多组学生物信息学分析服务

目录

转录组

一、项目基本信息

样品与数据信息	内容选项
数据类型	☐ GEO ☐ TCGA ☐ 自测数据 ☐ 其他（请注明：_____）
物种	请注明：_____
分组信息	请注明：_____
分析需求	参考以下内容，选择需要的分析内容

二、生物信息分析流程

2.1 分析流程

我们对公共数据库（GEO、TCGA 等）获得的 Count 或者进行标准化后的基因矩阵数据进行生物信息分析，包括数据下载和质控和差异基因功能分析。通过质控，以筛选高质量数据做进一步分析。根据项目方案设计情况，筛选目的差异表达基因，后续对差异基因进行 GO、KEGG 和 GSEA 富集分析以及 PPI 网络互作分析。

分析	结果	备注
质控	PCA，箱线图	数据量太大不宜做热图
差异分析	火山图，热图	可按需选择火山图上标注基因
GO 富集分析	柱状图，气泡图	环状图一般用编号表示
KEGG 富集分析	柱状图，气泡图	--
GSEA 分析	经典 GSEA 图	--
WGCNA 分析	聚类图，模块图	可提取关键模块基因
趋势分析	趋势图	按时间或周期分类
免疫浸润	箱线图，热图	查看免疫细胞比例
描述性统计分析	箱线图	查看基因在两组间的表达情况
	韦恩图	查看数据集间交集的基因
PPI	网络互作图	基因太多会看不清

生存分析	Kaplan-Meier 曲线图	有完整生存期数据才可以做
GO、KEGG 个性化分析	弦图、桑葚图	展示基因与通路之间的关联

2.2 分析内容

三、项目分析结果

3.1 原始数据质量评估

3.1.1 样本表达分布

获得均一化的基因表达矩阵后，我们对不同样品间基因整体表达分布进行可视化，一方面展示基因表达情况，还可以作为一种质量控制手段检查样品情况。

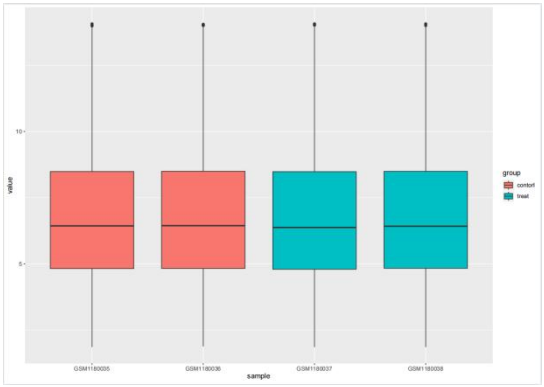


图 3-1-1 样本表达分布密度图

注：上图为不同样品表达分布密度图，可以直观地发现哪些类别中两组差异最大，哪些类别的数据变异最显著，为进一步的统计分析提供方向。

3.1.2 主成分分析

主成分分析(Principal Component Analysis， PCA)是一种利用正交变换将一组可能存在相关性的变量 转换为一组新的互相无关的几个综合变量(即主成分)的统计方法。这种分析方法可以降低数据的复杂性，深入挖掘样本之间的关系和变异大小。利用标准化的表达量对样本进行 PCA 分析，反映样本重复性，结果 如下 PCA 图：

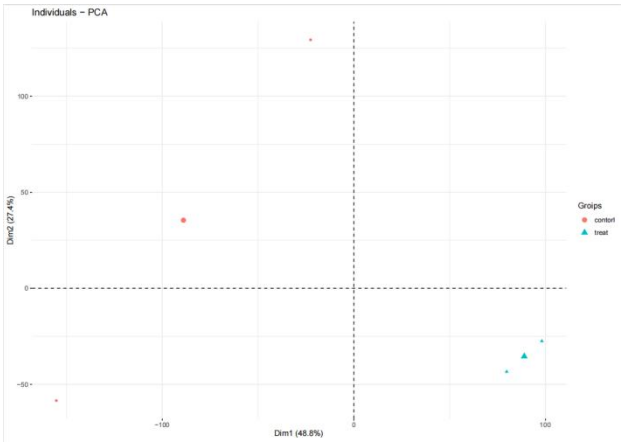


图 3-1-2 样本 PCA 得分散点图

注：每个散点代表一个样本，散点的颜色表示不同的分组，样本点分布越靠近，说明样本中基因的表达越相似；反之， 样本越远，其整体基因水平差异越大。

3.2 差异表达分析

基因表达具有时空特异性，不同试验处理基因表达丰度存在显著的变化，从而调控实现不同的生物学功能， 基因的差异表达就是比较两个样本或两组样本之间基因是否存在统计学差异，进而研究解释生物学现象

项目中，我们以基因表达矩阵为输入文件，应用 R 中 limma 包或 EdgeR 包进行差异分析。最终显著差异表达基因的筛选标准为：|log2FC|>1&P.value<0.05。

如下表为所有差异表达基因的统计结果

比较组间差异表达分析结果统计					
Dataset	**Pair**	**Species**	**DEG**	**Up-regulated**	**Down-regulated**
GSE20141	T 10 VS C 8	Homo sapiens	|log2FC|>1，P.Value<0.05	1127	319
GSE20163	T 8 VS C 9	Homo sapiens	|log2FC|>1，P.Value<0.05	339	299

注：Pair：格式为 T_vs_C，即基因在实验组的表达相对于对照组的变化情况

Up-regulated：表示实验组相对于对照组，基因表达显著上调

Down-regulated：表示实验组相对于对照组，基因表达显著下调

3.2.1 火山图

为了更形象展示每组样本比较的结果，我们采用火山图(Volcano-Plots)来展示两个(组)差异基因表 达状态，如下图：

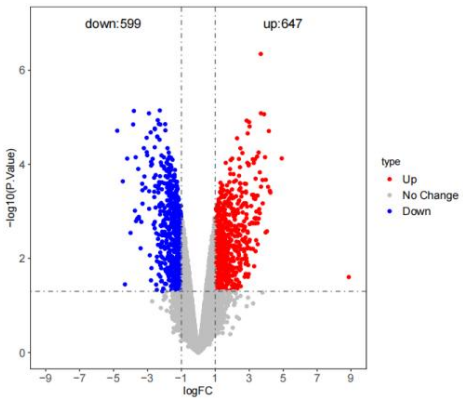


图 3-2-1 差异基因火山图

注：横坐标为基因在两组样本间表达量的倍数变化，即基因在实验组的表达量除以对照组的表达量的比值，然后对该比值进行以 2 为底数的对数处理。纵坐标为基因表达量变化差异的统计学检验值，即 logFC 值。对于-logl0(P.value)来说越高则表达差异越显著。 图中每个点代表一个特定的基因，红色点表示显著上调的基因，蓝色点表示显著下调的基因，灰色点为不显著差异的基因。

3.2.2 聚类分析

表达模式相似的基因通常具有功能相关性，聚类热图可以反映样本之间的相似性。我们将基因表达量为输入文件。如下的热图分别用所有基因的表达结果进行聚类作图，小于 3 个样本不展示聚类热图。

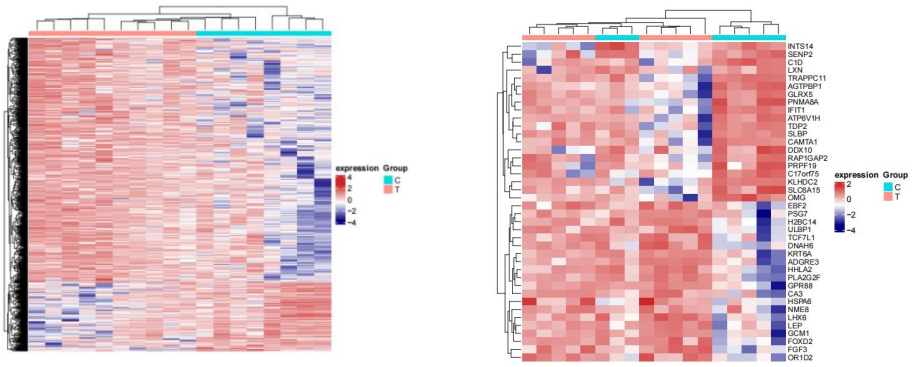


图 3-2-2 基因聚类热图

注：图中横向表示基因的聚类情况，而纵向表示样本的聚类结果，样本或基因在同一个树状分支下，表示它们就越相似。不同位置的色块代表对应位置基因的相对表达量，红色表示该基因高表达，蓝色表示该基因低表达。

3.3 富集分析

3.3.1 GO 富集分析

GO (Gene Ontology)是一种标准化的基因功能分类体系，目的在于标准化不同数据库中的关于基因和基因产物的生物学术语，对基因和蛋白功能进行限定和描述。GO 具体分为生物学过程(Biological Process,BP)、细胞组分(Cellular Component,CC)，分子功能(Molecular

Function,MF)三大类。

项目中，我们以差异基因作为输入的基因集，用 clusterProfiler 包进行富集分析，结果用 R 可视化。

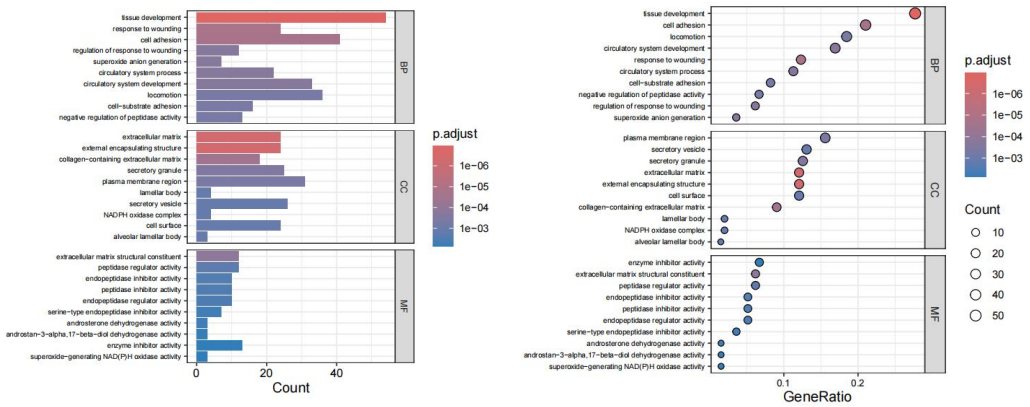


图 3-3-1 GO 富集分析条形图和气泡图

3.3.2 KEGG 通路富集分析

KEGG（Kyoto Encyclopedia of Genes and Genomes）是一个整合了基因、通路、疾病等信息的数据库，其核心是通路（Pathway）—— 由基因、蛋白质等分子相互作用形成的功能网络（如“PI3K-Akt 信号通路”）。

