of metabolomics analysis

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
-	-	CK-1	A
-	-	CK-2	A
-	-	CK-3	A
-	-	S1-1	B
-	-	S1-2	B
-	-	S1-3	B
-	-	S2-1	C
-	-	S2-2	C
-	-	S2-3	C
-	-	S3-1	D

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

2.3 Sample extraction process

2.3.1 Dry sample extraction

The samples were lyophilized using vacuum freeze-dryer and then ground in a ball mill grinder (30 Hz, 1.5 min) (MM 400, Retsch). 50 mg of the ground sample was mixed with 1200 μL of -20oC pre-cooled 70% methanol with internal standards. The mixture was mixed by vortex for 30 sec every 30 min for a total of 6 times, followed by centrifugation (12000 rpm, 3 min, 4oC). The supernatant was collected and filtered through a 0.22um filter membrane and kept for UPLC-MS/MS analysis.

2.4 Chromatography-mass spectrometry acquisition conditions

The data acquisition instruments consisted of Ultra Performance Liquid Chromatography (UPLC) (ExionLC™ AD, https://sciex.com/) and tandem mass spectrometry (MS/MS) (Applied Biosystems QTRAP 4500, https://sciex.com/).

Liquid phase conditions were as follows:

(1) Chromatographic column: Agilent SB-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) Elution gradient: 0.00 min, the proportion of B phase was 5 %, within 9.00 min, the proportion of B phase increased linearly to 95 %, and remained at 95 % for 1 min, 10.00-11.10 min, the proportion of B phase decreased to 5 %, and balanced at 5 % up to 14 min;

(4) Flow rate: 0.35 mL/min;

(5) Column temperature: 40 °C;

(6) Injection volume: 4 μL.

The mass spectrum conditions were as follows:

ESI (electrospray ionization) source temperature 550°C; Ion spray voltage (IS) 5500 V (positive ion mode) / -4500 V (negative ion mode); The ion source gas I (GSI), gas II (GSII) and curtain gas (CUR) were set to 50, 60 and 25 psi respectively, and the collision-induced ionization parameter was set to high. QQQ scan used MRM mode and collision gas (nitrogen) was set to medium. DP (declustering potential) and CE (collision energy) of each MRM ion pair were completed by further DP and CE optimization. A specific set of MRM ion pairs was monitored at each period based on the eluted metabolites in each period.

2.5 Qualitative and quantitative principles of metabolites

The metabolites were identified qualitatively based on their secondary spectrum information. Isotope signals, repeated signals containing K^+ ions, Na^+ ions and $NH4^+$ ions, as well as repeated signals of fragments of other substances with larger molecular weights were removed during analysis.

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 non-target ions. By selecting a particular fragment ion, quantification is more accurate and reproducible.

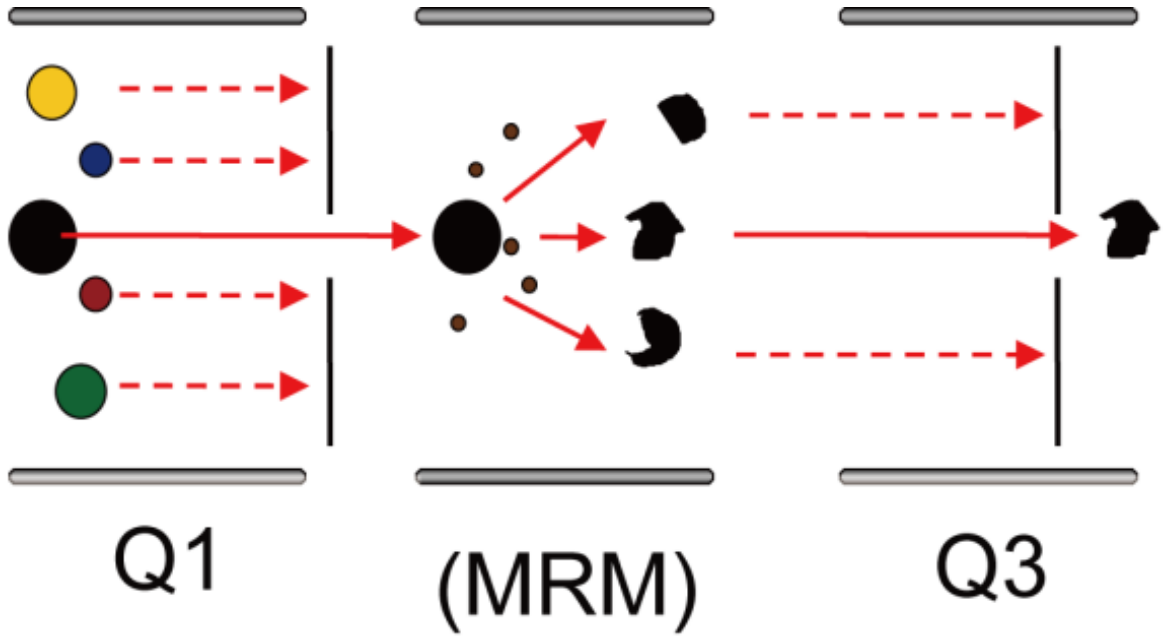


Fig 2:
Schematic diagram of multiple reaction monitoring mode by mass spectrometry

2.6 Data preprocessing

Based on the raw data file ALL_sample_data_raw.xlsx, the k-nearest neighbors algorithm (KNN) was first used to fill in the missing values, and then the CV value of the QC sample was calculated, and the metabolites with a CV value less than 0.5 were retained to obtain the final data file ALL_sample_data.xlsx.

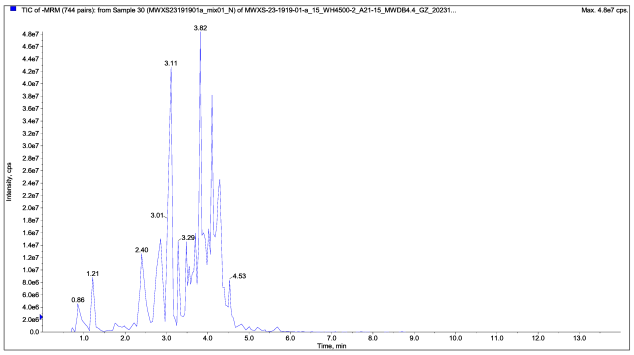
ALL_sample_data_raw.xlsx: Final report/1.Data_Assess/all_group/ALL_sample_data_raw.xlsx

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

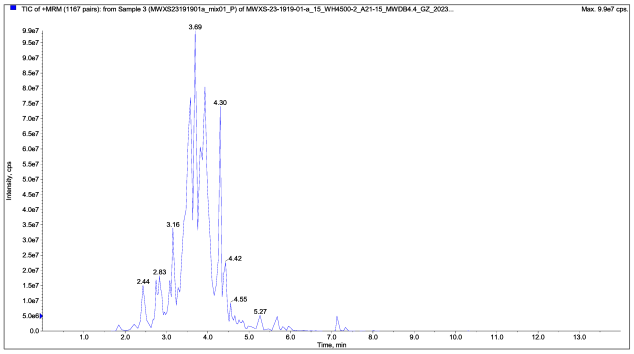
3 Data evaluation

3.1 Qualitative and quantitative analysis of metabolites

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) MWXS-23-1919-01-a_QC_MS_TIC-N

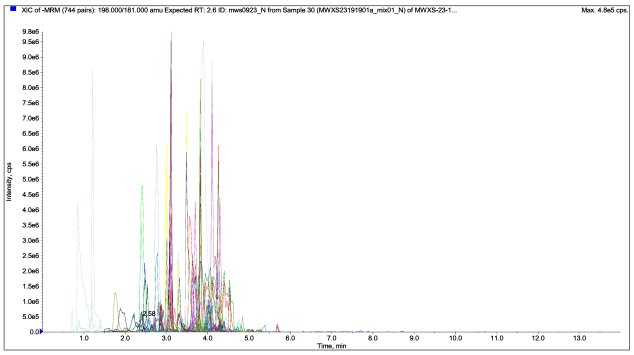


(b) MWXS-23-1919-01-a_QC_MS_TIC-P

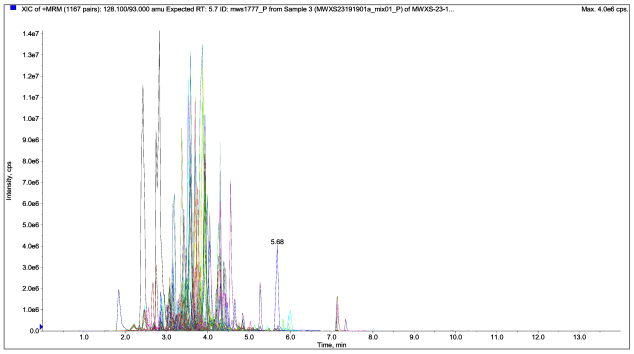
Fig 3: Total ion current diagram of mixed phase mass spectrum analysis

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) MWXS-23-1919-01-a_MRM_detection_of_multimodal_maps-N



(b) MWXS-23-1919-01-a_MRM_detection_of_multimodal_maps-P

Fig 4: Multi-peak diagram of MRM metabolite detection

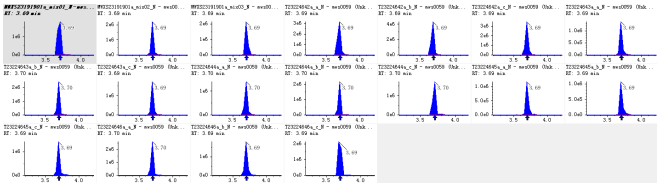
Note: N stands for negative ion mode, P for positive ion mode

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

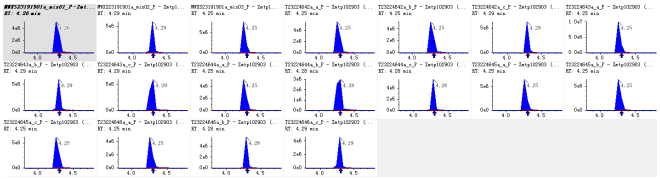
The MRM metabolite detection multi-peak diagram shows the compounds that were detected in the sample, with each mass spectrum peak color representing one detected metabolite. The characteristic ions of each compound were selected by triple quadrupole and measured for their signal intensity (CPS). The 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) MWXS-23-1919-01-a_Integral_correction_diagram-N



(b) MWXS-23-1919-01-a_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 4: Information of metabolite detected in sample

Index	CK-1	CK-2
ZBN0397	2686296.96	2421769.670
Lazn004839	189733.24	199418.524
Wbtp004753	326002.40	233521.248
Wbtn006715	438578.25	473758.633
Wbtp004961	315814.21	212573.596
Wbtp005759	780542.48	978694.225
Wbtn004362	158889.16	190737.419
Wbtn005483	39672.48	34274.560
Wbtn005631	23686.73	2636.967
Wbtn004861	278123.09	359014.780

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.

