及因生物
— TGENE —

非靶代谢组结题报告

及因（上海）生物科技有限公司

Tgene Biotech (Shanghai) Co.,Ltd.

目录

1. 项目总览

本项目样本所属物种为：植物；样本类型为：土壤；样本数目：12 例。

（1）代谢物鉴定

通过非靶向代谢组学技术检测样本中代谢物，并与参考库比对分析，进行代谢物注释。

表 1.1： 鉴定结果总表

IonType	Feature 数目	QC 样本 RSD≦30 峰数目比例	QC 样本 RSD≦30 峰数目	代谢物数目
正离子	3904	85%	3281	515
负离子	2240	80.4%	1782	

（2）差异代谢物统计

采用 T-test 检验和多变量统计分析 VIP 进行样本间差异代谢物分析，当 PValue < 0.05 且 VIP > 1 时，筛选得到差异代谢物。

表 1.2： 差异代谢物数目总表

比较组别	上调差异数目	下调差异数目	差异总数目
L vs CK	19	16	35
M vs CK	16	11	27
H vs CK	38	16	54
L vs M	26	14	40
L vs H	24	33	57
M vs H	11	43	54
L vs M vs H	/	/	54
L vs M vs H vs CK	/	/	50

结果路径：report/6_MetDiff/Sig.Count.Stat.xlsx

2. 项目流程

采用液质联用(LC-MS)技术对样本进行非靶向代谢组研究，从生物样本中检测并筛选出具有重要生物学意义和统计学显著性差异的代谢物，并以此为基础阐明生物体的代谢过程和变化机制。项目流程主要包括对样本进行精确称量、样本制备和质检，然后进行质谱上机检测，数据下机后进行数据分析，并生成结果分析报告。

项目流程如下：

图 2.1 项目流程

基于质谱检测得到原始文件，对原始文件进行数据预处理和质控分析，以保障数据的准确性与可靠性。采用参考谱图库对检测谱图进行比对分析，实现代谢物鉴定。对样本数据进行单变量分析和多元统计分析来揭示和筛选出不同组分间差异的代谢物，从而揭示代谢物与样本之间的关系，最后通过功能预测和分析来发现代谢物的生物学意义。

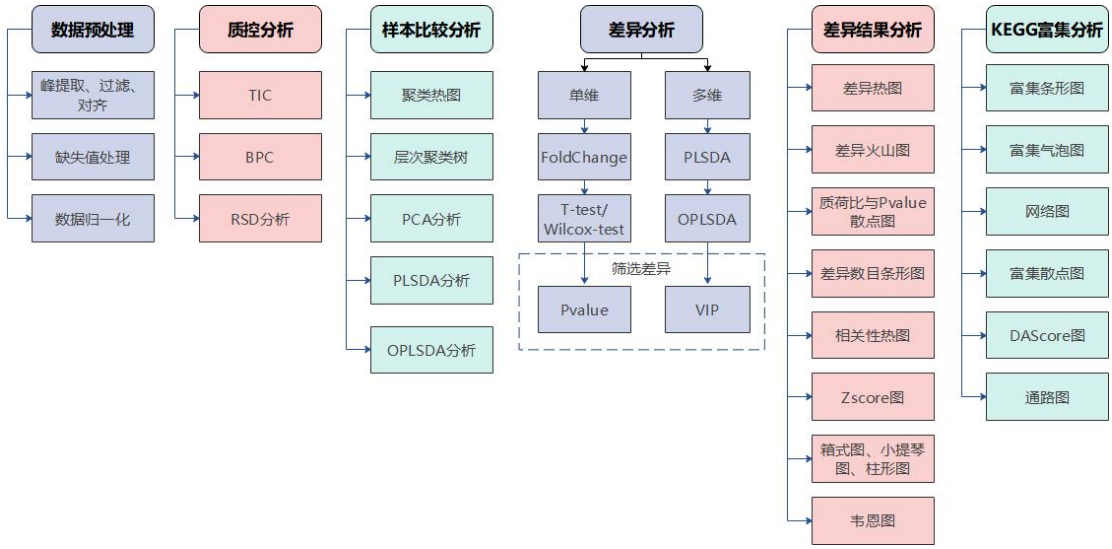

图 2.2 数据分析流程

2.1 样本基本信息

本实验共收集样本 8 个，分为 2 组，样本详细信息如下表：

表 2.1 样本详细信息(展示前 5 行)

Raw name	Sample name	Group name
MCK_1	MCK_1	CK

MCK_2	MCK_2	CK
MCK_3	MCK_3	CK
MH_1	MH_1	H
MH_2	MH_2	H

注：Raw name：编号；Sample name：样本名称；Group name：组别名称

结果路径：0_Project_Data/Sample.Info.xlsx

2.2 数据预处理

通过 Proteowizard 软件包（v3.0.8789）[1]中 MSConvert 工具将原始质谱下机文件转换为 mzXML 文件格式。采用 XCMS 软件包进行峰检测、峰过滤、峰对齐处理[2]，参数设置有 bw=2，ppm=15，peakwidth=c(5, 30)，mzwid=0.015，mzdiff=0.01，method=“centWave”，分别得到正负离子模式下的 Pos.AllSample.Quant.Raw.xlsx 和 Neg.AllSample.Quant.Raw.xlsx 定量文件。

采用基于 QC 样本的 LOESS 信号校正方法，消除系统误差。然后保留 QC 样本中变异系数（Coefficient of Variance, CV）小于 30%[3]的物质，得到 Pos.AllSample.Quant.Nor.xlsx 和 Neg.AllSample.Quant.Nor.xlsx 数据文件，进行后续分析。

表 2.2 定量结果(展示前 5 行)

ID	mz	rt	rtmin
M100T107	100.076	107.1	1.8
M100T128	100.0755	127.5	2.1
M100T147	100.0755	147.4	2.5
M100T162	100.0754	162.0	2.7
M100T178	100.0754	178.4	3.0

注：ID：Feature 编号；mz：质荷比；rt：保留时间，单位 s；rtmin：保留时间，单位 min；其他列为各个样品的相对定量信息。

结果路径：0_Project_Data/Pos.AllSample.Quant.Raw.xlsx
0_Project_Data/Pos.AllSample.Quant.Nor.xlsx
0_Project_Data/Neg.AllSample.Quant.Raw.xlsx
0_Project_Data/Neg.AllSample.Quant.Nor.xlsx

2.3 代谢物鉴定

物质鉴定使用 HMDB[8]、massbank[9]、LipidMaps[10]、mzcloud[11]、KEGG[12]以及自建代谢物标准品数据库等谱图数据库进行检索比对（搜库）。将定量列表中含有二级谱图的代谢物与数据库中每个二级谱图的碎片离子等信息进行比较、匹配，实现代谢物的二级定性鉴定。

代谢物鉴定结果列表见：5_MetIdent/MetAnnotation.xlsx

包含二级鉴定代谢物的质荷比（mass to charge ratio, m/z）和保留时间（retention time,rt）及峰面积（intensity）等信息的数据矩阵。

物质鉴定二级谱图匹配图见文件夹：5_MetIdent/SpectrumMatch

鉴定图文件夹里面包含每个物质的鉴定匹配图，红色质谱图为数据库收录信息，蓝色质谱图为实验检测信息。

3. 数据质量评估

质控 QC 样本通常是混合相同体积的所有待检测样本，然后按照与待测样本相同的前处理方法来处理 QC 样本，之后进样进行质谱分析，可对仪器的稳定性、实验的重复性、数据质量的可靠性进行全面评价。若仪器分析系统稳定性较好，则检测数据稳定可靠。在检测中获得的代谢谱差异能更好地反映样本间自身的生物学差异。

3.1 总离子流图

将不同质控 QC 样本检测的总离子流图（Total ion chromatogram，TIC）进行谱图重叠比较，可判断代谢物提取和检测的重复性，即技术性重复。各色谱峰的响应强度和保留时间重叠度越高，说明在整个实验过程中仪器误差引起的变异较小。

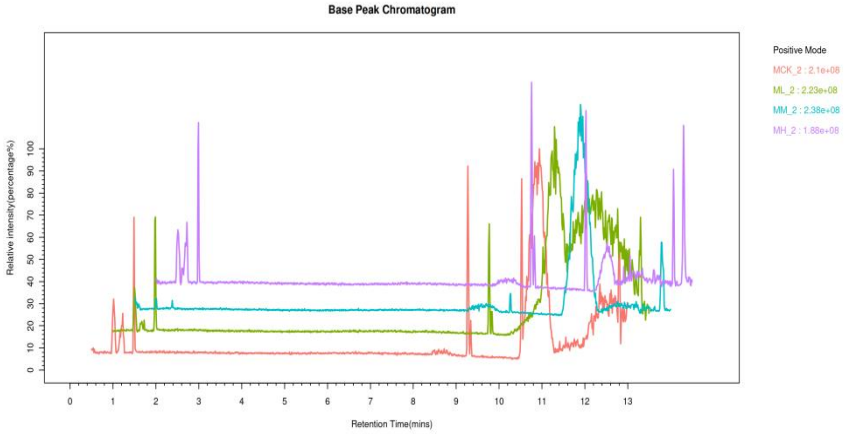

图 3.1 正离子模式 QC 样本 TIC 重叠图

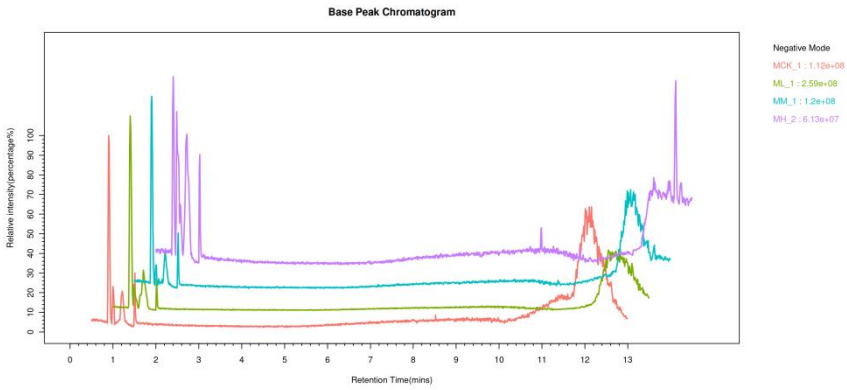

图 3.2 负离子模式 QC 样本 TIC 重叠图

说明：横坐标为保留时间，纵坐标为总离子强度，图示右上角代表每个样本的总离子强度。总离子流的曲线重叠度越高，即保留时间和峰强度均一致，表明质谱对相同样本不同时间检测时，信号稳定性较好。