KEGG 富集分析的目的是：找到差异基因显著参与的信号通路。通过分析差异基因在 KEGG 通路中的分布，我们可以揭示这些基因是否协同参与某一特定生物学通路，进而理解其在生理或病理过程中的分子机制（例如，癌症中异常激活的通路）。我们以差异基因作为输入的基因集，用 clusterProfiler 包进行 KEGG 富集分析，结果用 R 可视化。

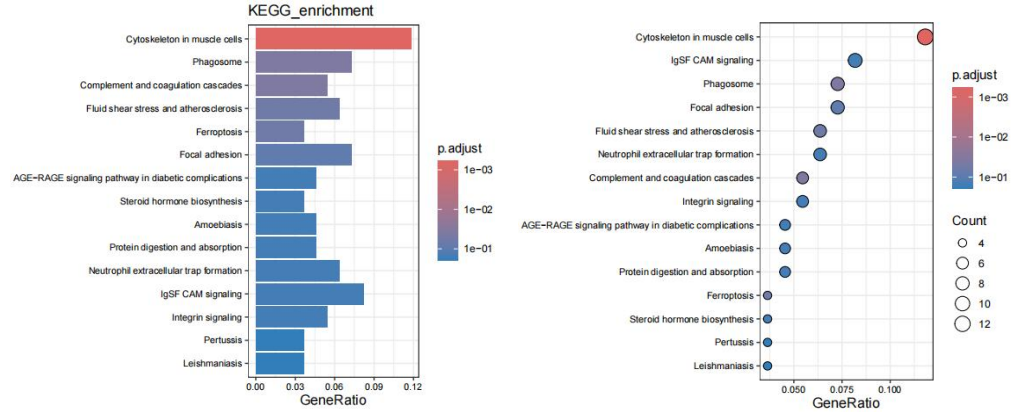


图 3-2-2 差异基因 KEGG 通路富集图

注：左图为 KEGG 通路富集条形图，右图为 KEGG 通路富集气泡图。图中 y 轴代表 KEGG 数据库中 pathway 名称；x 轴的 Gene Ratio 表示差异基因占注释到该通路所有基因的比例，该值越大表示差异基因在该通路中富集的越多；气泡的大小反映差异表达分析中差异基因与该通路基因的重叠情况，气泡越大代表重叠的差异基因越多；bar 和气泡的颜色反映通路富集的显著程度，从蓝到红，表示越来越显著。

3.3.3 GSEA 富集分析

基因集富集分析(Gene Set Enrichment Analysis,GSEA)是一种根据预定义的基因集来统计提供的 gene list 在两个生物状态之间的一致性和差异性，其目的是发现共有的生物功能或生物特性。GSEA 不关注某几个表达发生显著改变的基因，而是整个表达数据在特定功能基因集中的表达一致性，以此来解读数据中蕴含的生物学信息。因此 GSEA 可以避免差异表达分析中阈值筛选带来的问题。

MSigDB (Molecular Signatures Database)提供预定义基因集，根据蛋白定位、性质、功能、生物学意义等性质将基因分为八类，包括 H (hallmark gene sets)，C1 (positional gene sets)，C2 (curated gene sets) ，C3 (regulatory target gene sets)，C4 (computational gene sets)，C5 (ontology gene sets)，C6 (oncogenic signature gene sets)，C7 (immunologic signature gene sets)，C8 (cell type signature gene sets) 项目中，我们应用 MSigDB 数据库中的已知基因集进行 GSEA 显著性检验分析，

GSEA 富集分析的过程主要包括三个步骤:

1.计算富集分数(Enrichment Score)；

2.评估富集分数的显著水平

3.多重假设检验矫正。

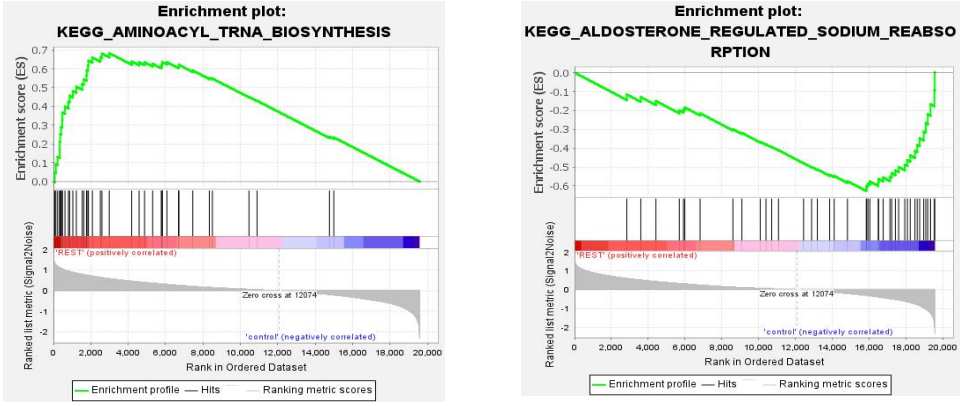


图 3-3-3 基因集富集图

注：ES 值计算的基本原理是扫描排序序列，当出现一个功能基因集中的基因时，就增加 ES 值，反之，就减少 ES 值， 所以在整个扫描过程中，ES 是一个动态的值(绿色曲线)。最终 ES 值的确定是将杂交数据排序序列所在位置定义为 0，ES 值定义为距离排序序列的最大偏差。当 ES 值为正，表示某一功能基因集富集在排序序列的前方，当 ES 值为负，表示某一功能基因集富集在排序序列的后方。

领头亚集(Leading Subset)中的基因是指对 ES 值贡献最大的基因集合。领头亚集的出现说明一方面这些基因在通路中有富集，非散在分布（在这里是指这些基因不是随机、零散地分布在基因组的不同位置，而是在功能上具有协同性、在调控上具有关联性的基因集合），另一方面，说明这些基因在通路中有共同的表达趋势。下方的每条竖线(黑色) 表示在该 gene set 中富集的基因。

排序得分(灰色)由大到小分布(即 Fold Change)，正值表示该通路基因多数上调，负值表示该通路基因多数下调，横坐标表示基因根据 Fold Change 由大到小的排序编号。

3.4 WGCNA 分析

加权基因共表达网络分析 (WGCNA,Weighted correlation network analysis)是用来描述不同样品之间基因关联模式的系统生物学方法，可以用来鉴定高度协同变化的基因集，并根据基因集的内连性和基因集与表型之间的关联鉴定候补生物标记基因或治疗靶点。

该分析方法旨在寻找协同表达的基因模块(module)，并探索基因网络与关注的表型之间的关联关系，以及网络中的核心基因。适用于复杂的数据模式，推荐 5 组(或者 15 个样品)以上的数据。

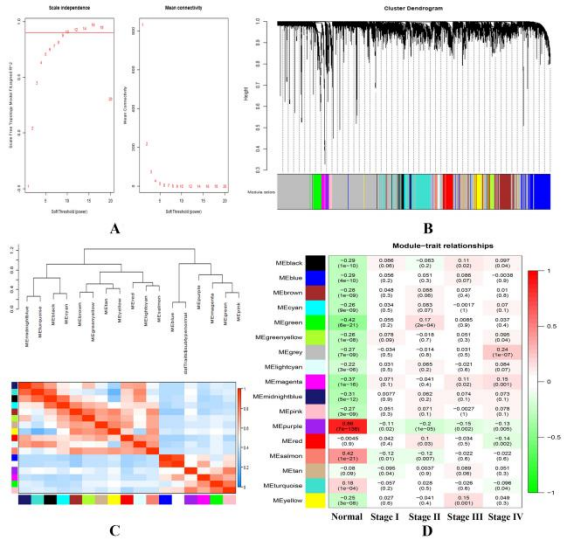


图 3-4 A 软阈值图；B 聚类图；C 模块间相关性图；D 基因与性状相关性热图。

注：WGCNA：Co-expression network（共表达网络） ：undirected, weighted gene networks，其点代表基因，边代表基因表达相关性，加权(weighted)是指对相关性值进行幂次运算。

Connectivity (连接度)：类似于网络中 “度”(degree)的概念。每个基因的连接度是与其相连的基因的边属性之和。

Module(模块)：高度内连的基因集。在无向网络中，模块内是高度相关的基因。

Module Eigengene(模块特征值)：模块内所有基因进行主成分分析（PCA），第一主成分的值即为 Eigengene。它代表该模块内基因表达的整体水平。

Module membership：给定基因表达谱与给定模型的 Eigengene 的相关性。

Hub gene：关键基因 (连接度最多或连接多个模块的基因)。

TOM (Topological overlapmatrix)：把邻接矩阵转换为拓扑重叠矩阵，以降低噪音和假相关，获得的新距离矩阵，这个信息可拿来构建网络或绘制 TOM 图。

3.5 趋势分析

趋势分析为梯度类文章的核心分析点，如实验设计（3-5 组）涉及梯度处理（时间变化、药物浓度/剂量变化、疾病程度、生长周期等）可利用趋势分析将表达模式相似的基因进行归类，从而找到实验变化过程中最具有代表性的基因集以及对应的趋势特征，揭示生物样本在变化过程中所特有的规律。后续可再与功能富集分析结合，更有效地挖掘数据内部的规律。

