



广泛靶向代谢组技术开发结题报告



武汉迈特维尔生物科技有限公司

www.metware.cn



目 录

1 差异代谢物分析结果	3
2 数据结果评估	3
2.1 代谢物定性定量分析	3
2.2 主成分分析（PCA）	4
2.3 聚类分析	5
2.4 重复相关性评估	6
3 数据结果分析	7
3.1 分组主成分分析	7
3.2 正交偏最小二乘法判别分析（OPLS-DA）	9
3.3 代谢物含量差异动态分布	13
3.4 差异代谢物筛选	13
3.5 差异代谢物 KEGG 功能注释及富集分析	25
4 参考文献	29
5 附录	30
5.1 分析方法英文版	30
5.2 分析软件列表以及版本	31
5.3 常见问题	31



result

1 差异代谢物分析结果

表 1: 差异代谢物数目统计表

group name	All sig diff	down regulated	up regulated
DoB_vs_DoS	286	159	127
DoS_vs_DoB	286	127	159

原始文件路径: 结题报告/2.Basic_analysis/Difference_analysis/significant_compound_count.*。

2 数据结果评估

2.1 代谢物定性定量分析

基于本地代谢数据库, 对样本的代谢物进行了质谱定性定量分析。图中多反应监测模式 MRM 代谢物检测多峰图展示了样本中能够检测到的物质, 每个不同颜色的质谱峰代表检测到的一个代谢物。通过三重四级杆筛选出每个物质的特征离子, 在检测器中获得特征离子的信号强度 (CPS), 用 MultiaQuant 软件打开样本下机质谱文件, 进行色谱峰的积分和校正工作, 每个色谱峰的峰面积 (Area) 代表对应物质的相对含量, 最后导出所有色谱峰面积积分数据保存。本实验中所检测到的部分代谢物的代谢物编号、积分数值以及对应代谢物名称等信息见下表:

表 2: 代谢物数量统计表

Index	Compounds	Class I	物质
M69T495	Propynoic acid	Organic acids and derivatives	-
M73T100	Propionic acid	Organic acids and derivatives	-
M85T70	Methyl acrylate	Organic acids and derivatives	-
M93T265	Phenol	Benzenoids	-
M101T64	Acrolein	Organic oxygen compounds	-
M111T67	2-Furancarboxylic acid	Organoheterocyclic compounds	-
M111T95	2-Furancarboxylic acid	Organoheterocyclic compounds	-
M113T496	Trifluoroacetic acid	Organic acids and derivatives	-
M113T51	Imidazole	Organoheterocyclic compounds	-
M115T61	Maleic acid	Organic acids and derivatives	-

原始文件路径: 结题报告/1.Data_assess/all_group/ALL_sample_data.xlsx。



2.2 主成分分析（PCA）

2.2.1 主成分分析原理

采用多元统计分析，可以在最大程度保留原始信息的基础上将高维复杂的数据进行“简化和降维”，建立可靠的数学模型对研究对象的代谢谱特点进行归纳和总结。其中，主成分分析（Principal Component Analysis, PCA）是一种无监督模式识别的多维数据统计分析方法，通过正交变换将一组可能存在相关性的变量转换为一组线性不相关的变量，转换后的这组变量叫主成分。这个分析方法常用来研究如何通过少数几个主成分来揭示多个变量间的内部结构，即从原始变量中导出少数几个主成分，使它们尽可能多地保留原始变量的信息，且彼此间互不相关，通常数学上的处理就是将原来多个指标作线性组合，作为新的综合指标（Eriksson et al., 2006）。

PCA 的数据处理原理：将原始数据压缩成 n 个主成分来描述原始数据集的特征，PC1 表示能描述多维数据矩阵中最明显的特征，PC2 表示除 PC1 之外的所能描述数据矩阵中最显著的特征，PC3……PCn 以此类推。PCA 用 R 软件（www.r-project.org/）的内置统计 prcomp 函数，设置 prcomp 函数参数 scale=True，表示对数据进行 unit variance scaling (UV) 归一化。

2.2.2 总体样本主成分分析

通过对样本进行主成分分析，以便初步了解各组样本之间的总体代谢差异和组内样本之间的变异度大小。PCA 结果显示各组之间代谢组分离趋势，提示样品组间代谢组是否存在差异（Chen et al., 2009）。PCA 得分图如下：

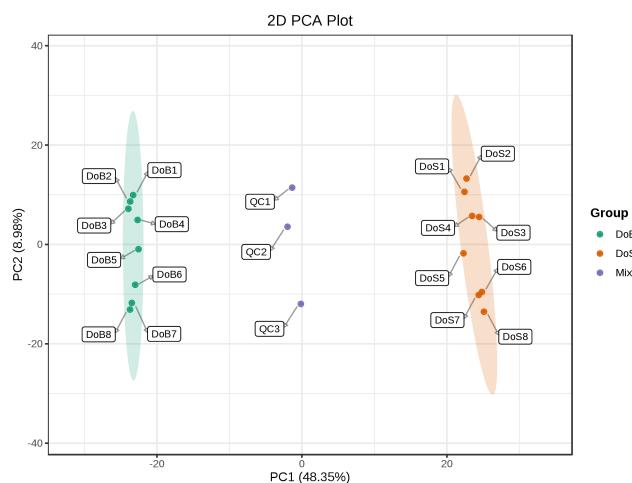


图 1: 各组样品质谱数据的 PCA 得分图

注：PC1 表示第一主成分，PC2 表示第二主成分，PC3 表示第三主成分，百分比表示该主成分对数据集的解释率；图中的每个点表示一个样品，同一个组的样品使用同一种颜色表示。



原始文件路径:

总体样本 PCA 二维结果图见: 结题报告/1.Data_assess/pca/*_PCA.*;

总体样本 PCA 三维结果图见: 结题报告/1.Data_assess/pca/*_PCA3D.*;

总体样本 PCA 前 5 个主成分的可解释变异见: 结题报告/1.Data_assess/pca/*_PCA_variance.*;

总体样本 PCA 所有主成分数据见: 结题报告/1.Data_assess/pca/*_PCA_components.xlsx;

总体样本 PCA 主成分方差的贡献率见: 结题报告/1.Data_assess/pca/*_PCA_variance_proportion.xlsx。

2.3 聚类分析

2.3.1 聚类分析原理

聚类分析（Cluster Analysis）是一种分类的多元统计分析方法。按照个体或样品（Individuals, Objects or Subjects）的特征将它们分类，使同一类别内的个体具有尽可能高的同质性（Homogeneity），而类别之间则应具有尽可能高的异质性（Heterogeneity）。聚类分析主要应用于探索性的研究，其分析的结果可以提供多个可能的解，选择最终的解需要研究者的主观判断和后续的分析。

代谢物含量数据采用归一化处理（Unit Variance Scaling, UV Scaling），通过 R 软件 Complex-Heatmap 包绘制热图，对代谢物在不同样本间的积累模式进行层次聚类分析（Hierarchical Cluster Analysis, HCA）。

