



及因生物
— TGENE —

风味组学结题报告



及因（上海）生物科技有限公司

Tgene Biotech (Shanghai) Co.,Ltd.

目录

1. 项目概况.....	1
1.1 项目信息.....	1
1.2 项目流程.....	2
1.3 产品介绍.....	3
2. 数据分析.....	4
2.1 样本详细分组信息.....	4
2.2 数据检查.....	5
2.2.1 总离子流色谱图.....	5
2.3 数据注释.....	6
2.3.1 鉴定物质数目统计.....	6
2.3.2 鉴定物质韦恩图.....	7
2.3.3 风味物质种类分析.....	8
2.3.4 挥发物阈值分析.....	9
2.3.5 感官风味特征分析.....	10
2.4 多元统计分析.....	11
2.4.1 主成分分析.....	12
2.4.2 偏最小二乘判别分析 (PLS-DA)	13
2.4.3 正交-偏最小二乘判别分析 (OPLS-DA)	14
2.5 差异分析.....	15
2.5.1 差异物质筛选.....	15
2.5.2 韦恩图.....	17
2.5.3 热图.....	18
2.5.4 火山图.....	19
2.5.5 感官风味特征分析.....	20
3. 实验.....	21
3.1 实验方法.....	21
4. 数据预处理.....	22
5. 参考文献.....	23
6. 附录.....	24
6.1 生物信息学常用软件及数据库.....	24
6.2 结果文件及路径说明.....	25