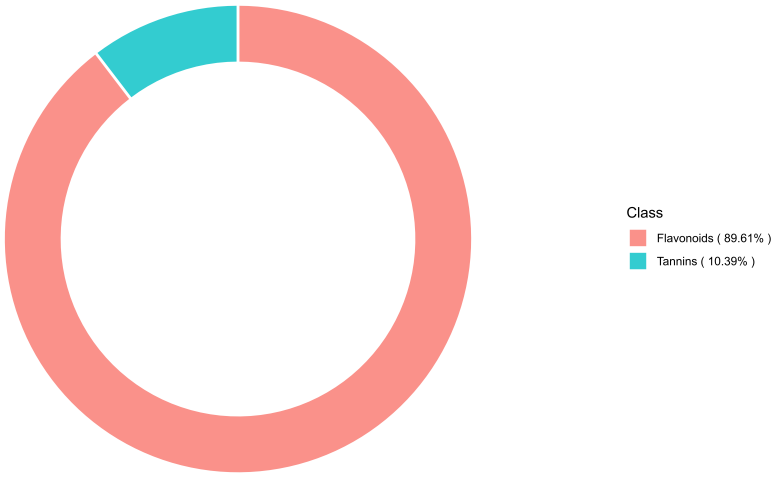


Fig 6: Ring plot of metabolite categories

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

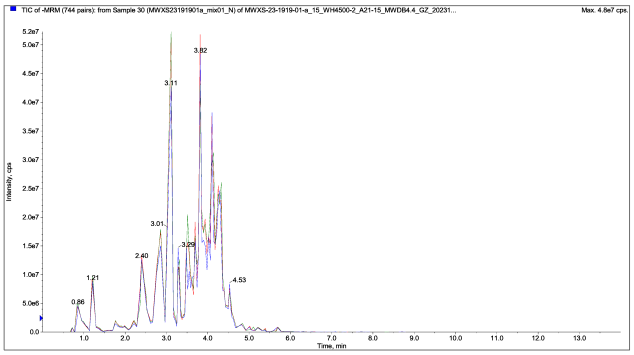
Ring plot of metabolite categories: Final report/1.Data_Assess/Class_Count/Class_Count_Ring.*

3.2 Quality control sample analysis

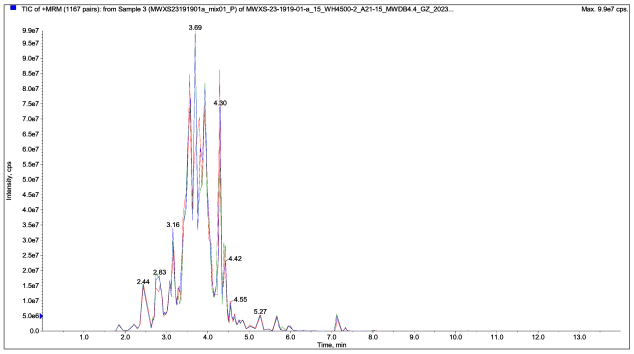
3.2.1 Total ion flow chromatogram

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) MWXS-23-1919-01-a_QC_MS_tic_overlap-N



(b) MWXS-23-1919-01-a_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*.*

3.2.2 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.

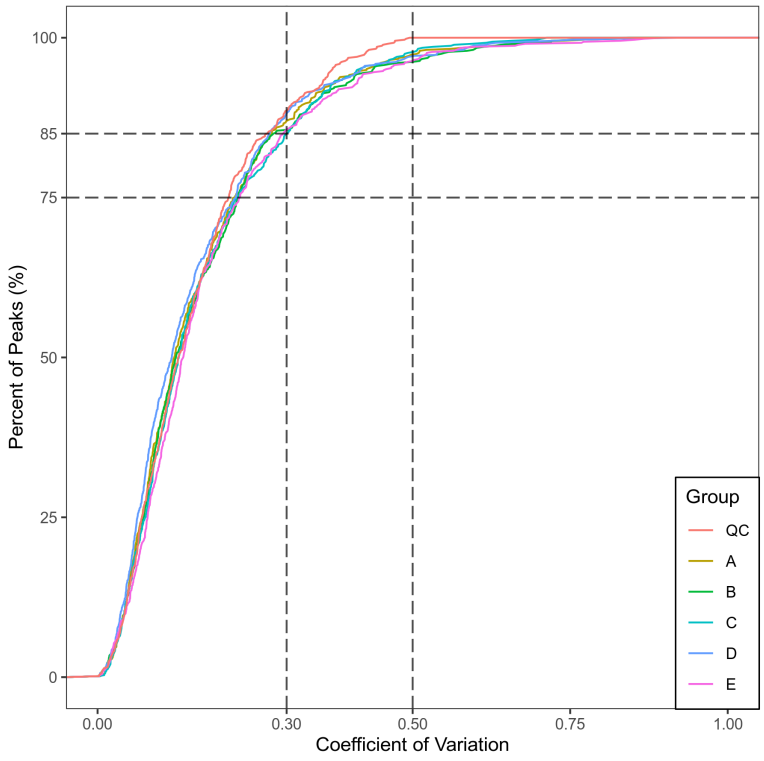


Fig 8: CV distribution of each group

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.*

3.3 Principal Component Analysis (PCA)

3.3.1 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 char-

acteristics 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.

3.3.2 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:

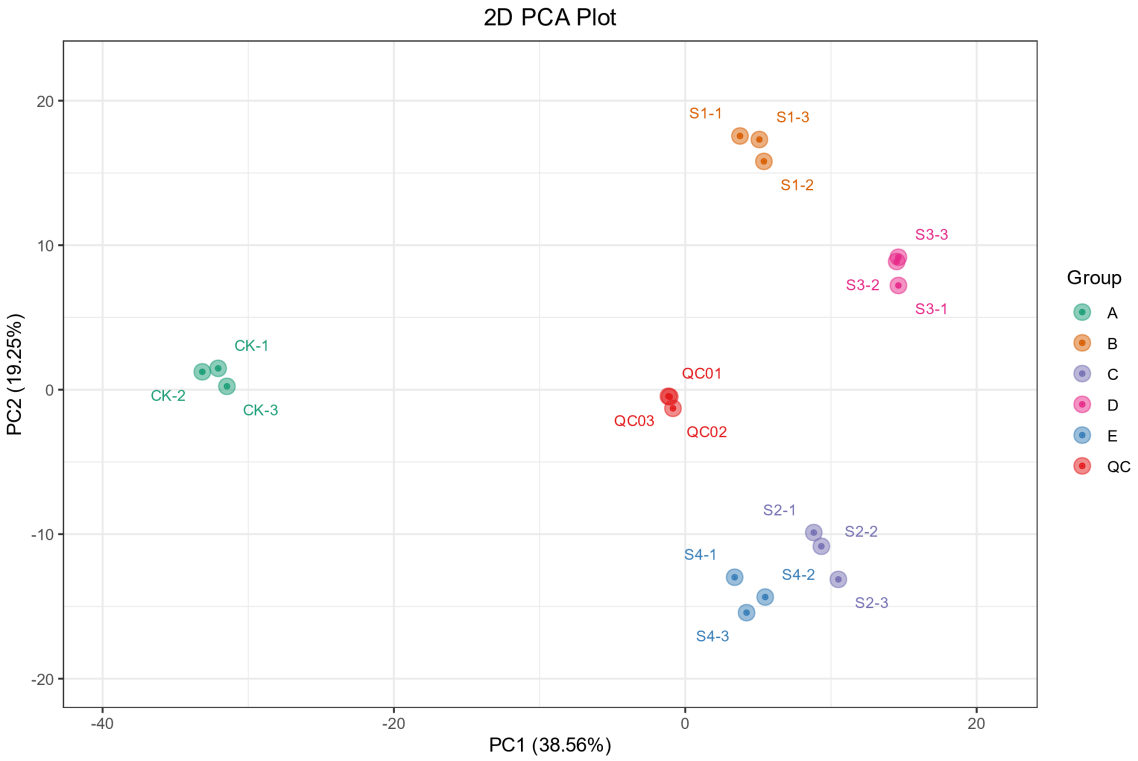


Fig 9: PCA score diagram of quality spectrum data of each group of samples and quality control samples

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/

3.3.3 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.

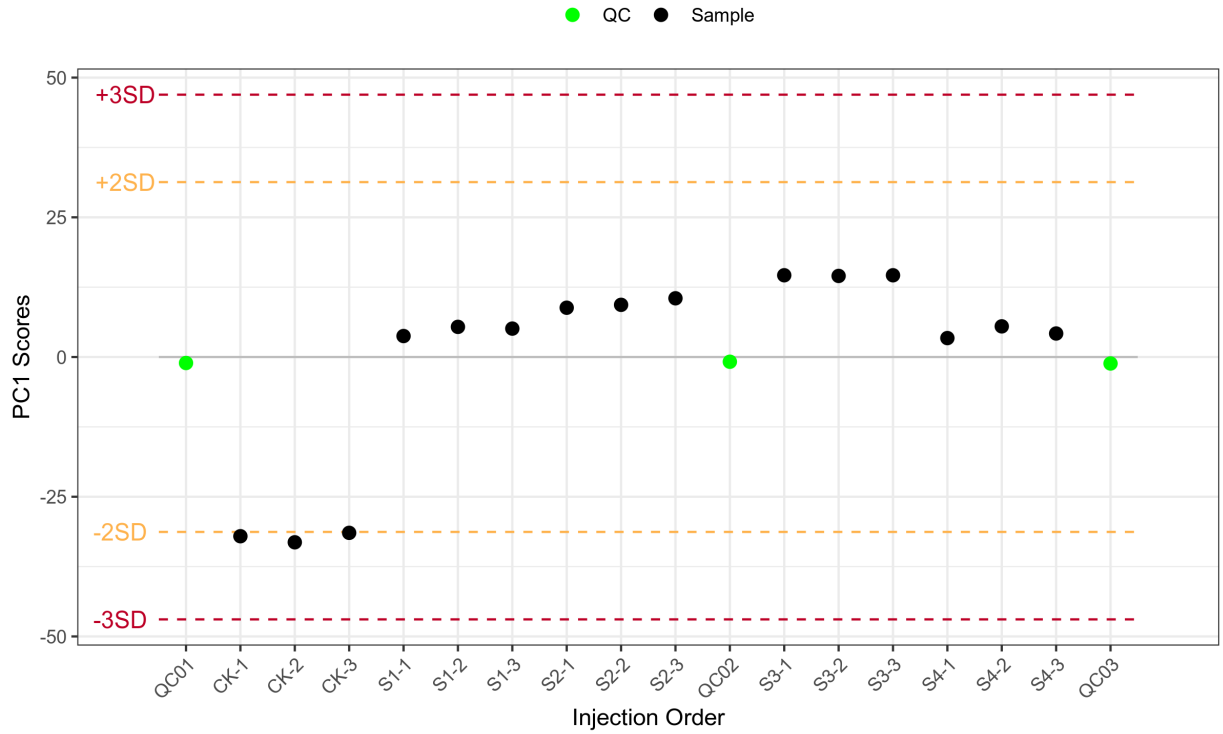


Fig 10: PC1 variation diagram of all the sample

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.*

3.4 Hierarchical Cluster Analysis (HCA)

3.4.1 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

3.4.2 Hierarchical Cluster Analysis results

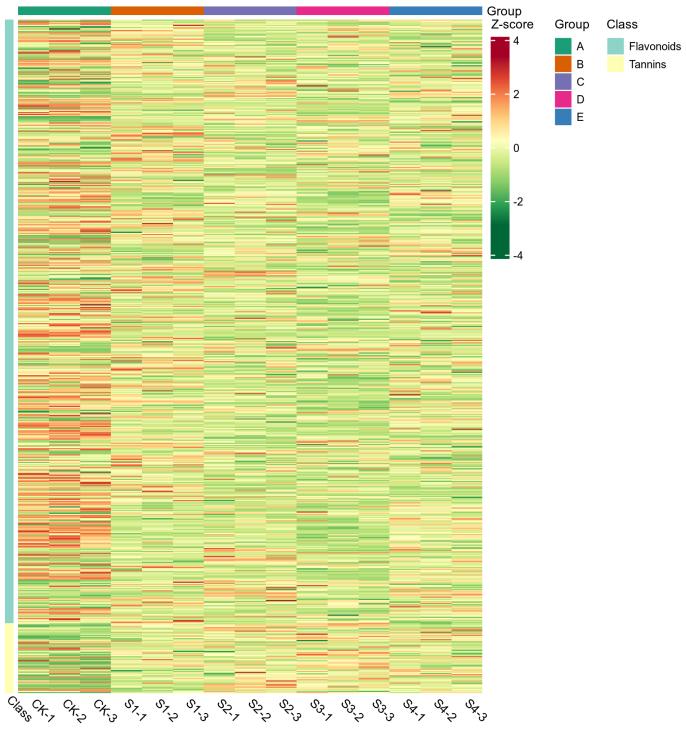


Fig 11: Sample clustering diagram

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/

3.5 Sample correlation assessment

Pearson's correlation analysis between samples are useful for examining the biological replicates in a group. The higher the correlation between samples in the same group compared with samples in a different

group, the more reliable the differential metabolite analysis. The cor function in R was used to calculate Pearson correlation between every pair of samples. The higher the correlation of QC samples (| r | closer to 1) means that the stability of the whole testing process is better and the data quality is higher. The correlation results can be seen in the figure below:

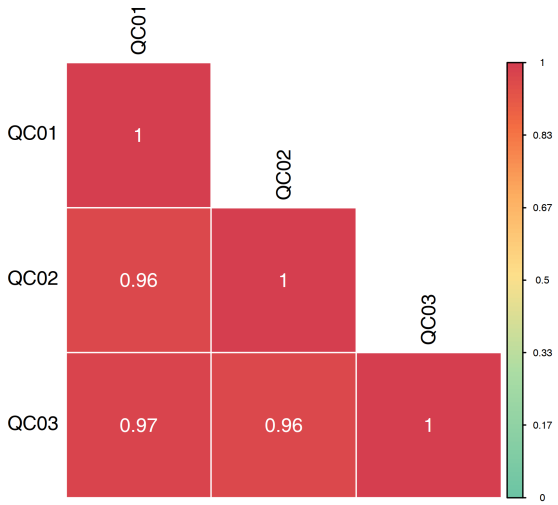


Fig 12: Correlation diagram between samples

Note: The labels on the left of each row and on the top of each column represent samples. Different colors represent different Pearson correlation coefficients. The darker the red, the stronger the correlation. The values in each box represent the correlation coefficients. "Correlation_expt" is the evaluation of the correlation between test samples; "correlation_QC" is the evaluation of the correlation between QC samples.

Sample correlation assessment: Final report/1.Data_Assess/correlation_analysis/

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.

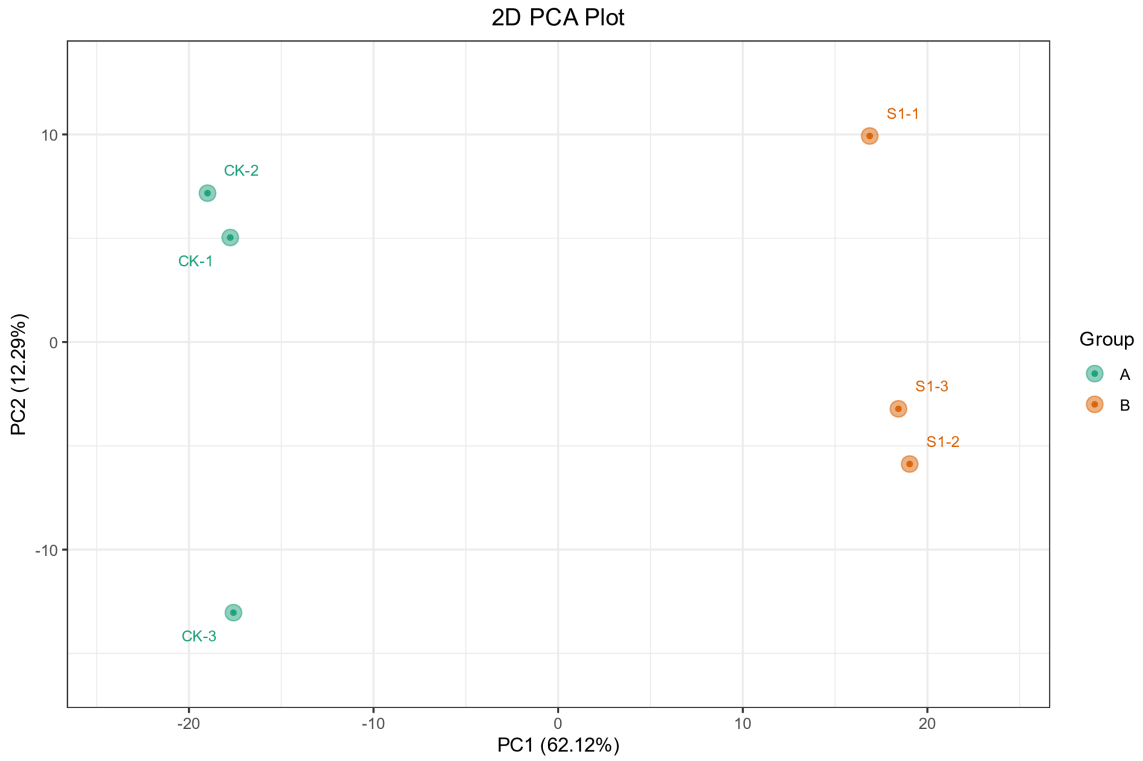


Fig 13: Principal component analysis of different groups

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

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

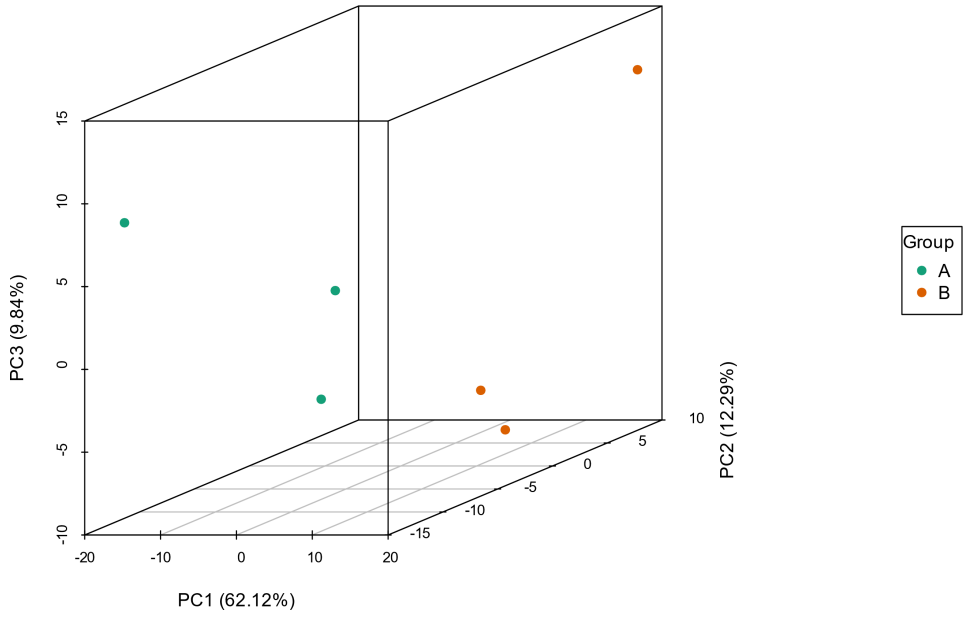


Fig 14: Three-dimensional PCA plot of different groups

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

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

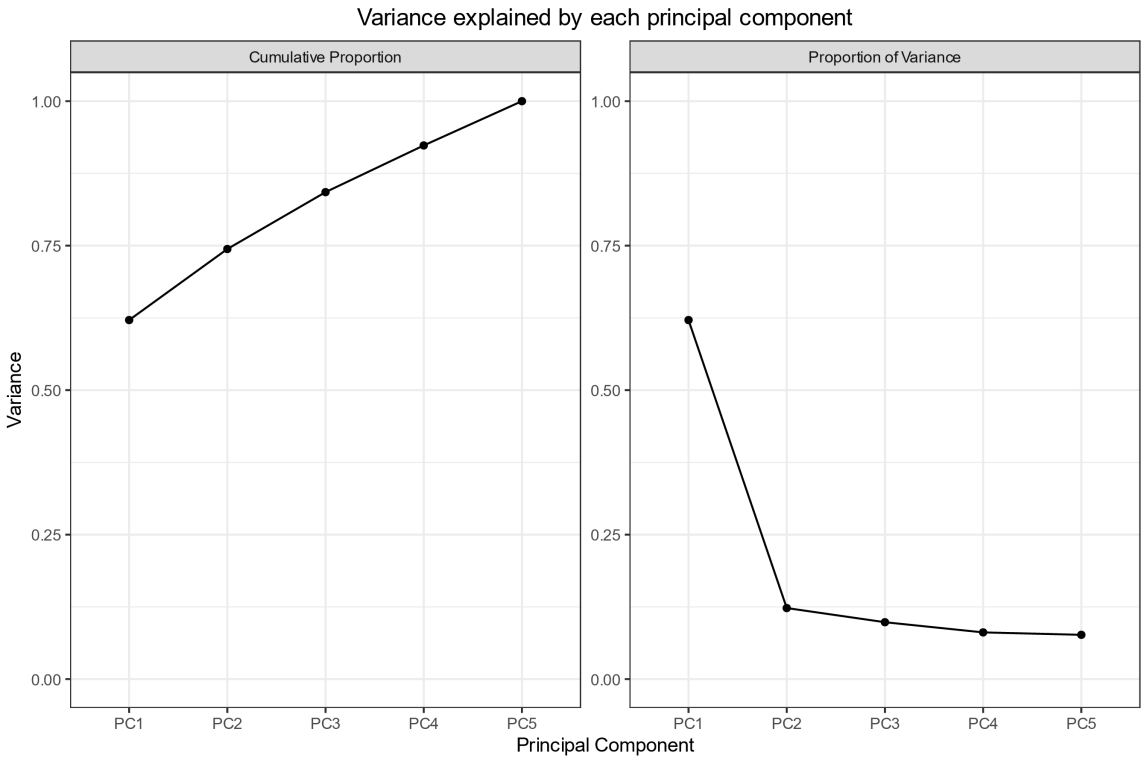


Fig 15: The explainable variation of the first five principal components

Note: the X-axis represents each principal component, the Y-axis represents the explainable variation, the left figure represents the cumulative explainable variation, and the right figure represents the explainable variation of each principal component

Principal component analysis of 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 5: Partial results of OPLS-DA

Index	Compounds	VIP
ZBN0397	(-)-Epicatechin-(4β->8)-(-)-epigallocatechin	1.1435258
Lazn004839	(2"'E,6"'S)-4"-(6-Hydroxy-2,6-dimethylocta-2,7-dienoyl)-vitexin	1.2552407
Wbtp004753	1,2,3,7,8-pentahydroxy-6-methylanthracene-9,10-dione*	1.1994801
Wbtn006715	1,2,4,5,8-pentahydroxy-6-methylanthracene-9,10-dione*	1.2593856
Wbtp004961	1,2,4,5-tetrahydroxy-7-(hydroxymethyl)anthracene-9,10-dione*	1.1293483
Wbtp005759	1,2,5,7,8-pentahydroxy-3-methylanthracene-9,10-dione*	0.3709564
Wbtn004362	1,3,6,7-tetrahydroxy-2-(3,4,5-trihydroxyoxan-2-yl)xanthen-9-one	1.2080352
Wbtn005483	1,3,6-trihydroxy-2,5,7-trimethoxyxanthen-9-one	0.9093782
Wbtn005631	1,3,7-trihydroxy-2-[3,4,5-trihydroxy-6-(hydroxymethyl)oxan-2-yl]xanthen-9-one	0.6721466
Wbtn004861	1,8-dihydroxy-2,6-dimethoxy-5-{[(2s,3r,4s,5s,6r)-3,4,5-trihydroxy-6-(hydroxymethyl)oxan-2-yl]oxy}xanthen-9-one	1.1764065

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).