2.3.2 聚类分析结果

先对数据进行归一化处理，对所有样品进行聚类热图分析，并使用 R 程序脚本绘制聚类热图。

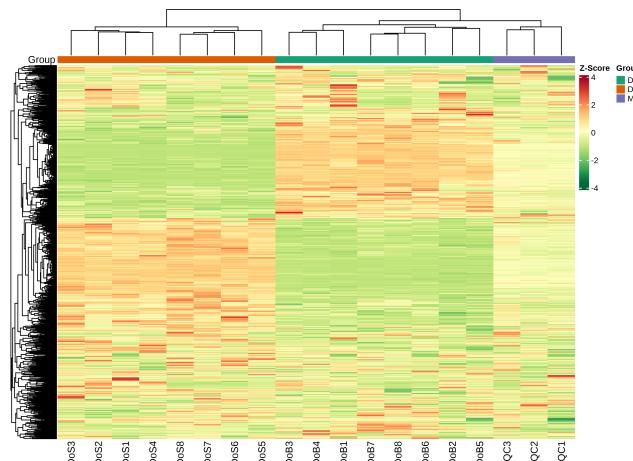


图 2: 样品总体聚类图

注: 横向为样品名称, 纵向为代谢物信息, Group 为分组, Class 为物质分类, 不同颜色为相对含量标准化处理后得到的数值 (红色代表高含量, 绿色代表低含量)。其中, all_heatmap: 对代谢物和样品均进行聚类分析, 图中左侧的聚类线为代谢物聚类线, 图中上方的聚类线为样品聚类线。all_heatmap_class: 按物质分类热图; all_heatmap_no_cluster: 热图。

原始文件路径:

代谢物含量热图 (代谢物及样本均聚类) 见: 结题报告/1.Data_assess/heatmap/all_heatmap.*;

代谢物含量热图 (按物质分类) 见: 结题报告/1.Data_assess/heatmap/all_heatmap_class.*;

代谢物含量热图见: 结题报告/1.Data_assess/heatmap/all_heatmap_no_cluster.*。

2.4 重复相关性评估

通过样品之间的相关性分析可以观察组内样品之间的生物学重复。同时组内样品相对组间样品的相关系数越高, 获得的差异代谢物越可靠。将皮尔逊相关系数 r (Pearson's Correlation Coefficient) 作为生物学重复相关性的评估指标。皮尔逊相关系数利用 R 软件的内置 cor 函数计算, $|r|$ 越接近 1, 说明两个重复样品相关性越强。见下图:

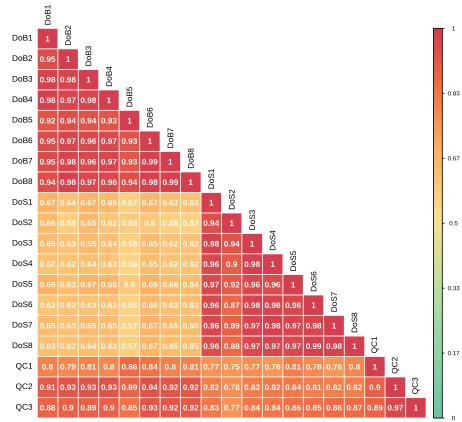


图 3: 样品间相关性图

注: 纵向和对角线上分别代表不同样品的样品名称, 不同的颜色代表不同的皮尔逊相关系数大小, 颜色越红代表正相关性越强, 颜色越绿代表相关性越差, 颜色越蓝代表负相关性越强, 同时两个样品之间的相关性系数大小标注在方格内。

原始文件路径:

所有样品间相关性图见: 结题报告/1.Data_assess/correlation_analysis/*_correlation.*;

所有样品皮尔逊相关系数表:结题报告/1.Data_assess/correlation_analysis/*_correlation_samples_pearson.xlsx;

所有代谢物相关系数表:结题报告/1.Data_assess/correlation_analysis/*_correlation_metabolites*.xlsx。

3 数据结果分析

3.1 分组主成分分析

在做差异分析前, 首先对进行差异比较的分组样品进行主成分分析, 观察差异分组之间和组内样本之间的变异度大小。

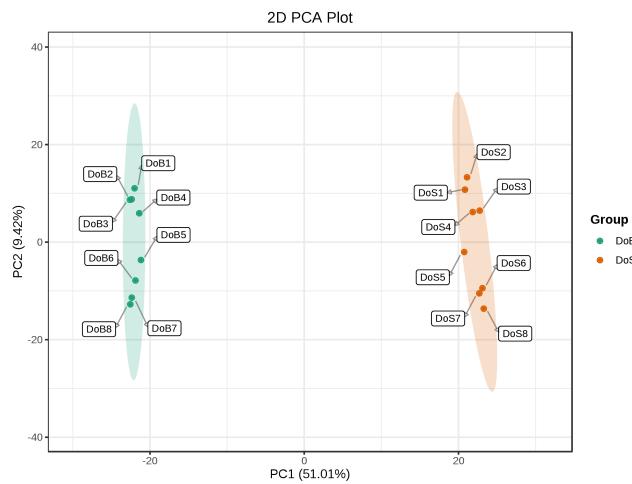


图 4: 分组主成分分析图

注: 每个分组一张 PCA 图, PC1 表示第一主成分, PC2 表示第二主成分, PC3 表示第三主成分, 百分比表示该主成分对数据集的解释率; 图中的每个点表示一个样品, 同一个组的样品使用同一种颜色表示, Group 为分组。

同时 PCA 的三维结果展示见下图:

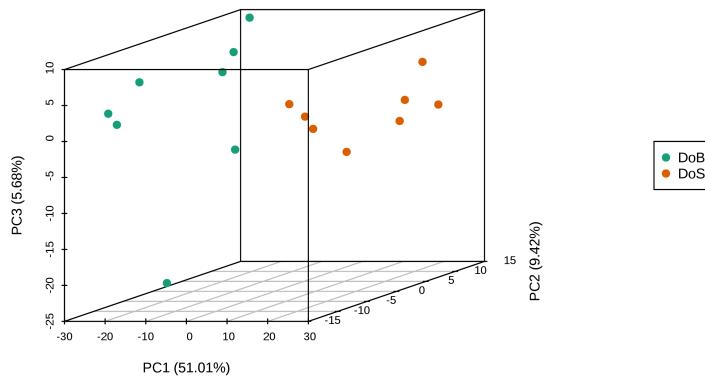


图 5: 分组主成分分析三维图

注: 其中 PC1 表示第一主成分, PC2 表示第二主成分, PC3 表示第三主成分

前 5 个主成分的可解释变异见下图:

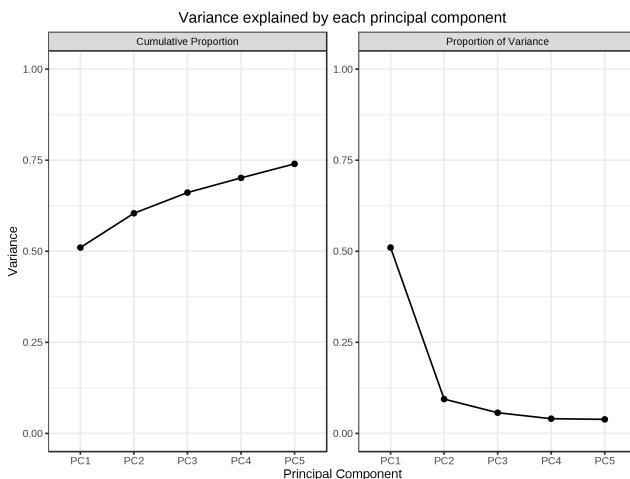


图 6: 分组主成分分析可解释变异图

注: 横坐标表示各个主成分, 纵坐标表示可解释变异, 左图为累计可解释变异, 右图为各个主成分的可解释变异

原始文件路径:

分组样品 PCA 图见: 结题报告/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/pca/group-ID*_vs_group-ID*_PCA.*;

分组样品 PCA 三维结果图见: 结题报告/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/pca/group-ID*_vs_group-ID*_PCA3D.*;

分组样品 PCA 前 5 个主成分的可解释变异: 结题报告/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/pca/group-ID*_vs_group-ID*_PCA_variance.*。

3.2 正交偏最小二乘法判别分析 (OPLS-DA)

