Innovative Metabolomics Insights for Better Health

# Quantitative Lipidomics Report

Metware Biotechnology Inc.

www.metwarebio.com

# Contents

# Quantitative Lipidomics Report for XXX

## 1 Abstract

Lipidomics is a branch of metabolomics that focuses on detecting and quantifying lipids. Lipid metabolism is a major biological process and is topic of intense research, involving research in energy transport, intercellular signaling and regulation, and other essential processes in growth and development. About 70% of the compounds in plasma are lipids. For this project:

(1) 36 samples were selected and divided into 6 groups for lipidomics and 803 lipids were detected. We performed qualitative and quantitative analysis using UPLC-MS/MS detection platform and in-house compound database. Differential analysis was performed using MetwareBio's bioinformatics pipeline.

(2) Results of differential lipid analysis:

Table 1: Statistical table of differential lipids

| group name | All sig diff | down regulated | up regulated |
|---|---|---|---|
| A_vs_B | 447 | 242 | 205 |

Original file path/2.Basic_analysis/Difference_analysis/sigMetabolitesCount.xlsx;

## 2 The experimental process

Ultra-performance liquid chromatography-tandem mass spectrometry (UPLC-MS/MS) is often used for accurate qualitative and quantitative compound analysis. The main purpose of lipidomics is to detect and identify lipids with important biological significance that show statistically significant differences between biological samples. The overall lipidomics process at MetareBio can be seen as follows.

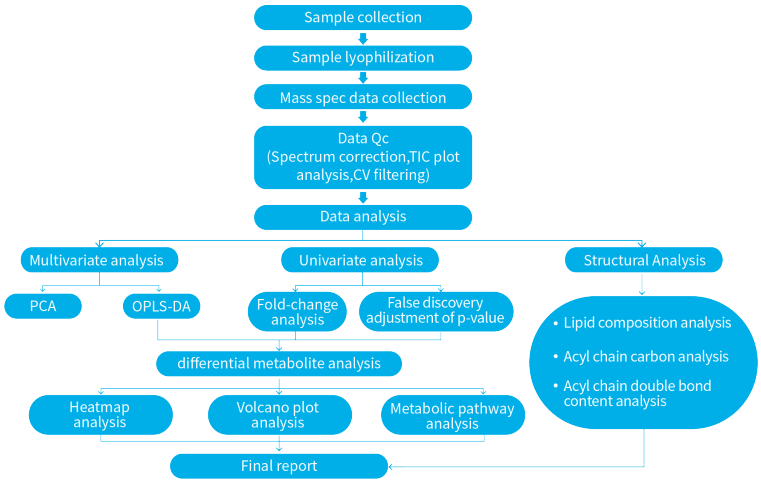


Fig 1: Flow chart of metabolomics analysis

## 2.1 Sample information

The samples in this project are grouped as follows:

Table 2: Sample information table

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

Original file path/0.data/sample_info.xlsx

## 2.2 Reagents and instruments

Table 3: Instrument information

| Instruments | Type | brand |
|---|---|---|
| LC-MS/MS | QTRAP 6500+ | SCIEX |
| Centrifuge | 5424R | Eppendorf |
| Electronic Balance | AS 60/220.R2 | RADWAG |
| Ball Mill | MM400 | Retsch |
| Centrifugal Concentrator | CentriVap | LABCONCO |
| Vortex Mixer | MI0101002 | Four E's |
| Ultrasonic Cleaner | CD-F15 | Olenyer |

Table 4: Information of standards and reagents

| Reagents | Level | Brand |
|---|---|---|
| Methanol | HPLC Grade | Thermo Fisher |
| Acetonitrile | HPLC Grade | Thermo Fisher |
| Formic acid | HPLC Grade | Sigma |
| Ammonium formate | Mass Spectrometry pure | Sigma |
| Isopropyl alcohol | HPLC Grade | Fisher |
| Methyl tert-butyl ether | HPLC Grade | Fisher |
| Standard | ＞ 99% | Avanti/zzstandard |

## 2.3 Sample extraction process

Take out the sample from the -80°C refrigerator and thaw it on ice. Weigh 20 mg of sample, then add 1mL of the extraction solvent (MTBE: MeOH =3:1, v/v) containing internal standard mixture. After whirling the mixture for 15 min, 200 µL of ultrapure water was added. Vortex for 1 min and centrifuge at 12,000 rpm for 10 min. 200 µL of the upper organic layer was collected and evaporated using a vacuum concentrator. The dry extract was dissolved in 200 µL reconstituted solution (ACN: IPA=1:1, v/v) to LC-MS/MS analysis.

## 2.4 Chromatography-mass spectrometry acquisition conditions

The data acquisition instruments consisted of Ultra Performance Liquid Chromatography (UPLC) (Nexera LC-40, https://www.shimadzu.com) and tandem mass spectrometry (MS/MS) (Triple Quad 6500+,https://sciex.com/ ).

Liquid phase conditions:

1) Chromatographic column: Thermo Accucore™C30 (2.6 µm, 2.1 mm×100 mm i.d.);
2) Mobile phase: A phase was acetonitrile /water (60/40, V/V) (0.1% formic acid added, 10 mmol/L ammonium formate); B phase was acetonitrile / Isopropyl alcohol (10/90, V/V) (0.1% formic acid added, 10 mmol/L ammonium formate);

3) Gradient program: 80:20(V/V) at 0 min, 70:30(V/V) at 2 min, 40:60(V/V) at 4 min , 15:85(V/V) at 9 min, 10:90(V/V) at 14 min, 5:95(V/V) at 15.5 min, 5:95(V/V) at 17.3 min, 80:20(V/V) at 17.5 min, 80:20(V/V) at 20 min;
4) Flow rate: 0.35 ml/min; Column temperature: 45°C; Injection volume: 2 μL.

Mass spectrometry conditions:

LIT and triple quadrupole (QQQ) scans were acquired on a triple quadrupole-linear ion trap mass spectrometer (QTRAP), QTRAP® 6500+ LC-MS/MS System, equipped with an ESI Turbo Ion-Spray interface, operating in positive and negative ion mode and controlled by Analyst 1.6.3 software (Sciex). The ESI source operation parameters were as follows: ion source, turbo spray; source temperature 500 °C; ion spray voltage (IS) 5500 V(Positive),-4500 V(Neagtive); Ion source gas 1 (GS1), gas 2 (GS2), curtain gas (CUR) were set at 45, 55, and 35 psi, respectively. Instrument tuning and mass calibration were performed with 10 and 100 μmol/L polypropylene glycol solutions in QQQ and LIT modes, respectively. QQQ scans were acquired as MRM experiments with collision gas (nitrogen) set to 5 psi. DP and CE for individual MRM transitions was done with further DP and CE optimization. A specific set of MRM transitions were monitored for each period according to the lipids eluted within this period.

## 2.5 Principles of lipid qualification and quantification

With our in-house database MWDB, lipids were annotated based on its retention time and ion-pair information from MRM mode. In MRM mode, the first quadrupole screens the precursor ions for target substance and excluded ions of other molecular weights. After ionization induced by the impact chamber, the precursor ions were fragmented, and a characteristic fragment ion was selected through the third quadrupole to exclude the interference of other non-target ions. By selecting a particular fragment, quantification is more accurate and reproducible.

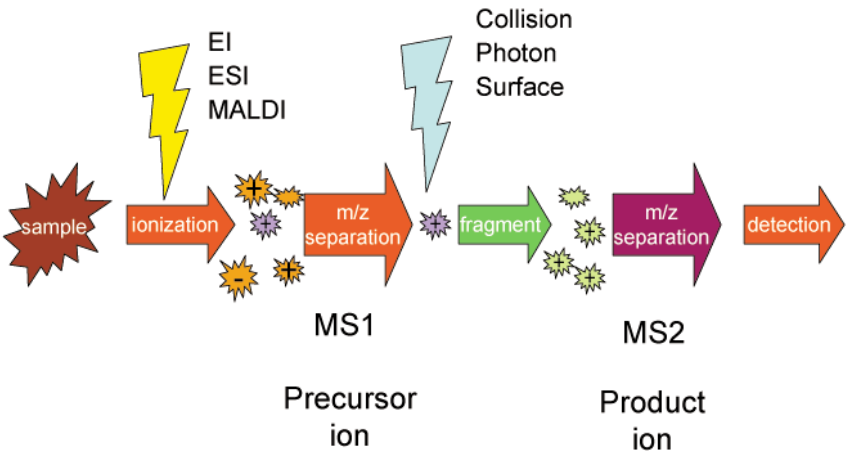


Fig 2:
Schematic diagram of multiple reaction monitoring mode by mass spectrometry