1. 项目概况

1.1 项目信息

合同编号：

项目客户：

样本数量：

启动日期：

报告完成日期：

技术支持负责人：

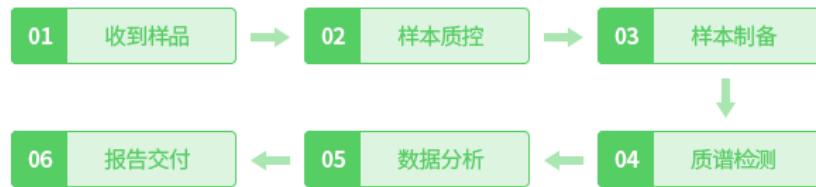
技术支持联系方式：

服务邮箱：

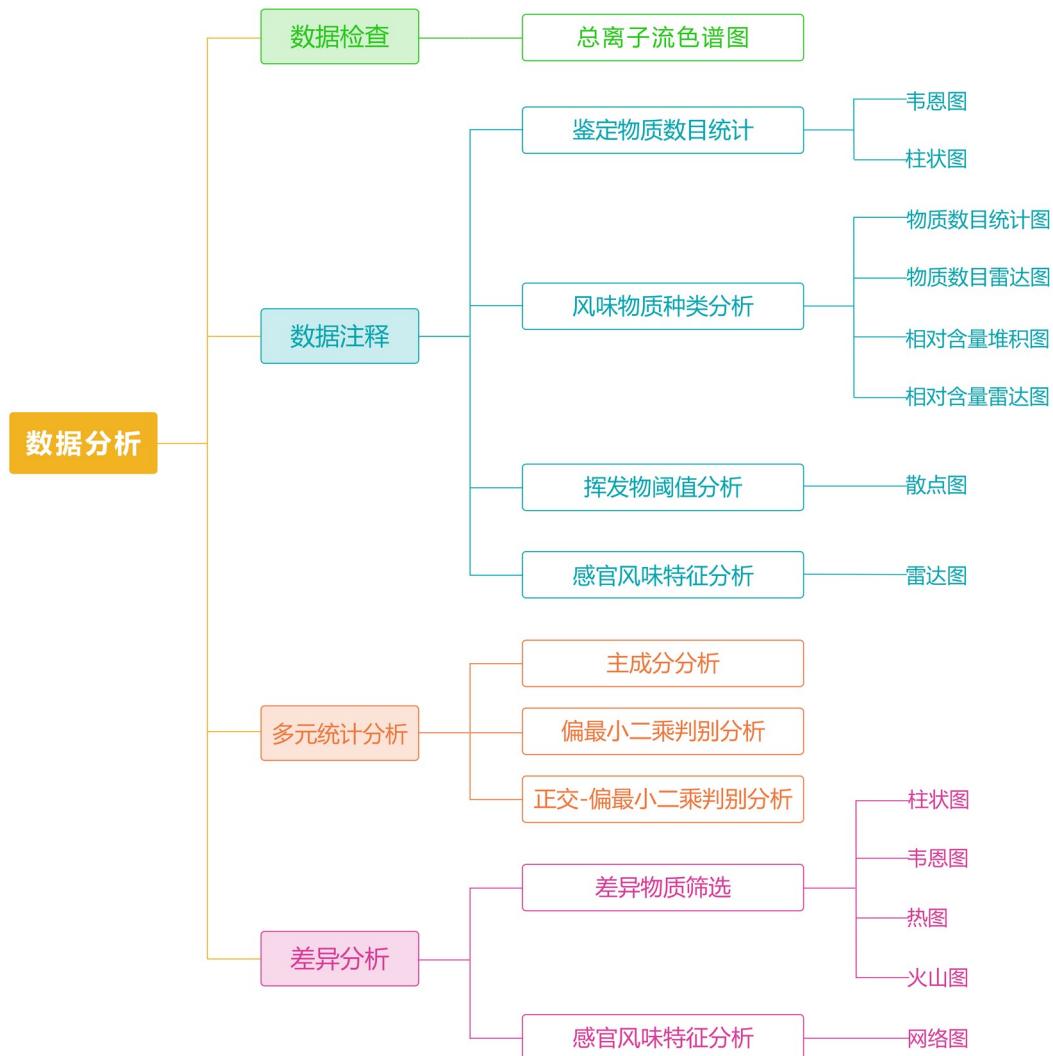
1.2 项目流程

及因生物自客户处收到样本以后，首先对样本进行质检，经确认符合送样要求后，进行质谱检测实验和生物信息学分析，最后生成最终的结题报告。

项目流程如下：



数据分析流程如下：



1.3 产品介绍

风味组学通常指挥发性代谢物检测，在常压下，沸点在 50-260°C 的各种有机化合物。及因生物风味代谢组基于 GC×GC TOF MS 全二维高端色谱质谱平台。具有高通量、高精度、高灵敏度及重现性的特点，且具有可参考的标准谱图数据库，易于定性；定性物质数量是常规 GC-MS 鉴定物质数量的几倍到几十倍。可检测样品中大多数有机分子。同时高分辨率、高灵敏度的全二维气相色谱技术及高采集速率（500 张全谱图/秒）飞行时间质谱技术，为复杂样品及未知物样品分析提供完美解决方案。全自动定性能力、定量能力超越同类飞行质谱。集成了气相和高采集速率的飞行时间质谱的所有优点。

基于定性定量结果，除提供差异风味物质筛选、韦恩图、火山图、热图等分析外。及因生物还提供感官风味特征分析，深度挖掘风味物质与风味特征的关联。

2. 数据分析

2.1 样本详细分组信息

研究对象：【植物组织】

样本数量：【6】例

样本详细信息如下：

表 2.1 样本详细信息

Raw name	Samples name	Group name
NP2405016015	A1	A
NP2405016015_001	A2	A
NP2405016018	B1	B
NP2405016018_001	B2	B
NP2405016021	C1	C

说明：报告只展示了部分结果，详细结果可见 data/sampleinfo.xlsx

(1) Raw name：及因生物提取编号

(2) Samples name：样本名称

(3) Group name：组别名称

2.2 数据检查

2.2.1 总离子流色谱图

经色谱分离流出的组分不断进入质谱，质谱连续扫描进行数据采集。每一次扫描得到一张质谱图，将每一张质谱图中所有离子强度相加，得到一个总的离子流强度。然后以离子强度为纵坐标、时间为横坐标，绘制的图即为总离子流色谱图（TIC）。相比一维气相色谱，全二维气相色谱的分辨力大大提升，峰容量从一两百飞跃到上万，灵敏度也显著提高，实现了高通量的广谱分析。通过 ChromaTOF 软件处理全二维气相数据，可展示多个维度的总离子流色谱情况。