上节所介绍的主成分分析法虽然能够有效地提取主要信息, 但是对于相关性较小的变量不敏感, 而偏最小二乘判别分析 (Partial Least Squares-Discriminant Analysis, PLS-DA) 可以解决此问题。PLS-DA 是一种有监督模式识别的多元统计分析方法, 具体做法是分别提取自变量 X 与因变量 Y 中的成分, 然后计算成分间的相关性。与 PCA 相比, PLS-DA 可以使组间区分最大化, 有利于寻找差异代谢物。正交偏最小二乘判别分析 (OPLS-DA) 结合了正交信号矫正 (OSC) 和 PLS-DA 方法, 能够将 X 矩阵信息分解成与 Y 相关和不相关的两类信息, 通过去除不相关的差异来筛选差异变量。OPLS-DA 在原始数据进行 \log_2 转换后, 再进行中心化处理 (Mean Centering), 公式如下:

$$x^* = x - \bar{x}$$



然后利用 R 软件中的 MetaboAnalystR 包 OPLSR.Anal 函数进行分析，下表为部分 OPLS-DA 模型计算结果：

表 3: 正交偏最小二乘法判别分析（OPLS-DA）部分计算结果

Index	Compounds	物质	VIP
M69T495	Propynoic acid	—	0.5601080
M73T100	Propionic acid	—	0.3601379
M85T70	Methyl acrylate	—	1.3494998
M93T265	Phenol	—	0.6578009
M101T64	Acrolein	—	1.3879273
M111T67	2-Furancarboxylic acid	—	0.8564483
M111T95	2-Furancarboxylic acid	—	1.3595391
M113T496	Trifluoroacetic acid	—	0.7851226
M113T51	Imidazole	—	0.5740628
M115T61	Maleic acid	—	1.3363757

原始文件路径：

分组比较 OPLS-DA 所有代谢物计算结果见：结题报告/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/group-ID*_vs_group-ID*_VIP.xlsx。

根据 OPLS-DA 模型分析代谢组数据，绘制各分组的得分图，进一步展示各个分组之间的差异 (Thévenot et al., 2015)。评价模型的预测参数有 R^2X , R^2Y 和 Q^2 ，其中 R^2X 和 R^2Y 分别表示所建模型对 X 和 Y 矩阵的解释率， Q^2 表示模型的预测能力，这三个指标越接近于 1 时表示模型越稳定可靠， $Q^2 > 0.5$ 时可认为是有效的模型， $Q^2 > 0.9$ 时为出色的模型。

3.2.1 OPLS-DA 模型概要

OPLS-DA 建模时，将 X 矩阵信息分解成与 Y 相关和不相关的两类信息，其中与 Y 相关的变量信息为预测主成分，与 Y 不相关的变量信息为正交主成分。

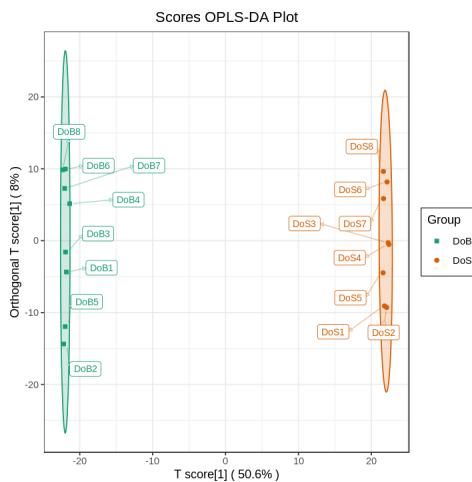


图 7: OPLS-DA 得分图

注: 横坐标表示预测主成分, 横坐标方向可以看出组间的差距; 纵坐标表示正交主成分, 纵坐标方向可以看出组内的差距; 百分比表示该成分对数据集的解释度。图中的每个点表示一个样品, 同一个组的样品使用同一种颜色表示, Group 为分组。

原始文件路径:

OPLS-DA 得分图见: 结题报告/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/opls/group-ID*_vs_group-ID*_opls_score.*。

3.2.2 OPLS-DA 模型验证

横坐标表示模型准确率, 纵坐标是模型分类效果出现的频数, 即本模型对数据进行 200 次随机排列组合实验, 若 Q^2 的 $p = 0.02$, 说明在此次 Permutation 检测中共有 4 个随机分组模型的预测能力优于本 OPLS-DA 模型, 若 R^2Y 的 $p = 0.545$, 说明在此次 Permutation 检测中共有 109 个随机分组模型其对 Y 矩阵的解释率优于本 OPLS-DA 模型。一般情况下, $p < 0.05$ 时模型最佳。

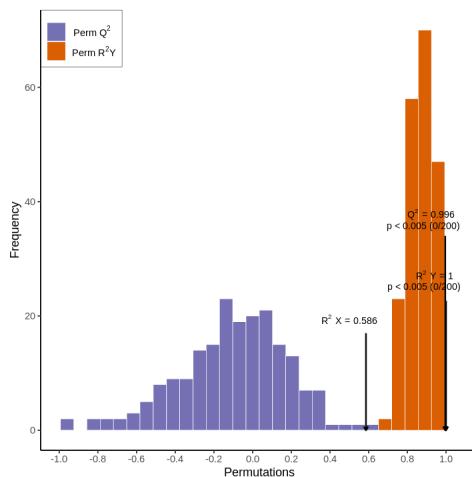


图 8: OPLS-DA 验证图

原始文件路径:

OPLS-DA 模型验证图见: 结题报告/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/opls/group-ID*_vs_group-ID*_opls_permutation.*。

3.2.3 OPLS-DA S-plot

下图为 OPLS-DA 的 S-plot 图, 横坐标表示主成分与代谢物的协方差, 纵坐标表示主成分与代谢物的相关系数, 越靠近右上角和左下角的代谢物表示其差异越显著, 红色的点表明这些代谢物的 VIP 值大于等于 1, 绿色的点表示这些代谢物的 VIP 值小于 1。

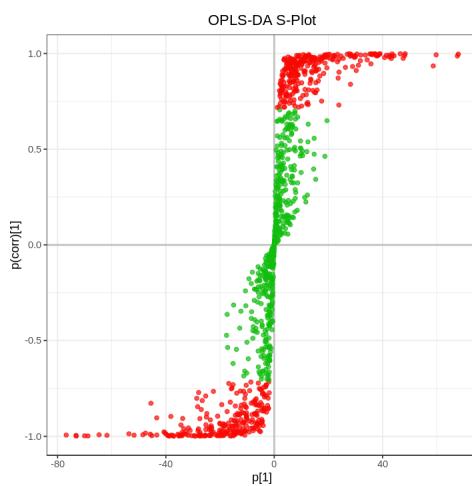


图 9: OPLS-DA S-plot

原始文件路径:



OPLS-DA S-plot 图见: 结题报告/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/opls/group-ID*_vs_group-ID*_opls_splot.*。

3.3 代谢物含量差异动态分布

为了更清楚、直观的展示总体代谢差异情况, 对比较组中代谢物进行差异倍数 (Fold Change, FC) 值计算, 计算之后根据 FC 值大小进行从小到大的排列, 绘制代谢物含量差异动态分布图, 对上调和下调前 10 个代谢物进行标注。