# 3 Evaluation of data results

## 3.1 Data pre-processing

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

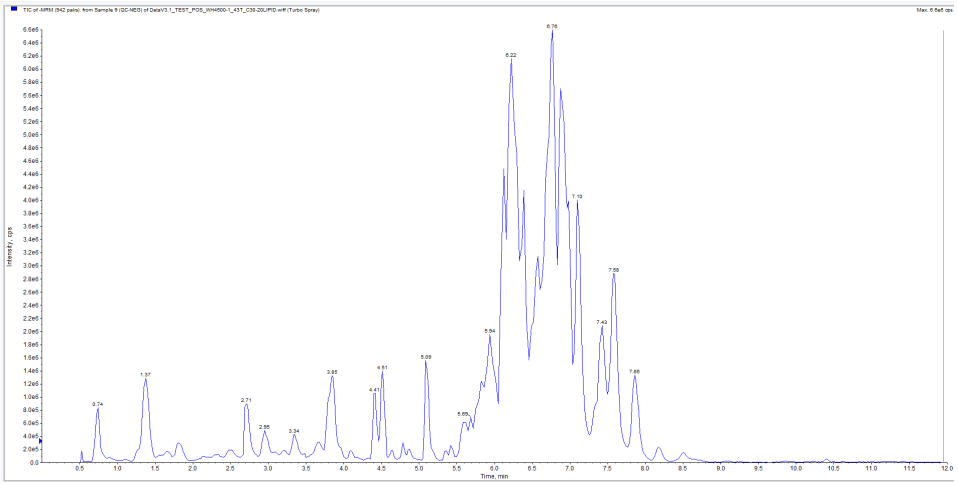

Fig 3: Total ion current diagram of one sample

Final Report/0.data/QC/*QC_MS_TIC.png

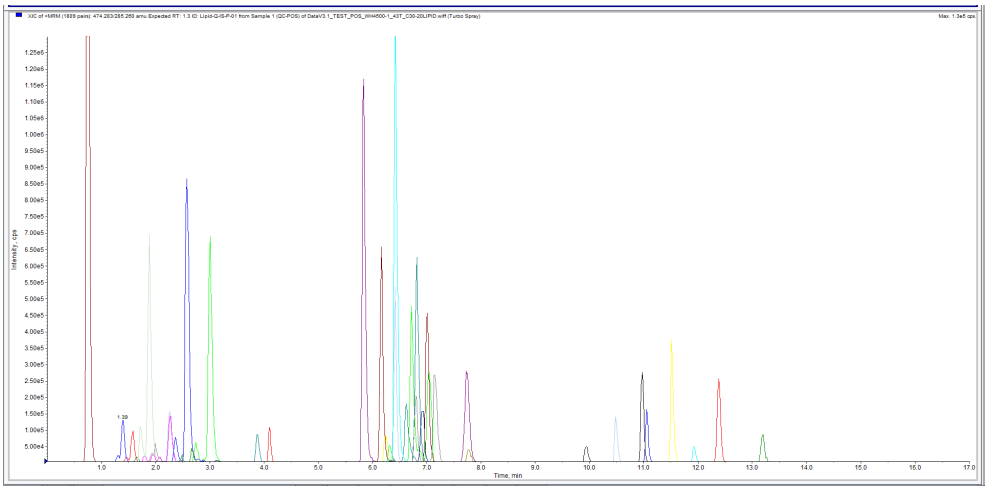

Fig 4: Multi-peak diagram of MRM lipid detection

Final Report/0.data/QC/*MRM_detection_of_multimodal_maps*

We corrected the mass spectrum peak of each lipid in different samples according to the retention time and peak distribution information to ensure the accuracy in the analysis. The following figure shows the correction results from a randomly selected lipid in different 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 lipid ion fragment.

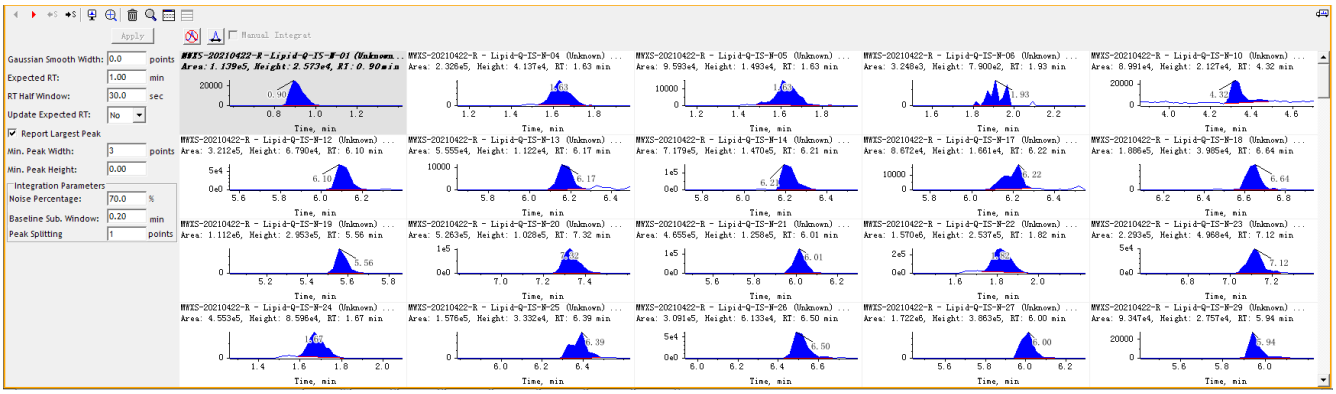

Fig 5: Scoring correction diagram for quantitative analysis of lipids
Note: The peak area represents the relative content of the substance in the sample.

Final Report/0.data/picture/*Integral_correction.png

### 3.1.1 Quantification Results

Quantification is calculated based on the calibration curve equation:

The lipid content in the sample was calculated by the formula: X = 0.001*R*c*F*V /m

X:Content of lipids in the sample (nmol/g);

R:The ratio of the peak area of the substance to be measured to the peak area of the internal standard (Area Ratio);

F:Internal standard correction factors for different types of substances;

c:Concentration of internal (μmol/L);

V:Extraction solution for samples (μL);

m:Weighed sample size (g);

The absolute quantitative results of a few samples in this project are shown in the following table.

Table 5: Statistical Table of lipid quantity

| Index | A2 | A3 |
|---|---|---|
| LIPID-N-0001 | 1.30845 | 0.929612 |
| LIPID-N-0010 | 0.719581 | 0.963375 |
| LIPID-N-0014 | 0.323411 | 0.675464 |
| LIPID-N-0015 | 0.758068 | 1.48059 |
| LIPID-N-0017 | 0.432447 | 0.271496 |
| LIPID-P-1600 | 22.5106 | 26.1192 |
| LIPID-P-1599 | 21.7128 | 14.4088 |
| LIPID-P-1598 | 27.8511 | 18.9142 |
| LIPID-P-1594 | 36.0775 | 34.4951 |
| LIPID-P-1593 | 6.62496 | 5.9616 |

Final Report/0.data/*level.xlsx

## 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 analyze the reproducibility of the entire lipidomics process. During data collection, one quality control sample was inserted for every 10 test samples.

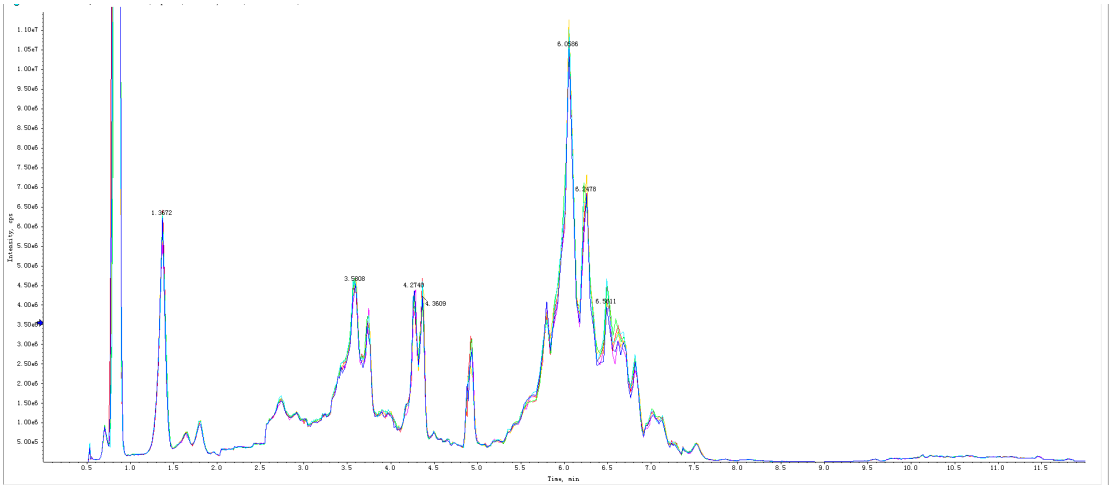

Fig 6: TIC overlap diagram detected by QC sample essence spectrum
Note: Superimposed spectrum from different QC samples. The results showed that the spectrum of total ion flow were highly consistent indicating that the signal stability was good when the same sample was detected at different times by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

Final report/0.data/picture/*QC_MS_tic_overlap*

### 3.2.2 QC Sample correlation assessment

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

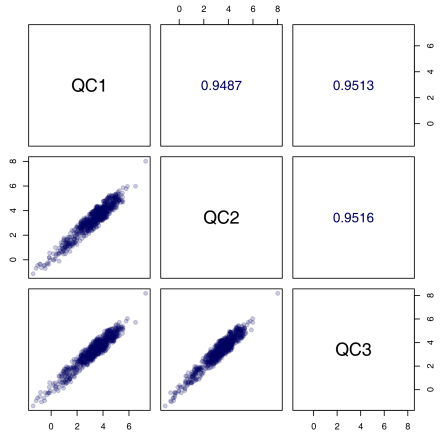


Fig 7: Correlation diagram between QC samples

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

Final report/1.Data_Assess/pcc/*mix*

### 3.2.3 CV value distribution of all samples

The Coefficient of Variation (CV) value is the ratio between the standard deviation of the original data and the mean of the original data, which can reflect the degree of data dispersion. The Empirical Cumulative Distribution Function (ECDF) 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.

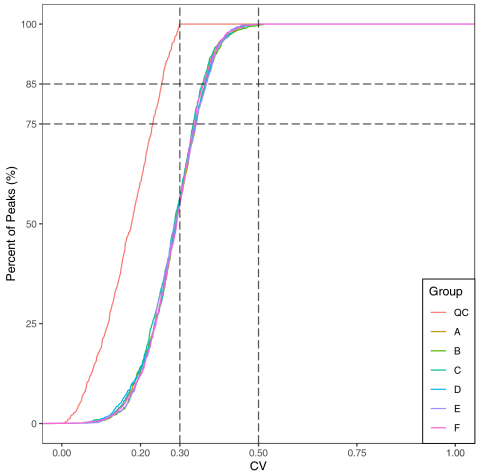


Fig 8: CV distribution of each group

Note: the X-axis represents the CV value, the Y-axis represents the proportion of compounds. 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%.