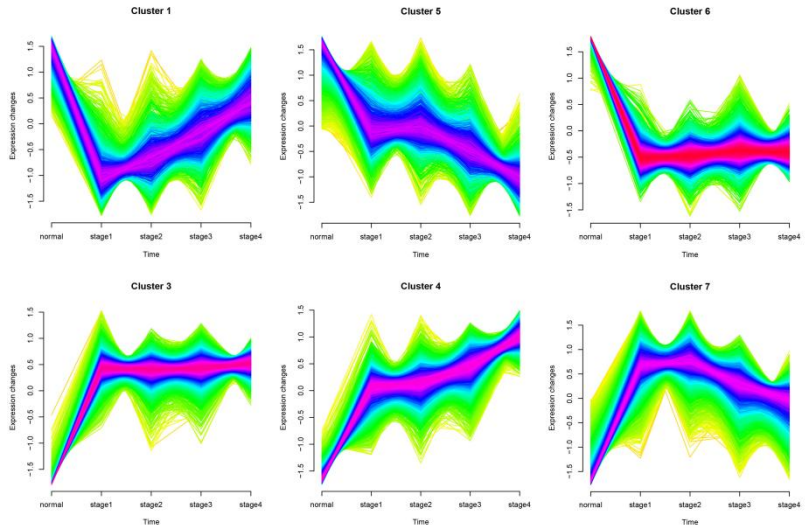


图 3-5 基因表达趋势图

3.6 免疫浸润

免疫浸润分析旨在探讨肿瘤微环境中免疫细胞的分布及其对肿瘤发生、发展和治疗的影响。肿瘤微环境由肿瘤细胞、免疫细胞、基质细胞及其他成分组成，其中免疫细胞在调控肿瘤免疫逃逸和抗肿瘤免疫中起着重要作用。通过分析免疫浸润特征，可以揭示免疫细胞的组成及其功能状态，从而为肿瘤的诊断、预后评估和个性化免疫治疗策略提供理论支持。

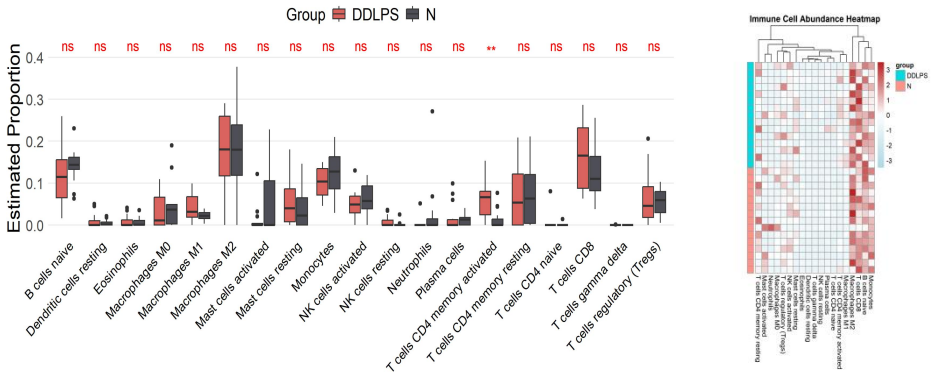


图 3-6 箱线图与热图

注：利用 CIBERSORT 包进行免疫浸润分析。提取每组的免疫细胞的比例数据进行免疫细胞比例热图、堆叠条形图的绘制。按细胞类型分组，用 Wilcoxon 秩和检验（非参数检验）比较两组差异并出局箱线图。最后进行目标基因与免疫细胞的相关性分析并出具热图。

3.7 描述性统计分析

3.7.1 箱线图

箱线图——既能展示数据分布特征，又能直观对比组间差异。箱线图的用途：展示每组数据的中位数、四分位范围；展示数据分布的离散程度；展示原始数据点的分布；展示组间差异显著性。箱线图可用于可视化并分析基因在两组或多组条件下的表达量分布。

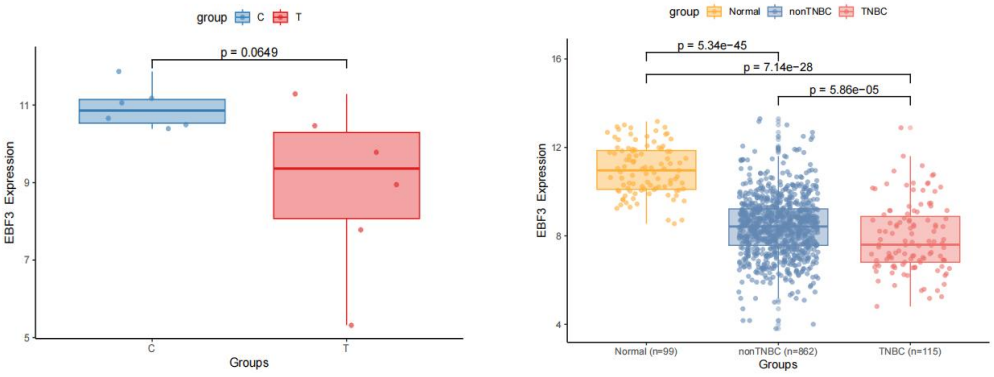


图 3-7-1 箱线图

3.7.2 韦恩图

韦恩图（Venn Diagram）是一种用图形表示集合之间关系的工具。在这里我们可以用于取多个数据集的交集。注意：韦恩图适合表示少量集合（通常 2-4 个），过多（如 4 个以上）会导致图形复杂难辨，可以使用 UpSetR 图来代替韦恩图。

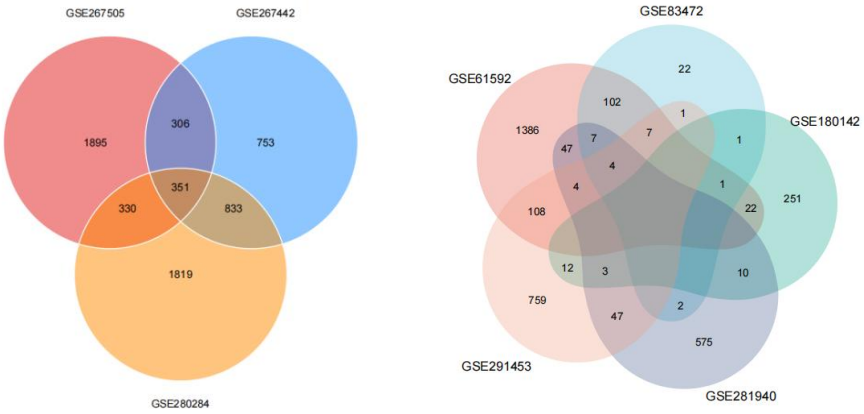


图 3-7-2 箱线图

3.8 PPI

蛋白质相互作用网络（Protein-Protein Interaction Network, PPI）为我们揭示细胞内蛋白质间的协同工作机制提供了一种直观的图模型。通过构建 PPI 网络，我们可以识别关键节点（hub 蛋白）、划分功能模块，并进一步结合富集分析推断相关生物学功能及疾病机制。

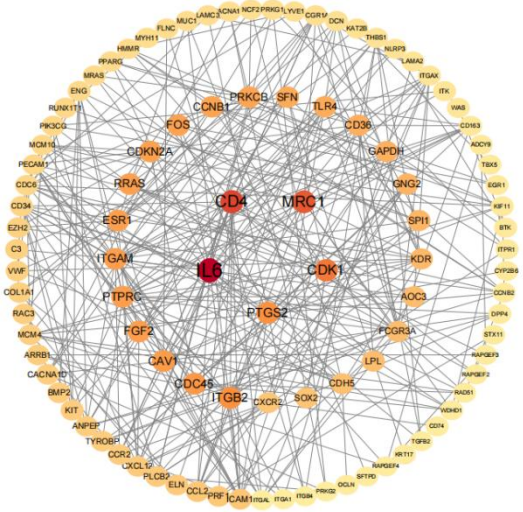


图 3-8 PPI 网络图

3.9 生存分析

Kaplan Meier 是一种单因素生存分析，它可用于研究 1 个因素（如基因表达量）对于生存时间的影响。

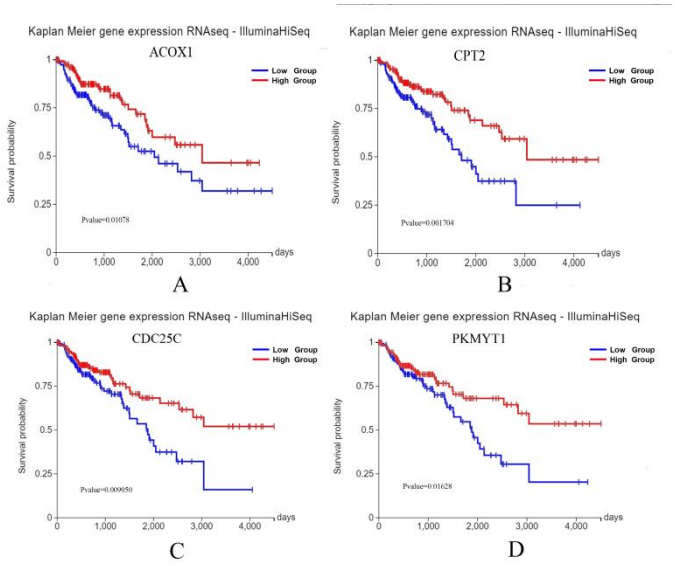


图 3-9 Kaplan-Meier 曲线图

注：一般而言，两条曲线之间的距离越大（分叉越大），说明两组患者预后（终点事件发生率）的差别越大，也越容易做出统计学差异。其实这个和 t 检验差不多，两组数据的均数差异越大，越容易有统计学差。随访时间越长，越容易做出统计学差异，样本量越大，越容易做出统计学差异。样本量越大，误差（标准误）越小，当然越有统计学意义。其实这相当于在 t 检验中，两组数据的标准差越小，当然越容易得到阳性统计学结果。