结果路径：1_QC/(pos|neg)_QC_BPC.png

1_QC/(pos|neg)_QC_TIC.png

3.2 QC 样本相关性

对 QC 样本进行 Pearson 相关性分析，QC 样本间的相关性系数越高（一般在 0.9 以上），说明实验重复性较好，数据质量越高。

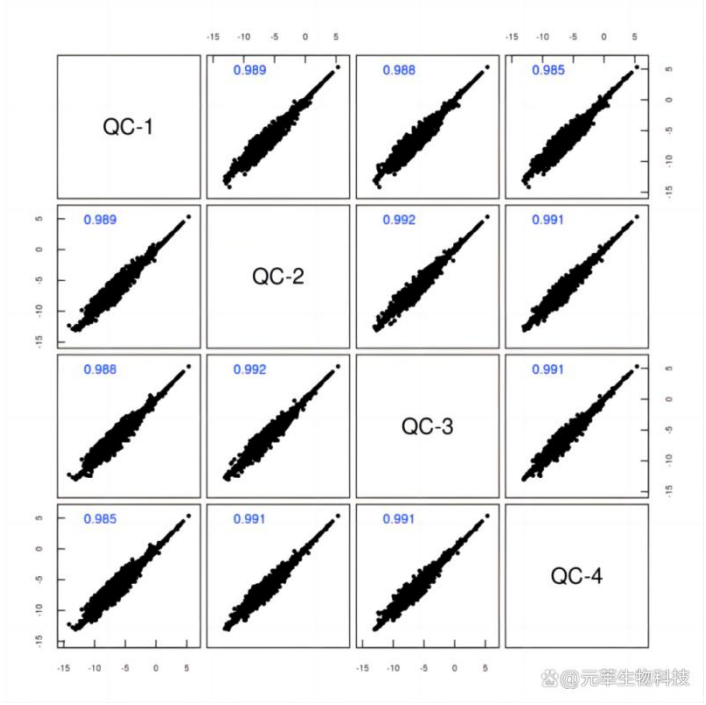

图 3.3 正离子模式 QC 样本相关性

说明：对角线表示 QC 样本；对角线下三角为 QC 相关性散点图，横纵坐标为代谢物丰度(Log10 处理)；对角线上三角为 QC 样本的相关性系数。

结果路径：1_QC/(pos|neg)_QC_cor.png

1_QC/(pos|neg)_QC_cor.xlsx

3.3 QC 样本 RSD 分析

标准偏差（RSD）是原始数据标准差与原始数据平均值的比，可反映数据的离散程度。质控 QC 样本中相对标准偏差较低的物质占比越高，表明仪器的稳定性越好，实验数据越稳定。QC 样本中 RSD≤50%的 Peak 数目占比高于 85%以上，表明实验数据较稳定；QC 样本中 RSD≤30%的 Peak 数目占比高于 70%以上，表明实验数据非常稳定[3]。

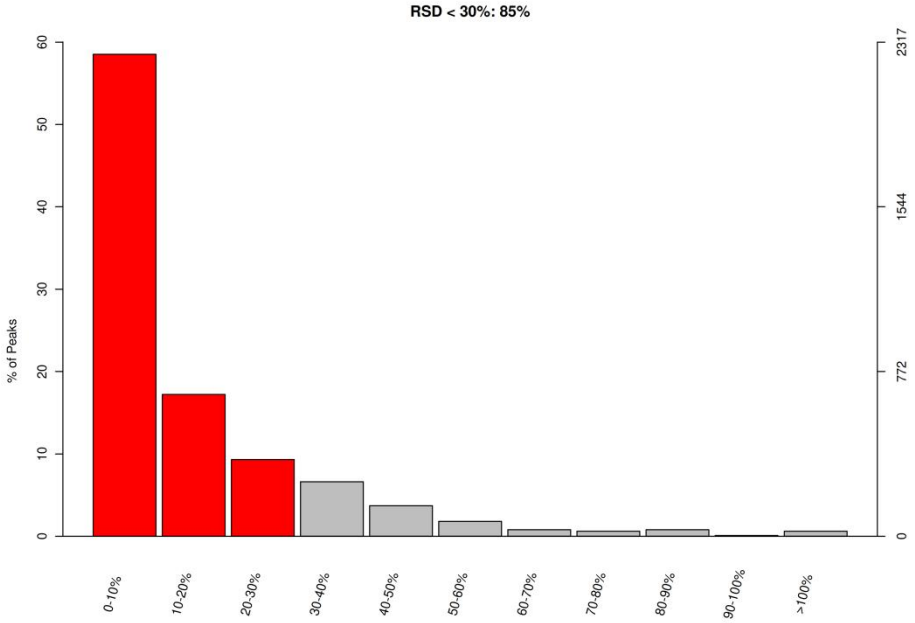

图 3.4 正离子模式 RSD 分布图

说明：纵坐标代表所占比例，纵坐标代表具体数量，横坐标为 RSD 值范围。

结果路径：1_QC/(pos|neg)_RSD.png

3.4 主成分分析

将所有实验样本和 QC 样本进行 PCA 分析，以便初步了解各组样本之间的总体代谢物差异和组内样本之间的变异度大小。PCA 得分图中 QC 样本紧密聚集在一起，说明实验的重复性好。

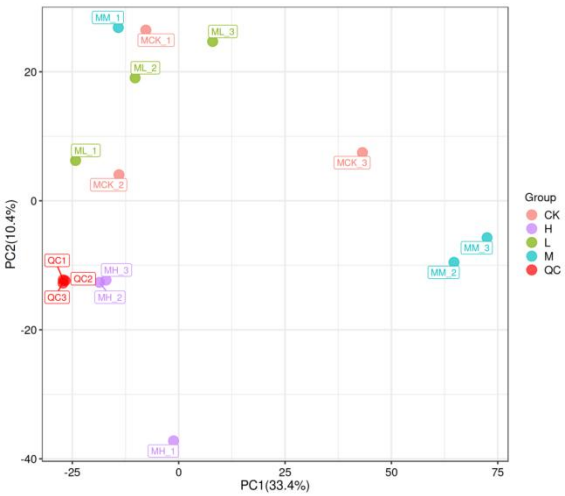

图 3.5 正离子模式总体样本的 PCA 分析

说明：红点代表 QC 样本，其余颜色点为样本。若 QC 样本具有聚集趋势，在 95%置信区间内，说明重复性良好。

结果路径：1_QC/(pos|neg)_QCQA/PCA_score.png

3.5 多变量控制图

多变量控制图（Multivariate Control Chart ，MCC）是基于所有样本检测到的离子峰建立的 PCA 多元变量统计学模型，是用于监控和判断仪器状态是否稳定的一种质量管理工具。

多变量控制图中的每个点代表一个样本，横坐标是所有样本的上机顺序。由于仪器状态的波动，图中的点呈现上下波动的情况，一般正负 3 个标准差范围内为正常范围。

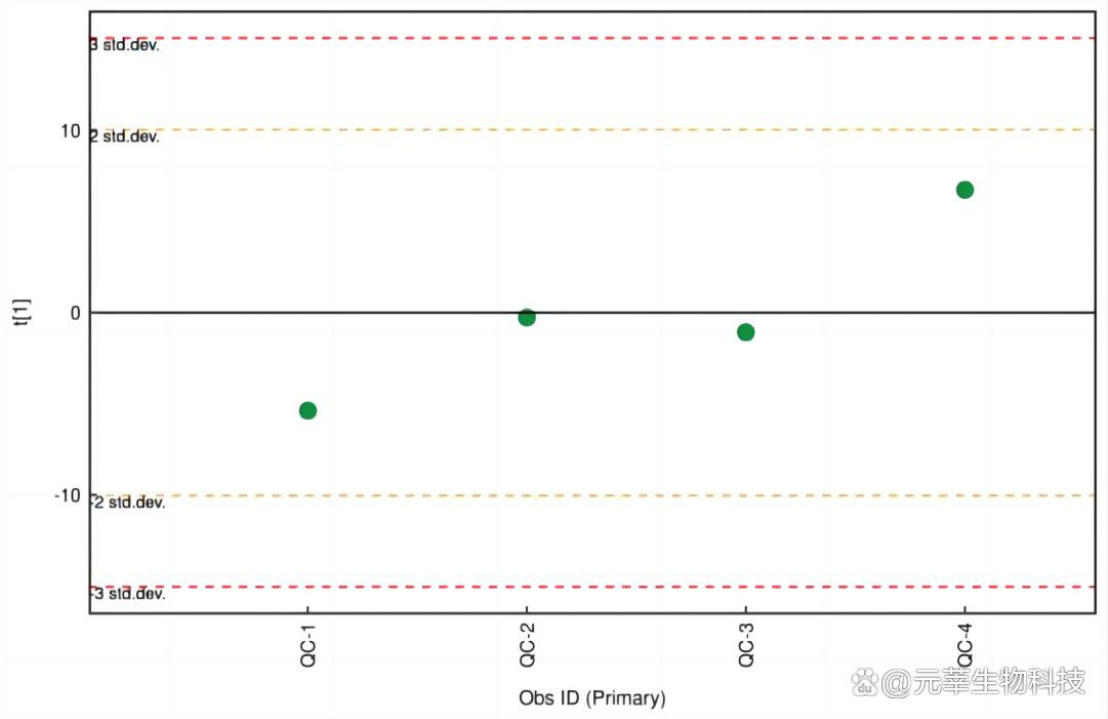

图 3.6 正离子模式 QC 样本 MCC 图

说明：横坐标代表各 QC 样本，纵坐标反映标准差，黄色和红色的线分别定义了正负 2 个、3 个标准差范围。

结果路径：1_QC/(pos|neg)_QCQA/MCC_plot.png

4. 数据分析

4.1 聚类热图

聚类分析是模式识别和数据挖掘中普遍使用的一种方法，可用于判断代谢物在不同实验条件下的代谢模式。代谢模式相似的代谢物具有相似的功能，或是共同参与同一代谢过程或者细胞通路。

通过 R 中 Pheatmap 程序包对代谢物定量值进行数据缩放（Scale），同时对样本和代谢物进行双向聚类，绘制聚类热图。

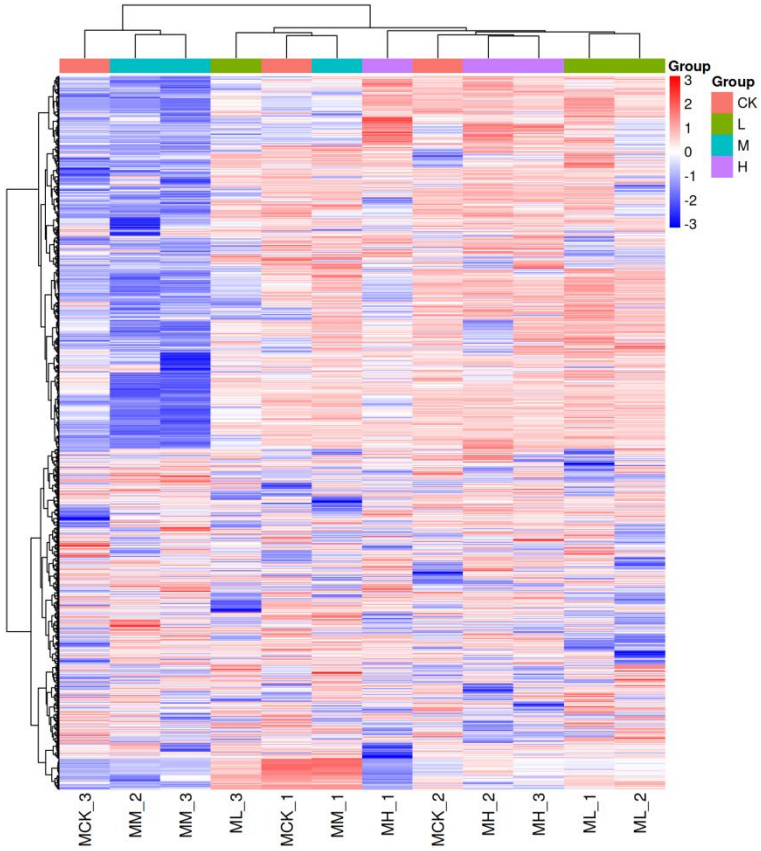

图 4.1 正离子模式总体代谢物聚类热图

说明：其中横坐标代表样本，纵坐标代表代谢物，颜色表示代谢物相对峰强度，颜色越红表达量越高，越蓝表达量越低。图中对代谢物和样本均进行聚类分析，左侧的聚类树为代谢物聚类树，代谢物数目超过 150 个则不显示代谢物名称，图中上方的聚类树为样本聚类树。