Final report/1.Data_Assess/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 characteristics of the metabolic spectrum. Among them, Principal Component Analysis (PCA) is an unsupervised pattern recognition method for statistical analysis of multidimensional data. Through orthogonal transformation, a group of variables that may be correlated are converted into a group of linear unrelated variables that are called principal components. This method is used to study how a few principal components may reveal the internal structure of between multiple variables, while keeping the original variable information (Eriksson et al., 2006). The first principal component (PC1) represents the most variable features in the multidimensional data matrix, PC2 represents the second most variable feature in the data, and so on. The prcomp function of R software (www.r-project.org/) was used with parameter scale=True indicating unit variance Scaling (UV) for normalizing the data. See appendix for details of PCA calculation.

### 3.3.2 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:

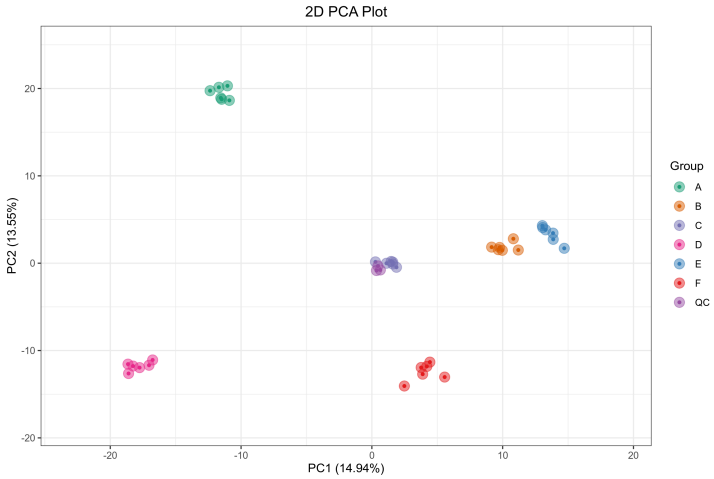


Fig 9: PCA score

diagram of quality spectrum data of each group of samples and quality control sample 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.

Final report/1.Data_Assess/*all_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.

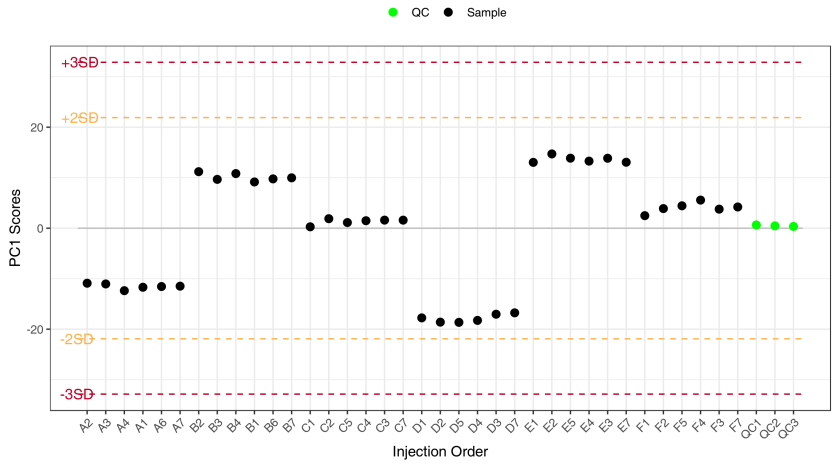


Fig 10: PC1 control diagram of population sample

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

Final report/1.Data_Assess/pca/*PC1_QCC*

## 3.4 Hierarchical Cluster Analysis

### 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

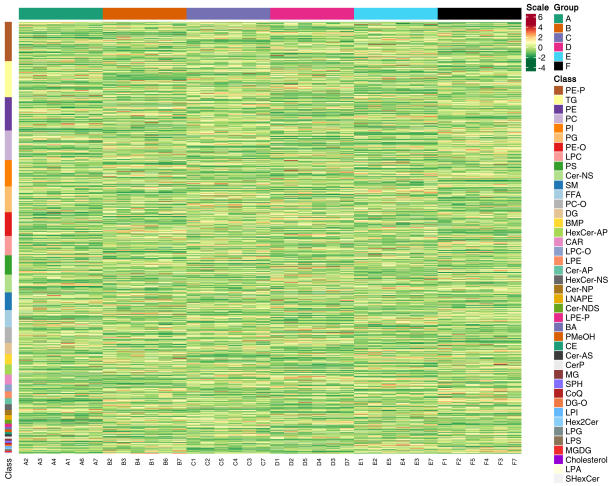


Fig 11: Sample clustering diagram

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

Final report /1.Data_Assess/*all_heatmap*

# 4 Analysis of data results

## 4.1 Lipid composition analysis

Lipid composition analysis is one of the main components of lipid data analysis. The Lipid Metabolic Pathways Research Program consortium classifies lipids into eight major groups: fatty acyl, glycerolipids, glycerophospholipids, sphingolipids, sterolipids, isopentenolipids, glycolipids, and polyketides. Lipids can be further classified into various subclasses depending on the polar head group or other properties. The following table shows the number of lipid compounds in each detected subclasses.

Table 6: Number of lipids identified in each subclass

| Class | all_sample | A | B |
|---|---|---|---|
| BA | 5 | 5 | 5 |
| BMP | 19 | 19 | 19 |
| CAR | 18 | 18 | 18 |
| CE | 4 | 4 | 4 |
| Cer-AP | 11 | 11 | 11 |
| Cer-AS | 4 | 4 | 4 |
| Cer-NDS | 7 | 7 | 7 |
| Cer-NP | 9 | 9 | 9 |
| Cer-NS | 33 | 33 | 33 |
| CerP | 4 | 4 | 4 |

Final report/1.Data_Assess/Class_Count.xlsx

The same statistics is shown in the following figure.

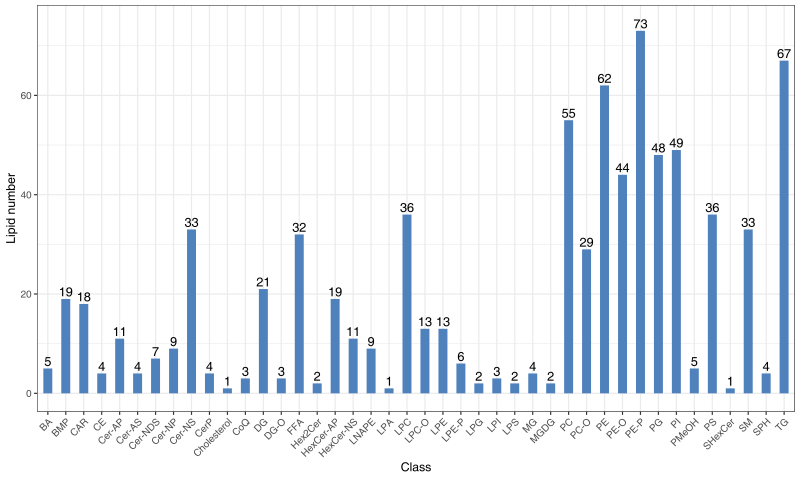


Fig 12: Histogram of the number of lipids identified in each subclass

Final report/1.Data_Assess/Class_Count/Class_Count_Bar.png

Lipid composition is sample-specific and varies between samples. The analysis of lipid composition ratios can examine the distribution of major lipids in the samples. The following ring figure shows the lipid subclass composition for each group.

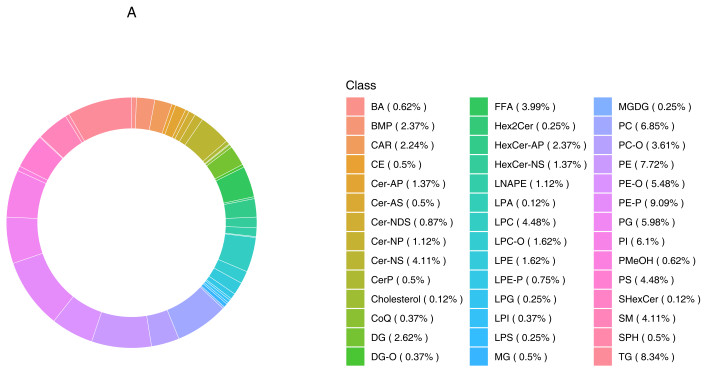


Fig 13: Loop diagram of lipid subclass composition

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

Final report/1.Data_Assess/Class_Count/Class_Count_Ring_*.png

## 4.2 Subclass level analysis

### 4.2.1 Changes in total lipid abundance

Lipid abundance is the total quantification of all lipids in the sample. The total lipid abundance between the samples from different groups is shown below.

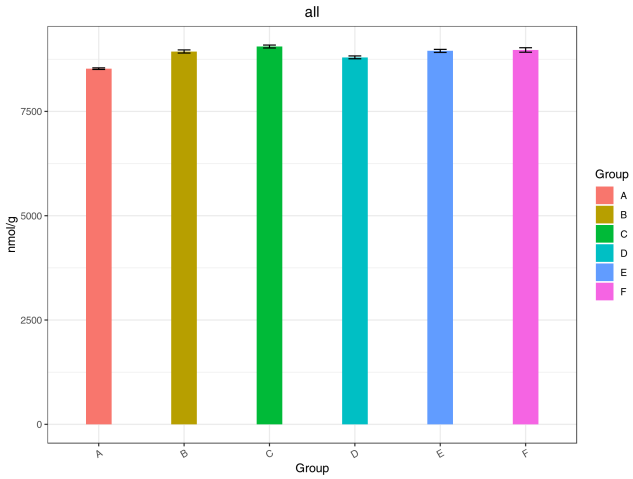


Fig 14: Change in total lipid molecule content

Note: The horizontal coordinates indicate the different groups; the vertical coordinates indicate the total lipid abundance in different groups.