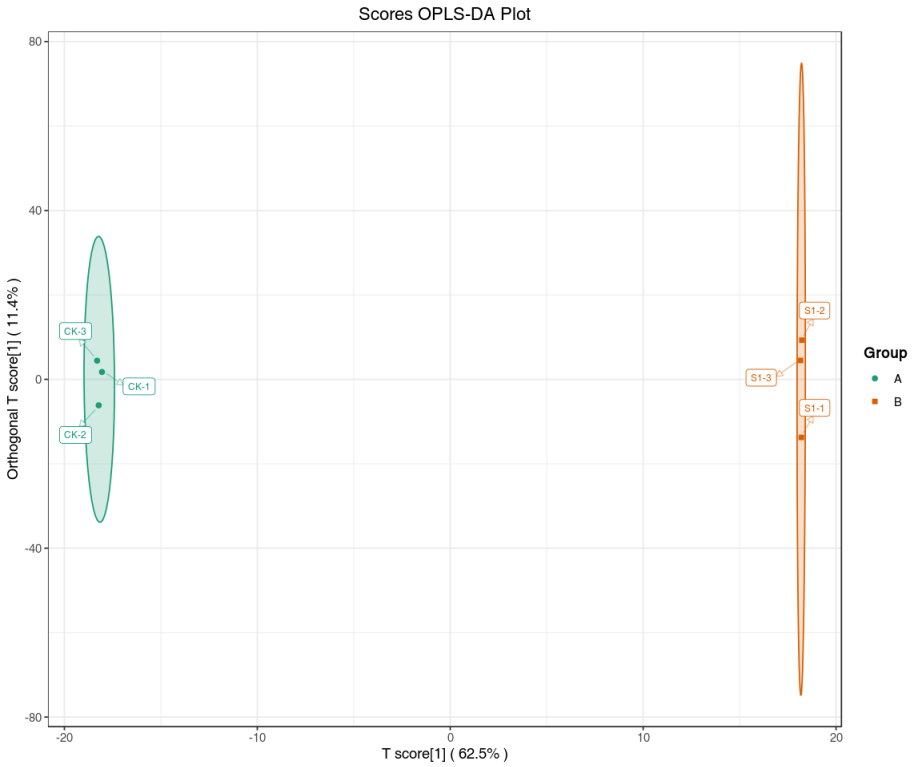


Fig 16: OPLS-DA score diagram

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.

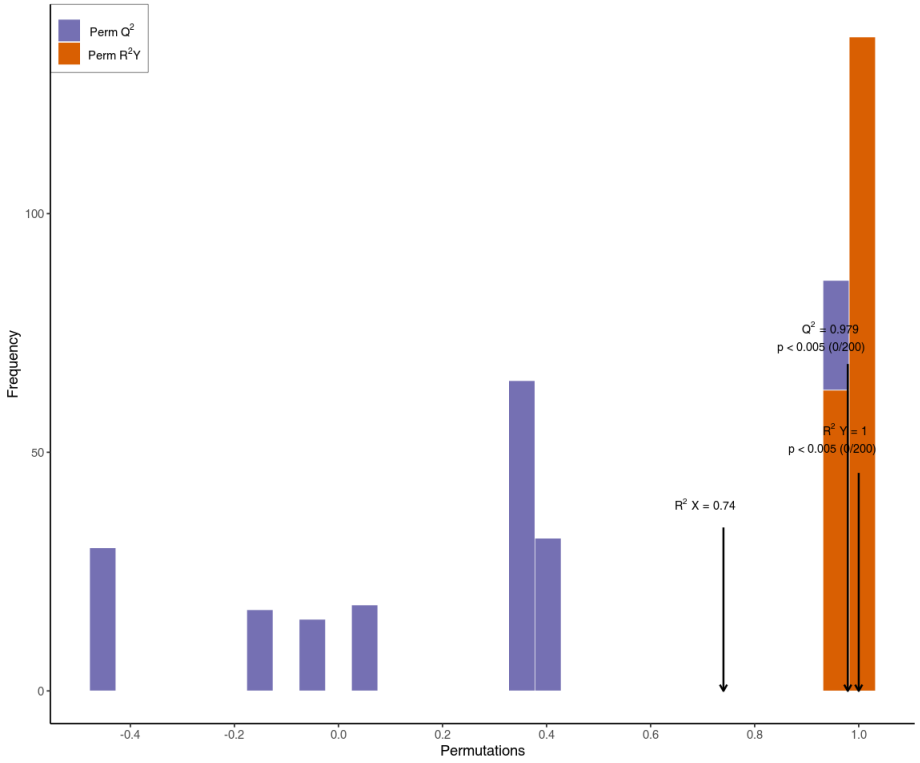


Fig 17: OPLS-DA verification diagram

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 18: 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.

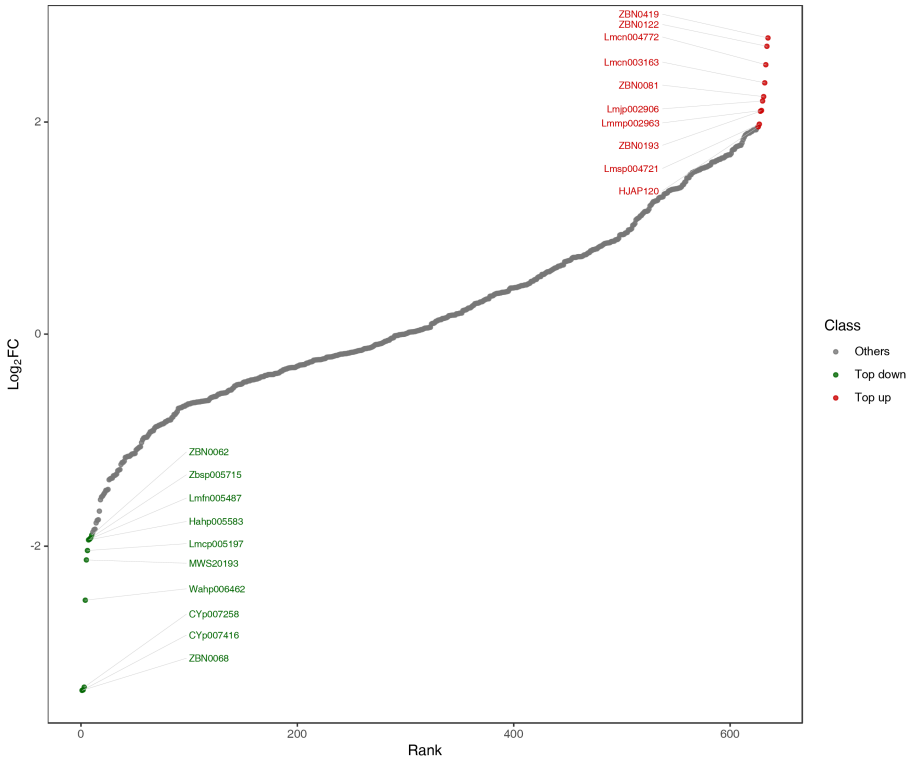


Fig 19: Dynamic distribution of metabolite content difference

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 fold change ≥ 2 or fold change ≤ 0.5 were considered as significant and selected.

Partial results from the screening criteria is shown below.

Table 6: Screening results of differential metabolites

Index	Compounds	Type
Lazn004839	(2"'E,6"'S)-4"-(6-Hydroxy-2,6-dimethylocta-2,7-dienoyl)-vitexin	down
Wbtp004753	1,2,3,7,8-pentahydroxy-6-methylanthracene-9,10-dione*	up
Zbsn002779	1-O-Galloyl-6-O-Luteoyl-Alpha-D-Glucose*	down
Wmmp000176	2,3-(s)-hexahydroxydibenzoyl glucose	down
Zbsp004450	3'-O-Methyltricetin-7-O-glucoside*	up
Zblp004717	3'-methoxyquercetin-3-O-L-rhamnosyl(1→2)-glucopyranoside*	up
Wayn007444	3,3'-Dimethylellagic acid 4'-sulfate	down
Lmjp004941	3,5,4'-Trihydroxy-7-methoxyflavone (Rhamnocitrin)*	up
Lmqn004838	3-O-Methylellagic acid	down
Lamn006359	4'-demethyl-3,9-dihydroeucomin glucoside	up

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.

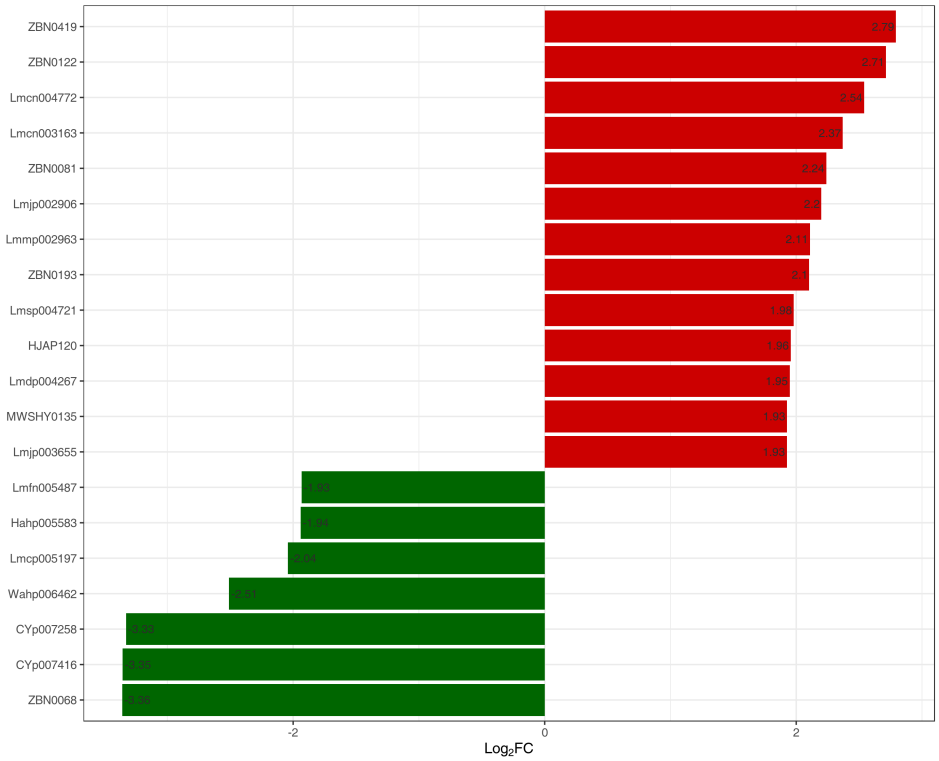


Fig 20: Bar chart of differential metabolites

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.