结果路径：2_BasicAnalysis/(pos|neg)_heatmap.png

4.2 层次聚类树

对所有样本进行层次聚类（Hierarchical Cluster）分析，生成样本间相似度的聚类树，

并用 R 程序包绘制聚类树状图。

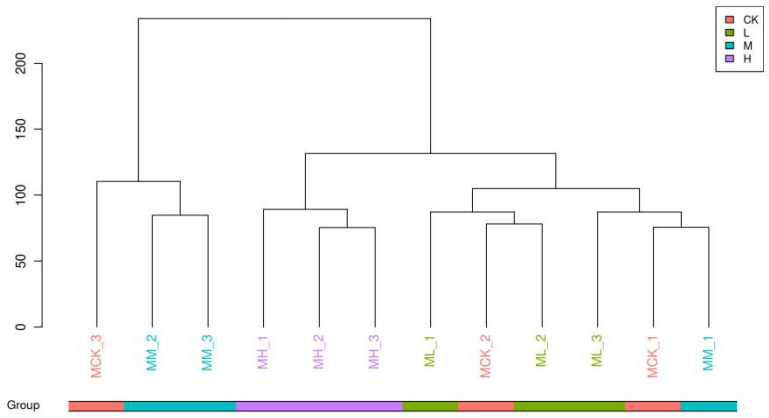

图 4.2 正离子模式样本层次聚类树

说明：每一个分支表示一个样本。样本之间相似度越高，则越聚集在一起。颜色表示样本分组。

结果路径：2_BasicAnalysis/(pos|neg)_clustree.png

4.3 多元统计分析

由于代谢组数据具有多维且某些变量间高度相关的特点，运用传统的单变量分析无法快速、充分、准确地挖掘数据内潜在的信息。因此在分析代谢组数据时需要运用多元统计的方法，如 PCA、PLS-DA、OPLS-DA，最大程度保留原始信息的基础上，对数据进行降维和回归分析，然后进行差异代谢物的筛选及后续分析。

本分析中使用 R 语言 Ropls 包[5]进行多元统计分析，方法包括：

（1）主成分分析（Principal Component Analysis, PCA）；

（2）偏最小二乘判别分析（Partial Least Squares-Discriminant Analysis, PLS-DA）；

（3）正交-偏最小二乘判别分析(Orthogonal Partial Least Squares Discriminant Analysis, OPLS-DA）。

在对代谢组学数据进行多元统计分析之前，需要将数据进行适当的权重转换，即标准化（Scaling）处理。本分析对数据采用“自适换算”处理，以获得更加直观的结果。

4.3.1 主成分分析

主成分分析（Principal Component Analysis，PCA）将代谢物变量按一定的权重通过线性组合后产生新的特征变量，通过主要新变量（主成分）对各组数据进行归类。作为无监

督学习方法，得到的 PCA 模型反映了代谢组数据的原始状态，有利于掌握数据的整体情况，尤其是有利于发现和去除重复性差的样本（离群样本）或异常样本，并提高模型的准确性。

采用 PCA 方法，观察样本之间的总体分布趋势。从 PCA 得分图可观察样本的聚集、离散程度。样本分布点越靠近，说明这些样本的组成和浓度越接近；反之，样本点越远离，其差异越大。模型的交叉验证主要参考 R2X 参数，表示模型的可解释度。通常情况下，R2 高于 0.5 较好。

表 4.1 正离子模式 PCA 模型验证参数

Group	pre	R2X(cum)
L vs M vs H vs CK	3	0.54
M vs H	2	0.634
L vs M	2	0.618
M vs CK	2	0.613
L vs CK	2	0.57

注：pre，主成分数；R2X，模型（对 X 变量数据集）可解释度。默认展示前五条，具体内容见结果文件。

结果路径：3_Ropls/(pos|neg)/PCA_summary.xlsx

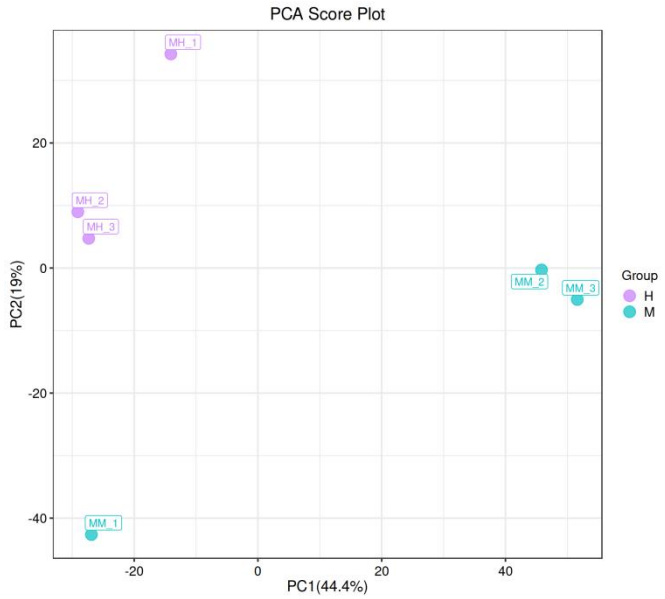

图 4.3 正离子模式 PCA 得分图

说明：横坐标 PC1 表示第一主成分得分值，纵坐标 PC2 表示第二主成分得分值。点表示样本，圈表示 95%置信区间，颜色表示不同分组。

结果路径：3_Ropls/(pos|neg)/groupID* vs groupID*/PCA/PCA_score.png

4.3.2 偏最小二乘判别分析（PLS-DA）

PLS-DA 是一种有监督的判别分析统计方法，通过建立代谢物表达量与样本类别之间的关系模型来实现对样品类别的预测，有利于发现不同组间的异同点。

为了避免模型过拟合，通常会采用置换检验（Permutation test）对模型进行检验，以保证模型的有效性。具体方法是将每个样本的分组标记随机打乱后再进行建模和预测，每次建模都对应着一组 R2 和 Q2 的值。通常来说，预测的 R2 和 Q2 小于真实分组的 R2 和 Q2，可以表明模型未“过拟合”[17]。

模型质量评估标准包括 R2X、R2Y 和 Q2 这三个指标，这些指标越接近 1 表示模型拟合数据效果越好。其中，R2X 和 R2Y：分别表示模型对自变量 X 和因变量 Y 的解释率；Q2 是通过对模型进行交叉验证计算得出的，用以评价模型的预测能力，通常 Q2> 0.5 被认为是有效模型，Q2> 0.9 则表示模型非常优秀。

表 4.2 正离子模式 PLS-DA 模型验证参数

Group	pre	R2X(cum)	R2Y(cum)	Q2(cum)
L vs M vs H vs CK	2	0.179	0.996	-0.0174
M vs H	2	0.629	0.997	0.841
L vs M	2	0.608	0.998	0.716
M vs CK	2	0.597	0.993	0.625
L vs CK	2	0.519	0.997	0.785

注：pre：主成分数；R2X：模型对 X 变量解释率；R2Y:模型对 Y 变量解释率；Q2：模型预测能力。具体内容见结果文件：

结果路径：3_Ropls/(pos|neg)/PLSDA_summary.xlsx

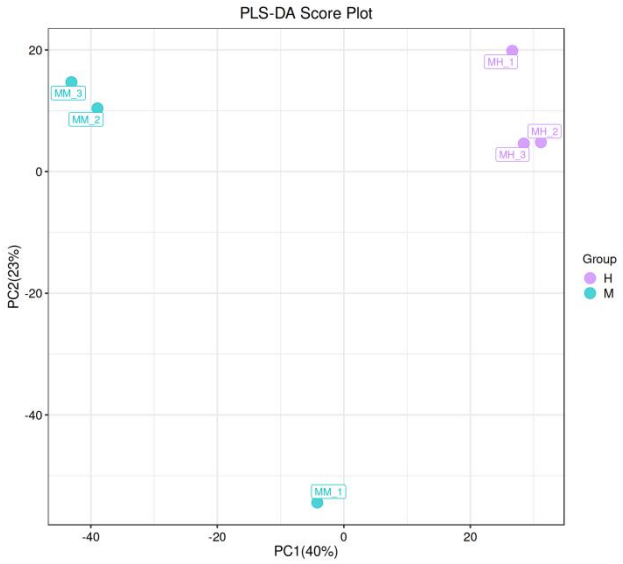

图 4.4 正离子模式 PLS-DA 得分图

说明：横坐标 PC1 表示第一主成分得分值，纵坐标 PC2 表示第二主成分得分值。点表示样本，圈表示 95%置信区间，颜色表示不同分组。

结果路径：3_Ropls/(pos|neg)/groupID* vs groupID*/PLSDA/PLSDA_score.png

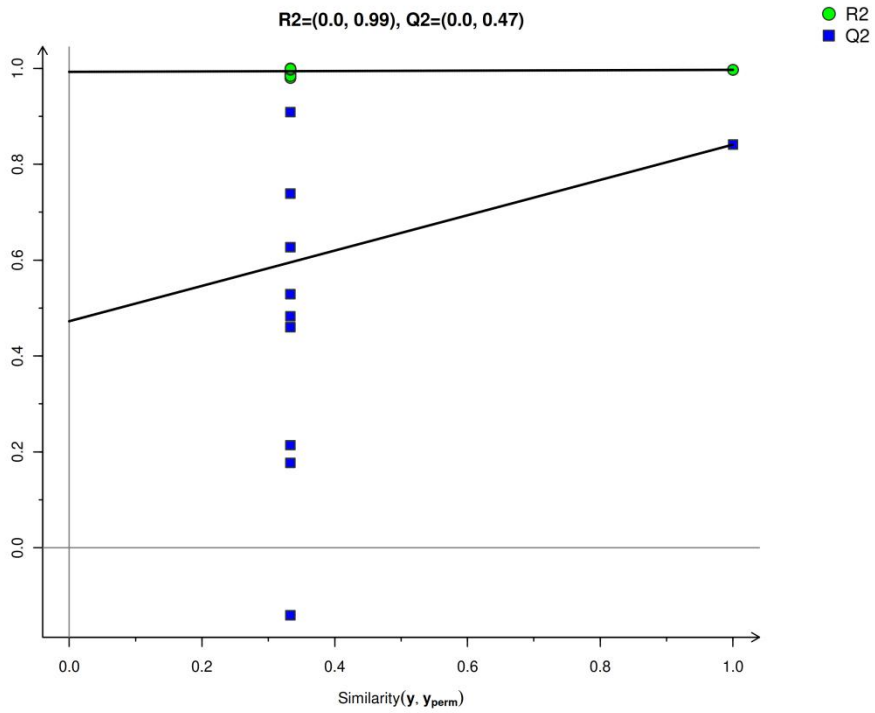

图 4.5 正离子模式 PLS-DA 置换检验图

说明：横坐标表示样本真实分组与 100 次随机分组的相似性，纵坐标表示模型评价参数，右上角 Q2 和 R2 点表示真实分组的模型评价参数。当满足左侧的所有蓝色 Q2 值都低于右侧的原始点，或者 Q2 点的蓝色回归线与垂直轴 （左侧）在零或以下相交

结果路径：3_Ropls/(pos|neg)/groupID* vs groupID*/PLSDA/PLSDA_permutation.png

4.3.3 正交-偏最小二乘判别分析（OPLS-DA）

正交-偏最小二乘判别分析(Orthogonal Projections to Latent Structures Discriminant Analysis, OPLS-DA)为 PLS-DA 的扩展。相比于 PLS-DA，该方法可以在不降低模型预测能力的前提下，有效减少模型的复杂性和增强模型的解释能力[6]，从而最大程度查看组间差异。正交信号校正技术，将 X 矩阵信息分解成与 Y 相关和不相关的两类信息，然后过滤掉与分类无关的信息，相关的信息主要集中在第一个预测主成分。

与 PLS-DA 模型相同，OPLS-DA 同样可以用 R2、Q2 来评价模型效果，并进行置换检验。同时，可以通过 VIP 值（Variable Importance for the Projection）来说明变量对模型的贡献度，可用于衡量各代谢物积累差异对各组样本分类判别的影响强度和解释能力。