生存曲线与 X 轴有交叉，并不意味着研究对象全部死亡（发生终点事件）。实际上，在生存曲线中，每一个时间点上只要有病人死亡（或者发生终点事件），曲线就会下降一定的幅度。下降的幅度具体有多大，取决于该时间点上病人的死亡例数和后续随访时间（该时间截点以后的时间）病人的样本量。

3.10 GO、KEGG 个性化分析

弦图（Gene Ontology Chord Diagram）是一种用于展示基因功能富集结果的可视化工具，通过弦状连接可以更直观的展示基因与 GO term（如生物过程、分子功能等）之间的关联。

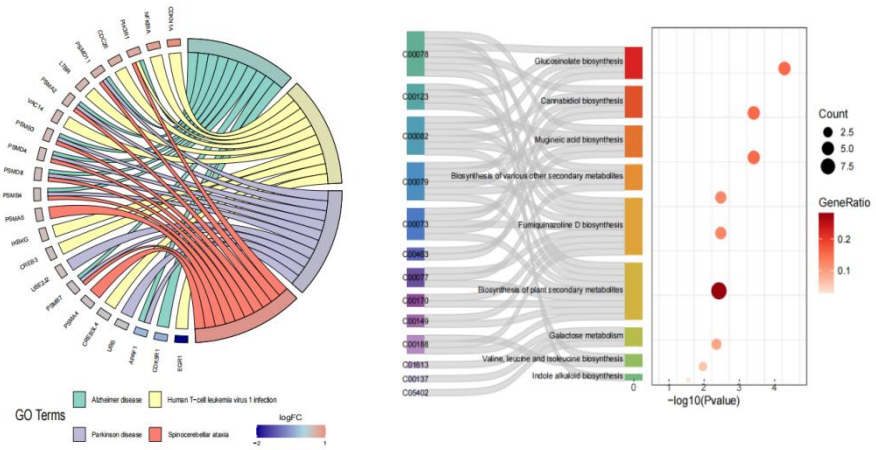


图 3-10 弦图和桑葚图

注：左图为弦图，它内圈连线表示基因和生物过程之间的关联，相同颜色的连线表示这几种基因富集到该 term 上；筛选出参与这些个生物过程的基因；标注显著富集的 GO term 的名称和颜色标识；不同颜色代表不同的 GO term。右侧为桑基图，它可以用于可视化生物体内代谢途径的不同代谢产物之间的流动。通过展示代谢物和这些代谢物所属分类以及富集在不同的 KEGG Pathway 之间的流动分布，帮助我们理解代谢途径中不同分子的相互关系和转化。

蛋白组

一、项目基本信息

样品与数据信息	内容选项
数据类型	☐ PDC ☐ 自测数据 ☐ 其他（请注明：_____）
物种	请注明：_____
分组信息	请注明：_____
分析需求	参考以下内容，选择需要的分析内容

二、生物信息分析流程

2.1 分析流程

缺失值处理：由于质谱检测的局限性，数据中常存在缺失值。DEP 包对缺失值进行填充。

标准化与归一化：对数据进行标准化和归一化处理，以消除实验过程中系统误差对数据的影响，使不同样本之间的数据具有可比性。常用方法包括中位数标准化、量值标准化、总和标准化等。

统计检验：使用合适的统计方法（如 t 检验、ANOVA、线性模型（limma 包）等）比较不同样本或组之间的蛋白质表达水平，筛选出显著差异表达的蛋白质。

差异阈值设定：根据生物学意义和研究目的，设定差异蛋白的筛选标准，如显著性 P 值（通常小于 0.05）和生物学差异倍数（如|log2 Fold Change|≥1、1.2 或 1.5）。

GO 富集分析和 KEGG 通路分析（见转录组）。还可进行蛋白相互作用网络分析（PPI），进一步挖掘蛋白质之间的相互作用和功能关联。

2.2 分析内容

分析	结果	备注
数据预处理	--	缺失值处理
PCA 主成分分析	PCA 图	样本需>3
样品间相关性分析	相关性热图	展示样本聚类情况
样品层次聚类分析	样品聚类树状图	

差异分析	火山图，热图	可按需选择火山图上标注蛋白
GO 富集分析	柱状图，气泡图	展示 GO 条目
KEGG 富集分析	柱状图，气泡图	展示 KEGG 条目
GO、KEGG 个性化分析	GO 弦图	展示通路到蛋白
描述性统计分析	箱线图	查看蛋白在两组间的表达情况
	韦恩图	查看数据集间交集的蛋白
PPI	网络互作图	蛋白太多会看不清
转录组与蛋白组联合分析	九象限图	展示基因与蛋白的变化情况

三、项目分析结果

3.1 PCA 主成分分析

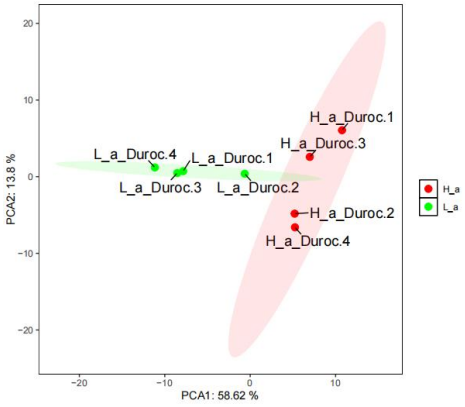


图 3-1 PCA

3.2 样品间相关性分析

通过计算样品之间的相关系数，量化样品在特征（如基因表达量、代谢物含量、物理化学性质等）上的相似性或关联程度。评估样品之间的关系，判断是否存在规律或趋势，辅助筛选异常样品、验证实验可靠性，或为后续分析（如聚类、差异分析）提供依据。常用方法为 Pearson 相关系数。

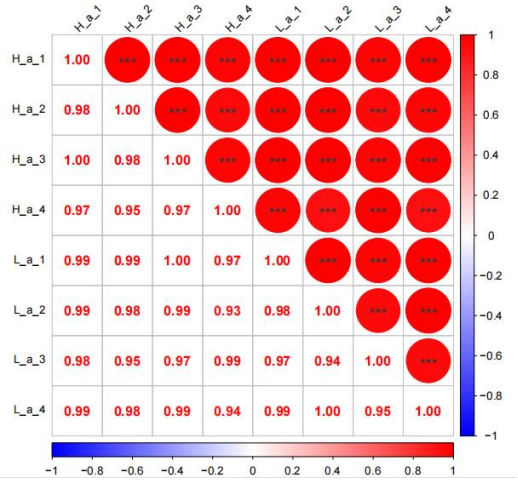


图 3-2 相关性热图

注：相关系数值：绝对值越接近 1，表明样品间相关性越强；越接近-1，相关性越弱。例如，相关系数为 0.8 表示强相关，-0.8 表示弱相关。

3.3 样品层次聚类分析

样品层次聚类分析是一种用于对样品（数据对象）进行分类的统计方法，通过计算样品之间的相似性或距离，逐步构建一个层次化的聚类结构，以揭示样品之间的内在关系和群体特征。样品层次聚类分析旨在将相似的样品归为一类，不同类的样品之间差异较大。

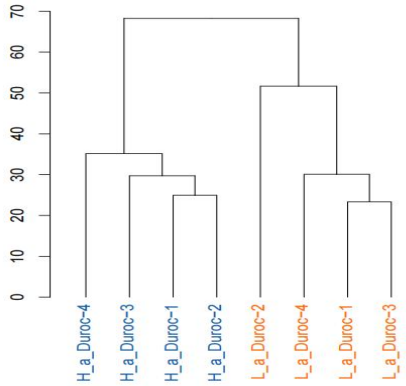


图 3-3 样品聚类树状图

注：横轴表示样品，纵轴表示合并或分裂时的距离。树的分支结构展示了样品的聚合过程，分支越短，说明对应的样品或簇越相似；分支越长，差异越大。

3.4 蛋白表达数据差异分析

同转录组（该步骤包含火山图、热图的绘制）。

3.5 GO、KEGG 富集分析

同转录组。

3.6 GO、KEGG 个性化分析

同转录组。

3.7 描述性统计分析

同转录组。

3.8 PPI 网络分析

同转录组。

3.9 转录组与蛋白组联合分析--九象限图

多组学关联分析可以使研究变得更加全面，整合蛋白组与转录组数据进行关联也是常见的分析方法之一。其中，我们通常可以使用九象限图直观地展示基因与蛋白的变化情况。

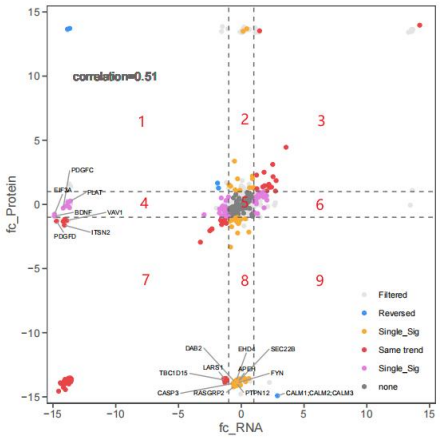


图 3-10 九象限图

注：九象限图顾名思义，该图形有九个象限（区域），如九象限图所示，横坐标是转录组的差异倍数（取了 log2），纵坐标是蛋白组的差异倍数（同样取了 log2），横坐标上的虚线表示转录组的差异倍数阈值，纵坐标上的虚线表示的蛋白组的差异倍数阈值，阈值线外表示显著差异的基因/蛋白，阈值线内则表示非显著差异的基因/蛋白，每个点代表一个基因/蛋白，深灰色的点表示非差异的蛋白和基因。象限从左上角开始，顺时针依次类推，第三和第七象限的点表示同时有显著差异的基因和蛋白（上调或下调），第二、四、六和八象限的点表示其中一个组学有显著差异而一个组学无显著差异的基因，浅灰色的点表示两个组学的差异倍数的绝对值大于 1 但 Q-value 大于 0.05 的基因和蛋白，第五象限为不显著的基因和蛋白。特别说明，生成的图片是没有的图中的象限数字，是便于理解而加上的。