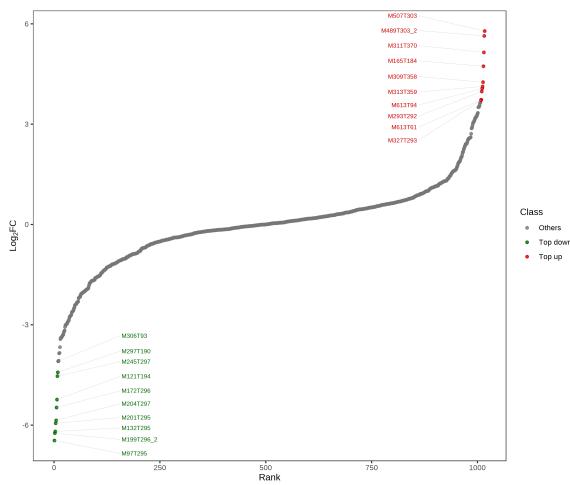


图 10: 代谢物含量差异动态分布图

注: 图中横坐标代表按差异倍数从小到大排列的累计物质数目, 纵坐标代表差异倍数以 2 为底的对数值, 每一个点代表一个物质, 绿色的点代表下调排名前 10 的物质, 红色的点代表上调排名前 10 的物质。

原始文件路径:

结题报告/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/group-ID*_vs_group-ID*_TopFcDistribution_*.*

3.4 差异代谢物筛选

代谢组学数据具有“高维、海量”的特点, 因此需要结合单变量统计分析和多元统计分析的方法, 并根据数据特性从多角度分析, 最终准确地挖掘差异代谢物。单变量统计分析方法包括参数检验和非参数检验。多元统计分析方法包括主成分分析、偏最小二乘法判别分析等。基于 OPLS-DA 结果, 从获得的多变量分析 OPLS-DA 模型的变量重要性投影 (Variable Importance in Projection, VIP), 可以初步筛选出不同品种或组织间差异的代谢物。同时可以结合单变量分析的 p-value 或者差异倍数值 (Fold Change) 来进一步筛选出差异代谢物。若为无生物学重复样本比较, 根据 Fold Change 值进行差异筛



选。若有生物学重复，则采取将 Fold Change、OPLS-DA 模型的 VIP 值相结合的方法来筛选差异代谢物。筛选标准：

1. 选取 Fold Change ≥ 2 和 Fold Change ≤ 0.5 的代谢物。代谢物在对照组和实验组中差异为 2 倍以上或 0.5 以下，则认为差异显著。
2. 选取 VIP ≥ 1 的代谢物。VIP 值表示对应代谢物的组间差异在模型中各组样本分类判别中的影响强度，一般认为 VIP ≥ 1 的代谢物则为差异显著。

部分差异代谢物筛选结果如下：

表 4: 差异代谢物筛选结果

Index	Compounds	物质	Type
M101T64	Acrolein	-	down
M115T99	Maleic acid	-	up
M119T185	4-Methylbenzaldehyde	-	up
M121T219_2	3-Hydroxybenzaldehyde	-	down
M121T194	3-Hydroxybenzaldehyde	-	down
M121T307	3-Hydroxybenzaldehyde	-	down
M121T464	3-Hydroxybenzaldehyde	-	down
M121T438	3-Hydroxybenzaldehyde	-	down
M121T280	2-Hydroxybenzaldehyde	-	down
M145T64	2-Oxopentanedioic acid	-	down

原始文件路径：

差异代谢物筛选结果表见：结题报告/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/group-ID*_vs_group-ID*_filter.xlsx。

3.4.1 差异代谢物条形图

在对所检测到的代谢物进行定性和定量分析后，结合具体样品的分组情况，比较在各分组中代谢物定量信息发生的差异倍数变化。下图为各分组比较中差异倍数 \log_2 处理后，将变化排在前面的差异表达代谢物结果展示：

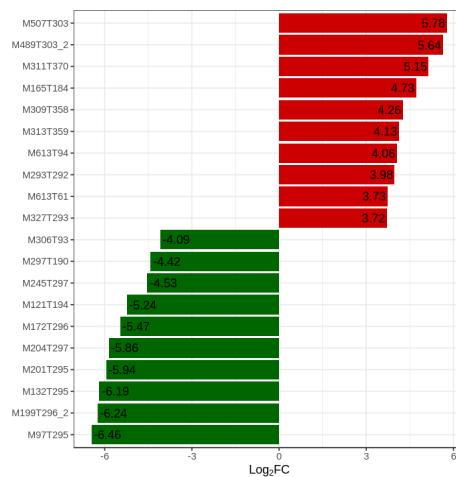


图 11: 差异倍数柱状图

注: 横坐标为差异代谢物的 $\log_2\text{FC}$, 即差异代谢物的差异倍数以 2 为底取对数的值, 纵坐标为差异代谢物。红色代表上调差异代谢物, 绿色代表下调差异代谢物。

原始文件路径:

结题报告/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/TopFcMetabolites/group-ID*_vs_group-ID*_TopFcBarChart_*.*。

3.4.2 差异代谢物雷达图

对不同分组代谢物定量结果计算差异, 基于筛选标准鉴定得到的差异代谢物中, 挑选差异变化最大的前 10 个代谢物进行雷达图的绘制, 雷达图即差异代谢物条形图的变形图。

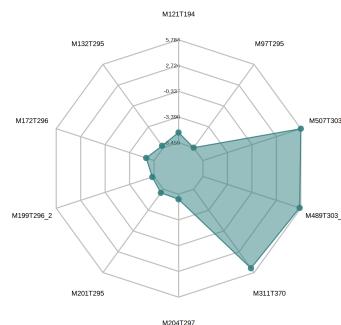


图 12: 差异代谢物雷达图

注: 网格线对应 $\log_2 FC$, 即差异代谢物的差异倍数以 2 为底取对数的值, 绿色阴影由每个物质的 $\log_2 FC$ 连线组成。

原始文件路径:

结题报告/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/TopFcMetabolites/group-ID*_vs_group-ID*_TopFcRadarChart_*.*。

3.4.3 差异代谢物 VIP 值图

对各分组比较中基于筛选标准鉴定得到的差异代谢物, 选择在 OPLS-DA 模型中 VIP 值最大的前 20 个代谢物进行展示:

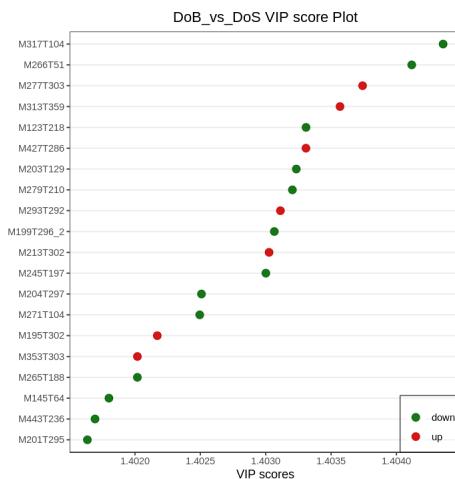


图 13: 差异代谢物 VIP 值图

注: 横坐标表示 VIP 值, 纵坐标表示差异代谢物, 红色代表上调差异代谢物, 绿色代表下调差异代谢物

原始文件路径:

结题报告/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/vipscore/group-ID*_vs_group-ID*_vipScore.*。

3.4.4 差异代谢物火山图

火山图 (Volcano Plot) 主要用于展示代谢物在两个 (组) 样品中的相对含量差异以及在统计学上差异的显著性。对每一个比较组中的代谢物及差异代谢物展示如下图: 对于该图, 我们同时提供不同于指定筛选标准的差异代谢物火山图供参考, 具体筛选条件见附件火山图目录下的 readme 文档; 此外, 附件还提供交互式网页版火山图, 可查阅代谢物的具体信息。