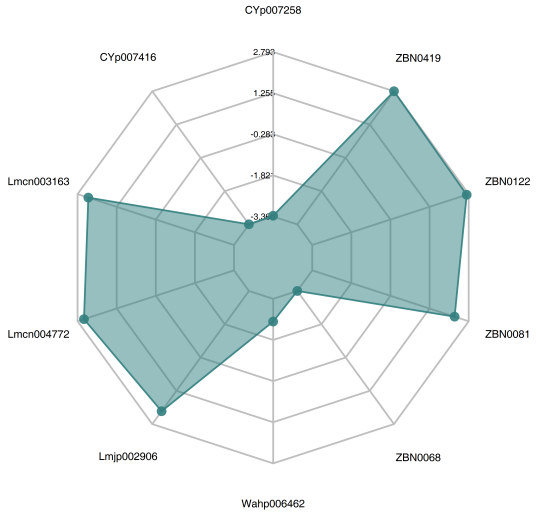


Fig 21: Radar map of differential metabolites

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.

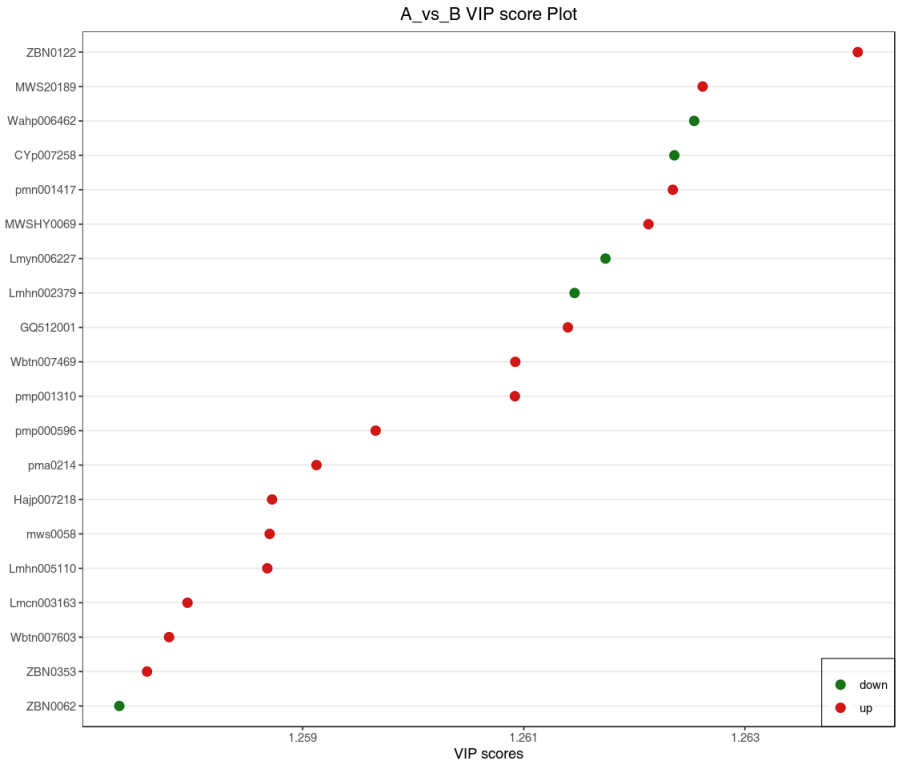


Fig 22: VIP values of differential metabolites

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.

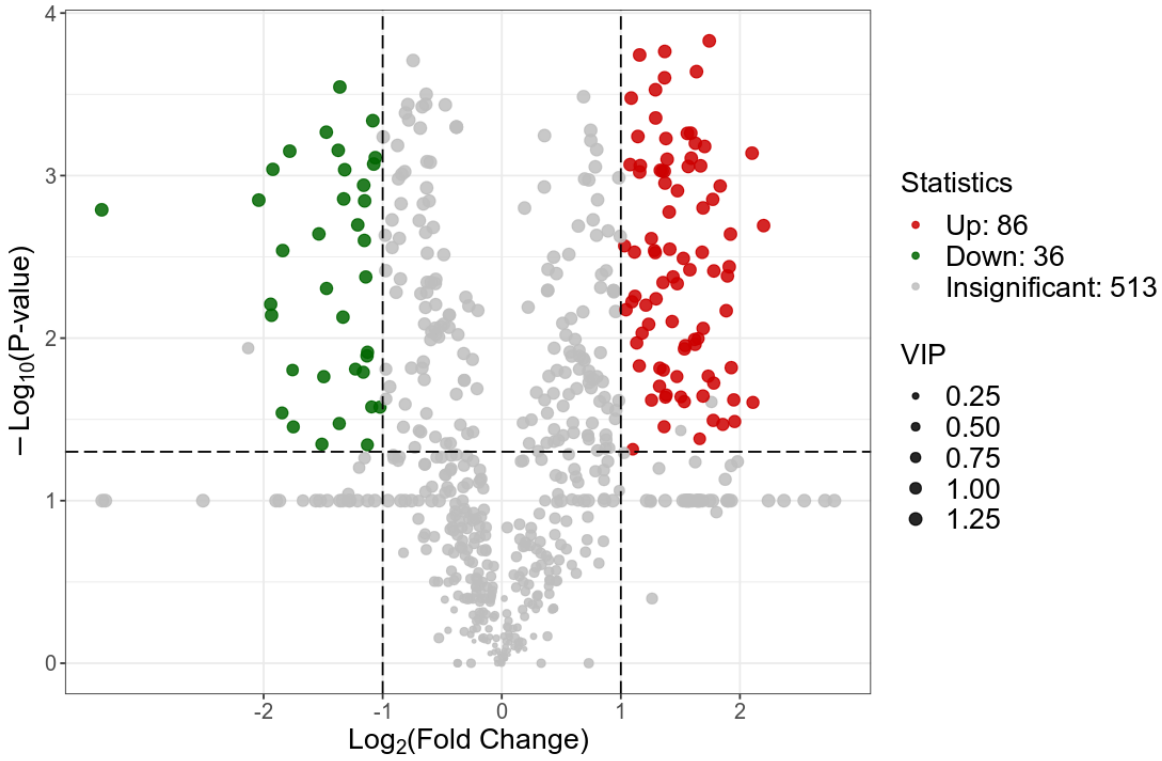


Fig 23: Volcano plot of differential metabolites

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, the Y-axis will represent VIP value. The larger the VIP value, the more significant the difference and the more reliable in the screening process.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.

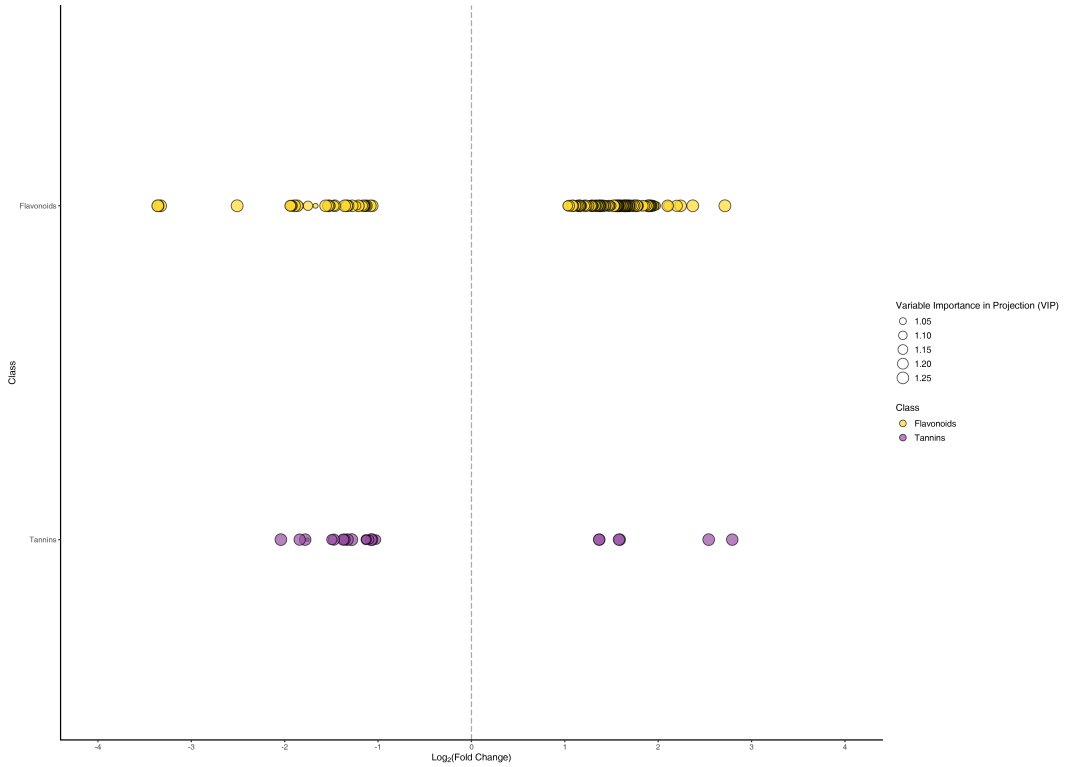


Fig 24: Scatter plot of differential metabolites

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.

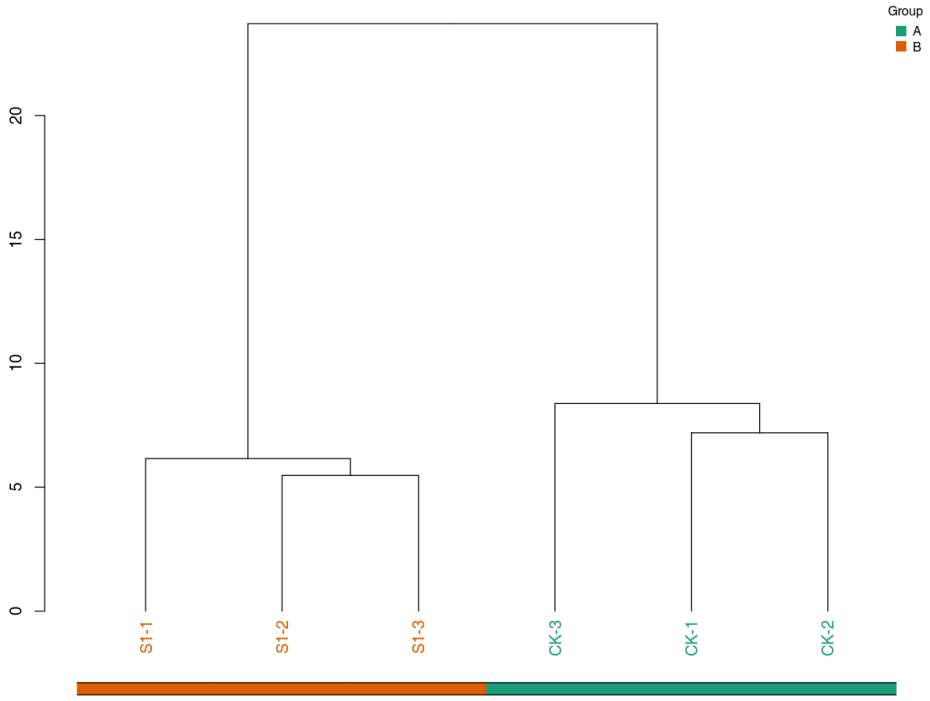


Fig 25: Hierarchical clustering tree

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.