本次分析中典型样本的基峰色谱图如下所示：

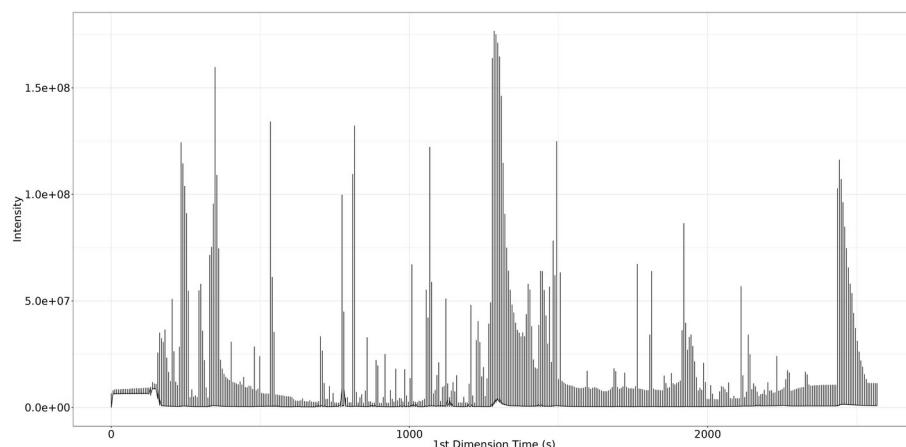


图 2.1 一维总离子流色谱图

注：横坐标为保留时间(s)，纵坐标为离子强度。

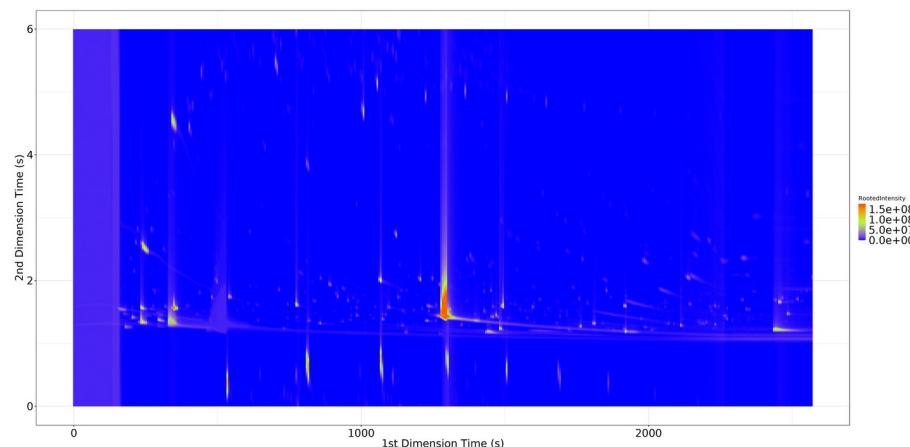


图 2.2 二维总离子流色谱图

注：横坐标为一维保留时间（s），纵坐标为二维保留时间（s）。颜色代表检测物质离子峰，颜色越红，表明响应强度越高。通过二维保留时间可查看有无共馏物。

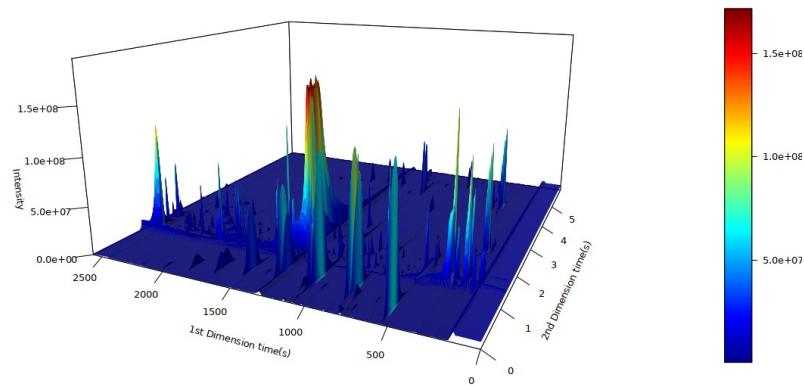


图 2.3 三维总离子流色谱图

注：横坐标为一维保留时间（s），纵坐标为二维保留时间（s）。颜色及峰高可知离子响应强度。颜色越红，表明响应强度越高。通过二维保留时间可查看有无共馏物。

2.3 数据注释

采用 NIST2020 数据库，使用 Chroma TOF 搜库软件对下机原始数据进行风味物质注释。

2.3.1 鉴定物质数目统计

通过对检测物质进行注释、除杂后，得到风味物质鉴定信息，采用表和柱形图来展示样本分组条件中鉴定到的总物质数目。具体如下。

表 2.2 物质鉴定信息

组别	鉴定数目
A	1053
B	1213
C	1238

注：具体鉴定信息可查看文件：metabolome.xlsx，归一化处理后数据可查看文件：metabolome_normalization.xlsx。

其中 OdorCharacter 和 flavor_profile 为物质属性相关描述；Retention Index 是实际保留指数；Lib_RI 是数据库的参考保留指数；RI_sd 是实际保留指数和参考保留指数差值；Identify_Level 是根据 RI_sd 值划分的物质鉴定等级，Level1 是指实际保留指数和参考保留指数差值<20，Level2 是指没有参考 Lib_RI 的物质。

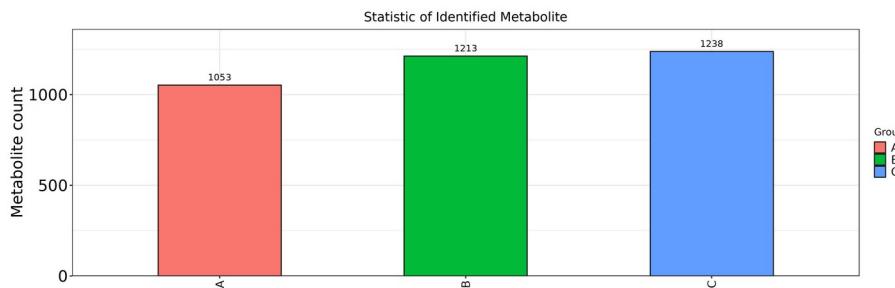


图 2.4 物质鉴定数目

注：横坐标表示样本分组条件，纵坐标表示鉴定到的风味物质数目。

2.3.2 鉴定物质韦恩图

对于多组比较分析获取到的风味物质列表，利用韦恩图（超过 6 组采用 UpSetPlot 图）展示风味物质在不同组间的鉴定情况。该图能够展示一个组中特异性鉴定到的风味物质数以及在多个组中共同鉴定到得风味物质数目。

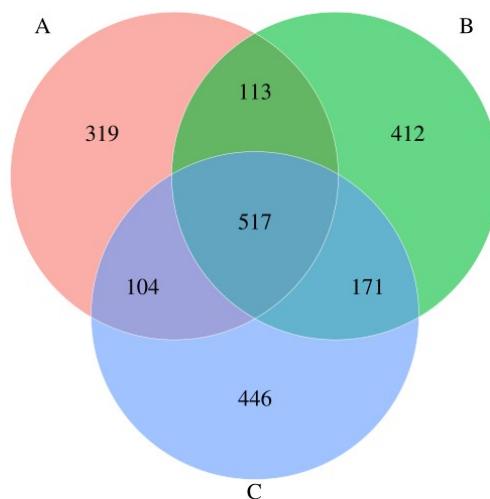


图 2.5 组别物质重叠数韦恩图

注：图中每个圈代表一个样本组鉴定风味物质数，重叠部分代表多个样本组共同鉴定到的风味物质数，非重叠部分代表相应样本组特异性鉴定到的风味物质数。

2.3.3 风味物质种类分析

风味物质种类主要包括烃类、醛类、酯类、酸类、酮类、醇类、醚类、酚类和杂环化合物等挥发性风味物质，由风味前体物质在加工过程中发生的一系列复杂生化反应所产生。

采用 PubChem 数据库及 Classyfire 软件^[1]对检测风味物质进行种类注释分析，并分析各种类对应风味物质数目、相对含量。

表 2.3物质种类相对含量

分组	相对含量/%							
	烃类	醇类	醛类	酮类	酯类	酸类	杂环化合物	其他
A	17.0461	8.9786	3.716	2.6226	4.0919	2.6606	2.6985	58.1858
B	10.4011	18.9404	3.8419	6.7434	13.5016	1.0923	0.7744	44.7048
C	12.2749	14.5503	2.9044	14.2012	1.4264	0.678	1.2898	52.675