Final report/1.Data_Assess/Class_Content/All_Class_Content.png

### 4.2.2 Changes in lipid abundance by lipid subclasses

The biological functions of different lipid subclasses are different. Changes in the lipid abundance of a subclasses between sample groups are important information to identify important lipid subclasses that may be involved in relevant biological processes for the observed phenotype. The following bar graph shows the differences in lipid abundance of each subclass between sample groups.

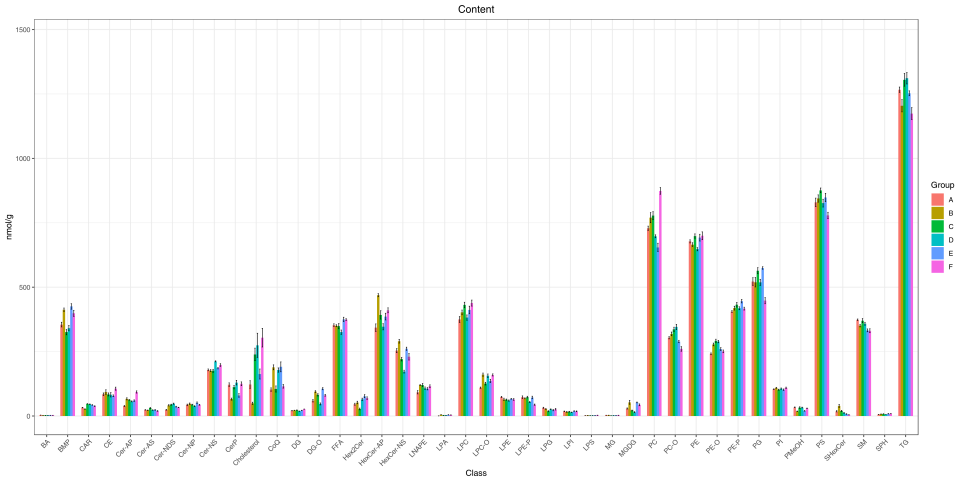


Fig 15: Changes in lipid subclass content

Final report/1.Data_Assess/Class_Content/Class_Content.png

To facilitate subsequent in-depth data analysis, each lipid subclass is plotted separately.

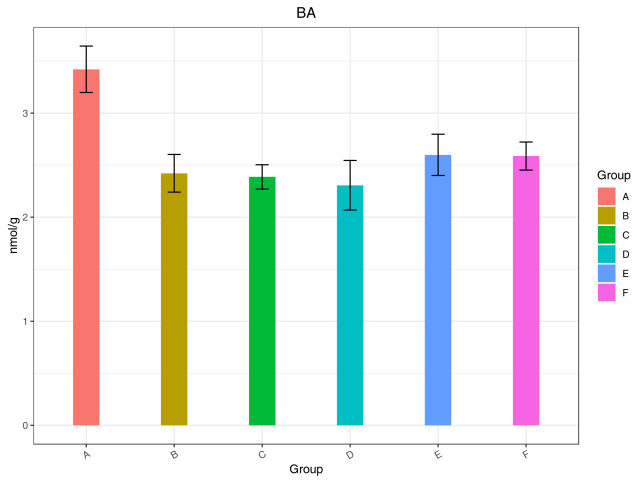


Fig 16: Differences in the content of lipid subclasses between groups

Final report/1.Data_Assess/Class_Content/Class_Content_*.png

A radar chart is a graphical method of displaying multivariate data in the form of a two-dimensional chart with three or more variables on an axis starting from the same point. The figure below shows the changes in

lipid abundance of each subclass.

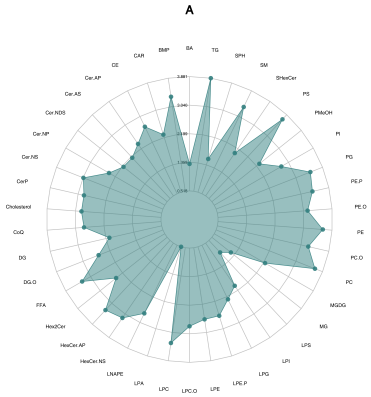


Fig 17: Radar of changes in lipid subclass content

Final report/1.Data_Assess/Radarchart/*

### 4.2.3 Dynamic distribution of lipids

The dynamic range of lipid distribution allows the examination of the least and the most abundant lipids in each group, as well as the variation of lipid abundances across the entire abundance range.

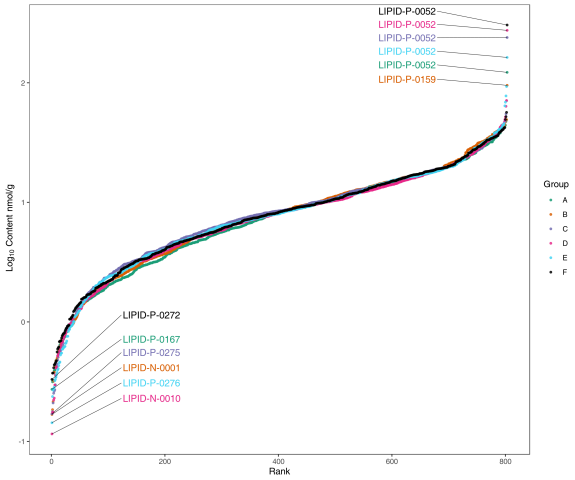


Fig 18: Dynamic distribution of lipid content

Note: Each point in the graph represents a lipid molecule. The vertical coordinate represents the corresponding abundance of each lipid molecule (log10), and the lipid molecule with the lowest and highest content is labeled. Different color represent different sample groups.

Original file path/1.Data_Assess/distribution/*

## 4.3 Analysis of lipid chain length and unsaturation

### 4.3.1 Chain length analysis

The chain length is the number of the carbon atoms in a fatty acid chain, and it is closely related to lipid function. Chain length can affect cell membrane thickness, fluidity of the cell membrane, and the activity and function of lipid transport proteins. We analyzed the abundance of lipids with the same chain length and examined the differences between lipids with different chain length.

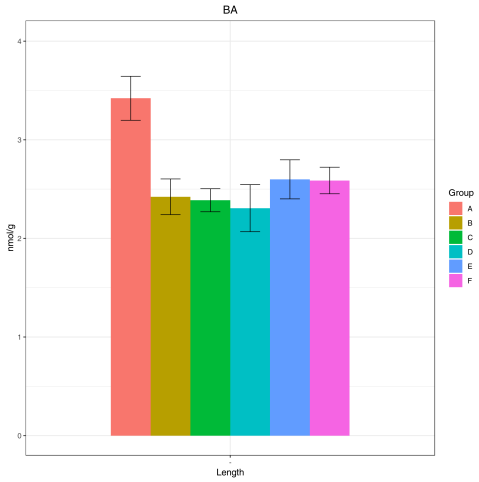


Fig 19: Analysis of chain length

Note: The horizontal coordinates indicate the different carbon chain lengths and the vertical coordinates indicate the abundance of lipid compounds.

Final report/1.Data_Assess/Class_Length/*

### 4.3.2 Differences in chain length

This analysis shows the fold change of lipids with different chain lengths between sample groups.

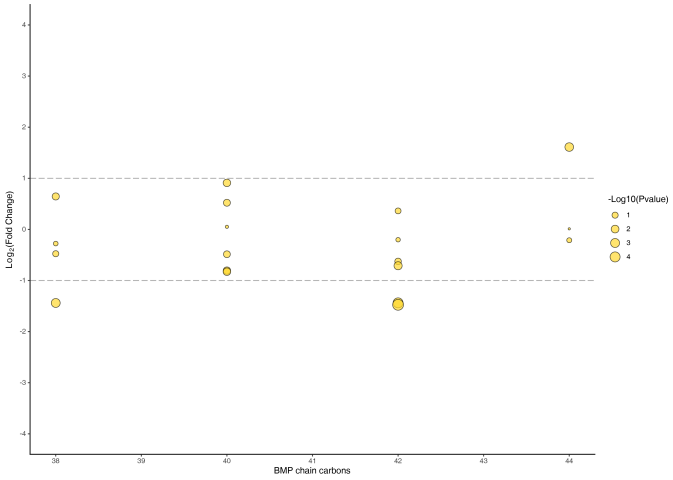


Fig 20: Analysis of differences in chain length

Note: The horizontal coordinates represent the carbon chain length, the vertical coordinates represent the differential expression multiples, each point represents a lipid, the size of the point represents the P-value, the larger the point means the smaller the P-value.

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

### 4.3.3 Chain unsaturation analysis

Chain unsaturation is the number of double bonds in the fatty acid chain. This analysis shows the abundance of lipid compounds with the same number of unsaturated bonds between sample groups.

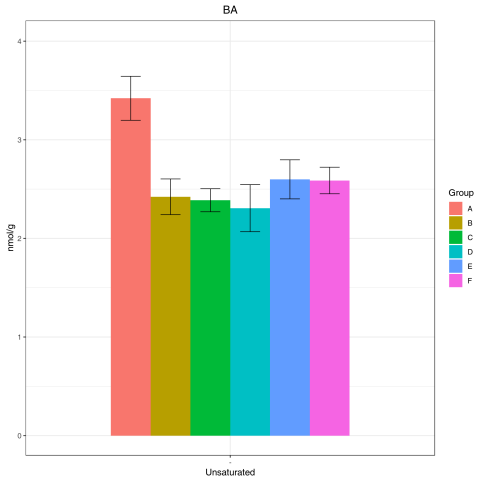


Fig 21: Analysis of chain unsaturation

Note: The horizontal coordinate indicates the number of unsaturated bonds and the vertical coordinate indicates the abundance of lipid compounds.