代谢组

一、项目基本信息

样品与数据信息	内容选项
数据类型	☐ 自测数据 ☐ 其他（请注明：_____）
物种	请注明：_____
分组信息	请注明：_____
分析需求	参考以下内容，选择需要的分析内容

二、生物信息分析流程

2.1 分析流程

代谢组学是研究生物体内所有内源性小分子物质（分子量<1000 Da）的系统性科学。它通过定性和定量分析生物体内代谢物组成，揭示生物体在病理生理刺激或基因修饰下的代谢动态变化规律。

代谢组学数据分析大概可分为以下几个步骤：首先对检测得到的原始数据进行数据预处理、对处理完的数据进行统计分析，其中包括无监督的主成分分析（PCA）用于观察样本聚类，以及有监督的偏最小二乘判别分析（PLS-DA）、正交偏最小二乘判别分析（OPLS-DA）用于分类模型构建。对差异代谢物的筛选后并进行通路分析与功能富集。

2.2 分析内容

分析	结果	备注
数据预处理	--	缺失值处理
主成分分析（无监督）	PCA 图	样本需>3
主成分分析（有监督）	PLS-DA、OPLS-DA	样本需>3
差异分析		可按需选择代谢物在图上进行标注
差异分析可视化	火山图、热图、载荷图、环状热图、差异代谢物分类圈图	
KEGG 富集分析	柱状图、气泡图	展示 KEGG 条目

KEGG 个性化分析	桑葚图、弦图	展示通路及代谢物信息
代谢通路影响图	Pathway impact 图	展示特定通路在生物过程中影响
描述性统计分析	箱线图	查看蛋白在两组间的表达情况
	韦恩图	查看数据集间交集的蛋白

三、项目分析结果

3.1 主成分分析

单变量与多变量分析：单变量分析采用 t 检验、Anova 等方法筛选差异代谢物；多变量分析包括无监督的主成分分析（PCA）用于观察样本聚类，以及有监督的偏最小二乘判别分析（PLS-DA）、正交偏最小二乘判别分析（OPLS-DA）用于分类模型构建。

差异代谢物筛选：结合 Fold Change（FC）>2 或<0.5、P 值（校正后，如 FDR）<0.05、VIP 值（来自 PLS-DA 模型）>1 等标准筛选差异代谢物，并通过火山图、热图等可视化方法展示结果。

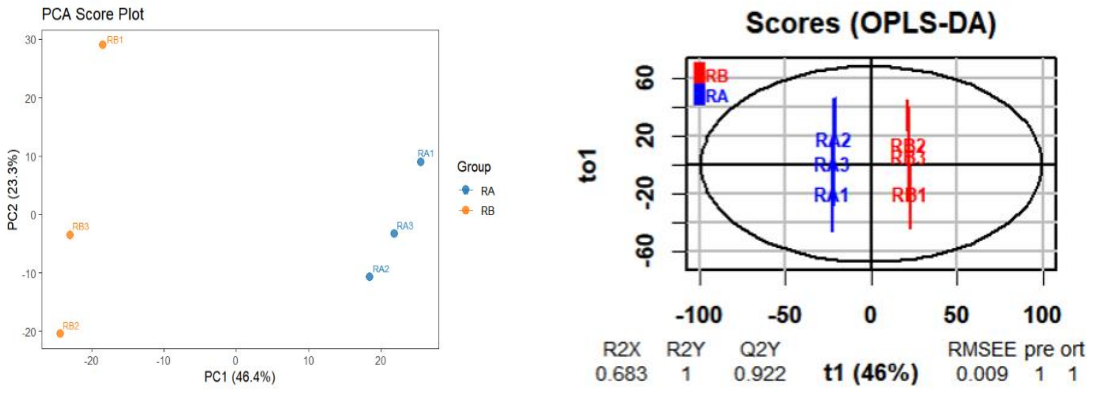


图 3-1 PCA 和 OPLS-DA

3.2 差异分析

结合 Fold Change（FC）>2 或<0.5、P 值（校正后，如 FDR）<0.05、VIP 值（来自 PLS-DA 模型）>1 等标准筛选差异代谢物。

3.3 差异分析可视化

通过载荷图、火山图、热图等可视化方法展示结果（火山图、热图同转录组）。

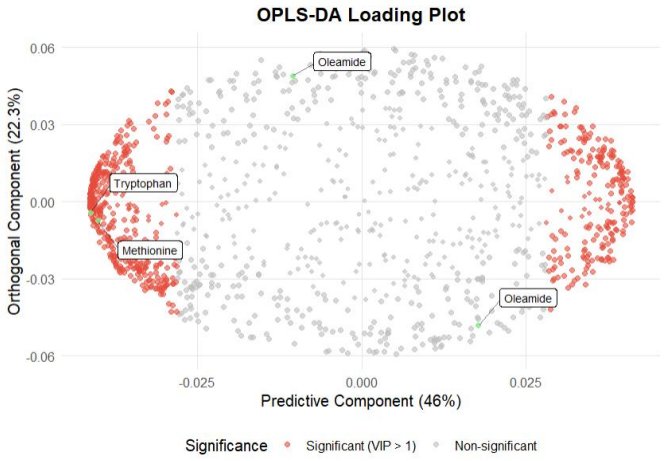


图 3-3 载荷图

注：载荷图展示了每个变量（代谢物或特征）如何影响模型的主成分。载荷（loadings）是变量在主成分方向上的投影，它们有助于了解哪些变量对样本的分类最为重要。

用途：通过载荷图，我们可以了解哪些变量在区分不同组别（例如饮食组）时最为关键。

载荷值较高的代谢物表明它们对主成分有较大的贡献，因此可能在分类过程中起到了重要作用。如果载荷图中的某些代谢物在主成分上的位置很远，说明它们对样本的分离具有很强的影响。

3.4 环状热图的绘制

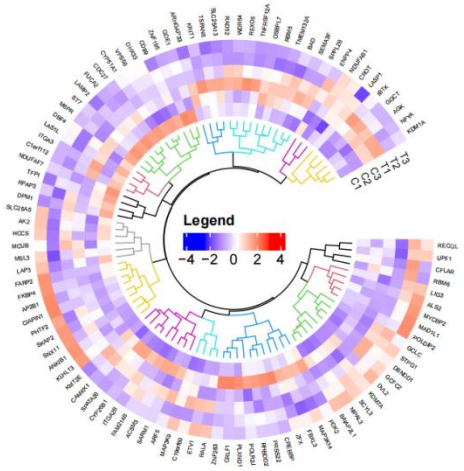


图 3-4 环状热图

3.5 差异代谢物分类

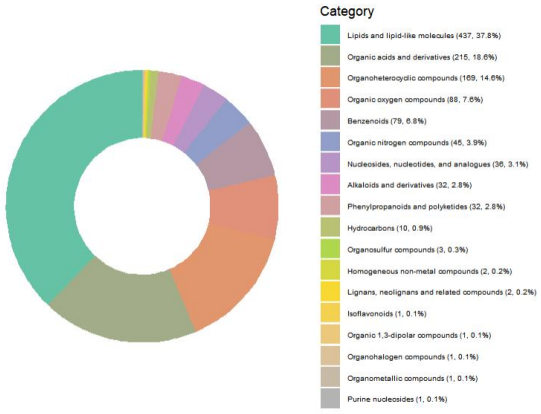


图 3-5 差异代谢物分类圈图

3.6 通路分析与功能富集

代谢通路富集分析：采用 KEGG 等数据库进行通路富集分析，识别显著的代谢通路。

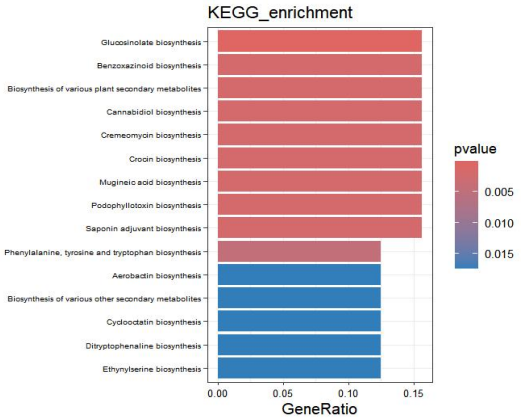


图 3-3 KEGG 条形通路图

3.7 KEGG 个性化分析

同转录组（弦图和桑葚图）。

3.8 Pathway impact 图

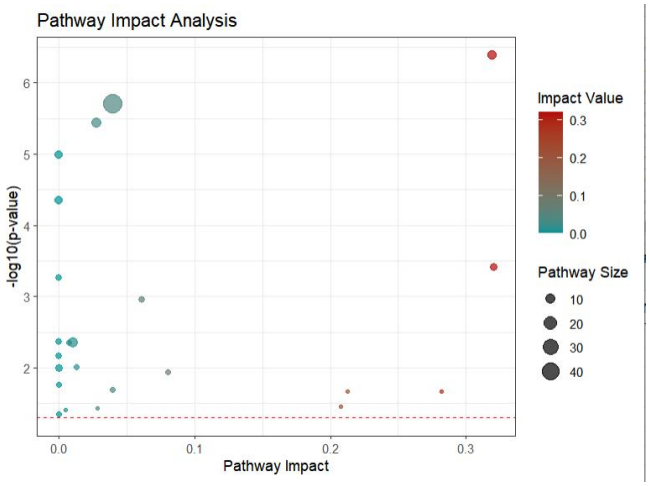


图 3-6 代谢通路影响图

注：Pathway Impact 图用于展示特定通路在生物过程中影响的可视化工具。它通过展示通路中各个组分之间的相互作用和影响关系，理解通路在生物学功能和疾病发生中的作用。点的大小和颜色可以表示不同的生物学意义，比如表达水平或统计显著性。