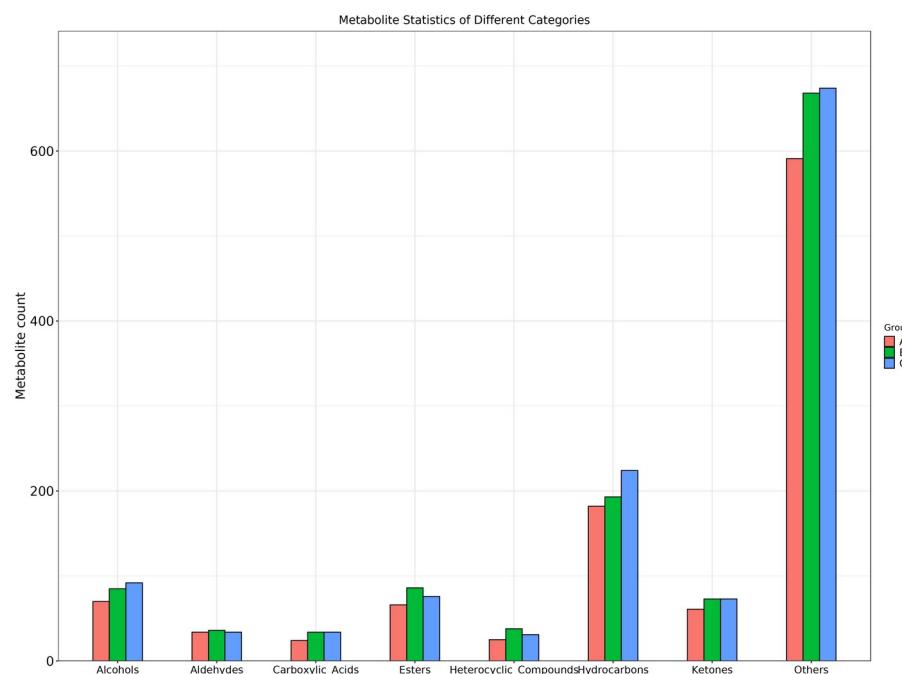


图 2.6物质种类数目统计柱形图

注:横坐标表示物质种类, 纵坐标表示风味物质数目, 颜色表示不同的分组。

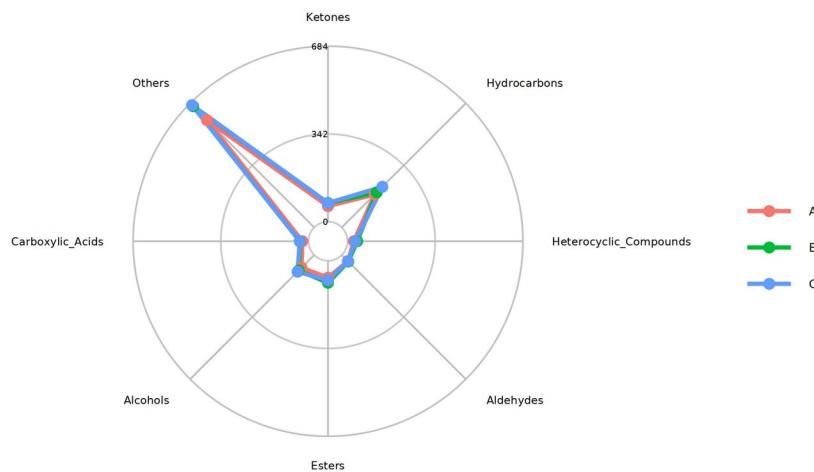


图 2.7 物质种类数目统计雷达图

注：最外圈名称表示物质种类，折线表示种类对应风味物质数目，颜色表示不同分组。

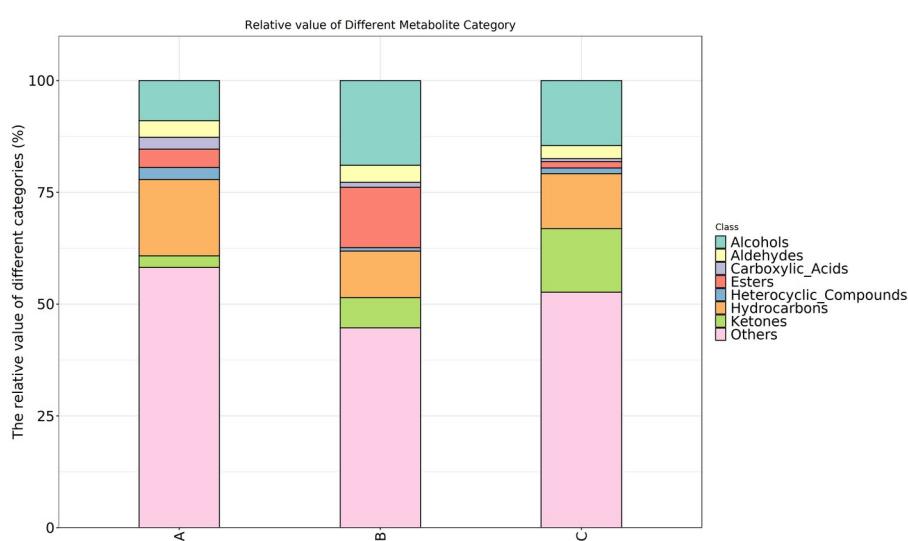


图 2.8 物质种类相对含量堆积图

注：横坐标表示样本分组，纵坐标表示种类相对含量，颜色表示风味物质种类。

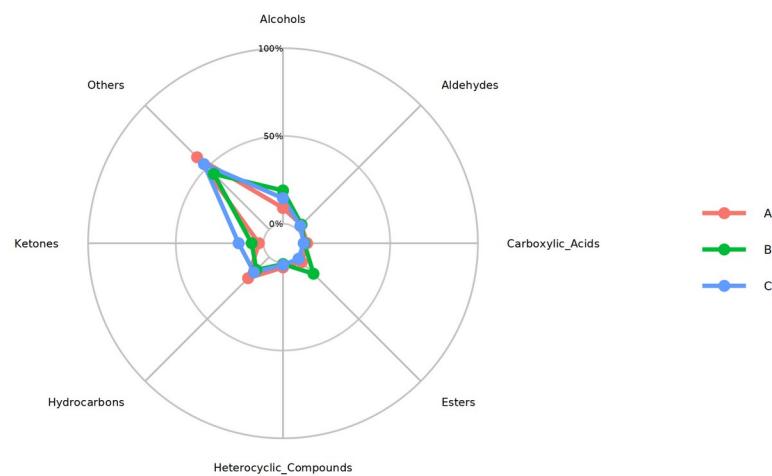


图 2.9 物质种类相对含量雷达图

注：最外圈名称表示种类，折线表示种类相对含量，颜色表示不同分组。

2.3.4 挥发物阈值分析

目前，常用相对气味活度法（ROAV）来评价样品的风味。一般 ROAV 越大，说明该物质对风味的贡献度越大^[2]。

表 2.4 气味活性物质信息表（部分结果）

Name	OdorCharacter	A_ROAV	B_ROAV	C_ROAV
3-Hexen-1-ol, (E)-	Fresh,Green,Raw	0.002611169646	0.001834983303	0.003808355115
	Fruity With A	211964	383337	1701
2-Hexen-1-ol, (E)-	Fruity,Fresh,Green	2.010406776949	0.000860770541	0.002695972678
		13e-07	59525	6782
1-Octen-3-ol		0.013949405283	0.005911004255	0.020373403130
	Mushroom	01314	66365	1164
1-Heptanol		0.008538600597	0.002146977032	0.013851571055