通常来说，当某个变量的 VIP > 1 时，说明该变量是重要的，可以作为潜在生物标记物的筛选条件之一。

表 4.3 正离子模式 OPLS-DA 模型验证参数

Group	pre	R2X(cum)	R2Y(cum)	Q2(cum)
L vs M vs H vs CK	1+1+o	0.179	0.996	0.367

M vs H	1+1+o	0.629	0.997	0.691
L vs M	1+1+o	0.608	0.998	0.56
M vs CK	1+1+o	0.597	0.993	0.267
L vs CK	1+1+o	0.519	0.997	0.345

注：pre，主成分数；R2X，模型（对 X 变量数据集）可解释度；R2Y，模型（对 Y 变量数据集）可解释度；Q2，模型可预测度。默认展示前五条，具体内容见结果文件。

结果路径：3_Ropls/(pos|neg)/OPLSDA_summary.xlsx

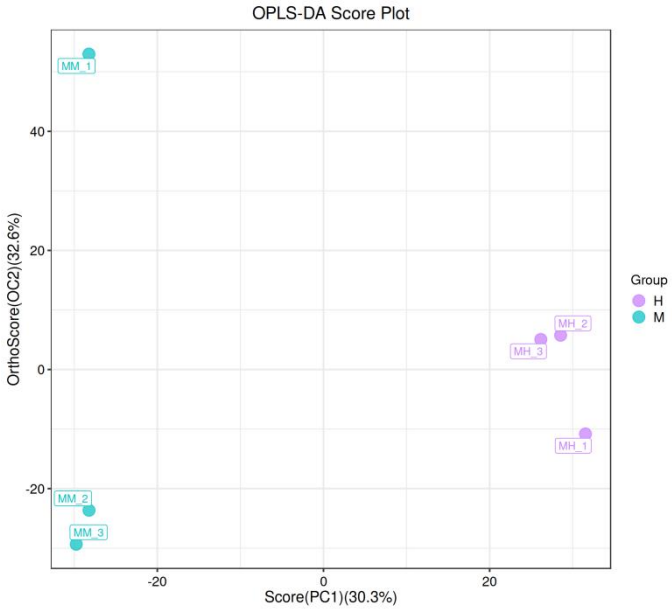

图 4.6 正离子模式 OPLS-DA 得分图

说明：横坐标 PC1 表示第一主成分得分值，纵坐标 OC2 表示第一正交成分得分值。点表示实验样本，颜色表示不同分组。横坐标看组间差异，纵坐标看组内差异，组内样本越聚集，组间样本越分散，说明结果越可靠。

结果路径：3_Ropls/(pos|neg)/groupID* vs groupID*/OPLSDA/OPLSDA_score.png

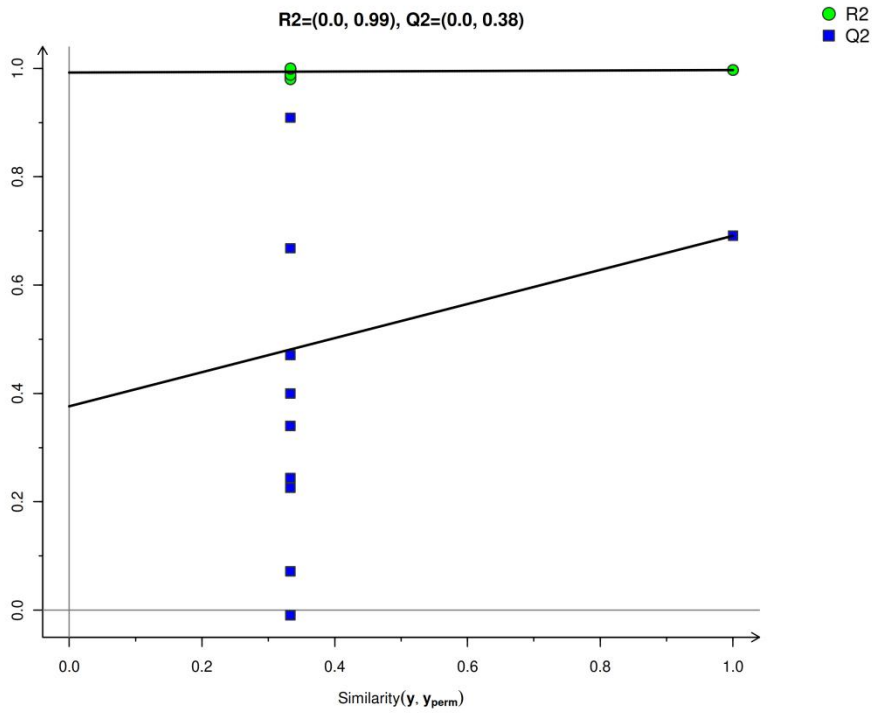

图 4.7 正离子模式 OPLS-DA 置换检验图

说明：横坐标表示样本真实分组与 100 次随机分组的相似性，纵坐标表示模型评价参数，右上角 Q2 和 R2 点表示真实分组的模型评价参数。当满足预测的 R2 和 Q2 点均低于右上的原始 R2 和 Q2 点(图中最右的蓝色 Q2 点有可能和绿色 R2 点重合在最右上角)，说明模型可靠。

结果路径：3_Ropls/(pos|neg)/groupID* vs groupID*/OPLSDA/OPLSDA_permutation.png

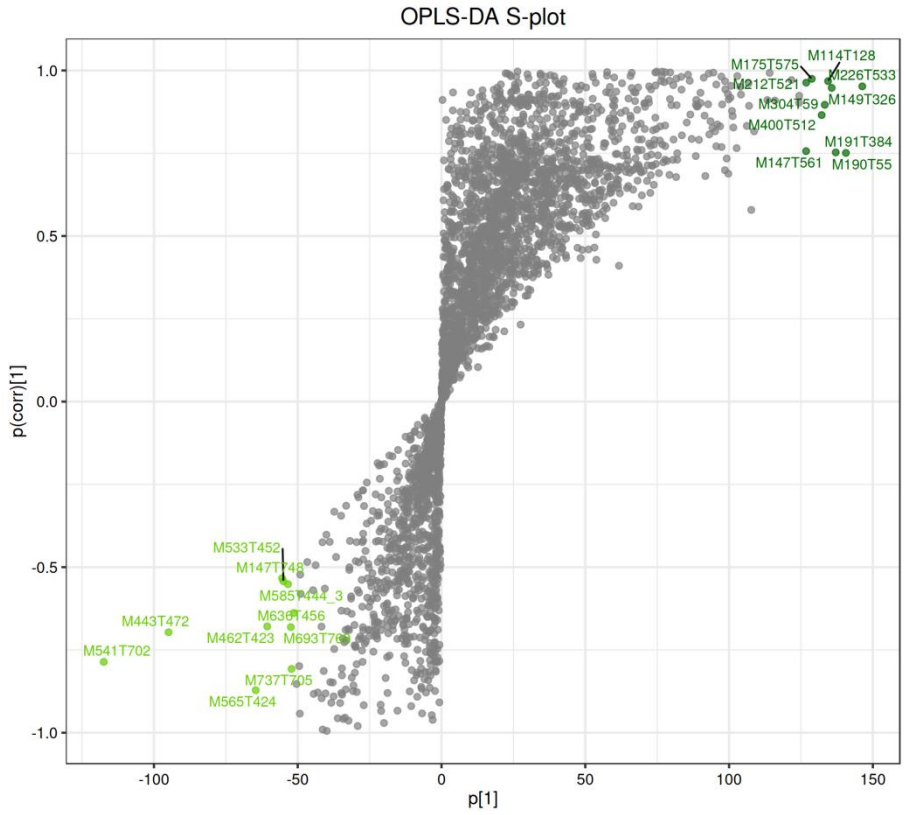

图 4.8 正离子模式 Splot 图

说明:横坐标表示主成份与代谢物的协相关系数,纵坐标表示主成份与代谢物的相关系数。S-plot 图一般用来挑选与正交过程中主成分的相关性比较强的代谢物。越靠近两个角的代谢物重要度越强。默认分别展示右上角与左下角前 10 个分子名称。

结果路径：3_Ropls/(pos|neg)/groupID* vs groupID*/OPLSDA/OPLSDA_Splot.png

4.4 差异代谢物分析

4.4.1 差异统计

基于参考谱图库比对分析得到的代谢物注释结果，以及一级特征峰差异分析结果，根据设置的 Pvalue 和 VIP 阈值进行筛选[7]，得到差异代谢物结果。

表 4.7 差异代谢物统计表

组别	总代谢物数	上调差异代谢物数	下调差异代谢物数	总差异代谢物数
L vs CK	515	19	16	35
M vs CK	515	16	11	27
H vs CK	515	38	16	54
L vs M	515	26	14	40
L vs H	515	24	33	57

注：仅两组比较有差异上下调代谢物。默认展示前五条，具体内容见结果文件。

结果路径： 6_MetDiff/SigDiff.Count.Stat.xlsx

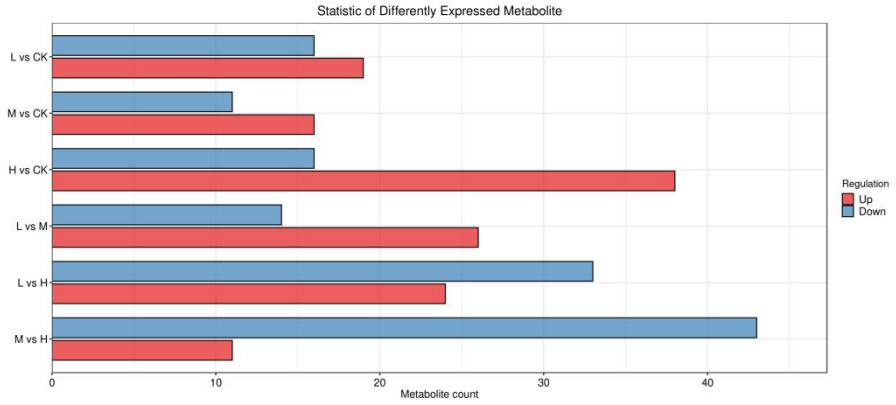

图 4.14 差异代谢物统计柱状图

说明：X 轴表示差异代谢物数目，Y 轴表示比较组别。红色表示上调数目，蓝色表示下调数目。

结果路径：6_MetDiff/SigDiff.Count.Stat.png

4.4.2 韦恩图

韦恩图（Venn Diagram）和 Upset 韦恩图可用于统计不同比对组别中所共有和特有的差异代谢物数目。（单个比较组别不进行韦恩图分析；如果比较组别超过六组，仅提供 Upset 韦恩图；Upset 韦恩图仅展示数目大于 10 的集合）

图 4.15 差异代谢物韦恩图

说明：不同颜色区域代表不同组别的差异代谢物，重叠区域为多组别共有的差异代谢物。默认比对方式小于等于五组时提供。

结果路径：6_MetDiff/VennPlot.png

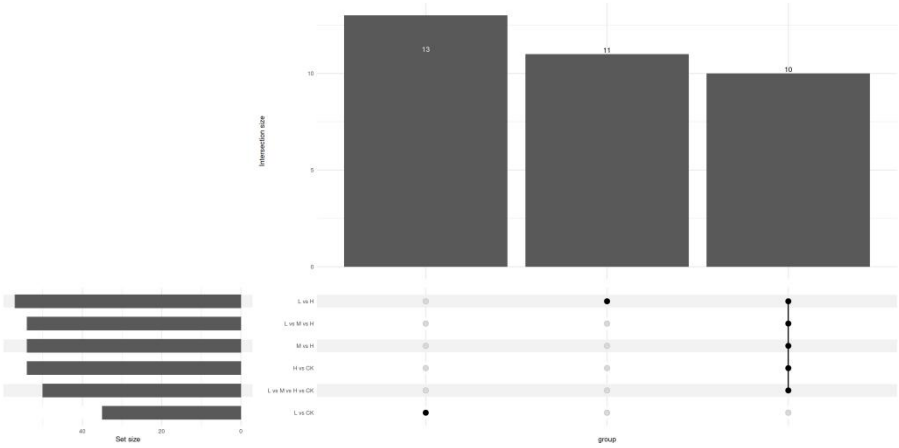

图 4.16 差异代谢物 Upset 韦恩图

说明：左侧柱形图表示比较组别中鉴定到的差异代谢物数目，下方的点和连线表示所选的比较组别集合，右侧柱形图表示仅在所选的比较组别集合中共同鉴定到的差异代谢物数目。默认显示数目大于 10 的交集。