横轴通常表示 Pathway Impact 值，这是基于拓扑分析进行的权重计算，即通过将匹配的代谢物的重要性度量相加并除以每个通路中所有代谢物的重要性度量之和来计算得到。值越大表示该通路在生物学过程中的影响越大。

单细胞组学

一、项目基本信息

样品与数据信息	内容选项
数据类型	☐ GEO ☐ 自测数据 ☐ 其他（请注明：_____）
物种	请注明：_____
样本个数	请注明：_____
分组信息	请注明：_____
分析需求	参考以下内容，选择需要的分析内容

二、生物信息分析流程

2.1 分析流程

首先我们从公共数据库或实验平台获取单细胞测序数据，经过严格的质量控制、标准化处理和批次效应校正，确保数据可靠性。接着进行特征选择与降维，筛选高变基因并通过 PCA、t-SNE 或 UMAP 等方法将高维数据投影到低维空间。基于降维结果进行细胞聚类与分群，利用 Louvain 等算法将相似细胞划分为不同簇。随后通过细胞注释与标记基因识别，结合已知标记基因和参考数据库确定各簇的细胞类型，并识别特异性标记基因。在此基础上开展富集分析，探索差异表达基因的生物学功能和信号通路。为揭示细胞动态变化，进行拟时序分析重建细胞发育轨迹，细胞通讯分析解析细胞间信号网络，并在肿瘤研究中应用 inferCNV 分析推断染色体拷贝数变异。整个流程从数据到知识，系统揭示细胞异质性、功能状态和互作关系，为理解组织微环境、发育过程和疾病机制提供多维度视角。

2.2 分析内容

分析	结果	备注
公共数据库挖掘	--	--
细胞质控与过滤	质控、标准化、数据校正、特征选择和降维	--
降维与聚类分析	marker 热图、Dotplot 图、UMAP、tSNE	展示细胞聚类

差异表达分析	差异分析结果、多组火山图	差异分析结果是全部细胞类型的结果，如需指定，可单独分析
GO 富集分析	柱状图，气泡图	展示 GO 条目
KEGG 富集分析	柱状图，气泡图	展示 KEGG 条目
轨迹分析	细胞轨迹图、基因拟时序点图	展示特定通路在生物过程中影响
BEAM 进行统计分析	热图	展示同一时间点两个谱系的变化
细胞通讯分析	细胞互作关系、number of interaction 图、层次图、热图和配体-受体层级的可视化	展示通路及代谢物信息
inferCNV 分析	表达强度热图	展示肿瘤基因组各染色体区域相对表达强度的热图

三、项目分析结果

3.1 可视化 QC 指标

我们一般使用这些指标来过滤细胞：过滤具有 UMI 计数超过 2500 或小于 200 的细胞；过滤具有>5%线粒体的细胞。

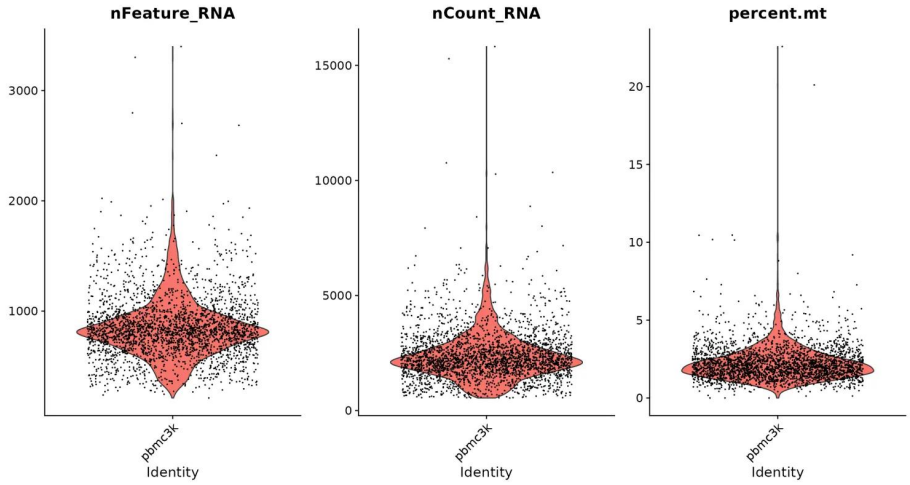


图 3-1 各个指标质控小提琴图

注：nFeature_RNA 代表检测到的基因数；nCount_RNA 为总 UMI 计数；percent.mt 线粒体基因百分比。每一个点代表一个个体细胞数据，identity 表示样本。

3.2 筛选高可变辑因

我们计算数据集中显示高变异的特征子集（即，它们在某些细胞中表达强烈，在另一些单元格中表达得很低）。在下游分析中关注这些基因有助于突出单细胞数据集中的生物信号。默认情况下，我们使用每个数据集的 2000 个基因。这些将用于下游分析，如 PCA。

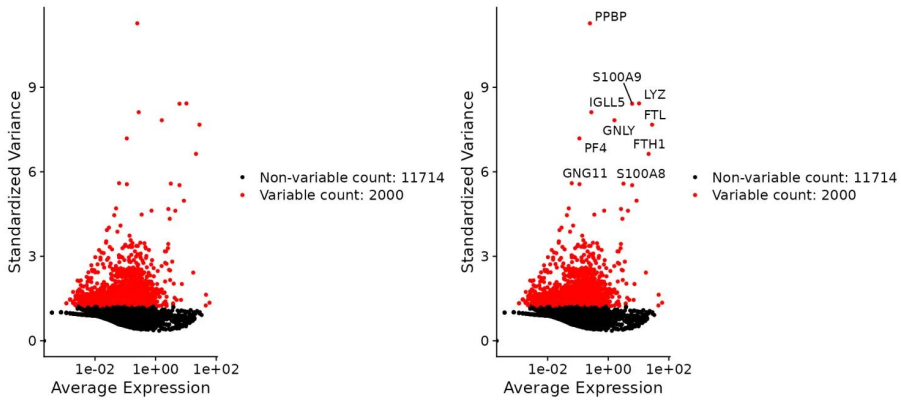


图 3-2 高变基因筛选图

注：左图为全部基因分布，黑点代表非高变基因，红点为高变基因。右图为高变基因突出显示，红点代表高变基因，标注为重要高变基因名称。

3.3 可视化 PCA

3.3.1 PCA

Seurat 提供了几个有用的方法来可视化定义 PCA 的单元格和功能，包括[DimPlot()]，[DimHeatmap()]等。

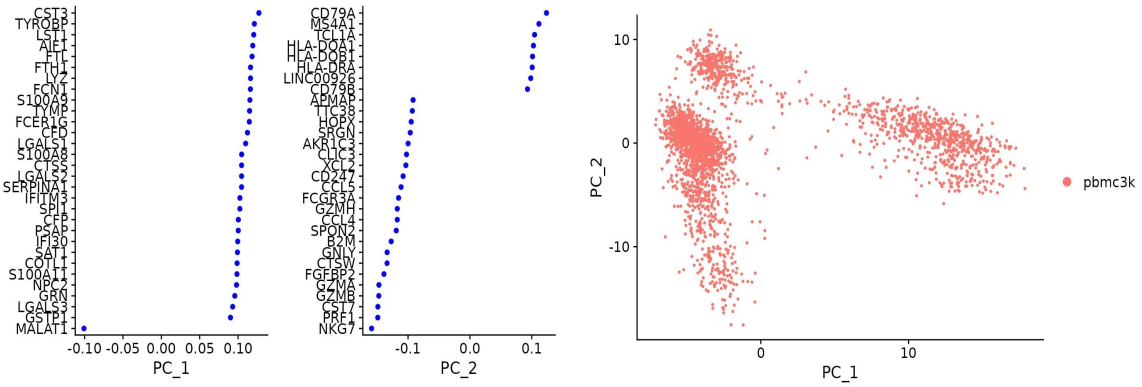


图 3-3-1 PCA

3.3.2 不同 PC 展示

我们通常选择包含前两个 PC 在内的一系列重要 PC（10-20）用于进一步下游分析。

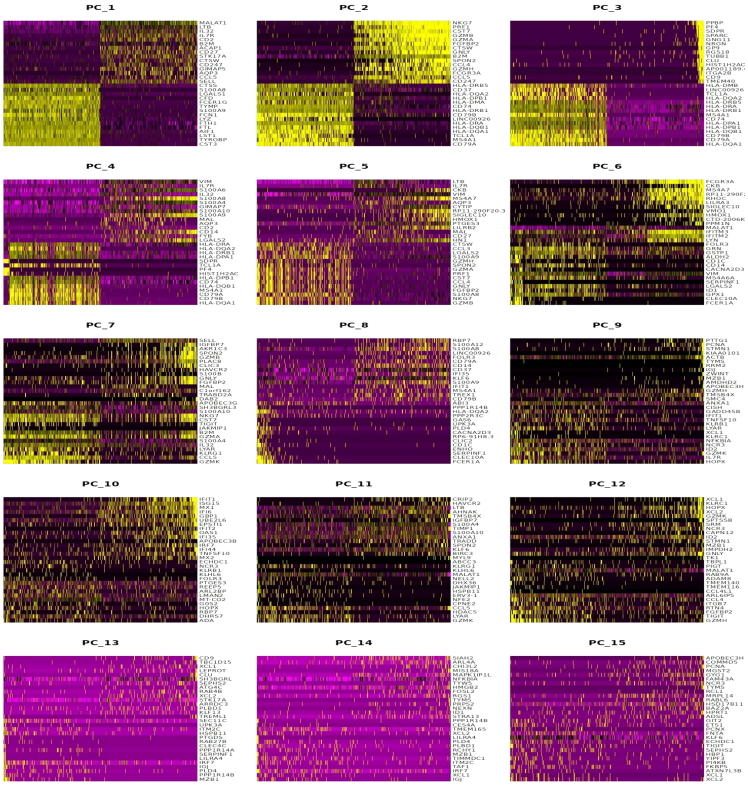


图 3-3-2 不同 PC 展示

3.4 确定数据集的“主成分个数”