	Grassy	886309	806488	06415
1-Butanol, 3-methyl-	sweet, malty, rancid,rubber,	0.027070459477	0.010787856026	0.012530972101
		7265	05455	17635

注：Name：化合物名称；Chinese_name：中文名称；Class：注释种类；Chinese_Class：中文注释分类；1st Dimension Time (s)：一维保留时间（单位秒）；2nd Dimension Time (s)：二维保留时间（单位秒）；CAS：cas 号；Formula：分子式；分组_ROAV：各组别相对香气值；Range of odor Min：风味特性最低浓度值；Range of odor Max：风味特性最高浓度值；Order character：感官风味特征。默认显示前 5 条。

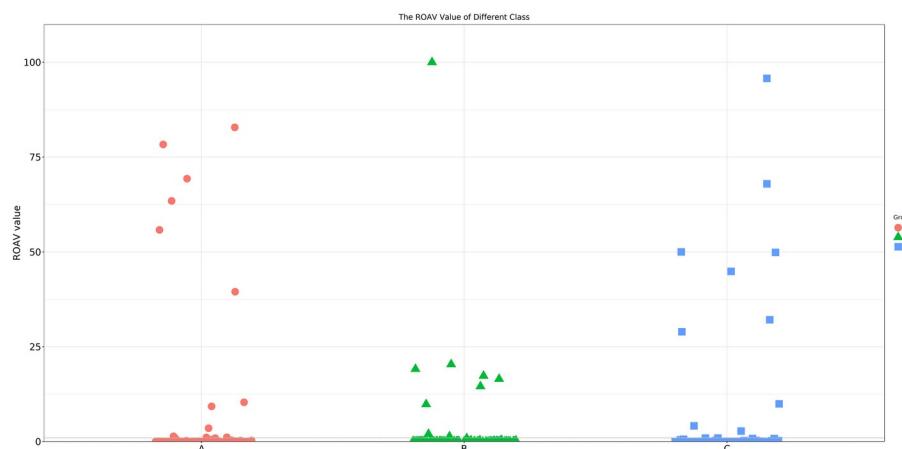


图 2.10 ROAV 气味活度值散点图

注：横坐标表示不同组别，纵坐标表示风味物质 ROAV 值，颜色表示不同的分组。

2.3.5 感官风味特征分析

产品的风味是由可识别的味觉和嗅觉特性，以及不能单独识别特性的复合体，两部分组成。本项目采用 Flavordb^[3]，对物质的感官风味进行了分析和比较。

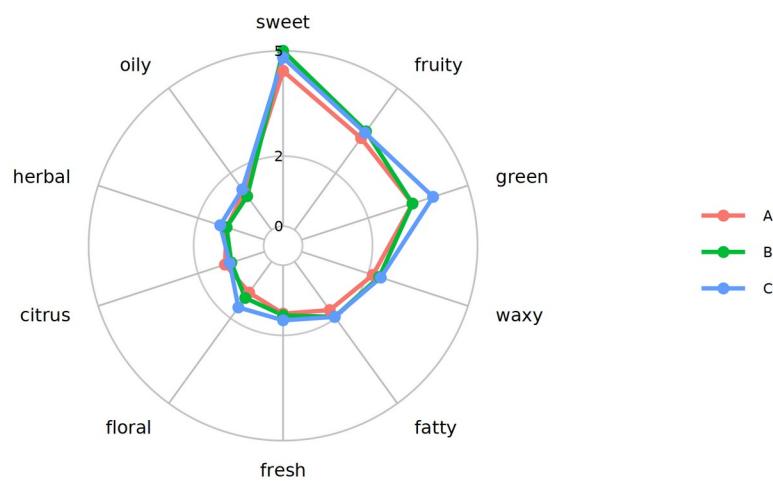


图 2.11 感官风味特征分析雷达图

注：最外圈名称表示感官风味特征，折线表示对应风味物质频次等级（检出频次从 1-5 进行等级划分，频次最高的为 5 级），颜色表示不同分组。

2.4 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需要运用化学计量学原理和多元统计的方法，对采集的多维数据进行降维和归类分析，从中挖掘提炼出最有用的信息。