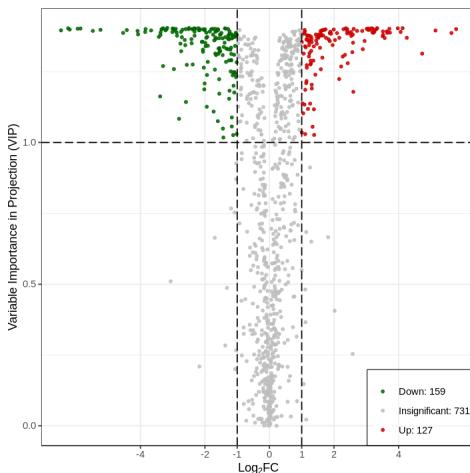


图 14: 差异代谢物火山图

注: 火山图中的每一个点表示一种代谢物, 其中绿色的点代表下调差异代谢物, 红色的点代表上调差异代谢物, 灰色代表检测到但差异不显著的代谢物; 横坐标表示某代谢物在两组样品中相对含量差异倍数的对数值 ($\log_2 FC$), 横坐标绝对值越大, 说明该物质在两组样品间的相对含量差异越大。VIP + FC 双重筛选条件下: 纵坐标表示 VIP 值, 纵坐标值越大, 表明差异越显著, 筛选得到的差异表达代谢物越可靠。VIP + FC + p-value 三重筛选条件下: 纵坐标表示差异显著性水平 ($-\log_{10} p\text{-value}$), 圆点的大小代表 VIP 值。

原始文件路径:

指定筛选条件下的差异代谢物火山图见: 结题报告/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/vol/group-ID*_vs_group-ID*_vol.*;

参考筛选条件下的差异代谢物火山图见: 结题报告/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/vol/group-ID*_vs_group-ID_*Pvalue_vol.*。

3.4.5 差异代谢物聚类热图

为了方便观察代谢物相对含量的变化规律, 我们对应用筛选标准鉴定得到的差异代谢物的原始相对含量按行采用归一化处理 (Unit Variance Scaling, UV Scaling), 通过 R 软件 ComplexHeatmap 包绘制热图, 结果如下图:

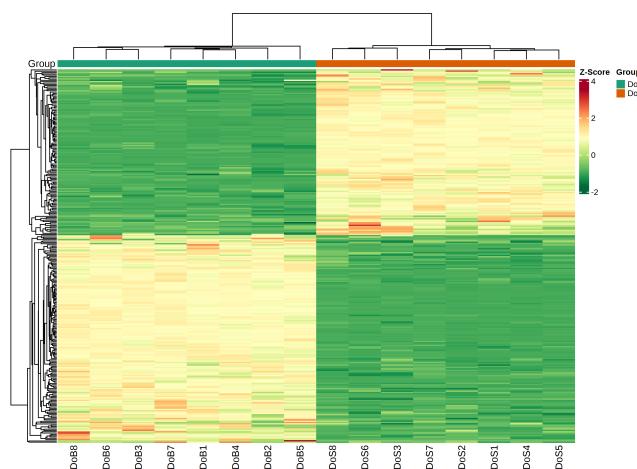


图 15: 差异代谢物聚类热图

注：横坐标为样品名称，纵坐标为差异代谢物，热图中不同颜色代表差异代谢物相对含量归一化处理后得到的数值，反映其相对含量的高低（红色代表高含量，绿色代表低含量），热图上方的注释条对应样品分组（Group）；对差异代谢物进行层次聚类（Hierarchical Clustering），则热图左侧的树状图代表差异代谢物聚类结果；或对差异代谢物进行分类，则热图左侧的注释条对应物质一级分类（Class），不同颜色代表不同的物质类别。

原始文件路径：

代谢物含量热图（代谢物聚类）见：结题报告/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/heatmap/group-ID*_vs_group-ID*_heatmap*.*;

代谢物含量热图（按物质分类）见：结题报告/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/heatmap/group-ID*_vs_group-ID*_heatmap_class_*.*。

3.4.6 差异代谢物 Z 值图

Z-score 图是在通过计算 Z 值来对不同样本中的差异代谢物做归一化处理，横坐标表示 Z 值，纵坐标表示差异代谢物，不同颜色的点表示不同组别的样本，可以非常直观的看到每个差异代谢物在不同组间的分布情况。具体公式为： $z = (x - \mu) / \sigma$ ；其中 x 为某一具体分数， μ 为平均数， σ 为标准差。

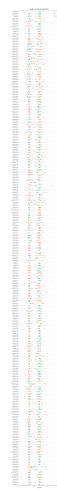


图 16: 差异代谢物 Z 值图

注: 横坐标为物质相对含量归一化处理后的数值, 纵坐标是代谢物编号, 不同颜色的点代表不同组的样品。

原始文件路径:

结题报告/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/zScore/group-ID*_vs_group-ID*_zScore*.*。

3.4.7 差异代谢物相关性分析

不同代谢物之间具有协同或互斥关系, 相关性分析可以帮助衡量显著性差异代谢物之间的代谢密切程度 (Metabolic Proximities), 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。通过皮尔逊相关分析方法对按照筛选标准鉴定得到的差异代谢物进行相关性分析, 结果如下图:

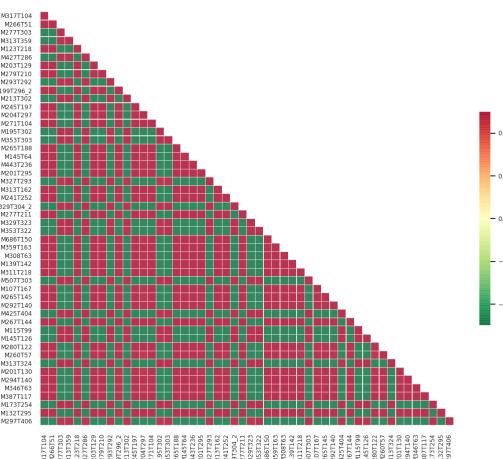


图 17: 差异代谢物相关性热图

注: 横向为差异代谢物名称, 纵向为差异代谢物名称, 不同颜色代表皮尔逊相关系数 r 的高低, 相关系数与颜色间的关系见右侧图例说明, 红色表示正相关性较强, 绿色表示负相关性较强, 颜色越深代表样品间相关系数的绝对值越大。默认对所有差异代谢物作图, 当差异代谢物数目超过 50 个时, 只展示 VIP 值最大的前 50 个差异代谢物。

原始文件路径:

差异代谢物相关性热图见: 结题报告/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/cpdCorr/group-ID*_vs_group-ID*_raw_cpdCorr_*.*;

VIP_top_* 差异代谢物相关性热图见: 结题报告/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/cpdCorr/group-ID*_vs_group-ID*_top*_VIP_cpdCorr_*.*;

差异代谢物相关系数表见: 结题报告/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/cpdCorr/group-ID*_vs_group-ID*_raw_cpdCorr_*.xlsx;

VIP_top_* 差异代谢物相关系数表见: 结题报告/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/cpdCorr/group-ID*_vs_group-ID*_top*_VIP_cpdCorr_*.xlsx。

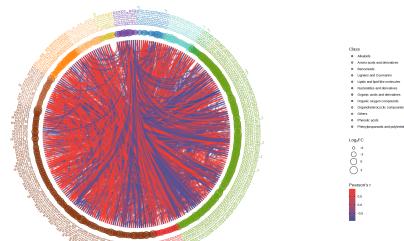


图 18: 差异代谢物和弦图

注: 图中最外层为差异代谢物名称, 中层点的大小代表 $\log_2\text{FC}$ 值大小, 点越大, 其对应的 $\log_2\text{FC}$ 值也就越大; 文字和点的颜色反映物质的一级分类, 不同的颜色代表不同代谢物来源分类 (Class); 内层连线反映对应位置代谢物之间的皮尔逊相关系数 r 大小, 红色线条代表正相关, 蓝色线条代表负相关。默认对 $|r| \geq 0.8$ 且 $p < 0.05$ 的差异代谢物对作图。

原始文件路径:

差异代谢物和弦图见: 结题报告/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/cpdCorr/group-ID*_vs_group-ID*_cpdCorrCir_*.*