结果路径：6_MetDiff/VennUpSet.png

4.4.3 差异代谢物详情

选择单个比较组间筛选的差异代谢物结果，示例如下表。

表 4.8 差异代谢物鉴定结果（前 5 行）

Name	mz	VIP	log2(FC)	P value	Formula	KEGG
Cyclohexylamine	100.0755	1.7	-6.29	3.0694E-02	C6H13N	C00571
L-Valine	100.0757	1.7	-1.14	5.2693E-03	C5H11NO2	C00183
2'-Aminoacetophenone	119.0732	1.7	-3.61	4.7838E-03	C8H9NO	
Carbamic acid	123.0405	1.6	-1.03	3.6422E-02	CH3NO2	C01563
Niacinamide	123.0555	1.8	-3.73	6.611E-04	C6H6N2O	C00153

注：Name：鉴定物质的名称；mz：质荷比；VIP，OPLS-DA 模型的重要性值投影；log2(FC)，差异倍数的 log2 值；P value，统计学 p 值，越小说明差异越显著；Formula：代谢物分子式；KEGG：KEGG 化合物编号；

结果路径： 6_MetDiff/groupID* vs groupID*/Sig.diff.table.xlsx

4.4.4 差异热图

采用 R Pheatmap 程序包对二级差异代谢物矩阵数据进行数据缩放（Scale），并对样本和差异代谢物进行双向聚类，绘制聚类热图。

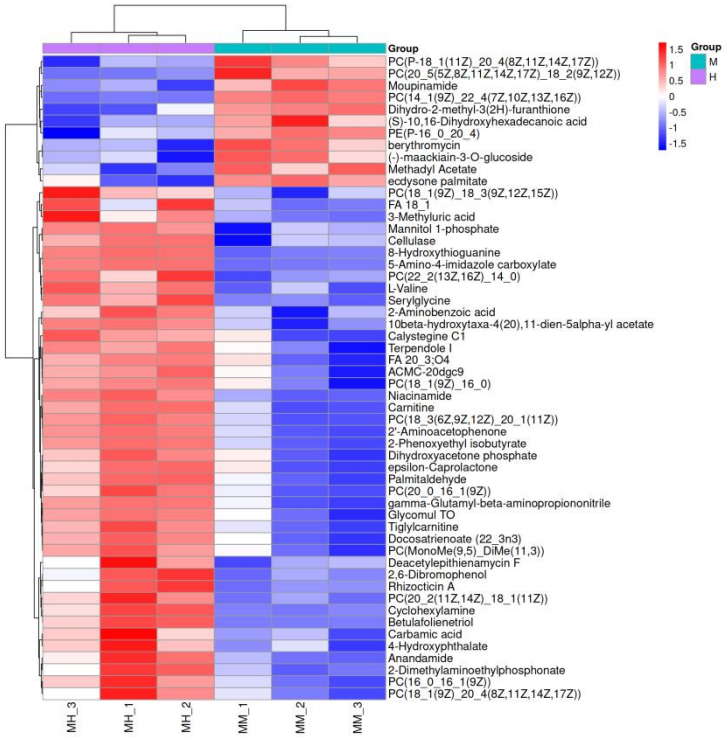

图 4.17 差异代谢物聚类热图

说明：其中横坐标代表样本，纵坐标代表代谢物，颜色表示代谢物相对峰强度，颜色越红表达量越高，越蓝表达量越低。图中对代谢物和样本均进行聚类分析，左侧的聚类树为代谢物聚类树，代谢物数目超过 150 个则不显示代谢物名称，图中上方的聚类树为样本聚类树。

结果路径：6_MetDiff/groupID* vs groupID*/heatmap/heatmap.png

4.4.5 火山图

火山图可直观的展示两组样本间的差异代谢物的分布情况和变化趋势。通常，横坐标用 log2（FC）表示，纵坐标用-log10（P value）表示，为统计检验的显著性 P value 的负对数。定量值变化越大，差异越显著的代谢物分布在左右两端。使用预设的 FC 值、P value 以及 VIP 等的差异代谢物筛选条件绘制代谢物火山图。（仅两个组之间比较时提供）

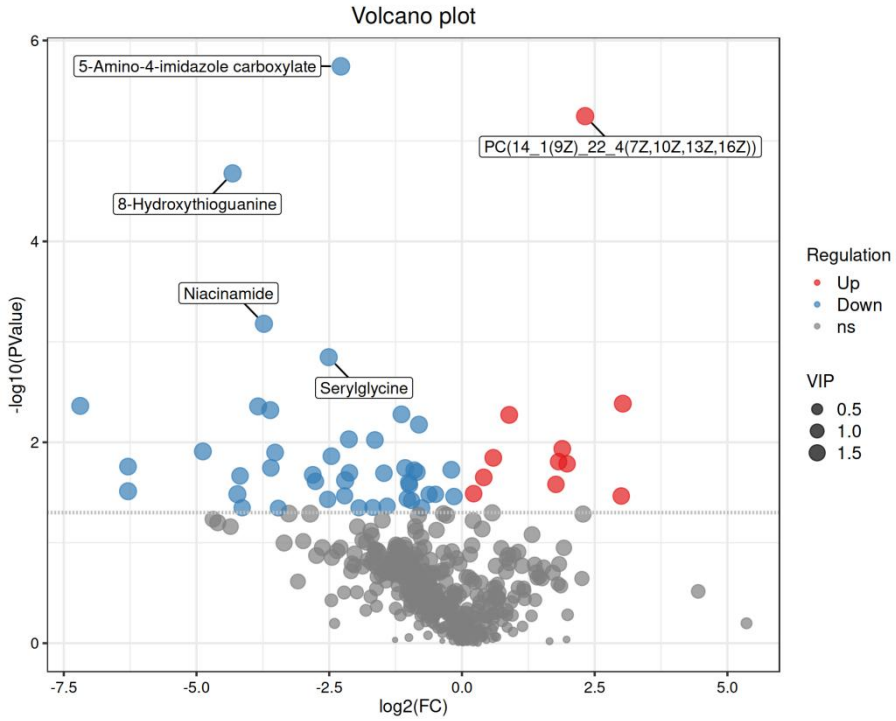

图 4.18 差异代谢物火山图

说明：横坐标表示代谢物在不同分组中的定量值倍数的 Log2 值；纵坐标表示差异显著性水平（P value 的-log10 值）。横坐标绝对值越大，说明代谢物在分组间的表达量倍数差异越大；纵坐标值越大，表明差异表达越显著。图中每一个点表示一种代谢物，红色点代表差异上调，蓝色点代表差异下调，灰色点则表示非差异代谢物，点的大小表示 VIP 值。默认显示 FC 绝对值最大的前 5 个代谢物名称。

结果路径：6_MetDiff/groupID* vs groupID*/volcanoplot/volcanoplot.png

4.4.6 Z 值图

Z-score（标准分数）是基于代谢物的相对含量转换而来的值[13]，用于衡量同一水平上代谢物的相对含量的高低。Z-score 的计算是基于参考数据集（对照组）的平均值和标准差进行的，具体公式表示为：$z = (x - \mu) / \sigma$。其中 x 为某一具体分数，μ 为平均数，σ 为标准差。

通过 Z-score 可以直观的展示实验组相比对照组，在不同分组和样本间，代谢物定量值的整体变化趋势和差异程度。默认对所有差异代谢物绘图，当差异代谢物超过 50 个时，展示 Pvalue 最显著的前 50 个差异代谢物。

图 4.19 Z-score 图

说明：横坐标为代谢物在样本中相对含量经过换算的 Z-score 数值，纵坐标为代谢物名称，点的颜色代表不同组别。越靠近右侧，说明当前代谢物在该样本中相对含量较高，越靠近左侧，说明当前代谢物含量较低。

结果路径：6_MetDiff/groupID* vs groupID*/zscore/zscorePlot.png

4.4.7 关联热图

进行差异代谢物相关性分析有利于进一步了解生物状态变化过程中代谢物之间的相互调节关系，代谢物相关性往往揭示了代谢物之间变化的协同性：与某类代谢物变化趋势相同，则为正相关；与某类代谢物变化趋势相反，则为负相关。通过计算差异代谢物两两之间的皮尔逊相关系数 r 来分析各个代谢物间的相关性。相关系数 r 的值在-1 和 1 之间，可以是此范围内的任何值。正相关时，r 值在 0 和 1 之间；负相关时，r 值在-1 和 0 之间。r 的绝对值越接近 1，两变量的关联程度越强，r 的绝对值越接近 0，两变量的关联程度越弱。同时

对代谢物相关性显著性进行统计检验分析[14]，选用显著性 P value < 0.05 为显著相关。

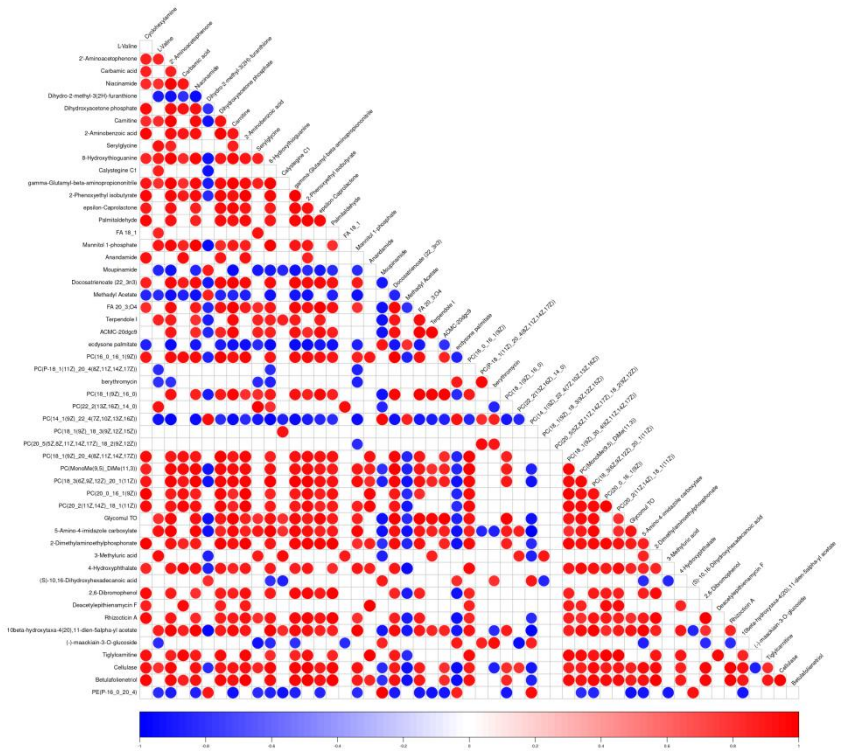

图 4.20 差异代谢物关联热图

说明：纵坐标和斜纵坐标都代表差异代谢物的名称，颜色代表相关性，红色表示正相关，蓝色表示负相关，颜色越深相关性越高。

结果路径：6_MetDiff/groupID* vs groupID*/cor/corplot.png

4.4.8 散点图

根据代谢物的质荷比与 P value 绘制差异散点图， 可在代谢物质荷比大小分布下展示差异特征情况。

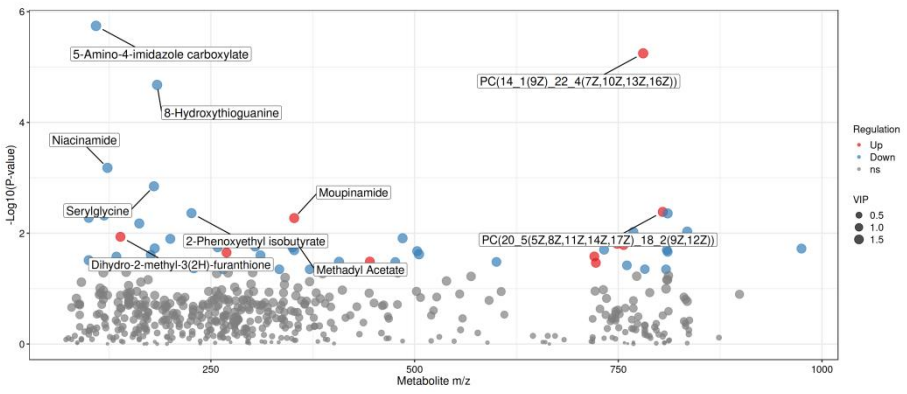

图 4.21 差异代谢物质荷比与 P 值散点图

说明：横坐标表示某代谢物的质荷比；纵坐标表示 P 值的-log10 的对数值，图中每一个点表示一种代谢物。红色点代表差异上调，蓝色点代表差异下调，灰色点表示不满足差异筛选条件的代谢物，点大小表示 VIP 值。三组及以上比较时无 FC 信息，用红点表示差异代谢物。默认显示 P value 最小的前 5 个代谢物名称。