Seurat 根据 PCA 分数对单元单元进行聚类，每个 PC 基本上代表一个“元结构”，该“元结构”将信息组合在相关功能集中。我们随机排列数据的子集（默认情况下为 1%）并重新运行 PCA，构建功能分数的“空分布”，并重复此过程。我们确定“重要”PC。

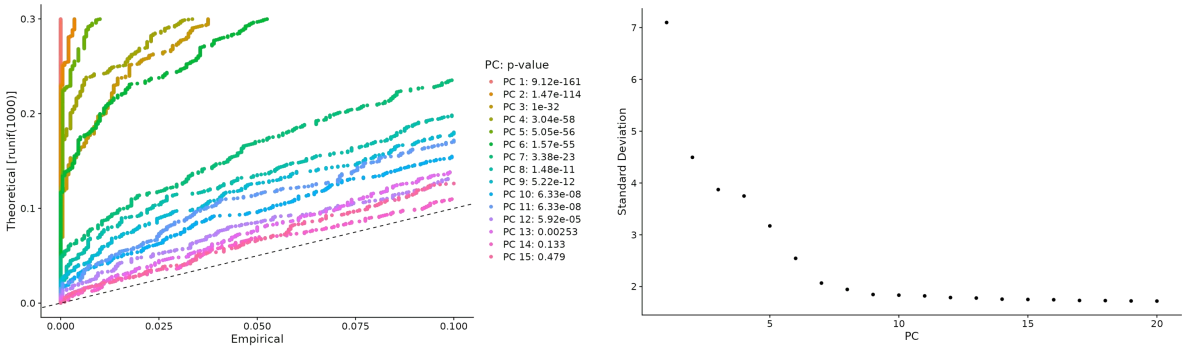


图 3-4 QQ Plot 和肘部图

注：每个 PC 的 p 值分布与统一分布（虚线）进行比较。“重要”PC 将显示在虚线上方。另一种方法生成“肘部图”：根据每个（函数）解释的方差百分比对原则组件进行排名。在此示例中，我们可以观察到 PC 9-10 周围的“肘部”，表明大多数真实信号在前 10 个 PC 中被捕获。

3.5 细胞聚类及可视化

Seurat 采用基于图形的聚类方法，简言之，这些方法将细胞嵌入到图形结构中，例如 K 最近的邻（KNN）图，在具有类似特征表达模式的单元之间绘制边缘，然后尝试将此图划分为高度互连的“集团”或“社区”。

与表象一样，我们首先根据 PCA 空间中的欧几里德距离构建 KNN 图，并根据当地社区的共享重叠（Jaccard 相似性）优化任意两个细胞之间的边缘权重。接下来将 Louvain 算法（默认值）或 SLM 等模块化优化技术应用于迭代组细胞，以优化标准模块化功能。该函数实现此过程，并包含一个分辨率参数，该参数设置下游聚类的“数量”，增加值导致更多群集。我们发现，将此参数设置在 0.4-1.2 之间通常会为大约 3K 细胞的单细胞数据集提供良好的结果。对于较大的数据集，最佳分辨率通常会增加。

Seurat 提供了多种非线性降维技术，如 tSNE 和 UMAP，以可视化和探索这些数据集。

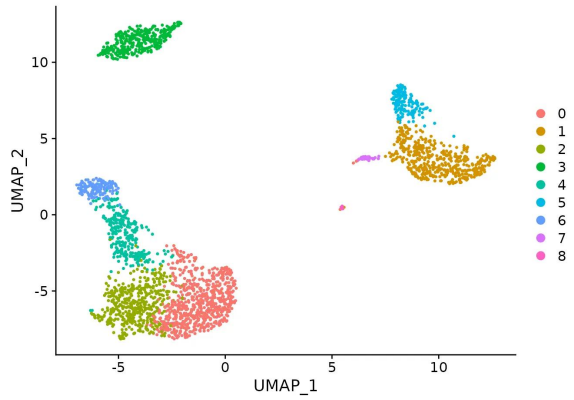


图 3-5 UMAP 降维散点图

注：9 个簇（Cluster 0-8），对应 9 种细胞类型；每个点代表一个单细胞。UMAP_1 和 UMAP_2 代表细胞类型分离维度。

3.6 查找不同表达的 marker

默认情况下，FindAllMarkers()与所有其他细胞相比，可识别单个群集的 marker。

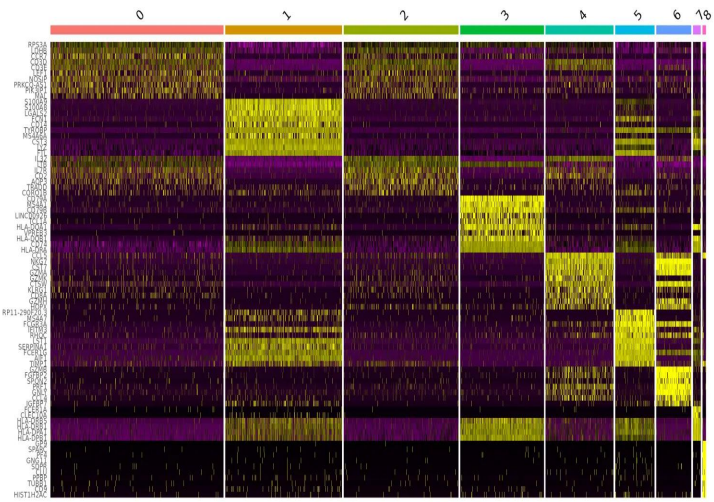


图 3-6 单个群集的 marker 热图

注：每一列的 cluster 代表了不同的细胞类群，每一行代表不同的基因。这里基因是取每个 cluster 的 top10 进行展示（可以进行调整）。

3.7 Dotplot

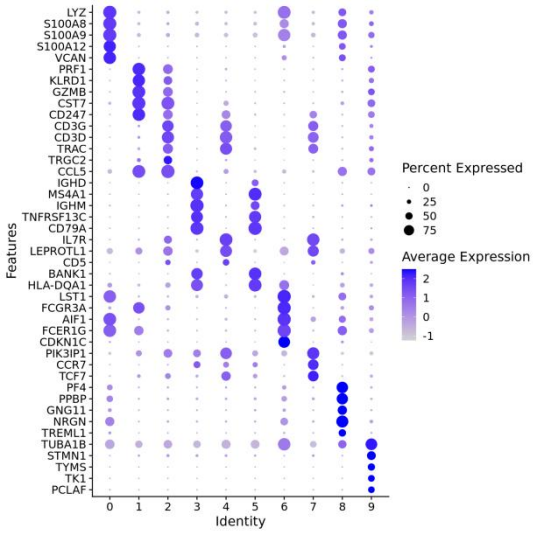


图 3-7 dotplot 图

注：点的大小代表在某个细胞类型中表达该基因的细胞比例，点的颜色代表该基因在表达细胞中的平均表达水平。

3.8 细胞注释

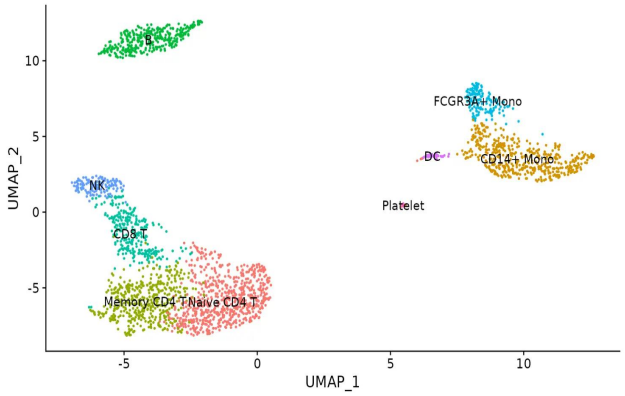


图 3-8 细胞注释后的 UMAP 图

注：通过 SingleR 包、cellmaker2.0 或者是根据已知的细胞标记基因和参考数据库，对聚类得到的细胞簇进行注释，确定细胞类型，替换之前的 Cluster 名称。

（以上为必选）

3.9 差异分析

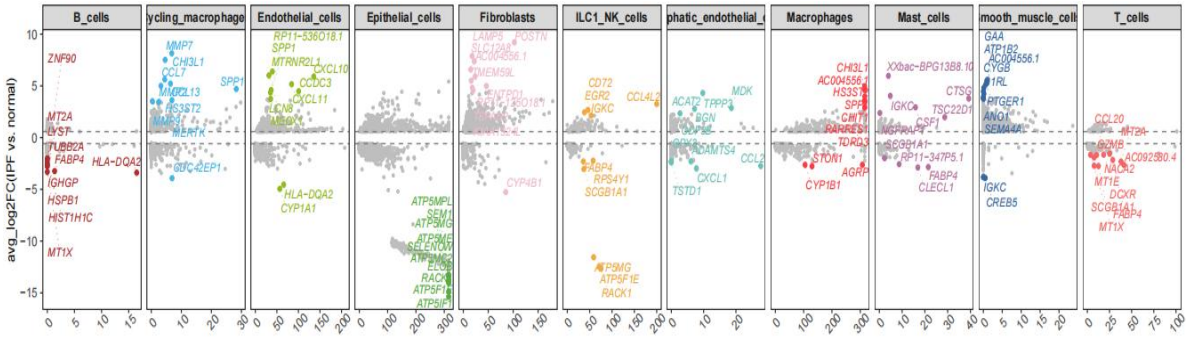


图 3-9 多组火山图

3.10 富集分析

出图同转录组。

3.11 拟时序分析

3.11.1 细胞轨迹分析

拟时序分析（Pseudotime Analysis），又称细胞轨迹分析（Cell Trajectory Analysis），是一种在单细胞转录组学中常用的方法，通过对细胞在特定生物过程中的转录组数据进行排序，推断其发展或变化的时间顺序。尽管单细胞数据通常是静态的（在一个时间点采集），拟时