差异代谢物皮尔逊相关系数和 P 值表: 结题报告/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/cpdCorr/group-ID*_vs_group-ID*_cpdCorr_Pvalue_*.xlsx。

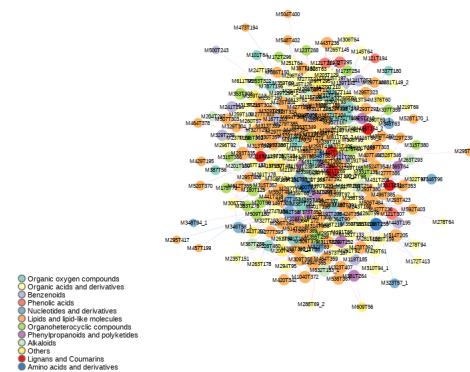


图 19: 差异代谢物相关性网络图

注: 图中点代表不同的差异代谢物, 点的大小与连接度 (Degree) 相关, 点越大连接度越大, 即与它连接的点 (邻居) 个数越多。红色线条代表正相关, 蓝色线条代表负相关。线条的粗细代表皮尔逊相关系数 r 的绝对值的大小, 线条越粗, $|r|$ 越大。默认对 $|r| \geq 0.8$ 且 $p < 0.05$ 的差异代谢物对作图。

原始文件路径:

相关性网络图见:结题报告/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/cpdCorr/group-ID*_vs_group-ID*_cpdCorrNet_*.*

3.4.8 K-Means 分析

为了研究代谢物在不同分组中的相对含量变化趋势, 将所有分组比较中按照筛选标准鉴定得到全部差异代谢物的相对含量进行 z-score 标准化, 随后进行 K 均值 (K-Means) 聚类分析。

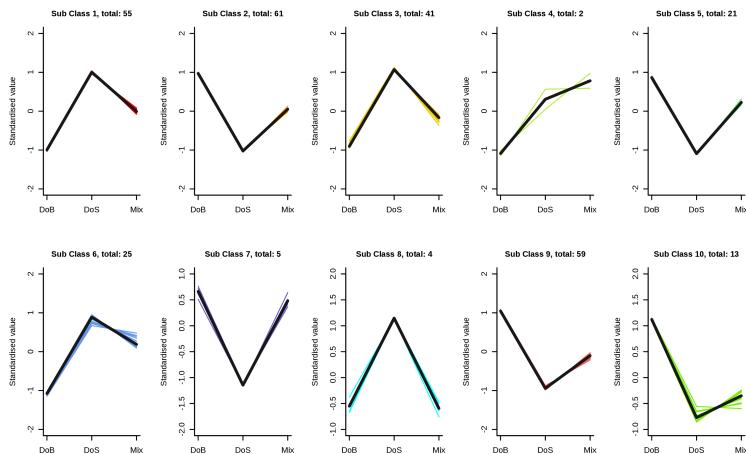


图 20: 差异代谢物 K-Means 图

注: 横坐标表示样品名称, 纵坐标表示标准化的代谢物相对含量, Sub Class 代表相同变化趋势的代谢物类别编号, total: * 代表该类别的代谢物的数目为 *。

原始文件路径:

K-means 聚类图见: 结题报告/2.Basic_analysis/kmeans/kmeans_cluster.*;

K-means 聚类具体代谢物信息见: 结题报告/2.Basic_analysis/kmeans/kmeans_group.*。

3.4.9 差异代谢物韦恩图

通过韦恩图的形式, 展示各组差异代谢物之间的关系。5 组以上展示花瓣图。结果见下图:

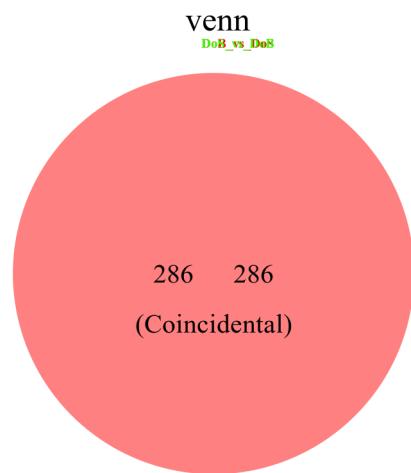


图 21: 各组差异韦恩图

注: 图中每个圈代表一个比较组, 圈和圈重叠部分的数字代表比较组之间共有的差异代谢物个数, 没有重叠部分的数字代表比较组特有差异代谢物个数。



原始文件路径:

结题报告/2.Basic_analysis/Venn。

3.5 差异代谢物 KEGG 功能注释及富集分析

KEGG (Kyoto Encyclopedia of Genes and Genomes) 数据库有助于研究者把基因、表达信息以及代谢物含量作为一个整体网络进行研究。作为有关 Pathway 的主要公共数据库，KEGG 提供的整合代谢途径 (Pathway) 查询，包括碳水化合物、核苷、氨基酸等的代谢及有机物的生物降解，不仅提供了所有可能的代谢途径，而且对催化各步反应的酶进行了全面的注解，包含有氨基酸序列、PDB 库的链接等等，是进行生物体内代谢分析、代谢网络研究的强有力工具。

3.5.1 差异代谢物功能注释

差异代谢在生物体内相互作用，形成不同的通路。利用 KEGG 数据库 (Kanehisa et al., 2000) 对差异代谢物进行注释并展示。各组详细信息在结果附件中可查询。部分结果如下：

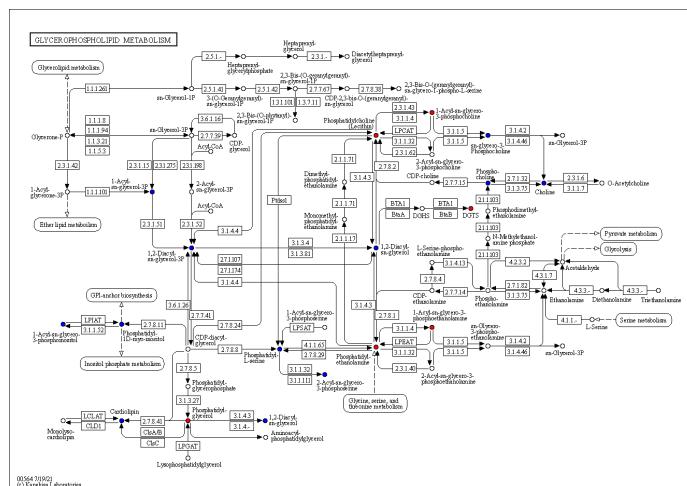


图 22: 差异代谢物 KEGG 通路图

注：红色表示代谢物含量在实验组中显著上调，蓝色代表该代谢物被检测到但未发生显著变化，绿色表示代谢物含量在实验组中显著下调。通过代谢通路寻找研究对象中表型差异的原因。

原始文件路径:

结题报告/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/enrichment/Graph/ko*****。

统计筛选出的差异显著的代谢物中 KEGG 数据库的注释情况。部分结果如下：



表 5: 差异代谢物 KEGG 注释表格

Index	Compounds	物质	Type	cpd_ID
M101T64	Acrolein	—	down	C01471
M115T99	Maleic acid	—	up	C01384
M119T185	4-Methylbenzaldehyde	—	up	C06758
M121T219_2	3-Hydroxybenzaldehyde	—	down	C03067
M121T194	3-Hydroxybenzaldehyde	—	down	C03067
M121T307	3-Hydroxybenzaldehyde	—	down	C03067
M121T464	3-Hydroxybenzaldehyde	—	down	C03067
M121T438	3-Hydroxybenzaldehyde	—	down	C03067
M121T280	2-Hydroxybenzaldehyde	—	down	C06202
M145T64	2-Oxopentanedioic acid	—	down	C00026