Final report/1.Data_Assess/Class_Unsaturated/Class_Unsaturated_*.png

### 4.3.4 Differences in chain unsaturation

This analysis shows the fold change of lipids with different chain unsaturations between sample groups

.

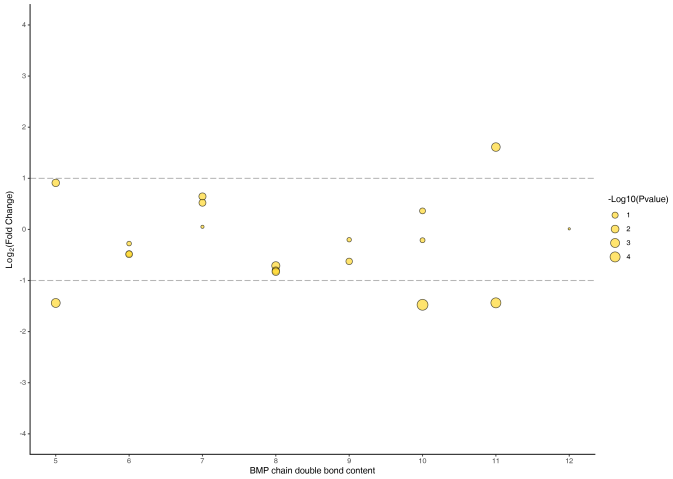


Fig 22: Analysis of differences in chain unsaturation

Note: The horizontal coordinate represents the carbon chain unsaturation, the vertical coordinate represents the differential expression multiplier, each point represents a lipid, the size of the point represents the P value, the larger the point means the smaller the P value.

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

## 4.4 Principal component analysis of sample groups

### 4.4.1 Principal component analysis between sample groups

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

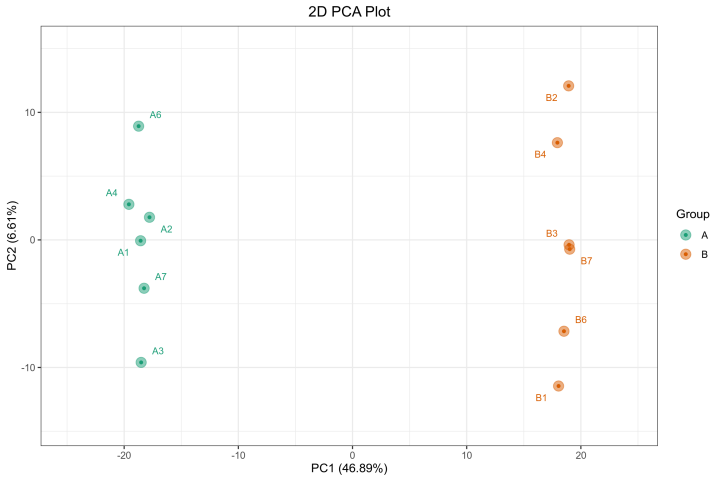


Fig 23: Principal component analysis of different groups

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-dimentional PCA result is shown in the figure below:

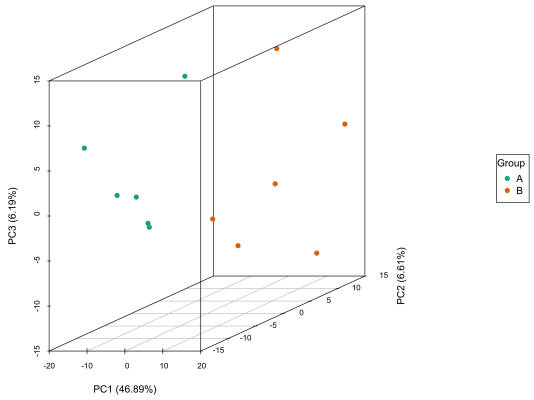


Fig 24: Three-dimensional PCA plot of different groups

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:

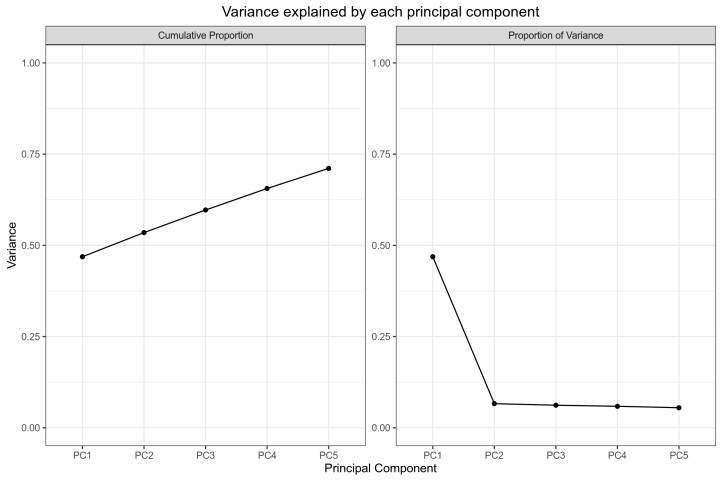


Fig 25: The explainable variation of the first five principal components
Note: The X-axis represents each principal component, the Y-axis represents the explainable variation, the left figure represents the cumulative explainable variation, and the right figure represents the explainable variation of each principal component.

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

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

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

## 4.5 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 7: Partial results of OPLS-DA

| Index | Compounds | Type |
|---|---|---|
| LIPID-N-0001 | taurolithocholicacid-3-sulfate | up |
| LIPID-N-0010 | Ursocholicacid | up |
| LIPID-N-0014 | lithocholicacid-3-sulfate | insig |
| LIPID-N-0015 | Glycocholicacid | up |
| LIPID-N-0017 | Taurocholicacid | down |
| LIPID-P-1600 | BMP(18:2_22:5) | insig |
| LIPID-P-1599 | BMP(18:1_22:6) | insig |
| LIPID-P-1598 | BMP(18:2_20:5) | insig |
| LIPID-P-1594 | BMP(18:1_22:5) | insig |
| LIPID-P-1593 | BMP(18:2_20:4) | insig |

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

### 4.5.1 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).

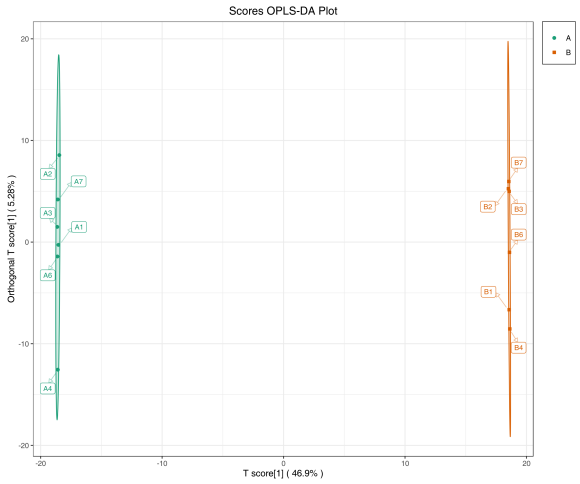


Fig 26: OPLS-DA score diagram

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

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

### 4.5.2 OPLS-DA model validation

The OPLS-DA model was used to analyze metabolome data and draw score charts for each group to further show the differences between each group (Thevenot et al., 2015). 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 prediction ability 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 horizontal coY-axis represents the model accuracy, and the vertical coY-axis is the frequency of the model classification effect. The model performs bootstrapping 200 times and if $Q^2$'s P = 0.02, it indicates that the prediction ability of four random grouping models is better than that of the OPLS-DA model in the Permutation detection. If $R^2Y$'s P = 0.545, it indicated that there were 109 random grouping models in the Permutation detection, whose explanation rate of Y matrix was better than that of the OPLS-DA model. In general, $p < 0.05$ is the best model.

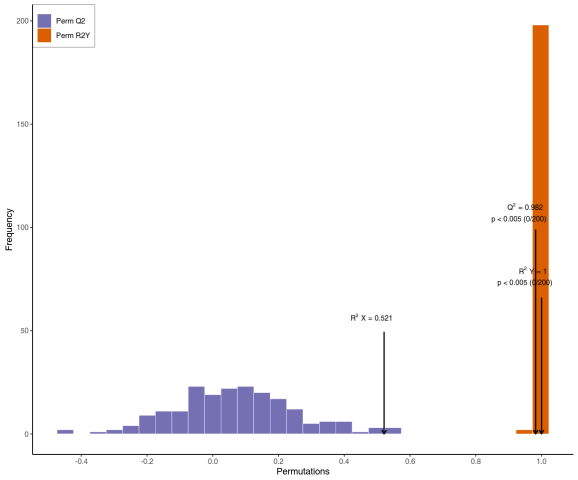


Fig 27: OPLS-DA verification diagram

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

### 4.5.3 OPLS-DA S-plot

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).