通常，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。

目前代谢组学研究常用的数据标准化方式有中心化处理、自适换算、帕莱托换算等。

本分析在进行此样品分型前对数据采用【自适换算】处理，以获得更加可靠且更加直观的结果。

本分析中使用的多元统计分析（软件包 SIMCA-P(v13.0)和 R 语言 rpls 包^[4]）方法有：

- (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）。

2.4.1 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，去除重复性差的样本（离群样本）和异常样本（在置信区间——Hotelling T₂ 椭圆外的样本）。

因无外加人为因素，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并对数据从整体上进行把握，尤其是有利于发现和剔除异常样品，并提高模型的准确性。经过 PCA 计算出的数学模型是否可靠需要进行严格的验证——不可靠的数学模型不仅不能很好地描述代谢组学数据特点，还可能严重影响正确结果的获得甚至误导分析结果。

模型的交叉验证主要参考 R_{2X} 等参数，R_{2X} 是模型的可解释度。通常情况下，R₂ 高于 0.5 较好。

各个样品在各个主成分的得分就是其在计算的数学模型中的空间坐标，直观地反映了各个样品在数学模型空间中的分布情况。从 PCA 得分图可观察样品的聚集、离散程度。样品分布点越靠近，说明这些样品中所含有的变量/分子的组成和浓度越接近；反之，样品点越远离，其差异越大。

表 2.5 PCA 模型验证参数

Comparison	pre	R _{2X(cum)}

注：pre，主成分数；R_{2X}，模型（对 X 变量数据集）可解释度。

无分析结果

图 2.12 PCA 得分图

注：横坐标表示第一主成分解释度，横坐标方向第二主成分解释度。点表示样本，颜色表示不同分组。组内样本越聚集，组间样本越分散，说明结果越可靠。

2.4.2 偏最小二乘判别分析 (PLS-DA)

无监督分析方法 (Unsupervised Analysis, 如 PCA) 不能忽略组内误差、消除与研究目的无关的随机误差, 过分关注于细节、忽略了整体和规律, 最终不利于发现组间差异和差异化合物。在这种情况下, 就需要利用样本的先验知识将数据分析, 进一步聚焦到我们要研究的方面, 采用有监督模式识别方法 (Supervised Analysis), 如偏最小二乘法-判别分析 (Partial Least Squares-Discriminate Analysis, PLS-DA)。

与 PCA 只有一个数据集不同, PLS-DA 在分析时必须对样品进行指定并分组, 这样模型会自动加上另外一个隐含的数据集 Y, 该数据集变量数等于组别数。PLS-DA 是目前代谢组学数据分析中最常使用的一种分类方法, 它在降维的同时结合了回归模型, 并利用一定的判别阈值对回归结果进行判别分析。

PLS 与 PCA 不同之处在于 PLS 即分解自变量 X 矩阵, 也分解因变量 Y 矩, 并在分解时利用其协方差信息, 从而使降维效果较 PCA 能够更高效的提取组间变异信息^[5]。

模型的交叉验证主要参考 R_{2X}、R_{2Y}、Q₂ 等参数, R_{2X} 是模型 X 变量 (自变量) 的可解释度, R_{2Y} 为模型 Y 变量 (因变量) 的可解释率, Q₂ 是模型的可预测度 (理论上, R_{2X}、Q₂ 数值越接近 1 说明模型越好, 越低说明模型的拟合准确性越差, 通常情况下, R₂、Q₂ 高于 0.5 较好, 且两者差值不应过大, R₂ 和 Q₂ 最大值为 1)。当 R₂ 值较小时, 往往意味着测试集中重复性较差 (背景噪音高时); Q₂ 值较小时, 表示测试集中具有较高的背景噪音, 或者模型具有较多的异常样本 (Outlier)。

表 2.6 PLS-DA 模型验证参数

Comparison	pre	R2X(cum)	R2Y(cum)	Q2(cum)

注: pre, 主成分数; R_{2X}, 模型 (对 X 变量数据集) 可解释度; R_{2Y}, 模型 (对 Y 变量数据集) 可解释度; Q₂, 模型可预测度。

无分析结果

图 2.13 PLS-DA 得分图

注: 横坐标表示第一主成分解释度, 横坐标方向第二主成分解释度。点表示样本, 颜色表示不同分组。组内样本越聚集, 组间样本越分散, 说明结果越可靠。

无分析结果

图 2.14 PLS-DA 置换检验

注: 当本图满足以下任意一点时, 说明结果可靠有效: 1.所有蓝色的 Q₂ 点从左到右均低于最右的原始的蓝色的 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); 2.点的回归线与纵坐标交叉或者小于 0。

2.4.3 正交-偏最小二乘判别分析 (OPLS-DA)