原始文件路径:

结题报告/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/enrichment/group-ID*_vs_group-ID*_filter_anno.xlsx。

3.5.2 差异代谢物 KEGG 分类

对差异显著代谢物 KEGG 的注释结果按照 KEGG 中通路类型进行分类，分类图如下图所示：

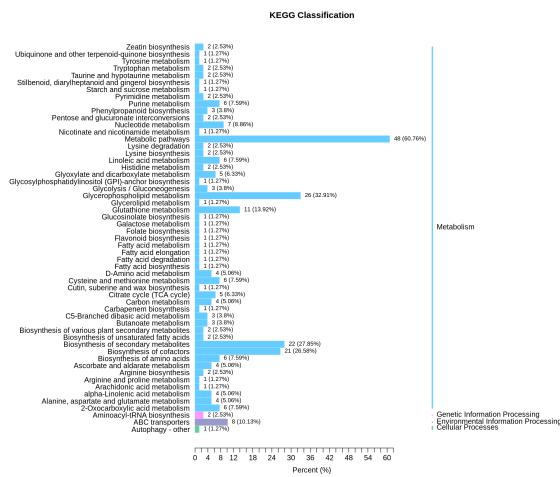


图 23: 差异代谢物 KEGG 分类图

注：纵坐标为 KEGG 代谢通路的名称，横坐标为注释到该通路下的代谢物个数及其个数占被注释上的代谢物总数的比例。

原始文件路径:



结题报告/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/enrichment/group-ID*_vs_group-ID*.KEGG.barplot.*。

3.5.3 KEGG 信号通路差异代谢物聚类分析

利用按照筛选标准鉴定得到的差异代谢物的 KEGG 注释信息，选择至少含有 5 个差异代谢物的 KEGG 代谢通路，对这些通路中的所有差异代谢物的相对含量进行聚类分析，以便更好地研究潜在重要代谢通路中的物质含量在不同分组中的变化规律。

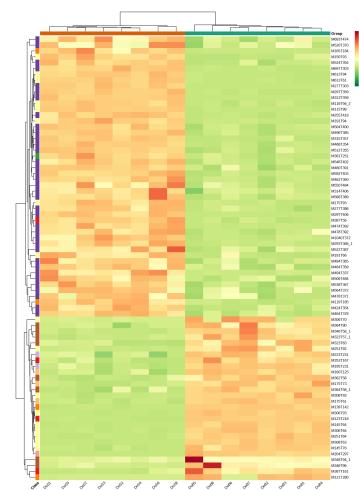


图 24: KEGG 通路的差异代谢物聚类热图

注：横坐标为样品名称，纵坐标为差异代谢物，热图中不同颜色代表差异代谢物相对含量归一化处理后得到的数值，反映其相对含量的高低（红色代表高含量，绿色代表低含量），热图上方的注释条对应样品分组（Group），热图左侧的树状图代表差异代谢物层次聚类结果，聚类树右侧的注释条对应物质一级分类（Class），不同颜色代表不同的物质类别。

原始文件路径：

结题报告/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/enrichment/group-ID*_vs_group-ID*.KEGG.heatmap*.*。

3.5.4 差异代谢物 KEGG 富集分析

根据差异代谢物结果，进行 KEGG 通路富集分析，其中 Rich Factor 为差异表达的代谢物中在对应通路中的个数与该通路检测注释到的代谢物总数的比值，该值越大表示富集程度越大。p-value 为超几何检验 p 值，超几何分布的计算公式如下所示：



$$P = 1 - \sum_{i=0}^{m-1} \frac{\binom{M}{i} \binom{N-M}{n-i}}{\binom{N}{n}}$$

其中，N 代表全部代谢物中具有 KEGG 注释的代谢物数量，n 代表 N 中差异代谢物的数量，M 代表 N 中某 KEGG 通路的代谢物数量，m 代表 M 中某 KEGG 通路的差异代谢物数量。p-value 越接近于 0，表示富集越显著。图中点的大小代表富集到相应通路上的差异显著代谢物个数。结果展示如下：

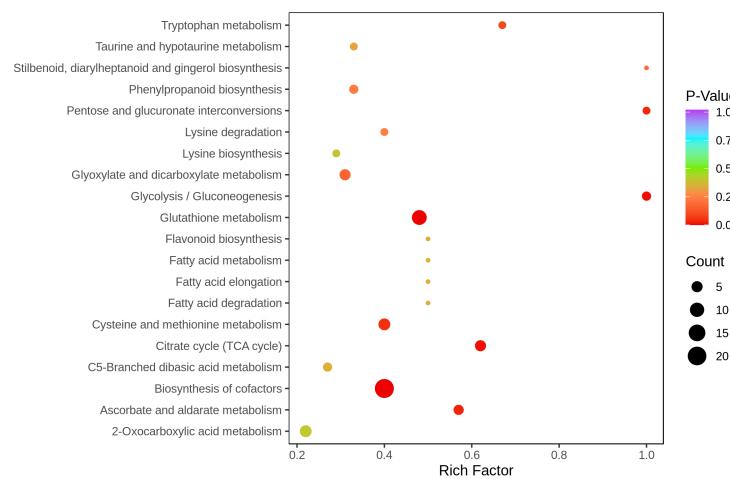


图 25: 差异代谢物 KEGG 富集图

注：横坐标表示每个通路对应的 Rich Factor，纵坐标为通路名称，点的颜色反映 p-value 大小，越红表示富集越显著。点的大小代表富集到的差异代谢物的个数多少。

原始文件路径：

结题报告/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/enrichment/group-ID*_vs_group-ID*.KEGG.Enrichment.*。

3.5.5 KEGG 代谢通路整体变化分析

差异丰度得分 (Differential Abundance Score, DA Score) 是一种基于通路的代谢变化分析方法，差异丰度得分可以捕捉到某一途径中所有差异代谢物的总体变化，其计算公式为：

$$\text{差异丰度得分} = \frac{\text{该通路上调差异代谢物个数} - \text{该通路下调差异代谢物个数}}{\text{注释到该通路的所有代谢物个数}}$$

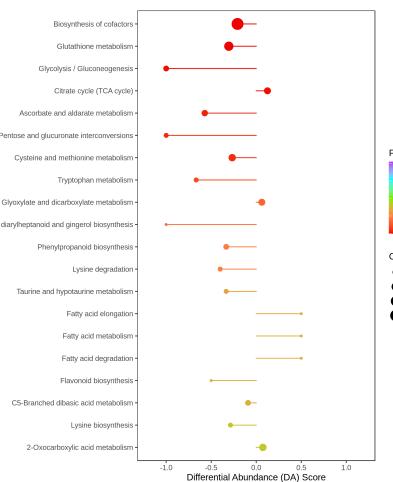


图 26: 差异丰度得分图

注: 纵坐标表示差异通路名称, 横坐标表示差异丰度得分 (DA Score)。DA Score 反映代谢途径所有代谢物的整体变化, 得分 1 表示该通路中所有鉴定到的代谢物表达趋势上调, -1 该通路中所有鉴定到的代谢物表达趋势下调。线段的长度表示 DA Score 的绝对值, 线段端点的圆点大小表示该通路中差异代谢物的个数, 圆点分布在中轴左侧且线段越长, 表示该通路整体表达情况越倾向于下调, 圆点分布在中轴右侧且线段越长, 表示该通路整体表达情况越倾向于上调, 圆点越大表示代谢物数目越多。线段和圆点颜色反映 p-value 大小, 越接近红色表示 p-value 越小, 越接近紫色表示 p-value 越大。

原始文件路径:

差异丰度得分图见:结题报告/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/enrichment/group-ID*_vs_group-ID*.KEGG.DA_score.*;

差异丰度统计表见:结题报告/2.Basic_analysis/Difference_analysis/group-ID*_vs_group-ID*/enrichment/group-ID*_vs_group-ID*_KEGG_DA_score.xlsx。

4 参考文献

- 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.
- Fraga, C.G., et al., Signature-discovery approach for sample matching of a nerve-agent precursor using liquid chromatography-mass spectrometry, XCMS, and chemometrics. Anal Chem, 2010. 82(10): p. 4165-73.