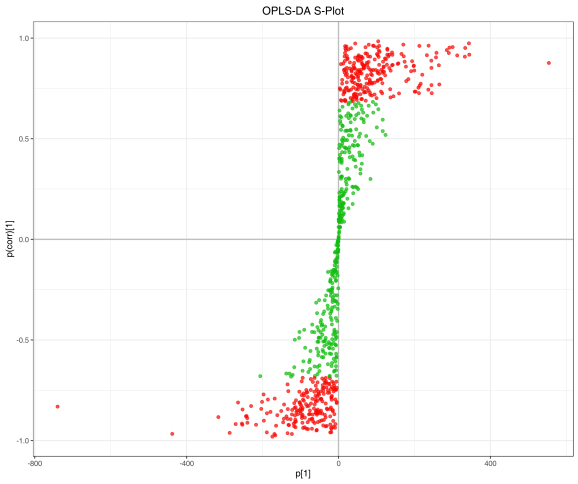


Fig 28: OPLS-DA S-plot

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

### 4.5.4 Dynamic distribution of lipid abundance differences

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

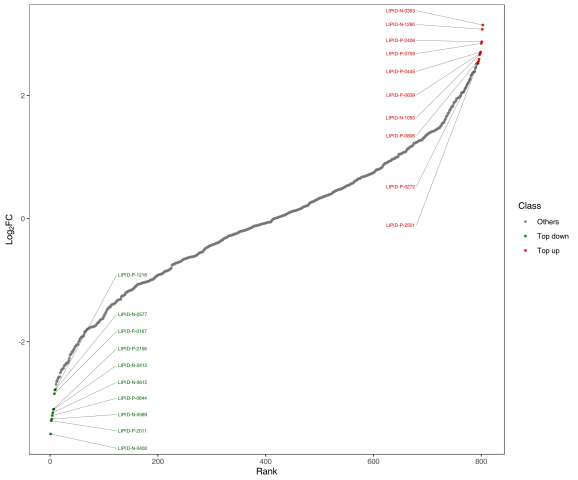


Fig 29: Dynamic distribution of lipid content differences

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

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

## 4.6 Differential lipid 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 lipids. 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 lipids from different samples. Differential lipids 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:

**For two sets of comparisons:**

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

**2.Metabolites with P-value < 0.05 were considered as significant.**

Partial result from the screening are shown below:

Table 8: Screening results of differential lipids

| Index | Compounds | Type |
|---|---|---|
| LIPID-N-0001 | taurolithocholicacid-3-sulfate | up |
| LIPID-N-0010 | Ursocholicacid | up |
| LIPID-N-0015 | Glycocholicacid | up |
| LIPID-N-0017 | Taurocholicacid | down |
| LIPID-P-1577 | BMP(22:5_22:6) | up |
| LIPID-P-1574 | BMP(20:5_22:6) | down |
| LIPID-P-1572 | BMP(20:4_22:6) | down |
| LIPID-P-1589 | BMP(18:1_20:4) | down |
| LIPID-P-1591 | BMP(18:1_22:4) | up |
| LIPID-P-1604 | BMP(20:4_22:4) | down |

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

### 4.6.1 Bar chart of differential lipids

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

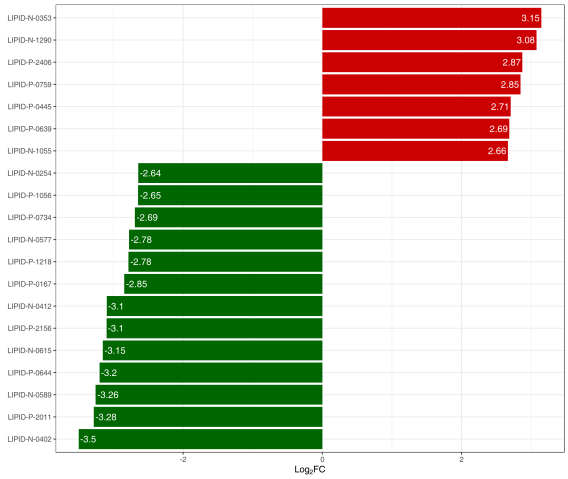


Fig 30: Bar chart of differential lipids

Note: X-axis refers to log_2 values of top differential lipids, the Y-axis refers to lipids. Red bars represent up-regulated differential lipids and green bars represent down-regulated differential lipids.

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

### 4.6.2 Radar map of differential lipids

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

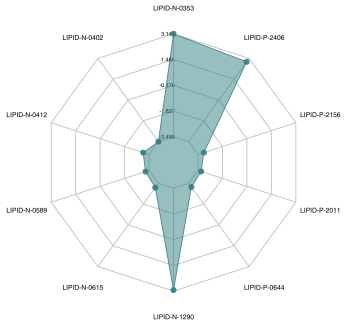


Fig 31: Differential lipid radar map

Note: The grid lines correspond to the log_2FC,The green colored area are formed from the lines connecting the dots

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

### 4.6.3 VIP values of differential lipids

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

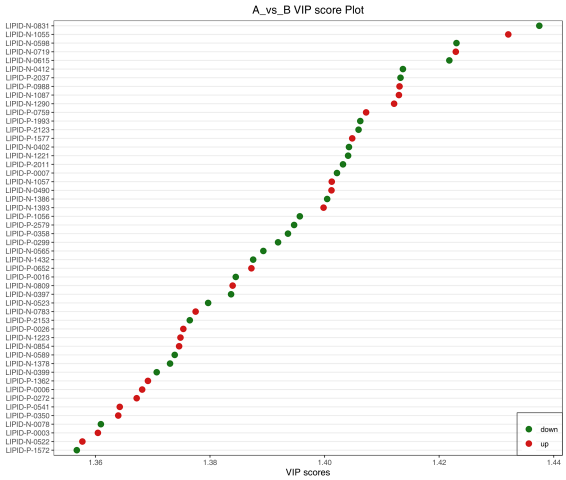


Fig 32: VIP values of differential lipids

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

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

### 4.6.4 Volcanic plot of differential lipids

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

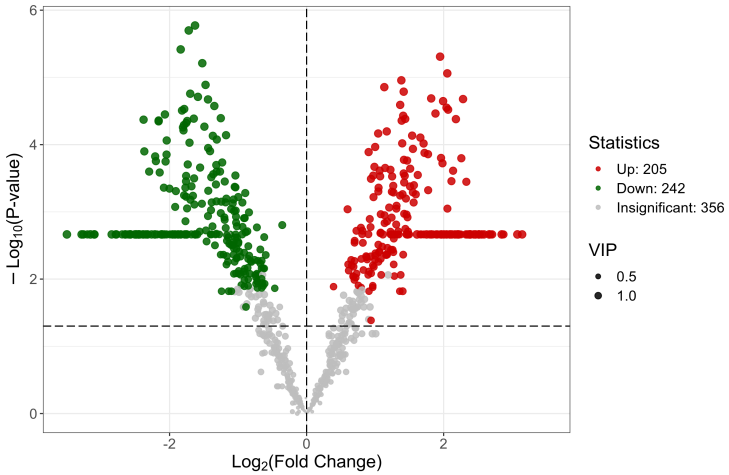


Fig 33: Volcanic plot of differential lipids

Note: Under VIP + FC + Pvalue/FDR triple screening conditions,the horizontal coordinate represents the multiple change of difference of lipids in different groups (log_2FC), and the vertical coordinate represents the significance level of difference (-log_10P-value).Each point in the volcano diagram represents a lipid. Significantly up-regulated lipids were represented by red dots,significantly down-regulated lipids were represented by green dots, and the size of the dots represented VIP values

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

### 4.6.5 Scatter plot of differential lipids

The differential lipids scatter plot is mainly used to show the abundance differences in compound subclasses between two groups.

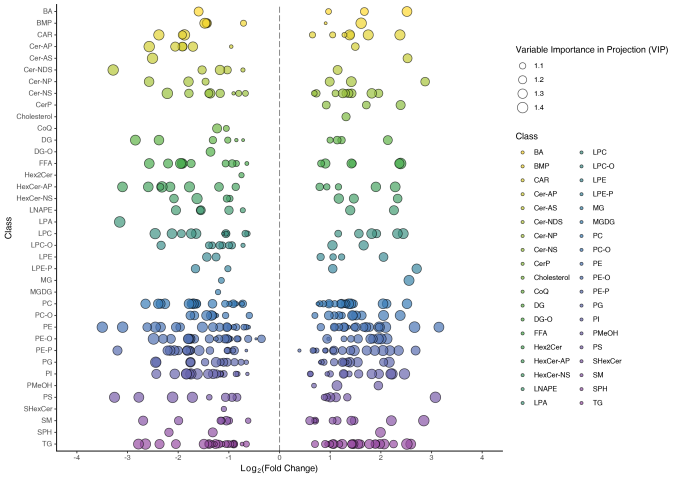


Fig 34: Scatter plot of differential lipids

Note: Each dot in the graph indicates a lipid, and different colors indicate different lipid 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.

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

### 4.6.6 Hierarchical clustering tree of samples

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.

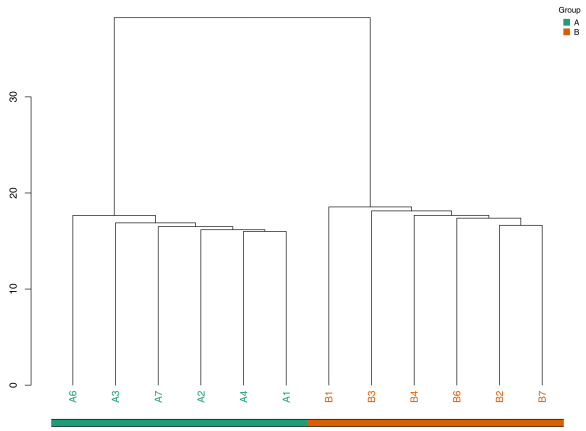


Fig 35: Hierarchical clustering tree of samples

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

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

### 4.6.7 Heatmap of differential lipids

In order to observe the fold-change of differential lipids 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 pheatmap in R.