代谢组学数据分析中另一种常用的方法是正交-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。

OPLS-DA 使用正交信号校正技术, 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成分。

与 PLS-DA 模型相同, OPLS-DA 同样可以用 R2X、R2Y、Q2Y、CV ANOVA 和 OPLS-DA 得分图来评价模型的分类效果。

通常, 根据 VIP (Variable Importance for the Projection) 值来说明变量 (特征峰) 能解释 X 数据集和关联 Y 数据集的重要性。所有 VIP 值的平方之和与模型中的变量总数相等, 因此, 其平均值为 1。

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

表 2.7 OPLS-DA 模型验证参数

Comparison	pre	R2X(cum)	R2Y(cum)	Q2(cum)

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。

无分析结果

图 2.15 OPLS-DA 得分图

注: 横坐标表示第一主成分解释度, 横坐标方向第二主成分解释度。点表示样本, 颜色表示不同分组。组内样本越聚集, 组间样本越分散, 说明结果越可靠。

无分析结果

图 2.16 OPLS-DA 置换检验

注: 当本图满足以下任意一点时, 说明结果可靠有效: 1.所有蓝色的 Q2 点从左到右均低于最右的原始的蓝色的 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); 2.点的回归线与纵坐标交叉或者小于 0。

2.5 差异分析

2.5.1 差异物质筛选

通过对物质进行筛选，从样本物质列表中寻找差异物质。可选择的相关差异物质筛选条件为：在 t test 或单因素方差分析检验中 pvalue < 0.05 + VIP > 1^[7]。

表 2.8 参数详细含义

筛选值	筛选值含义	筛选阈值	备注
pvalue	统计学上显著差异	pvalue 小于 0.05	无
VIP	OPLS-DA 第一主成分变量重要性值投影	VIP 大于 1.0	无

表 2.9 差异物质数目统计

Comparison	Total	Up	Down	Total_DE

注：Total：共有物质总数；Up：两组相比，相对含量较高物质数；Down：两组相比，相对含量较低物质数；Total_DE：总差异物质数。

最终获得的差异物质见下表（由于篇幅问题，显示部分数据结果）：

表 2.10 差异物质筛选结果列表（部分数据）

Name	Chinese_name	CAS	1st Dimension Time (s)	2nd Dimension Time (s)	VIP	Log2(FC)

注：Name：化合物名称；Chinese_name：中文名称；Class：注释种类；Chinese_Class：中文注释分类；1st Dimension Time (s)：一维保留时间（单位 s）；2nd Dimension Time (s)：二维保留时间（单位 s）；CAS：cas 号；Formula：分子式；分组_Mean：物质在各组别中相对含量均值，分组_Median：物质在各组别中相对含量中位数；分组_RSD：各组别 RSD 值；FC：Fold Change，为差异倍数；log2FC：log2 为底的差异倍数大小；P.value，统计学 p 值，越小说明差异越显著；-log10(P.value)：-log10 为底的 P.value；FDR，假阳性校正后值，采用 BH(Benjamini and Hochberg) 法计算，数值越大假阳性的可能越高；VIP，OPLS-DA 第一主成分变量重要性值投影；Regulation：该物质为两组相比，相对含量较高或相对含量较低。该表默认显示前 5 条。

以下为差异鉴定物质统计结果：

无分析结果

图 2.17 差异物质数目

注：横坐标表示差异物质数量；纵坐标表示对比组，其中红色表示两组相比，相对含量较高物质数量；蓝色表示两组相比，相对含量较低物质数量。

2.5.2 韦恩图

韦恩图（Venn Diagram），是显示元素集合重叠区域的图示，可用于统计不同比对组别中所共有和独有的差异物质数目，可以比较直观地显示不同比对组别中差异物质组成的相似性和重叠情况。（如果分组超过六组，改用 UpSetPlot 来画。默认显示数目大于 10 的交集）。

无分析结果

图 2.18 分组比较韦恩图

注：不同颜色表示两组共有差异物，重叠区域为多组共有差异物。

2.5.3 热图

聚类分析被用于判断代谢物在不同实验条件下的代谢模式。以不同实验条件下代谢物的相对值为代谢水平，做层次聚类（Hierarchical Clustering）分析，结果以热图表示^{[8][9]}^[10]。热图表现的是一个数据矩阵，通过使用颜色梯度使数据间的差异实现可视化，通过数据缩放，保留较大差异，同时也能突显较小差异。不同颜色的区域代表不同的聚类分组信息，同组内的代谢模式相似，可能具有相似的功能或参与相同的生物学过程。因此通过将代谢模式相同或者相近的代谢物聚成类，可以用来推测已知或未知代谢物的生物学功能。

本实验采用凝聚层次聚类（Agglomerate Hierarchical Clustering）：即将每个对象归为一类，合并这些类成为越来越大的对象，直到终结。通过 R (v3.3.2) 中 Pheatmap 程序包对数据集进行缩放，得到物质相对定量值层次聚类图如下：

无分析结果

图 2.19 差异分子热图

注：图中相对含量的大小通过颜色的不同来显示，颜色越红相对含量越高，越蓝相对含量越低。其中列代表样本，行代表物质，图中左侧的聚类树为差异物质种类聚类树。物质数目超过 150 不显示物质名称。

2.5.4 火山图

火山图（Volcano Plot）可直观的表现两组样品的差异代谢物的分布情况。通常横坐标用 \log_2 (FC) 表示，差异越大的代谢物分布在两端，纵坐标用 $-\log_{10}$ (pvalue) 表示，为统计检验的显著性 P 值的负对数。

图中 VIP 的过滤参数为 1，pvalue 的过滤参数为 0.05。

无分析结果

图 2.20 差异火山图

注：图中每一个点表示一种物质，横坐标表示某物质在两样品中定量差异倍数的对数值；纵坐标表示 P 值的 $-\log$ 对数值。横坐标绝对值越大，说明某物质在两样品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，筛选得到的差异表达物质越可靠。图中红色点代表上调差异表达物质，蓝色点代表下调差异表达物质，灰色点代表检测到但差异不显著的物质。