结果路径：6_MetDiff/groupID* vs groupID*/scatterplot/scatterplot.png

4.4.9 柱状图

柱状图是一种以长方形的长度为变量的表达图形统计报告图，由一系列高度不等的纵向条纹表示数据分布的情况，并添加误差线展示标准误差范围。箱式图和小提琴图用作绘制一组数据整体分布情况的统计图，展示数据分布特点。对每一个代谢物的定量值计算显著性 Pvalue 值，不同颜色表示样本的不同分组，并分别以柱状图、箱形图、小提琴图进行展示，可以很方便地进行多组间数据分布特征的比较。

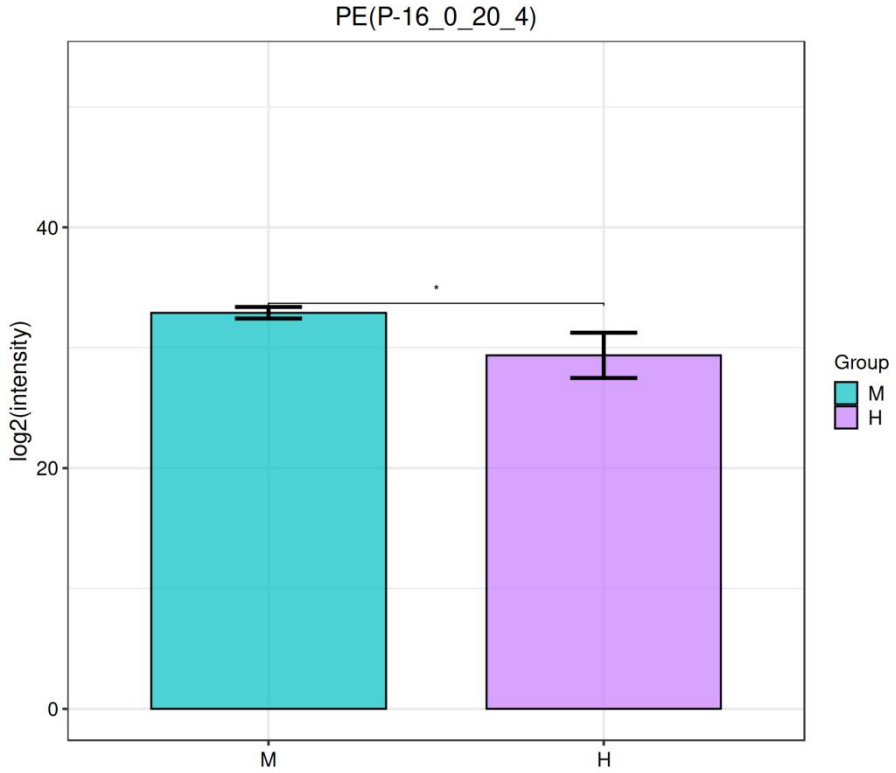

图 4.22 柱状图

说明：横坐标为不同组别，纵坐标为代谢物定量值范围。两组间星表示两组间的差异显著性，*P＜0.05，** P＜0.01，*** P＜0.001，**** P＜0.0001。三组及以上比较额外显示整体 P value。

结果路径：6_MetDiff/groupID* vs groupID*/barplot/

结果路径：6_MetDiff/groupID* vs groupID*/boxplot/

结果路径：6_MetDiff/groupID* vs groupID*/violin/

4.6 KEGG 富集分析

KEGG 数据库是连接已知分子间相互作用的信息网络，如代谢通路、复合物、生化反应。KEGG 途径主要包括：代谢、遗传信息处理、环境信息处理、细胞过程、人类疾病、药物开发等。通过 KEGG 显著性富集分析能确定差异代谢物参与的最主要代谢途径等。

4.6.1 富集分析

KEGG 显著性富集分析方法[15]是以 KEGG 通路为单位，基于超几何分布检验，拓扑分析采用中介中心性（betweenness），拓扑分析旨在根据给定基因或代谢物在途径中的位置来评估其是否在生物学反应中起重要作用。

表 4.9 富集结果列表（前 5 行）

Pathway ID	Pathway name	Total	Hits	Pvalue	FDR	Impact
ko04714	Thermogenesis	23	2	4.9328E-03	1.8219E-01	0.087
ko00930	Caprolactam degradation	26	2	6.2822E-03	1.8219E-01	0.1507
ko00460	Cyanoamino acid metabolism	45	2	1.8123E-02	2.6409E-01	0.0014
ko00760	Nicotinate and nicotinamide metabolism	55	2	2.6419E-02	2.6409E-01	0.0623
ko00051	Fructose and mannose metabolism	55	2	2.6419E-02	2.6409E-01	0.0715

注：Pathway ID，目标通路在 KEGG 数据库中的 ID 号；
Pathway name，目标通路名称；
Total，目标代谢通路中代谢物的总数；
Hits，目标代谢通路中总体差异代谢物数量；
Pvalue，超几何分布检验的 P 值，P 值越小，代表检测到的差异代谢物对该通路影响越显著；
FDR，假阳性校正后值，采用 BH（Benjaminiand Hochberg）法计算，数值越大假阳性的可能越高；
Impact，代谢通路影响值，越大说明本次检测出的差异代谢物对目标通路的影响越大。
结果路径：6_MetDiff/groupID* vs groupID*/KEGG_enrich/enrichtable.xlsx

为了便于查看差异代谢物在通路图中的分布情况，将差异代谢物标注到通路图中，在 KEGG 通路图中，圆圈代表代谢物，其中蓝色圆圈点表示下调差异代谢物，红色圆圈点表示

上调差异代谢物。

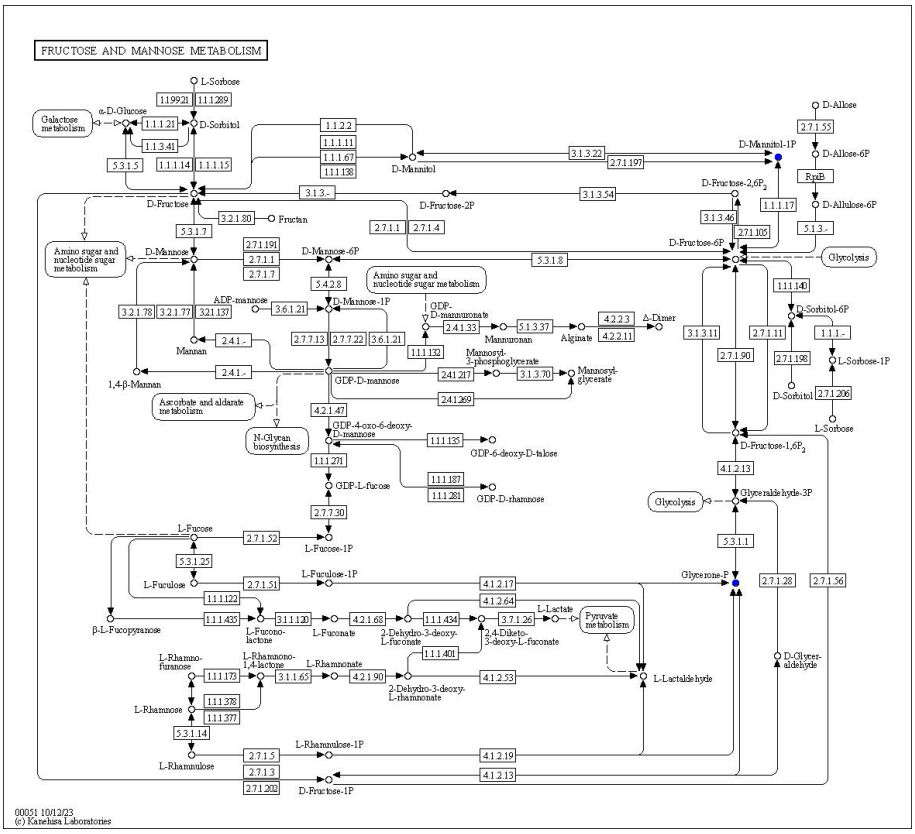

图 4.23 通路图

说明：方框表示蛋白分子，圆圈表示代谢分子。红色表示该差异上调，蓝色表示差异下调。多组比较无上下调信息，默认用蓝色表示代谢物。

文件路径：6_MetDiff/groupID* vs groupID*/KEGG_enrich/pathway_graph/

4.6.2 差异数目条形图

统计富集 KEGG 通路对应的上下调差异代谢物数目，并绘制柱形图。默认绘制所有富集通路，当富集通路超过 20 条时，展示 Pvalue 最显著的前 20 条富集通路。