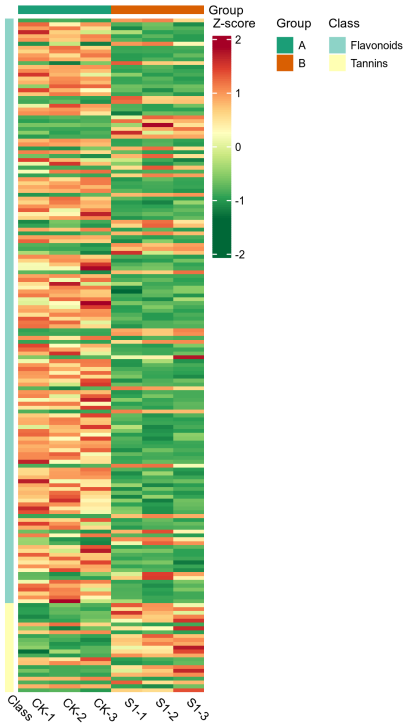


Fig 26: Heatmap of differential metabolites

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.

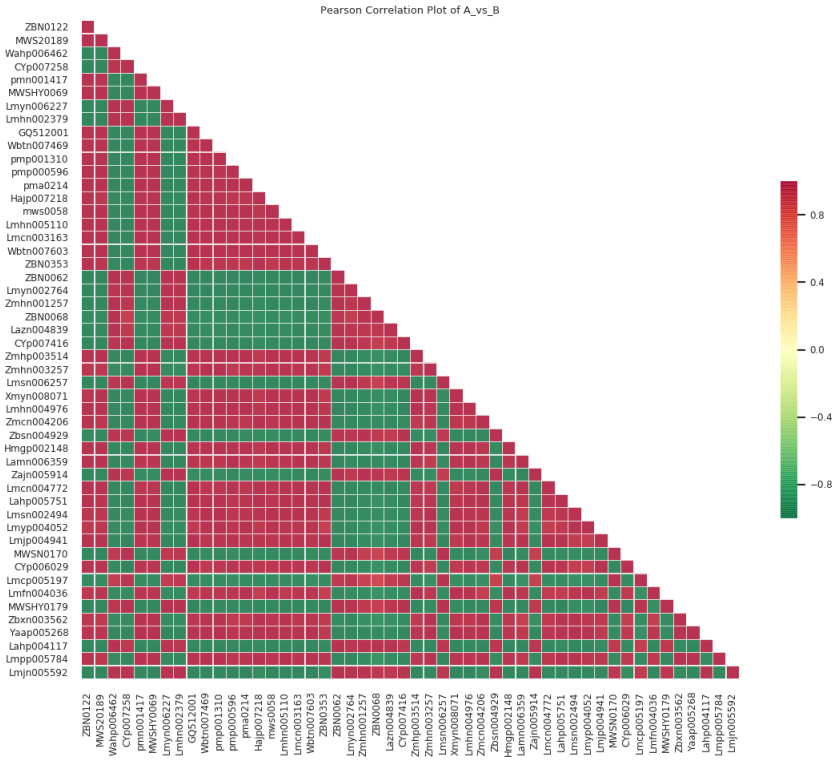


Fig 27: Heatmap of correlation of different metabolites

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_*.*

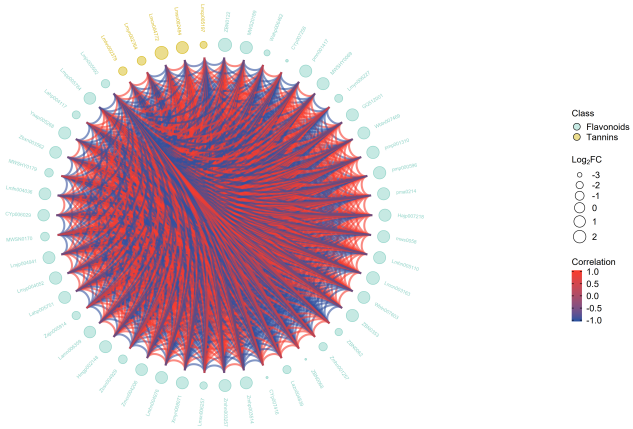


Fig 28: Chord diagram of differential metabolites

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.

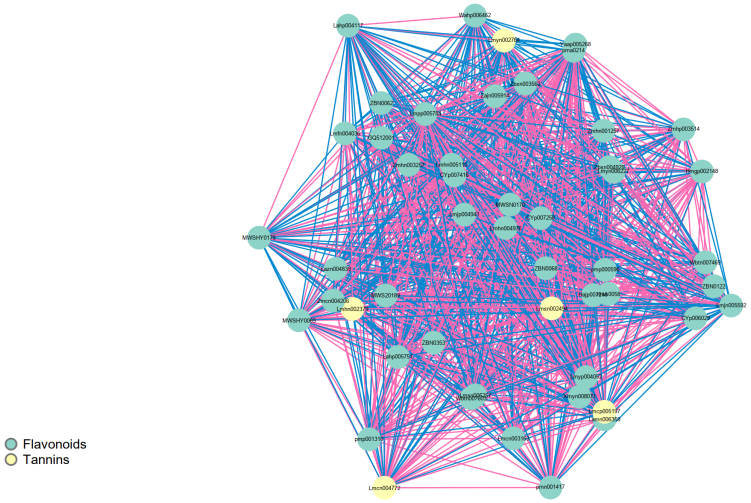


Fig 29: Correlation network diagram of differential metabolites

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.

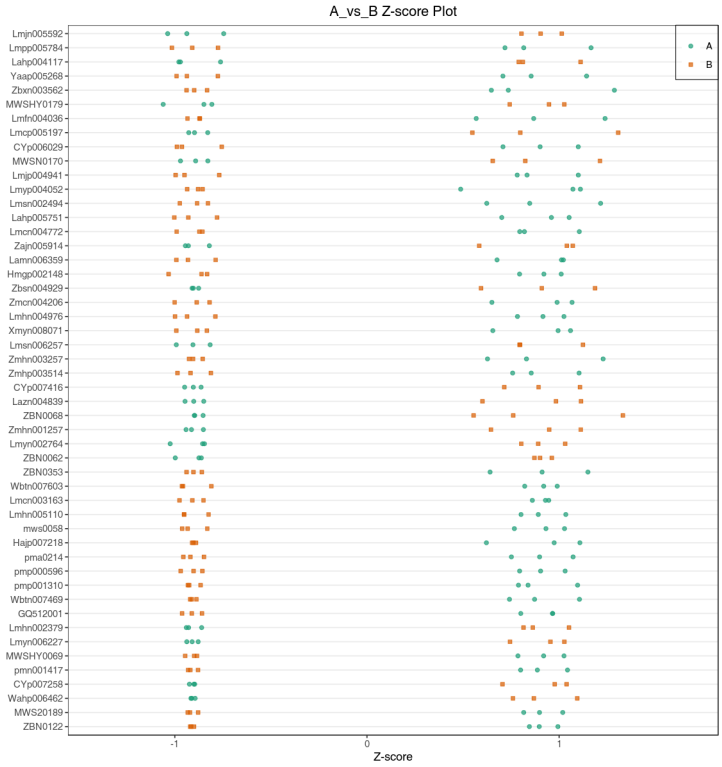


Fig 30: Z-value map of differential metabolites

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.

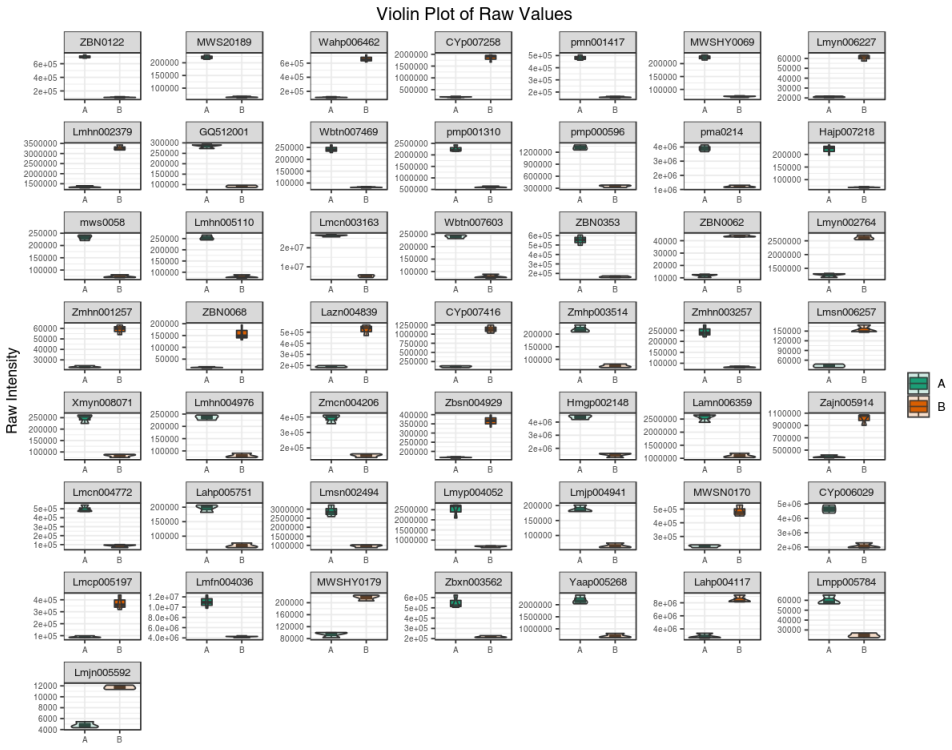


Fig 31: Violin plot of differential metabolites

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.

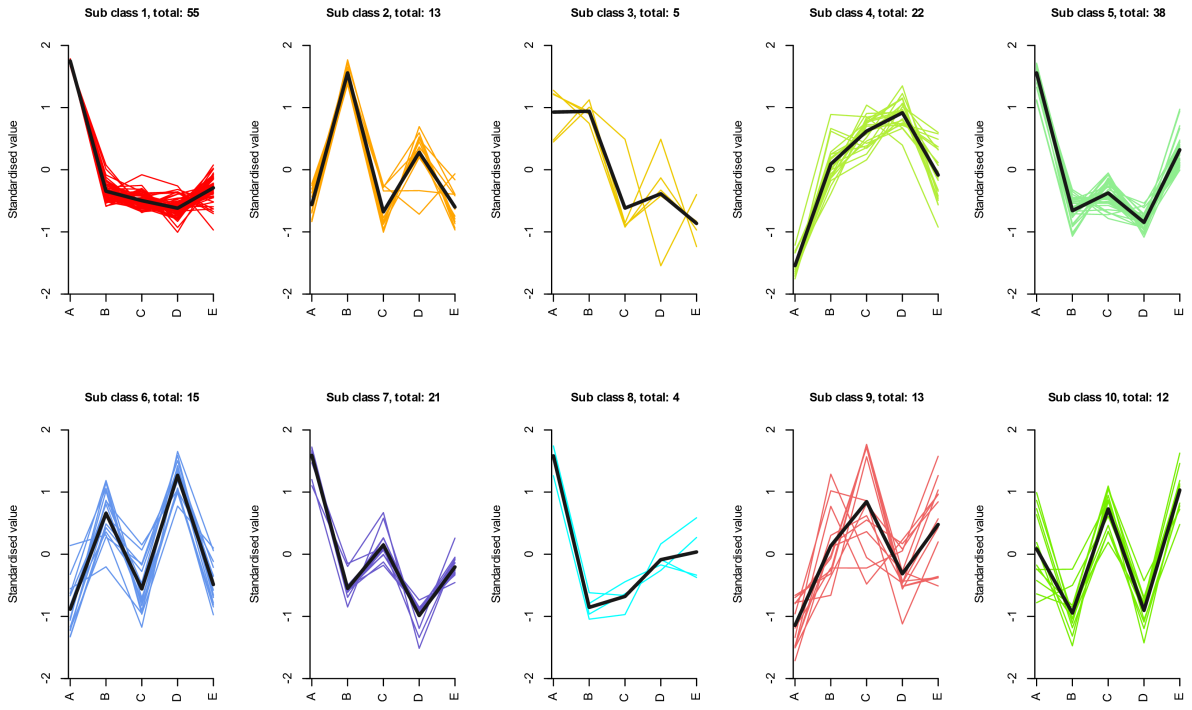


Fig 32: K-Means diagram of differential metabolites

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.