2.5.5 感官风味特征分析

基于数据库 Flavordb^[3]，针对风味物质特有的感官风味特征，采用 igraph 构建它们之间的网络图关系。

无分析结果

图 2.21 感官风味特征与风味物质关联网络图

注：绿色圆圈表示感官特征，红色圆圈表示风味化合物，绿色圆圈越大，说明与该感官特征连接的风味化合物种类越多，该感官特征越重要；红色圆圈越大，说明与该风味化合物相连的感官特征种类越多，该风味物质越重要。默认取感官特征 top10 制作网络图。

3. 实验

3.1 实验方法

样本处理及上机检测方法见 [《实验报告》](#)。

4. 数据预处理

1. 使用 ChromaTOF 软件进行数据分析后得到各个样本物质名称、保留时间、CAS 号、数据库 RI 信息、正构烷烃 C7-C30 计算的实际 RI 及其样本峰面积等信息，把每个样本的这些注释信息进行整合后得到最终的物质信息表格。

2. 为使不同量级的数据能够进行比较，对数据进行峰面积的总峰面积归一化或者内标归一化^[11]，该项目中使用 【】 进行归一化。

1. After data analysis using ChromaTOF software, the name of each compound, its retention time, CAS number, database RI information, actual RI calculated by normal alkanes C7-C30 and its sample peak area and other information are obtained, The final analysis result was obtained by integrating these interpretation information together.

2. In order to enable data of different magnitudes to be compared, the total peak area normalization or internal standard normalization was performed on the raw data^[11]. This project uses for normalization.

5. 参考文献

- [1] Djoumbou Feunang, Y., Eisner, R., Knox, C. et al. ClassyFire: automated chemical classification with a comprehensive, computable taxonomy[J]. Journal of Cheminformatics, 2016, 8:61.
- [2] Zhu, Yifan et al. Use of relative odor activity value (ROAV) to link aroma profiles to volatile compounds: application to fresh and dried eel (*Muraenesox cinereus*). International Journal of Food Properties 23 (2020): 2257 - 2270.
- [3] Garg N, Sethupathy A, Tuwani R, et al. FlavorDB: a database of flavor molecules[J]. Nucleic Acids Res, 2018, 46(D1):D1210-D1216.
- [4] EA Thévenot, 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, 2015:3322-35.
- [5] Anne-Laure B, Korbinian S. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data[J]. Briefings in Bioinformatics, 2007, 8(1):32-44.
- [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.
- [8] Rao J, Cheng F, Hu C, et al. Metabolic map of mature maize kernels[J]. Metabolomics, 2014, 10(5):775-787.
- [9] 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.
- [10] 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.
- [11] 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):1060-1083.

6. 附录

6.1 生物信息学常用软件及数据库

组学常用软件及数据库信息

数据库及相关软件	说明或网址连接
NIST2020	系统自带数据库，通过 LECO 供应商开发的 Chroma TOF 搜库软件 对下机原始数据进行物质注释分析
PubChem 数据库	有机活性小分子数据库 网址： https://pubchem.ncbi.nlm.nih.gov/
Classyfire 软件	对物质进行分类注释
Odor 数据库	基于参考文献整理 Odor Thresholds for Chemicals with Established Health Standards, 2nd Edition
Flavordb 数据库	风味物质数据库 网址： https://cosylab.iiitd.edu.in/flavordb/

6.2 结果文件及路径说明

分析内容	文件夹路径	结果文件说明
	metabolome.xlsx metabolome_normalization.xlsx	metabolome.xlsx 为经过除杂等筛选后的预处理数据； metabolome_normalization.xlsx 经缺失值填充，归一化处理后的数据。
2.2 数据检查	1.数据检查	包含一维、二维、三维总离子色谱图
		包含相关风味物质注释鉴定信息。其中： 鉴定物质数目统计文件夹，包含：物质鉴定数目、组别物质重叠数韦恩图
2.3 数据注释	2.数据注释	物质种类分析文件夹，包含：物质种类数目统计雷达图、物质种类数目统计柱形图、物质种类相对含量堆积图、物质种类相对含量雷达图 挥发物阈值分析文件夹，包含：ROAV 气味活度值散点图、ROAV 数据整合...
2.4 多元统计分析	3.多元统计分析	包含 PCA 分析综合结果、PLS-DA 分析综合结果、OPLS-DA 分析综合结果，以及各比较组的多元统计分析结果。其中： 各比较组多元统计分析包括 PCA 主成分分析、PLS-DA 偏最小二乘判别分析和 OPLS-DA 正交偏最小二乘判别分析。其中： PCA 主成分分析包含 PCA 得分图、PCA 载荷图和 PCA 分析结...
2.5 差异分析	4.差异物质分析	为所有差异分析结果文件夹，包含差异物质柱状图文件夹、分组比较韦恩图文件夹、各比较组差异分析文件夹。其中： 差异物质柱状图文件夹，包含：差异物质数目、差异物质数目统计.xlsx； 分组比较韦恩图文件夹，包含：各分组比较的韦恩图、及用于制作韦恩图数据表； 比较组差异分析文件夹，包含：差异物质筛选结果...
	data	包括 conditionDEM.xlsx (代谢物筛选条件)、datainfo.xlsx (项目及售后相关信息)、sampleinfo.xlsx (样本信息)
	png、static	生成网页端所需文件