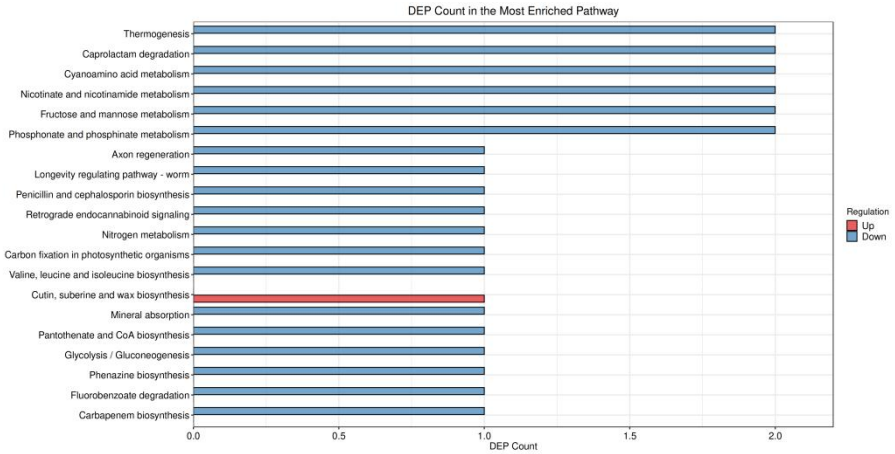

图 4.24 差异数目条形图

说明：横坐标表示差异代谢物数目，纵坐标表示代谢通路。红色表示上调数目，蓝色表示下调数目。三组及以上比较无此分析结果。

文件路径：6_MetDiff/groupID* vs groupID*/KEGG_enrich/kegg.diff.count.stat.png

4.6.3 富集条形图

根据上述富集结果，绘制富集通路的条形图。默认绘制所有富集通路，当富集通路超过20条时，展示Pvalue最显著的前20条富集通路。

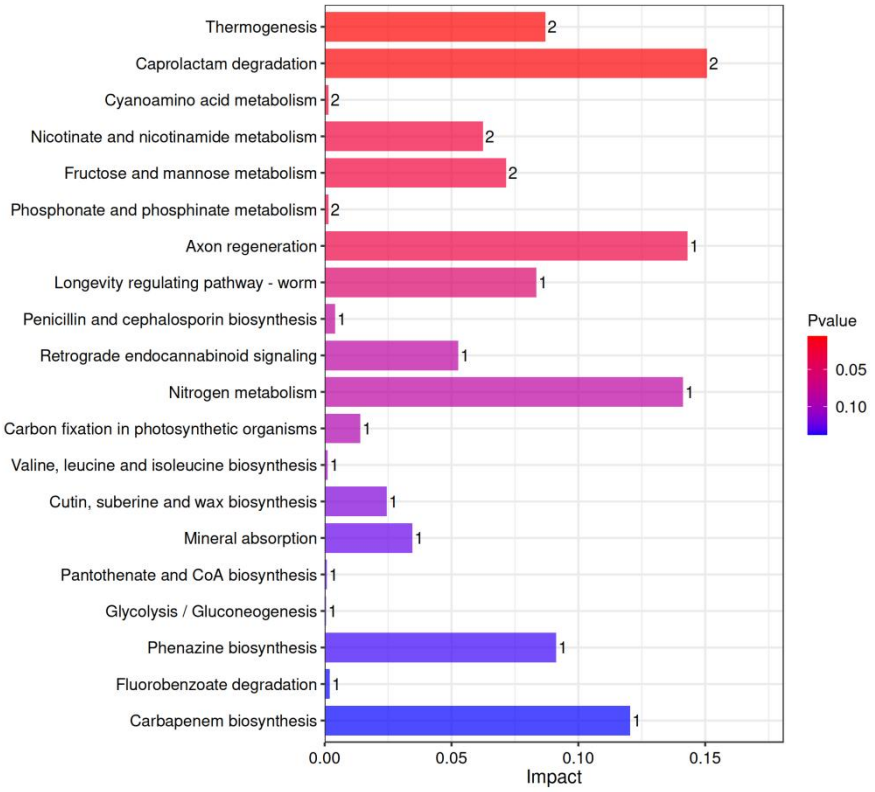

图 4.25 富集条形图

说明：横坐标代表富集到不同代谢通路中的Impact值，纵坐标代表代谢通路。数字表示通路上对应的差异代谢物数目。颜色越红，Pvalue值越小，颜色越蓝，Pvalue值越大。

文件路径：6_MetDiff/groupID* vs groupID*/KEGG_enrich/enrich_barplot.png

4.6.4 富集气泡图

根据上述富集结果，绘制富集通路的气泡图。默认绘制所有富集通路，当富集通路超过20条时，展示Pvalue最显著的前20条富集通路。

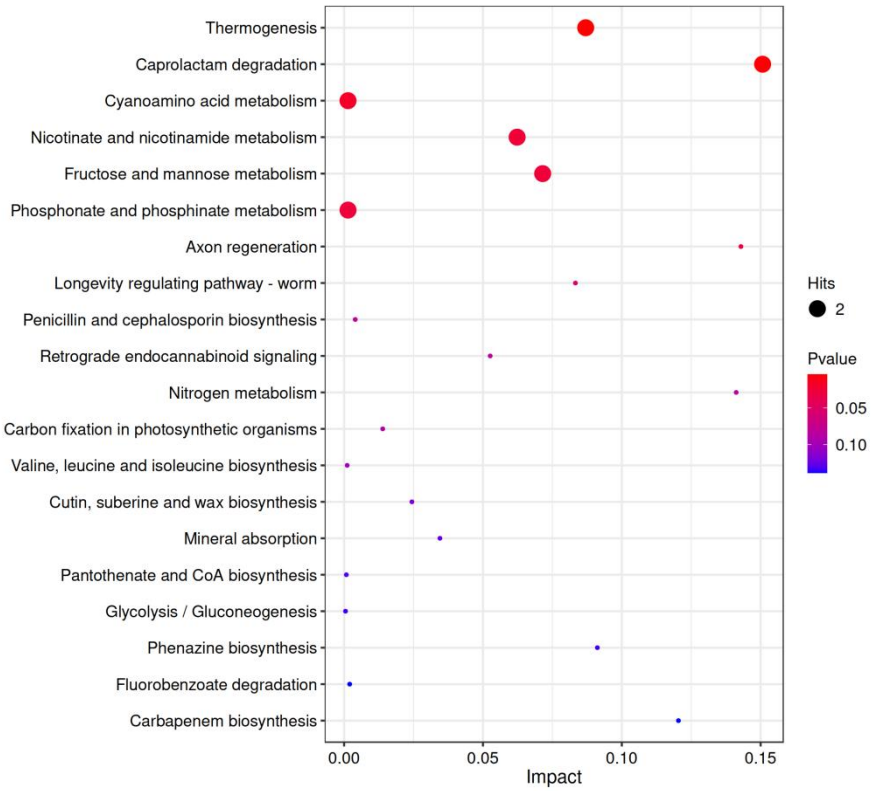

图 4.26 富集气泡图

说明：横坐标是富集到不同代谢通路中的 Impact 值，纵坐标是富集通路。点大小表示通路上对应的差异代谢物数目。颜色越红，Pvalue 值越小，颜色越蓝，Pvalue 值越大。

文件路径：6_MetDiff/groupID* vs groupID*/KEGG_enrich/enrich_bubbleplot.png

4.6.5 富集散点图

根据上述富集结果，取全部的富集通路绘制散点图，按照 Pvalue 值从小到大排序，默认展示前 5 条富集通路名称。

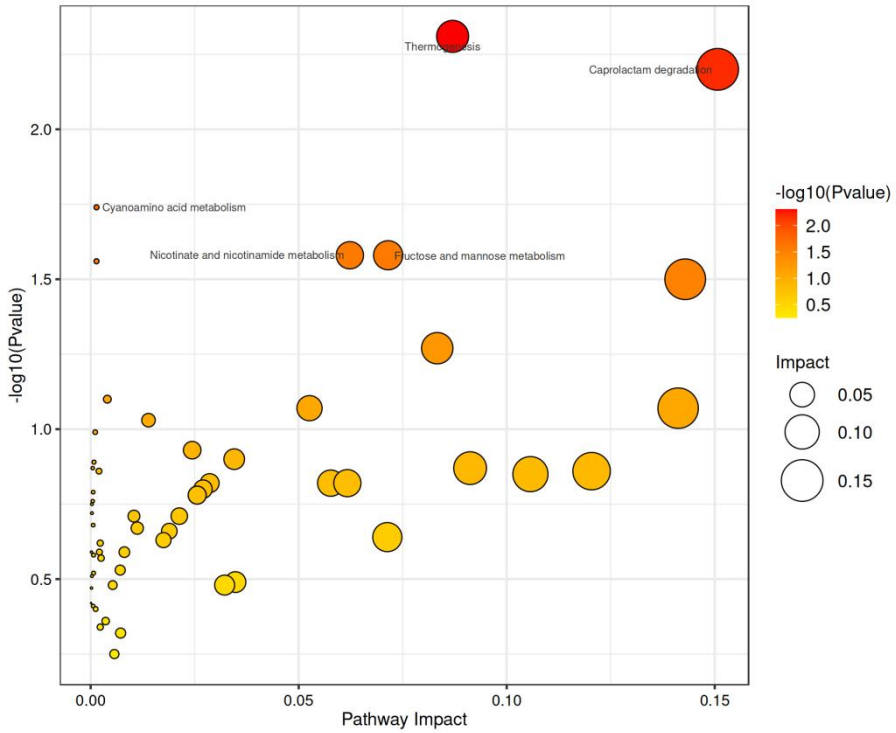

图 4.27 富集散点图

说明：横坐标是富集到不同代谢通路中的 Impact 值，纵坐标是-log10（Pvalue）值。颜色越深，Pvalue 值越小，颜色越浅，Pvalue 值越大。通路越靠近右上角，说明差异代谢物在该通路富集显著，对该通路影响越大。

文件路径：6_MetDiff/groupID* vs groupID*/KEGG_enrich/enrich_scatterplot.png

4.6.6 富集网络图

取富集通路结果，构建通路与差异代谢物之间的关系网络图，可以展示差异代谢物与通路之间的所属关系。默认绘制所有富集通路，当富集通路超过 10 条时，展示 Pvalue 最显著的前 10 条富集通路。

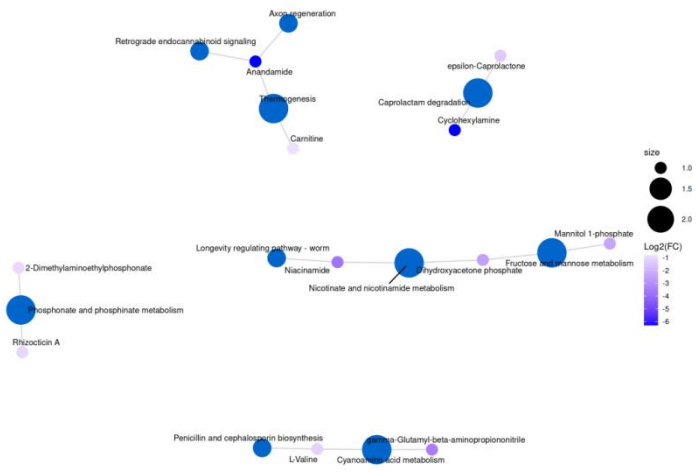

图 4.28 富集网络图

说明：蓝色点表示通路，其他点表示代谢物。通路点的大小表示与之相连的代谢物数目，相连数目越多，点越大，代谢物点颜色表示 log2(FC)值的大小，红色表示差异上调，蓝色表示差异下调，颜色越深，表示差异程度越大，三组及以上比较组方式差异代谢物不显示差异倍数信息(即颜色)。

文件路径：6_MetDiff/groupID* vs groupID*/KEGG_enrich/enrich_network.png

4.6.7 DAscore 图

通过分析富集通路上调与下调代谢物数目之差，计算 DA score 得分，用于评估通路可能存在激活或抑制的状态。默认绘制所有富集通路，当富集通路超过 20 条时，展示 Pvalue 最显著的前 20 条富集通路。

计算公式 DA-score =（上调物质数-下调物质数）/该通路上差异总物质数[16]