3. L. Eriksson, E.J., N. Kettaneh-Wold, J.Trygg, C. Wikström, and S. Wold, Multi- and Megavariate Data Analysis Part I Basic Principles and Applications, Second edition Umetrics Academy:Sweden, 2006.
4. Chen, Y., et al., RRLC-MS/MS-based metabonomics combined with in-depth analysis of metabolic correlation network: finding potential biomarkers for breast cancer. Analyst, 2009.134(10): p. 2003-11.
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. Kanehisa, M. and S. Goto, KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res, 2000. 28(1): p. 27-30.
7. Chong, J. and Xia, J., MetaboAnalystR: an R package for flexible and reproducible analysis of metabolomics data. Bioinformatics, bty528.

5 附录

5.1 分析方法英文版

5.1.1 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.

5.1.2 Hierarchical Cluster Analysis and Pearson Correlation Coefficients

The HCA (hierarchical cluster analysis) results of samples and metabolites were presented as heatmaps with dendograms, 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 Complex-Heatmap. For HCA, normalized signal intensities of metabolites (unit variance scaling) are visualized as a color spectrum.

5.1.3 Differential metabolites selected

Significantly regulated metabolites between groups were determined by $VIP \geq 1$ and absolute $\log_2 FC$ (fold change) ≥ 1 . 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.

5.1.4 KEGG annotation and enrichment analysis

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

5.2 分析软件列表以及版本

表 6: 软件列表信息

Analysis	Software	Version
PCA	R (base package)	3.5.1
皮尔逊相关系数	R (base package; Hmisc)	3.5.1; 4.4.0
样品间相关性图	R (corrplot)	0.84
热图	R (ComplexHeatmap)	2.8.0
OPLS-DA	R (MetaboAnalystR)	1.0.1
雷达图	R (fmsb)	0.7.0
和弦图	R (igraph; ggraph)	1.2.4.2; 2.0.2
相关性网络图	R (igraph)	1.2.4.2

5.3 常见问题

1.TIC 图和 MRM 图各有一张，分别用“N”和“P”表示，指的是什么意思？

答：“N”代表的是 negative-负离子模式；“P”代表 positive-正离子模式。

2. 对样品进行相关性分析后的皮尔逊相关性系数文件有吗？

答：原始文件路径：

结题报告/1.Data_assess/correlation_analysis/all_cor_metabolites_pearson*;

同时也有进行斯皮尔曼相关性分析的结果文件，原始文件路径：

结题报告/1.Data_assess/correlation_analysis/all_cor_metabolites_spearman*。

3.ALL_sample_data 表格中的数值怎么看呢，代谢物质的具体含量和单位是什么？

答：表中的数据使用科学计数法，如 1.22E+02 表示 1.22×10^2 ，即 122，如果您不习惯，可以更改数字格式来查看；这个数值是代谢物的相对含量，没有单位，是通过计算每个物质的特征离子在检测



器形成的峰面积，（虽然不能定量物质的绝对含量，但检测条件一致，可用于比较同一物质在不同样本中的差异）。

4. 广靶是怎么定性物质的，可否提供检测物质的打分值？

答：广靶定性有3个级别，Level：物质鉴定级别，1：样本物质二级质谱、RT与数据库物质匹配得分为0.7分以上；2：样本物质二级质谱、RT与数据库物质匹配得分为0.5-0.7分；3：样本物质Q1、Q3、RT、DP、CE与数据库物质核对一致。

在广泛靶向代谢组没有打分值的说法；非靶涉及到打分问题。可以从检测方法学上进行理解，广靶是基于MRM扫描，利用物质检测5个参数(DP、CE、RT、Q1、Q3)，检测不同样本中的物种相对含量，获得物质的定性定量数据。老师可以参考广靶方法学文献：Chen, W., Gong, L., Guo, Z., et al., A Novel Integrated Method for Large-Scale Detection, Identification, and Quantification of Widely Targeted Metabolites: Application in the Study of Rice Metabolomics. Molecular Plant, 2013, 6(6):1769-1780.

5. 质谱峰是如何校正的？

答：AB Sciex配备的调谐液，仪器自动根据调谐液中的标准离子校准质量轴。

6. 可以解释一下主成分分析的概念吗？在主成分中，组间变异度越大，组内变异度越小，是不是越好。那有没有个度呢，就是说在什么范围内比较好？

答：假设样品数为n，检测到的物质种类数为m。

那么原始数据代表了在一个m维度空间中分布着n个点。

PCA主成分分析就是首先计算时利用最小二乘法原理找到一条直线，所有样品到该直线的距离最小。该直线方向也就体现了样品间最大差异。

在此基础上，在该直线的垂直方向上可以找到其次最显著的直线，如此反复。

主成分图并非看主成分占的百分比，而是看组内3个生物学重复是否能够较好的聚集在一块，组间（不同分组之间）是否能很好的分开。这个主要是从整体上判断组内生物学重复性的好坏，和组间差异的大小。

7. 我重点关注样本中的A和B这两个代谢物的含量，我在计算样本间含量差异的时候能否将这两个代谢物的数据加和在一起进行比较呢？

答：不能。广靶是相对定量，首先提取方法没有对单个代谢物进行方法优化，不能保证将样本中特定代谢物完全提取。第二，由于代谢物本身的理化特性，它们的离子化效率也不尽相同，导致最终检测时信号强度也不一样。比如A和B两个物质，同样拿1nmol的量用LC-MS分析，它们的信号响应值差异可能会很大，这就是物质本身灵敏度不同造成的，物质的检测灵敏度跟自身的化学性质有关，化学性质对物质检测灵敏度的影响主要表现在离子化效率和质谱碎裂行为两方面。因此对于不同代谢物之间不能进行加和或者比较，只有不同样本间的同一代谢物之间可以进行比较。



8. 在进行分组差异代谢物筛选时, A_vs_B 中代谢物的上下调具体指的是在哪个组中含量高?

答: 在进行差异代谢物筛选时, 若老师给出的分组信息为 A_vs_B, 则表明以 A 为对照组, B 为实验组进行数据分析, 最终筛选的差异代谢物上调表示: 该代谢物在 A 中相对含量低, 在 B 中相对含量高。

9. 差异代谢物聚类热图可以标注上具体物质名称吗?

答: 可以的, 受图片大小和清晰度限制, 报告中不展示物质名称, 请查看:

结题报告/Basic_analysis/Difference_analysis/A_vs_B/A_vs_B_heatmap_*_Compounds.*。

10. 怎么找到 venn 图里的共有差异代谢物, 表格在哪里看? 如何看?

答: 可以打开文件夹 Basic_analysis/Venn/CF_vs_MT_QG_vs_WT_venn_result, 打开 excel 表格后, 可以看到有两列 TRUE/FALSE, 2 个 TRUE 的是共有的 (并且对代谢物上调和下调进行标注, up 或者 down), 一个 TRUE 一个 FALSE 代表一个分组有该代谢物, 一个分组没有该代谢物。

11. KEGG 通路中被注释到的差异代谢物个数之和, 与 KEGG 被注释到的差异代谢物个数不一致?

答: 被 KEGG COMPOUND 注释 (编号为 C******) 的代谢物并不一定有 KEGG PATHWAY 注释 (ko*****), 所以会有不一致。