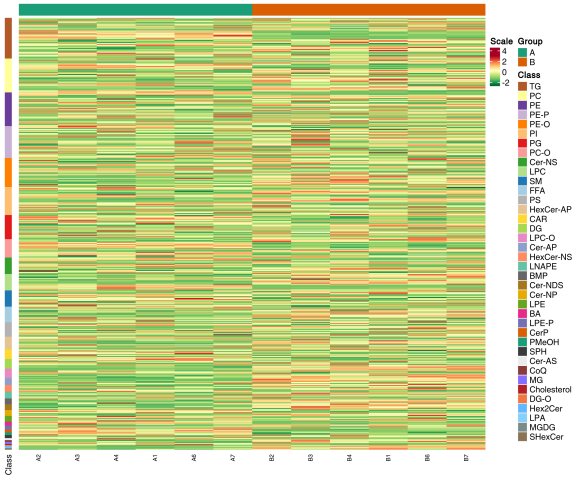


Fig 36: Heatmap of differential lipids

Note: The X-axis shows the name of the samples and the Y-axis shows the differential lipids. 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.

Final report/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/heatmap/group-ID*_vs_group-ID*_heatmap.*;

### 4.6.8 Z-value map of differential lipids

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

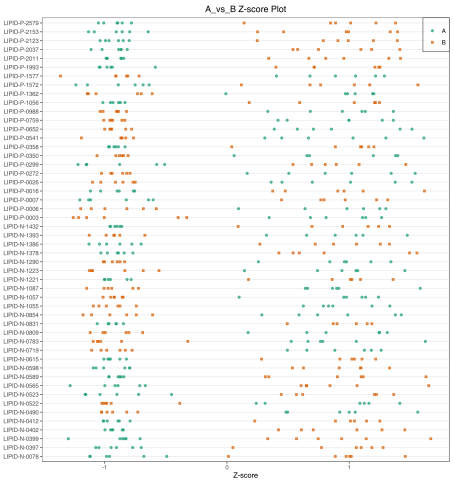


Fig 37: Z-value map of differential lipids

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

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

### 4.6.9 Correlation analysis of differential lipids

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

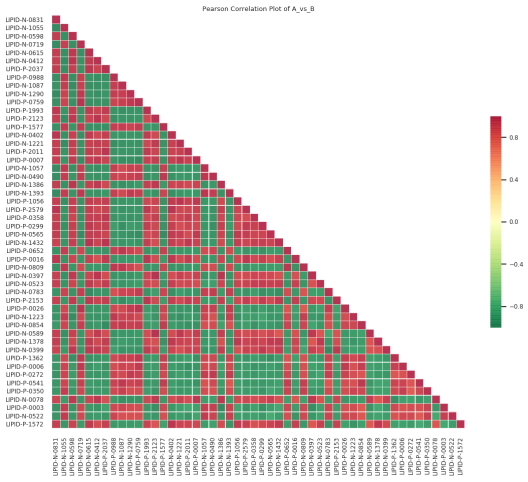


Fig 38: Heat map of correlation of different lipids

Note: The ID of the lipids 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 lipids, the figure will only show the top 50 lipids based on VIP.

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

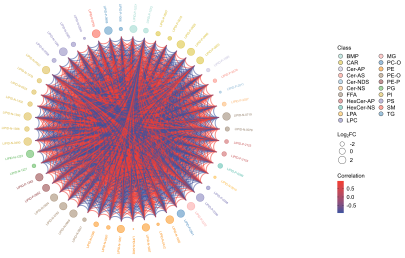


Fig 39: Chord diagram of differential lipids

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

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

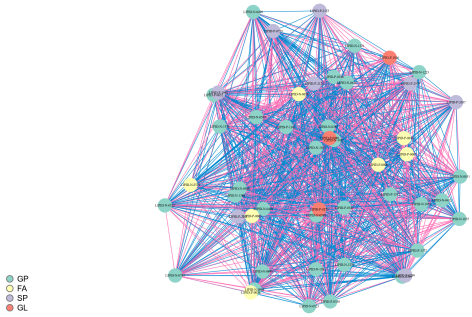


Fig 40: Correlation network diagram of differential lipids

Note: The dots in the figure represent the various differential lipids, and the size of the dot is related to the Degree of connection. The larger the dot, the greater the Degree of connection, i.e. the more dots (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 lipids, the figure will only show the top 50 lipids based on VIP.

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

### 4.6.10 Violin plot of differential lipids

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 plot in the middle show 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. The following figure shows the result of top 50 differentially compounds.

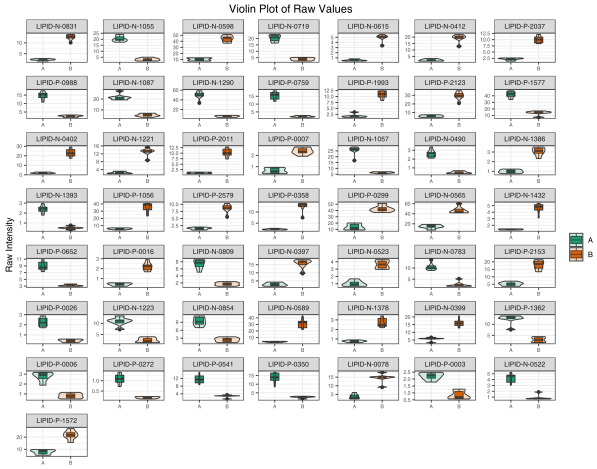


Fig 41: Violin plot of differential lipids

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

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

### 4.6.11 K-means analysis

K-means analysis is a method to examine the trend of relative quantification changes of a lipid in different sample groups. K-means is performed based on the UV（Z-score) standardized relative quantification value.

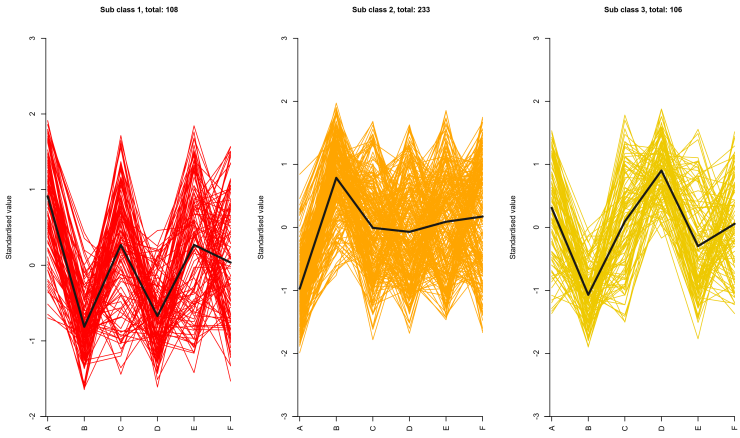


Fig 42: K-Means diagram of differential lipids

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

Final report/2.Basic_analysis/kmeans/kmeans_cluster.*

## 4.7 Functional annotation and enrichment analysis of differential lipids 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.

### 4.7.1 Functional annotation of differential lipids

Lipids are annotated using the KEGG database (Kanehisa et al., 2000), and only metabolic pathways containing differential lipids are shown. Detailed results are found in the attached results. A portion of the results is shown below:

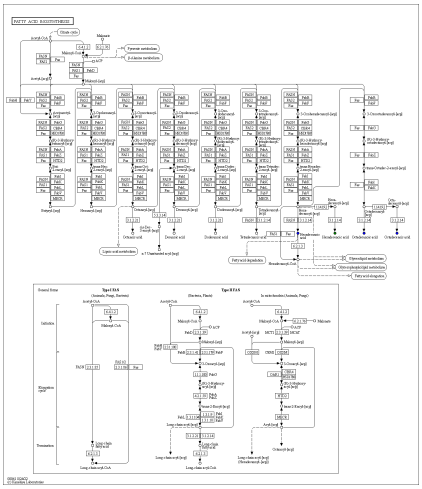

Fig 43: KEGG pathway of lipids

Note: Red circles indicate that the lipid content was significantly up-regulated in the experimental group; blue circles indicate that the lipid content was detected but did not change significantly; green circles indicate that the lipid content was significantly down-regulated in the experimental group; and orange circles indicate a mixture of both up- and down-regulated lipids.

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

Table 9: KEGG annotations for differential lipids

| Index | Compounds | Type | cpd_ID |
|---|---|---|---|
| LIPID-N-0001 | taurolithocholicacid-3-sulfate | up | C03642 |
| LIPID-N-0010 | Ursocholicacid | up | C17644 |
| LIPID-N-0015 | Glycocholicacid | up | C01921 |
| LIPID-N-0017 | Taurocholicacid | down | C05122 |
| LIPID-P-1577 | BMP(22:5_22:6) | up | - |
| LIPID-P-1574 | BMP(20:5_22:6) | down | - |
| LIPID-P-1572 | BMP(20:4_22:6) | down | - |
| LIPID-P-1589 | BMP(18:1_20:4) | down | - |
| LIPID-P-1591 | BMP(18:1_22:4) | up | - |
| LIPID-P-1604 | BMP(20:4_22:4) | down | - |

Table 10: Enrichment Statistics of KEGG annotations for differential lipids

| ko_ID | Sig_compound | compound | Sig_compound_all | compound_all |
|---|---|---|---|---|
| ko04976 | 4 | 8 | 275 | 499 |
| ko00120 | 3 | 3 | 275 | 499 |
| ko01100 | 233 | 419 | 275 | 499 |
| ko04979 | 45 | 74 | 275 | 499 |
| ko00430 | 1 | 1 | 275 | 499 |
| ko04714 | 57 | 91 | 275 | 499 |
| ko00600 | 39 | 66 | 275 | 499 |
| ko04071 | 35 | 57 | 275 | 499 |
| ko04217 | 29 | 51 | 275 | 499 |
| ko04722 | 15 | 21 | 275 | 499 |