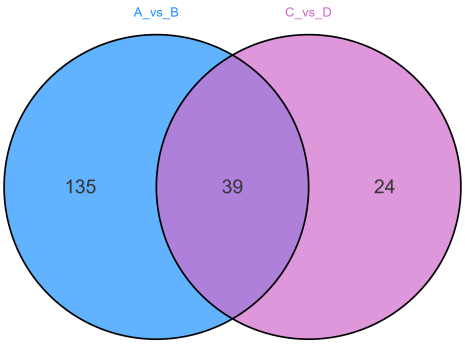


Fig 33: Venn diagram of differences among groups

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.

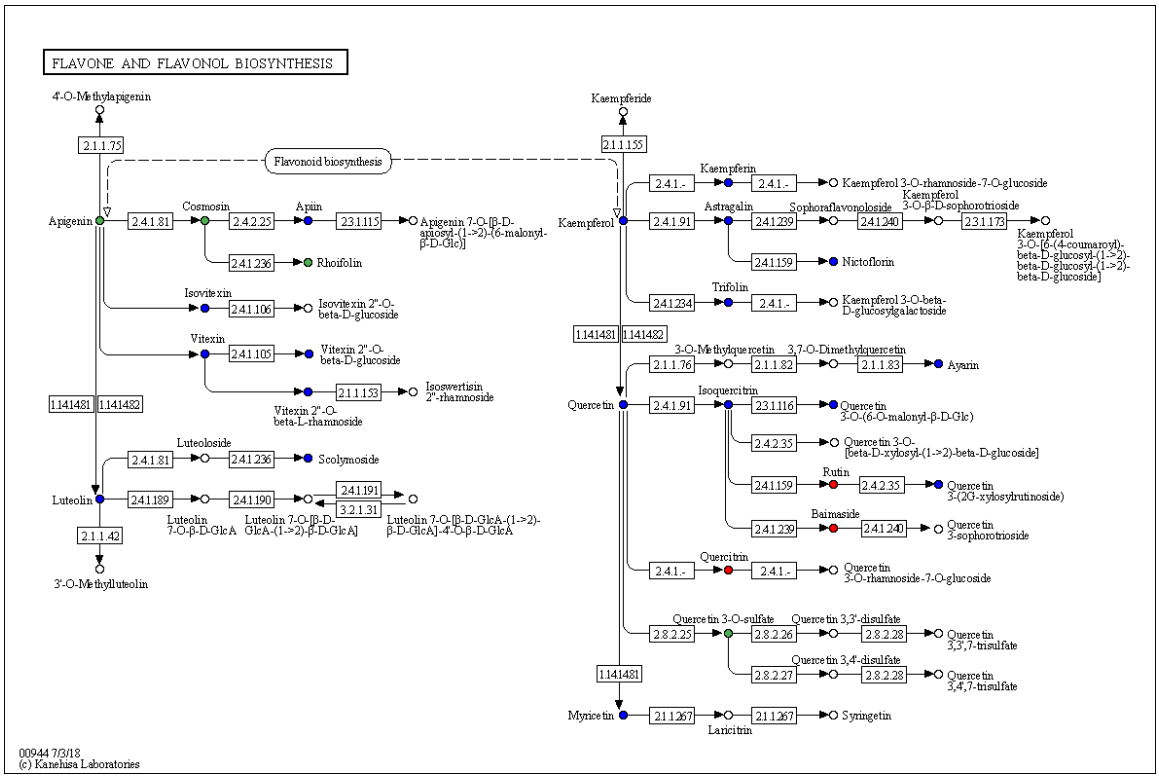


Fig 34: KEGG pathway with detected metabolites

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 7: KEGG annotations for differential metabolites

Index	Compounds	Type	cpd_ID
Lazn004839	(2"'E,6"'S)-4"-(6-Hydroxy-2,6-dimethylocta-2,7-dienoyl)-vitexin	down	-
Wbtp004753	1,2,3,7,8-pentahydroxy-6-methylanthracene-9,10-dione*	up	-
Zbsn002779	1-O-Galloyl-6-O-Luteoyl-Alpha-D-Glucose*	down	-
Wmmp000176	2,3-(s)-hexahydroxydibenzoyl glucose	down	-
Zbsp004450	3'-O-Methyltricetin-7-O-glucoside*	up	-
Zblp004717	3'-methoxyquercetin-3-O-L-rhamnosyl(1→2)-glucopyranoside*	up	-
Wayn007444	3,3'-Dimethylellagic acid 4'-sulfate	down	-
Lmjp004941	3,5,4'-Trihydroxy-7-methoxyflavone (Rhamnocitrin)*	up	C17059
Lmqn004838	3-O-Methylellagic acid	down	-
Lamn006359	4'-demethyl-3,9-dihydroeucomin glucoside	up	-

Table 8: Enrichment statistical of KEGG annotations for differential metabolites

ko_ID	Sig_compound	compound	Sig_compound_all	compound_all
ko00944	7	25	12	59
ko00941	4	29	12	59
ko00943	2	8	12	59
ko01100	3	16	12	59
ko01110	5	27	12	59
ko00942	1	5	12	59

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:

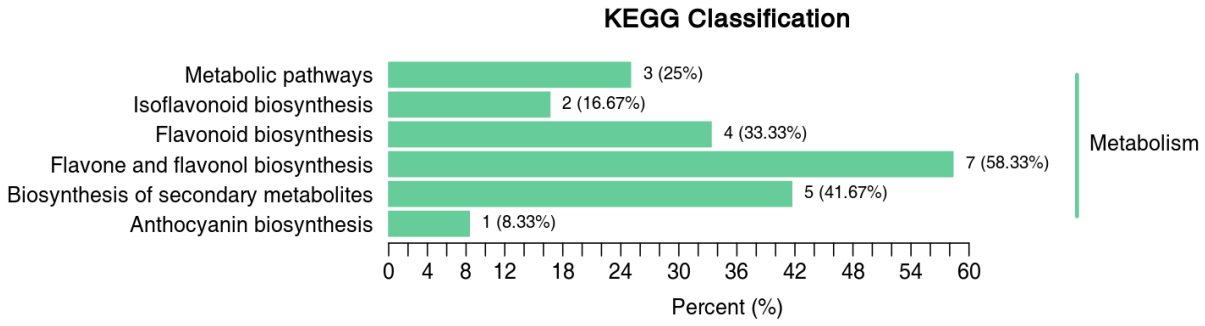


Fig 35: KEGG classification of differential metabolites

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.

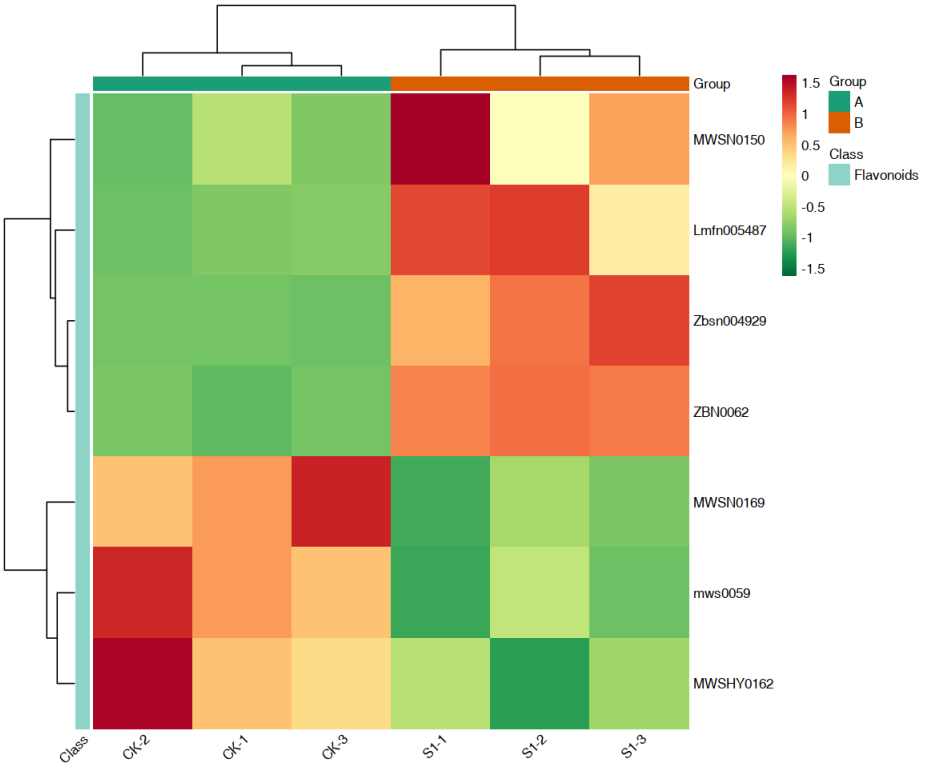


Fig 36: Clustering heat map of differential metabolites in KEGG pathway

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.