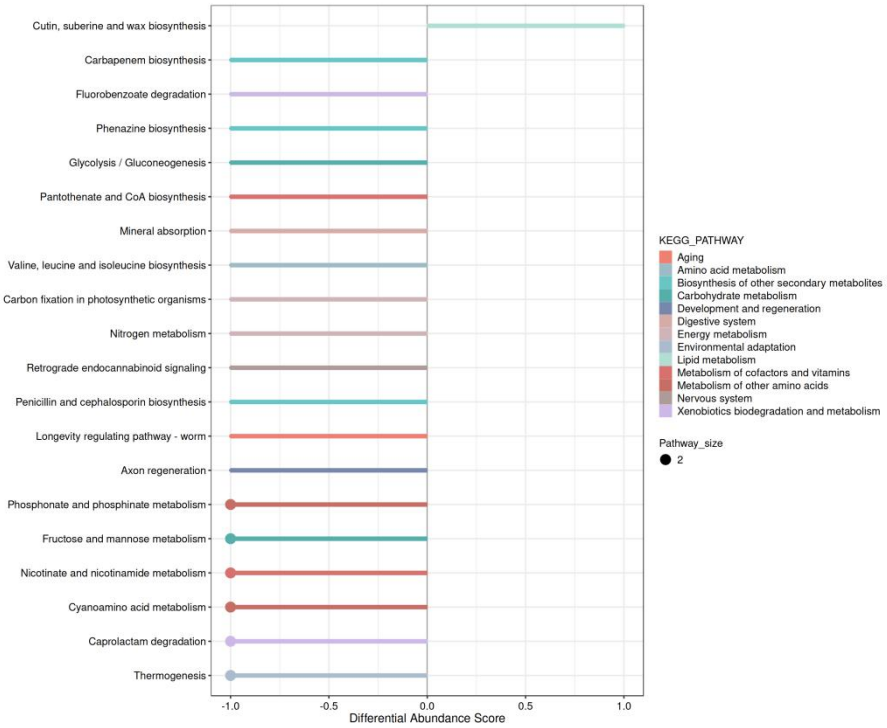

图 4.29 差异丰度得分图

说明：横坐标是 DA-score 值，纵坐标是代谢通路，柱形顶部点大小表示该通路上富集的差异代谢物数目，点越大，表示差异代谢物数目越多。颜色表示通路属于二级分类。

文件路径：6_MetDiff/groupID* vs groupID*/KEGG_enrich/enrich_dascore.png

4.6.8 调控网络分析

在生物化学领域，代谢通路是指细胞中代谢物质在酶的作用下转化为新的代谢物质过程中所发生的一系列生物化学反应。而代谢网络则是指由代谢反应以及调节这些反应的调控机

制所组成的描述细胞内代谢和生理过程的网络。基于前期分析得到的差异代谢物来构建基于网络的富集分析。结果包括代谢通路、模块、酶、反应及代谢物。可以反映特定研究条件下代谢通路之间的交集以及靶向潜在的酶和代谢物，表明扰动如何在通路水平上传播以及通路如何相互影响，从而增强了结果的可解释性。

在取得每组对比差异代谢物的匹配信息后，对对应物种 Homo sapiens (human)的 KEGG 数据库进行通路搜索和调控互作网络分析 16，以网络图（network plot）展示。此分析仅支持人、大鼠、小鼠。

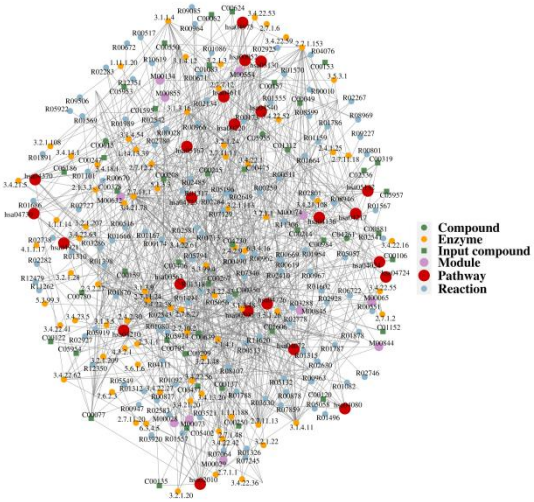

图 4.30 调控网络图

说明：图中，红色圆点代表一条代谢通路，黄色圆点代表一种物质相关调控酶信息，绿色圆点代表一个代谢通路的背景物质，紫色圆点代表一类物质分子模块信息，蓝色圆点代表一种物质化学互作反应，绿色方块代表此次对比得到的差异物质

文件路径：6_MetDiff/groupID* vs groupID*/KEGG_enrich/enrich_regulation_network.png

4.7 MSEA 富集分析

4.8 疾病关联分析

5. 参考文献

[1] Rasmussen JA, Villumsen KR, Ernst M, et al. A multi-omics approach unravels metagenomic and metabolic alterations of a probiotic and synbiotic additive in rainbow trout (Oncorhynchus mykiss)[J]. Microbiome. 2022, 10(1):21.

[2] Navarro-Reig M, Jaumot J, García-Reiriz A, et al. Evaluation of changes induced in rice metabolome by Cd and Cu exposure using LC-MS with XCMS and MCR-ALS data analysis

strategies[J]. Analytical and Bioanalytical Chemistry, 2015, 407(29):8835-47.
[3] Want E J, Masson P, Michopoulos F, et al. Global metabolic profiling of animal and human tissues via UPLC-MS[J]. Nature Protocols, 2013, 8(1):17-32.
[4] Dunn W B, Broadhurst D, Begley P, et al. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry.[J]. Nature protocols, 2011, 6(7):1060-83.
[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.[J]. Journal of Proteome Research, 2015, 14(8):3322-35.
[6] Trygg J, Wold S. Orthogonal projections to latent structures (O-PLS)[J]. Journal of Chemometrics, 2010, 16(3):119-128.
[7] Kieffer D A, Piccolo B D, Vaziri N D, et al. Resistant starch alters gut microbiome and metabolomic profiles concurrent with amelioration of chronic kidney disease in rats[J]. Am J Physiol Renal Physiol, 2016, 310(9):F857-71.
[8] Wishart D S, Dan T, Knox C, et al. HMDB: The human metabolome database[J]. Nucleic Acids Research, 2007, 35(Database issue):D521-6.
[9] Horai H, Arita M, Kanaya S, et al. MassBank: a public repository for sharing mass spectral data for life sciences[J]. Journal of Mass Spectrometry, 2010, 45(7):703-14.
[10] Manish S, Eoin F, Dawn C, et al. LMSD: LIPID MAPS structure database[J]. Nucleic Acids Research, 2007, 35(Database issue):527-32.
[11] Abdelrazig S, Safo L, Rance G A, et al. Metabolic characterisation of Magnetospirillum gryphiswaldense MSR-1 using LC-MS-based metabolite profiling[J]. RSC Advances, 2020, 10(54): 32548-60.
[12] Ogata H, Goto S, Sato K, et al. KEGG: kyoto Encyclopedia of Genes and Genomes[J]. Nucleic Acids Research, 1999, 27(1):29-34.
[13] Sreekumar A, Poisson L M, Rajendiran T M, et al. Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression[J]. Nature, 2009, 457(7231):910-914.
[14] Rao G, Sui J, Zhang J. Metabolomics reveals significant variations in metabolites and correlations regarding the maturation of walnuts (Juglans regia L.)[J]. Biology Open, 2016, 5(6):829-836.
[15] Xia J, Wishart D S. Web-based inference of biological patterns, functions and pathways from metabolomic data using MetaboAnalyst[J]. Nature Protocols, 2011, 6(6):743-60.
[16] Prabhu A H, Kant S, Kesarwani P, et al. Integrative cross-platform analyses identify enhanced heterotrophy as a metabolic hallmark in glioblastoma[J]. Neuro Oncol, 2019, 21(3):337-347.
[17] XIANGLI, KONG, XIAOQIN, et al. Analysis of plasma metabolic biomarkers in the development of 4-nitroquinoline-1-oxide-induced oral carcinogenesis in rats.[J]. Oncology Letters, 2015.

6. 附录

6.1 分析方法中文版

数据预处理

首先，通过 Proteowizard 软件包（v3.0.8789）中 MSConvert 工具将原始质谱下机文件转换为 mzXML 文件格式[1]。采用 R XCMS（v3.12.0）软件包进行峰检测、峰过滤、峰对齐处理[2]，得到代谢物定量列表，参数设置有 bw=2，ppm=15，peakwidth=c(5, 30)，mzwid=0.015，mzdiff=0.01，method="centWave"。然后，通过基于 QC 样本的支持向量回归校正，消除系统误差，并在质控与质保过程中过滤掉 QC 样本中 RSD > 30%的物质，用于后续的数据分析。

采用公共数据库 HMDB[8]、massbank[9]、LipidMaps[10]、mzcloud[11]、KEGG[12]及自建标准品库等谱图数据库进行物质鉴定（搜库），参数设置为 ppm < 30 ppm，得到代谢物定性结果。具体原理是根据一级质谱中母离子的质荷比（m/z）确定代谢物的分子量，通过质量数偏差（ppm）以及加合离子等信息进行分子式预测，然后与数据库进行匹配，同时，在定量列表中，检测到二级谱图的代谢物与数据库中每个代谢物的碎片离子等信息进行匹配，实现代谢物的二级鉴定。

数据分析

采用 R 软件包 RopIs[5]分别对样本数据进行主成分分析（PCA）、偏最小二乘判别分析（PLS-DA）、正交偏最小二乘判别分析（OPLS-DA）降维分析。用置换检验方法对模型进行过拟合检验。R2X 和 R2Y 分别表示所建模型对 X 和 Y 矩阵的解释率，Q2 标示模型的预测能力，它们的值越接近于 1 表明模型的拟合度越好，训练集的样本越能够被准确划分到其原始归属中。根据统计检验计算 P value 值、OPLS-DA 降维方法计算变量投影重要度（VIP）、fold change 计算组间差异倍数，来衡量各代谢物含量对样本分类判别的影响强度和解释能力，辅助标志代谢物的筛选。当 P value 值 < 0.05 和 VIP 值 > 1 时，认为代谢物分子具有统计学显著差异。

通路分析

通路富集分析，基于超几何分布的富集分析方法[15]。富集得到的通路采用 KEGG Mapper 可视化工具进行差异代谢物与通路图的浏览。

6.2 分析方法英文版

Data preprocessing

The raw data were firstly converted to mzXML format by MSConvert in ProteoWizard software package (v3.0.8789)[1] and processed using R XCMS(v3.12.0) for feature detection[2], retention time correction and alignment. Key parameters settings were set as follows: ppm=15, peakwidth=c(5, 30), mzdiff=0.01, method=centWave. The batch effect was then eliminated by correcting the data based on QC samples. Metabolites with RSD > 30% in QC samples were filtered and then used for subsequent data analysis.

The metabolites were identified by accuracy mass and MS/MS data which were matched with HMDB (http://www.hmdb.ca)[8], massbank (http://www.massbank.jp/)[9], KEGG

(https://www.genome.jp/kegg/)[12], LipidMaps (http://www.lipidmaps.org)[10], mzcloud (https://www.mzcloud.org)[11] and the metabolite database bulid by Panomix Biomedical Tech Co., Ltd. (Shuzhou, China). The molecular weight of metabolites was determined according to the m/z (mass-to-charge ratio) of parent ions in MS data. Molecular formula was predicted by ppm (parts per million) and adduct ion, and then matched with the database. At the same time, the MS/MS data from quantitative table of MS/MS data, were matched with the fragment ions and other information of each metabolite in the database, so as to realize the MS/MS identification of metabolites.

Data analysis

Two different multivariate statistical analysis models, unsupervised and supervised, were applied to discriminate the groups (PCA; PLS-DA; OPLS-DA) by R ropls (v1.22.0) package[5]. The statistical significance of P.value was obtained by statistical test between groups. Finally, combined with P.value, VIP (OPLS-DA variable projection importance) and FC (multiple of difference between groups) to screen biomarker metabolites. By default, when P value < 0.05 and VIP value > 1, we think that metabolite were considered to have significant differential expression.

Pathway analysis

Pathway enrichment analysis used the hypergeometric distribution enrichment analysis method to perform functional pathway enrichment and topological analysis of metabolites[15]. The identified metabolites in metabolomics were then mapped to the KEGG pathway for biological interpretation of higher-level systemic functions. The metabolites and corresponding pathways were visualized using KEGG Mapper tool.

6.3 软件列表

分析	软件/方法	版本
数据转置	proteowizard - MSConvert	V3.0.8789
解卷积	XCMS	V3.12.0
聚类热图	R/pheatmap	V1.0.12
多元统计分析	R/ropls	V1.22.0
层次聚类树	R/dendextend	v1.15.2
关联热图	R/cor	v4.0.3
统计检验	R/Wilcox.test/t.test	v4.0.3