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

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

### 4.7.2 KEGG classification of differential lipids

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

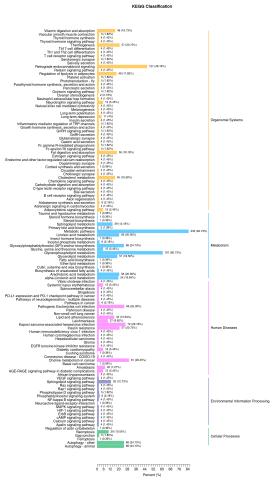


Fig 44: KEGG classification of differential lipids

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

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

### 4.7.3 Hierarchical Cluster Analysis of differential lipids in KEGG signaling pathway

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

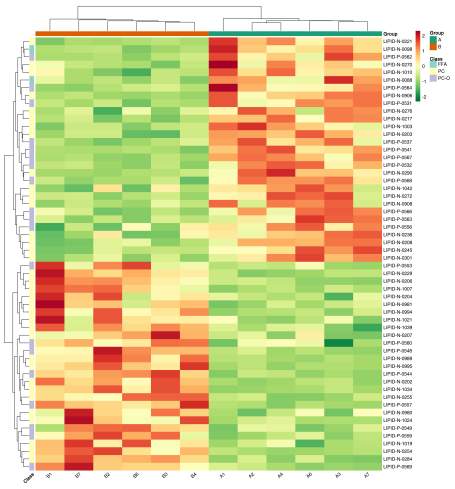

Fig 45: Clustering heat map of differential lipids in KEGG pathway
Note: The X-axis shows the name of the samples and the Y-axis shows the differential lipids. Different colors in the heatmap represent the values obtained after normalization and reflects the level of relative quantification. The darker 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 lipids, a colored bar will be shown on the left to depict compound classifications.

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

### 4.7.4 KEGG enrichment analysis of differential lipids

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 lipids in the corresponding pathway to the total number of lipids 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 lipids with KEGG annotation, n represents the number of differential lipids in N, M represents the number of lipids in a KEGG pathway in N, and m represents the number of differential lipids 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 lipids enriched in the corresponding pathway. The top 20 pathways in terms of P-value are plotted.

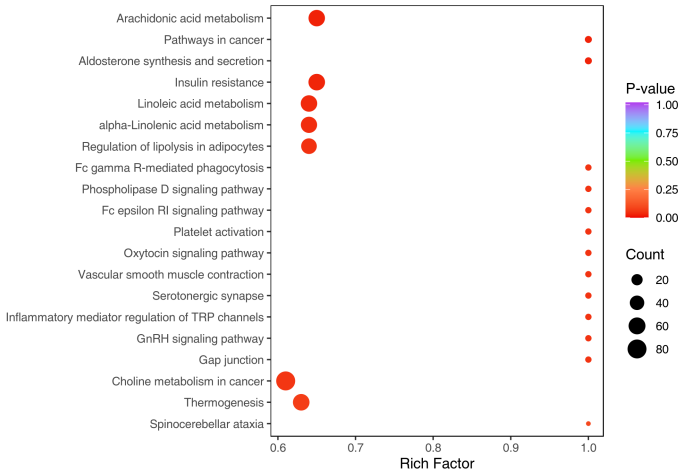


Fig 46: KEGG enrichment diagram of differential lipids

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 lipids.

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

### 4.7.5 Overall changes in KEGG metabolic pathway

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

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

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

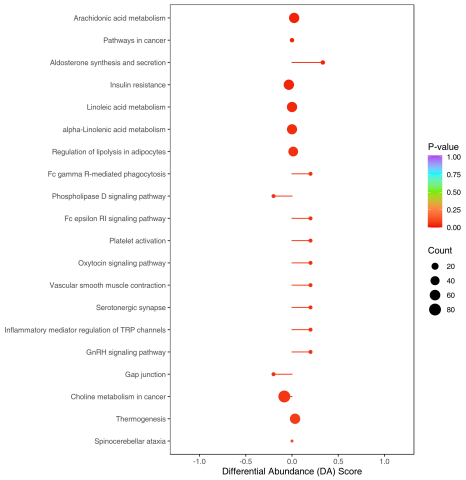


Fig 47: Difference abundance score

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 lipids in the metabolic pathway. A Score of 1 indicates that the expression trend of all identified lipids in this pathway is up-regulated, and -1 indicates that the expression trend of all identified lipids 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 lipids. 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.

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

## 4.8 ROC curve analysis of differential lipids

The ROC curve (Receiver Operating Characteristic Curve) is a quantitative method to measure the performance of a classification model. By default, ROC curve analysis is performed when the sample size is greater than 30.

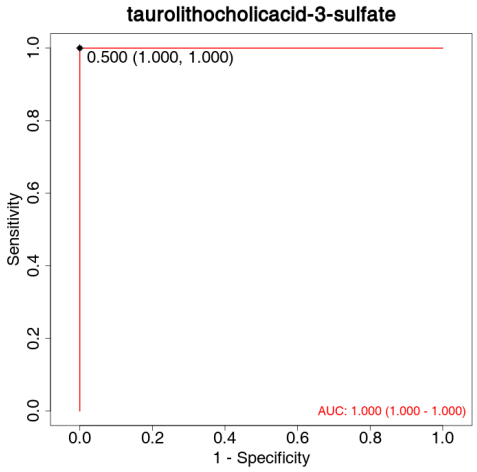


Fig 48: ROC curves for differential lipids

Note: The horizontal coordinate is 1 - specificity, i.e. false positive rate, false positive rate = false positive/(false positive + true negative); the vertical coordinate is sensitivity, i.e. true positive rate, true positive rate = true positive/(true positive + false negative).The area between the ROC curve and the horizontal coordinate is the Area Under Curve (AUC), which is the quantitative evaluation index of the ROC curve.The range of AUC is (0.5, 1], and the closer it is to 1, the better the prediction of the model. The text in red is the AUC value and 95% confidence interval of the curve; the text in black is the optimal threshold value, and the specificity and sensitivity are in parentheses.

Original file path/2.Basic_Analysis/Difference_analysis/NC_vs_BT/ROC/*ROC*

# 5 References

1. Matyash V , Liebisch G , Kurzchalia T V , et al. Lipid extraction by methyl-tert-butyl ether for high-throughput lipidomics[J]. The Journal of Lipid Research, 2008, 49(5):1137-1146.
2. Quehenberger O , Armando A M , Brown A H , et al. Lipidomics reveals a remarkable diversity of lipids in human plasma[J]. The Journal of Lipid Research, 2010, 51(11):3299-3305.
3. Harmonizing Lipidomics: NIST Interlaboratory Comparison Exercise for Lipidomics using Standard Reference Material 1950 Metabolites in Frozen Human Plasma[J]. Journal of Lipid Research, 2017:jlr.M079012.
4. Narvaez-Rivas, Monica, Zhang, et al. Comprehensive untargeted lipidomic analysis using core-shell C30 particle column and high field orbitrap mass spectrometer[J]. Journal of chromatography, A: Including electrophoresis and other separation methods, 2016.
5. Xuan, Qiuhui, Hu, et al. Development of a High Coverage Pseudotargeted Lipidomics Method Based on Ultra-High Performance Liquid Chromatography-Mass Spectrometry[J]. Analytical chemistry, 2018.

6. Chen, W., Gong, L., Guo, Z., et al., A Novel Integrated Method for Large-Scale Detection, Identification, and Quantification of Widely Targeted Metabolites: Application in the Study of Rice Metabolomics[J]. Molecular Plant, 2013, 6(6):1769-1780.
7. Fraga C.G., Clowers B.H., Moore R.J., et al., Signature-discovery approach for sample matching of a nerve-agent precursor using liquid chromatography-mass spectrometry, XCMS, and chemometrics[J]. Anal Chem, 2010. 82(10): p. 4165-73.
8. 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[J], Second edition Umetrics Academy:Sweden, 2006.
9. Chen, Y., et al., RRLC-MS/MS-based metabonomics combined with in-depth analysis of metabolic correlation network: finding potential biomarkers for breast cancer[J]. Analyst, 2009.134(10): p. 2003-11.
10. 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.
11. Kanehisa, M. and S. Goto, KEGG: kyoto encyclopedia of genes and genomes[J]. Nucleic Acids Res, 2000. 28(1): p. 27-30.

# 6 Appendix

## 6.1 分析方法英文版

1.PCA

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

2.Hierarchical Cluster Analysis and Pearson Correlation Coefficients

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

3. Differential lipids selected

For two-group analysis, differential lipids were determined by VIP (VIP > 1) and P-value (P-value < 0.05, Student's t test). VIP values were extracted from OPLS-DA result, which also contain score plots and permutation plots, was generated using R package MetaboAnalystR. The data was log transform ($log_2$) and mean centering before OPLS-DA. In order to avoid overfitting, a permutation test (200 permutations) was performed.

4.KEGG annotation and enrichment analysis

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

## 6.2 List of software and versions

Table 11: Software used

| Analysis | Software | Version |
|---|---|---|
| PCA | R (base package) | 3.5.1 |
| Pearson Correlation | R (base package; Hmisc) | 3.5.1; 4.4.0 |
| Inter-sample correlation plots | R (corrplot) | 0.84 |
| Heatmap | R (heatmaply; ComplexHeatmap) | 1.2.1; 2.7.1.1009 |
| OPLS-DA | R (MetaboAnalystR) | 1.0.1 |
| Radar map | R (fmsb) | 0.7.0 |
| Chord diagram | R (igraph; ggraph) | 1.2.4.2; 2.0.2 |
| Correlation network diagram | R (igraph) | 1.2.4.2 |
| Modulation network diagram | R (FELLA) | 1.10.0 |

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:

$$x' = x - \mu$$