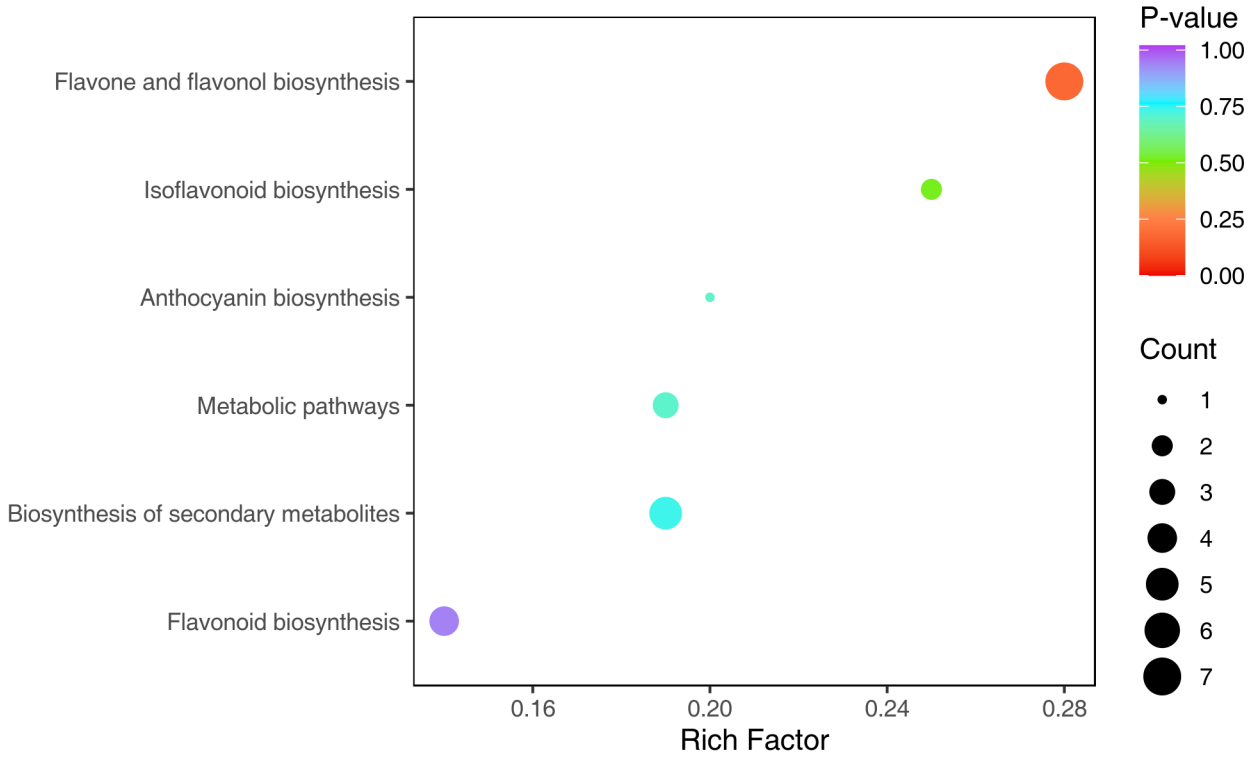


Fig 37: KEGG enrichment diagram of differential metabolites

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.

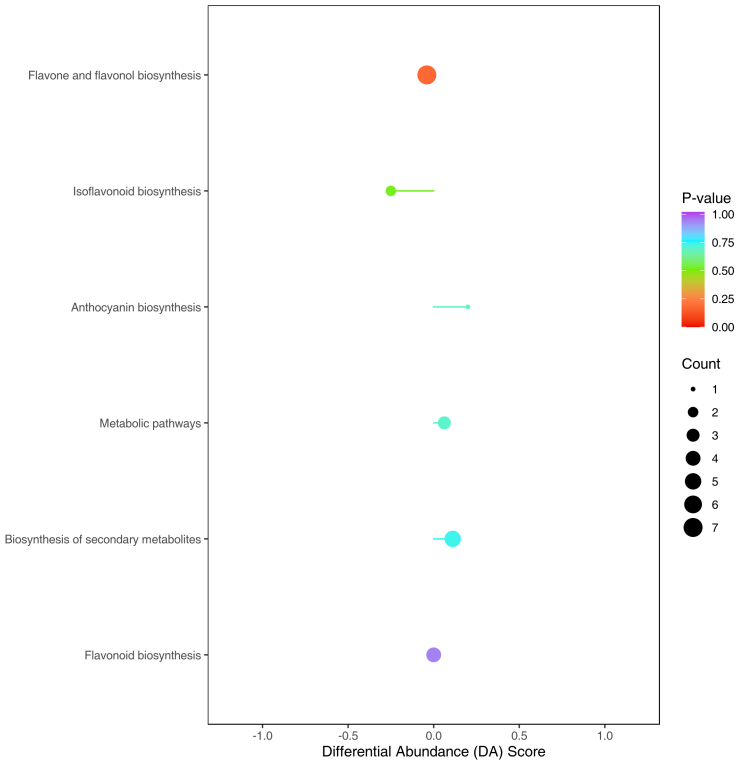


Fig 38: Difference abundance score

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

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. 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. 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
4. 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
5. 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. Journal of Proteome Research, 2015, 14(8):3322-35. https://dx.doi.org/10.1021/acs.jproteome.5b00354
6. 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/
7. 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
8. Gong L, Chen W, Gao Y, et al. Genetic analysis of the metabolome exemplified using a rice population. Proceedings of the National Academy of Sciences, 2013, 110(50): 20320-20325. https://www.pnas.org/content/110/50/20320.short
9. Ono E, Hatayama M, Isono Y, et al. Localization of a flavonoid biosynthetic polyphenol oxidase in vacuoles. The Plant Journal, 2006, 45(2): 133-143. https://pubmed.ncbi.nlm.nih.gov/16367960/
10. Shimada N, Aoki T, Sato S, et al. A cluster of genes encodes the two types of chalcone isomerase involved in the biosynthesis of general flavonoids and legume-specific 5-deoxy (iso) flavonoids in Lotus japonicus. Plant Physiology, 2003, 131(3): 941-951. https://pubmed.ncbi.nlm.nih.gov/12644647/

11. Nakayama T, Yonekura-Sakakibara K, Sato T, et al. Aureusidin synthase: a polyphenol oxidase homolog responsible for flower coloration. Science, 2000, 290(5494): 1163-1166. https://science.sciencemag.org/content/290/5494/1163

12. Wen W, Li D, Li X, et al. Metabolome-based genome-wide association study of maize kernel leads to novel biochemical insights. Nature communications, 2014, 5(1): 1-10. https://www.nature.com/articles/ncomms4438

13. Ono E, Fukuchi-Mizutani M, Nakamura N, et al. Yellow flowers generated by expression of the aurone biosynthetic pathway. Proceedings of the National Academy of Sciences, 2006, 103(29): 11075-11080. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1544175/

14. Unno H, Ichimaida F, Suzuki H, et al. Structural and mutational studies of anthocyanin malonyltransferases establish the features of BAHD enzyme catalysis. Journal of Biological Chemistry, 2007, 282(21): 15812-15822. https://pubmed.ncbi.nlm.nih.gov/17383962/

15. Ogata J, Kanno Y, Itoh Y, et al. Anthocyanin biosynthesis in roses. Nature, 2005, 435(7043): 757-758. https://doi.org/10.1038/nature435757a

16. Bowles D, Isayenkova J, Lim E K, et al. Glycosyltransferases: managers of small molecules. Current opinion in plant biology, 2005, 8(3): 254-263. https://pubmed.ncbi.nlm.nih.gov/15860422/

17. Sawada S, Suzuki H, Ichimaida F, et al. UDP-glucuronic acid: anthocyanin glucuronosyltransferase from red daisy (Bellis perennis) flowers: enzymology and phylogenetics of a novel glucuronosyltransferase involved in flower pigment biosynthesis. Journal of Biological Chemistry, 2005, 280(2): 899-906. https://pubmed.ncbi.nlm.nih.gov/15509561/

6 Appendix

6.1 Software list and version

Table 9: 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:

$$x' = x - \mu$$

6.2 FAQ

1.There are two TIC diagrams and two MRM diagrams, each with an "N" and a "P" designation. What do they mean?

A: "N" stands for negative-ion mode; "P" stands for positive-ion mode.

2.Where are the Pearson correlation coefficient data files?

A: Original file path

Final reports/1.Data_Assess/correlation_analysis/all_correlation_metabolites_pearson*;

There is also the result file from conducting Spearman correlation analysis.Original file path:

Final reports/1.Data_Assess/correlation_analysis/all_correlation_metabolites_spearman*.

3.How to interpret the values in All_sample_data? What is the quantification and the unit of measurement for a metabolite?

A: The data in the table uses scientific counting method, for example, 1.22e+02 means 1.22×10^2, which is also 122. If you are not used to it, you can change the number formatting. This value is the relative content of metabolites without units. It is calculated by calculating the peak area formed by the characteristic ions of each substance in the detector (although the absolute content of the substance cannot be quantified, the detection conditions are consistent, which can be used to compare the differences of the same substance in different samples)

4.How does the widely targeted metabolomics identify the metabolite? Can you provide the scoring value of the detected metabolite?

A: There are three levels of target identification. Level 1: concordance of MS/MS Spectrometry and retention time to metabolites in the database with a score of 0.7 or above; Level 2: concordance of MS/MS Spectrometry and retention time to metabolites in the database with a score between 0.5 and 0.7; Level 3: The Q1, Q3, RT, DP, and CE of the metabolite matches to a metabolite in the database.

There is no scoring value in the widely targeted metabolomics; Scoring only applicable to non-targted metabolomics. It can be understood from the perspective of detection methodology. Widely targeted metabolomics is based on MRM scanning of 5 parameters (DP, CE, RT, Q1, Q3) to detect the relative content of metabolites in different samples. You may refer to the literature on widely targeted metabolomics

methodology: 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. Molecular Plant, 2013, 6(6):1769-1780.

5.How are mass spectra peaks corrected?

A: The instrument automatically calibrates the mass axis according to the ions in the standard chemical equipped with AB Sciex.

6.Can you explain the concept of principal component analysis? In the principal component analysis, the greater the inter-group variation and the smaller the intra-group variation, the better? Is there a limit?

A: Assuming that the number of samples is N and the number of metabolites detected is M.

The raw data represents n points distributed in an m-dimensional space.

PCA is to use the least square method to find a straight line such that the distance between all samples and the straight line is the smallest. The linear direction also reflects the largest difference between samples.

On this basis, the next most significant line can be found in the orthogonal direction of the line, and so on.

The principal component diagram does not look at the percentage of the principal component, but whether the biological replicates within the group cluster well together, and whether the groups (different groups) separate well. This is mainly to judge the overall quality of the biological replicates within the group and the size of the difference between groups.

7.I want to focus on the content of A and B metabolites in the samples. When calculating the differences between samples, can I sum the value of these two metabolites together for comparison?

A: No. Widely targeted metabolomics obtains relative quantification. There are no methods to optimize for a single metabolite in the extraction method, and thus we cannot guarantee the complete extraction of specific metabolites in the sample. Furthermore, due to the physical and chemical characteristics of metabolites themselves, their ionization efficiency are not the same, resulting in the difference in final detection of signal. For example, two substances, such as A and B with 1 nmol each analyzed by LC – MS may show drastically different signal response values caused by the sensitivity of detection for different material, which in turn impacted by their chemical properties such as ionization efficiency and fragmentation. Therefore, the values between different metabolites cannot be added together or compared directly. Only the same metabolite from different samples can be used for comparison.

8.In screening for differential metabolites in a comparison of A_vs_B, the up-regulated and down-

regulated metabolites descdribes changes in which group?

A: In the process of differential metabolite screening, if the grouping information given is A_vs_B, the comparison calculation formula defaults to A/B. The upregulation of the final differential metabolite indicates that the metabolite has A high relative content in A and a low relative content in B.

9.Can you label the metabolite names on the heatmap?

A: Yes. However, due to the size of the image, such figure is not included in the report. Please check the actual data file:

Final reports/Basic_analysis/Difference_analysis/A_vs_B/heatmap/A_vs_B_heatmap_*_Compounds.*.

10.Where can I find the common and unique metabolites from the Venn diagram?

A: You can find the data in Basic_analysis/Venn/CF_vs_MT__QG_vs_WT_venn_result, If you open the Excel table, you can see that there are two TRUE/FALSE columns. Metabolites with two "TRUE" represent common metabolites. Metabolites with one TRUE and one FALSE indicate metabolite is present in only one group.

11.The number of annotated differential metabolites in KEGG pathway is inconsistent with the number of annotated differential metabolites in KEGG database?

A: Metabolites annotated by KEGG COMPOUND (numbered C*****) do not necessarily have KEGG PATHWAY annotations (ko*****)