序分析通过基因表达的变化模式来重建细胞的发育轨迹或动态过程，从而模拟细胞随时间推移的变化，揭示细胞在不同状态下的演变路径和关键转折点。

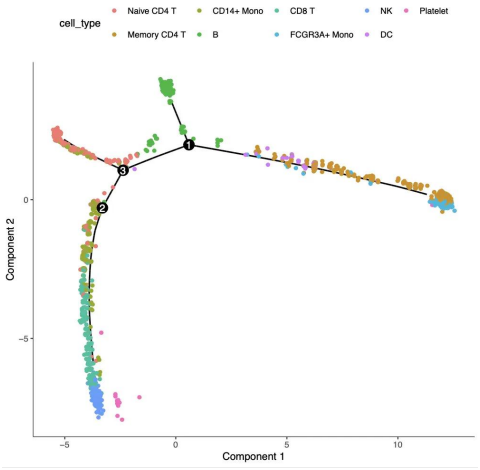
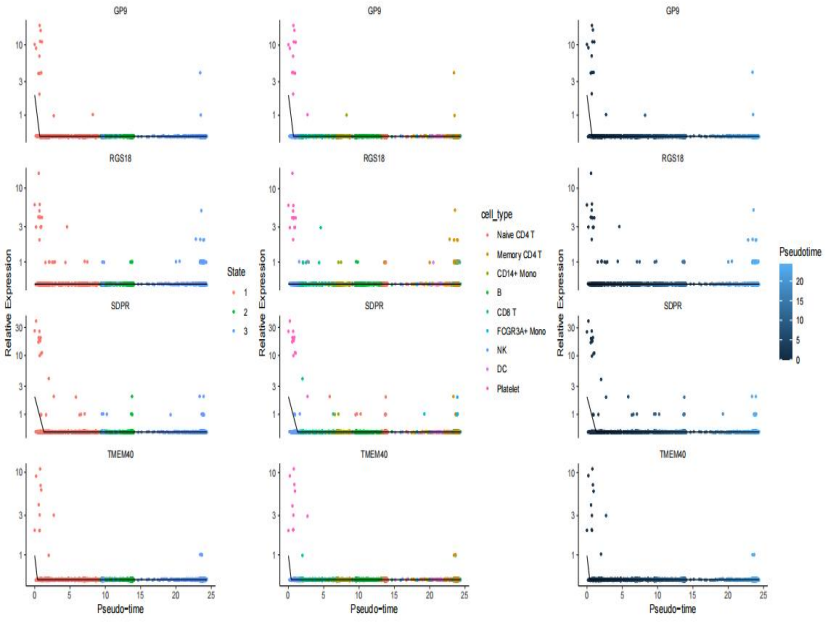


图 3-9 细胞轨迹图

注：图中 123 是分叉点，不代表特别意义，也不代表时间先后。这张图按照 sudotime，时间先后是从左往右，左边是起点(root)。起点的设置可以使用 orderCells 的 root_state 参数，将右侧设置为起点。

3.11.2 基因拟时序点图



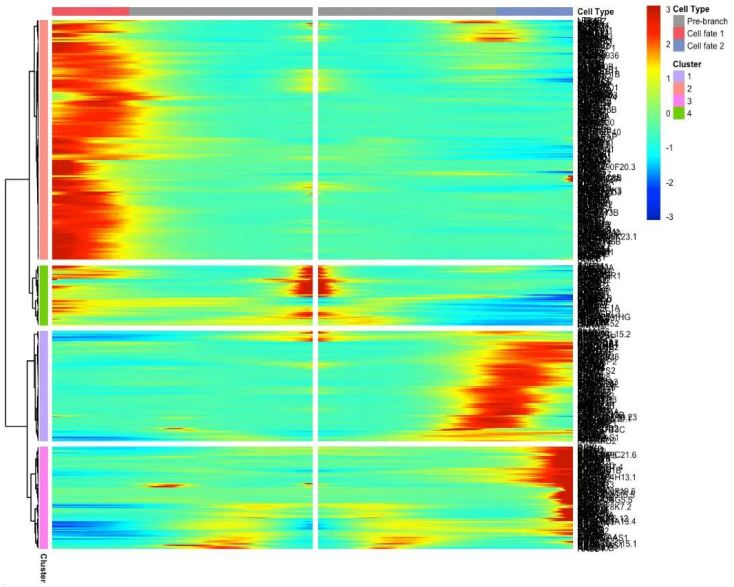
注：横轴表示拟时序(Pseudotime)，反映了细胞在发育或状态转换中的相对顺序。数值越大，表示细胞处于发育或分化的更晚阶段。

纵轴表示基因的相对表达水平。采用对数尺度(Log-scale)，更好地展示基因表达的动态范围（等同于 logFC），尤其是低表达和高表达的差异。不同的基因有不同的表达水平范围，从 1 到 100 或更高，显示了基因在不同阶段的活跃程度。

点的颜色代表不同的细胞类型。通过点的颜色，可以区分不同细胞类型中基因的表达情况，观察基因在不同细胞类型中的变化模式。

黑色曲线表示基因表达随拟时序变化的平滑趋势线，展示基因在细胞发育过程中的整体表达趋势。

3.11.3 BEAM 进行统计分析



注：该热图显示的是同一时间点两个谱系的变化，热图的列是伪时间的点，行是基因。这张图最上面的条条，灰色的代表分叉前，左边红色代表左边这个 cell fate，右边蓝色代表右边这个 cell fate，从热图中间往右读，是伪时间的一个谱系，往左是另一个谱系。

3.12 细胞通讯分析

3.12.1 细胞通讯分析

细胞通讯分析（Cell Communication Analysis）是一种用于研究不同细胞类型之间通过细胞间信号传导和分子互作进行通讯的生物学方法。该分析帮助理解细胞如何通过分泌的信号分子（如细胞因子、趋化因子、配体-受体对等）来协调其功能、调控免疫反应、维持组织稳态或参与疾病过程。

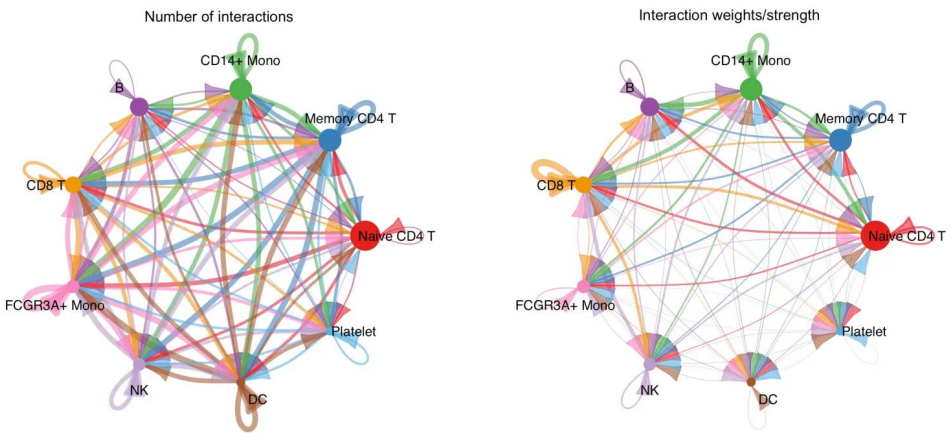


图 3-10-1 细胞互作关系

注：左图：外周各种颜色圆圈的大小表示细胞的数量，圈越大，细胞数越多。发出箭头的细胞表达配体，箭头指向的细胞表达受体。配体-受体对越多，线越粗。右图：互作的概率/强度值（强度就是概率值相加）。

3.12.2 每种细胞发出的信号

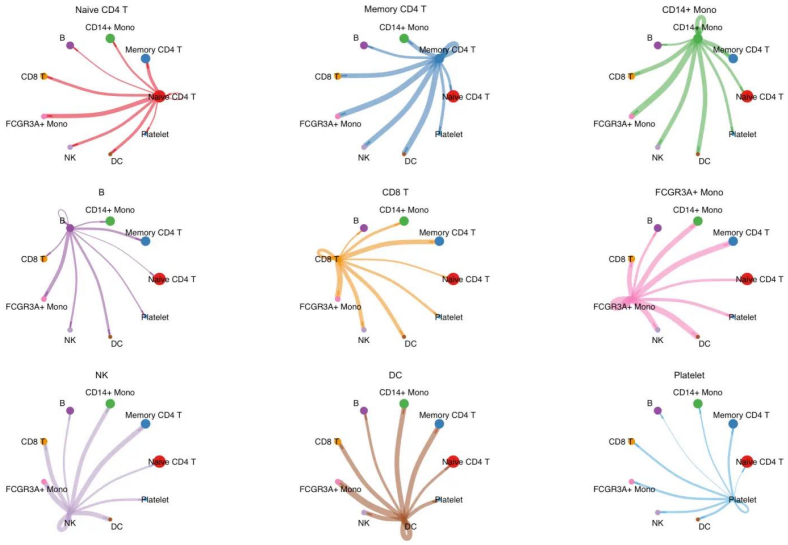


图 3-10-2 number of interaction 图

注：展示每个细胞如何跟别的细胞互作。

3.12.3 单个信号通路或配体-受体介导的细胞互作可视化（层次图、网络图、和弦图、热图）

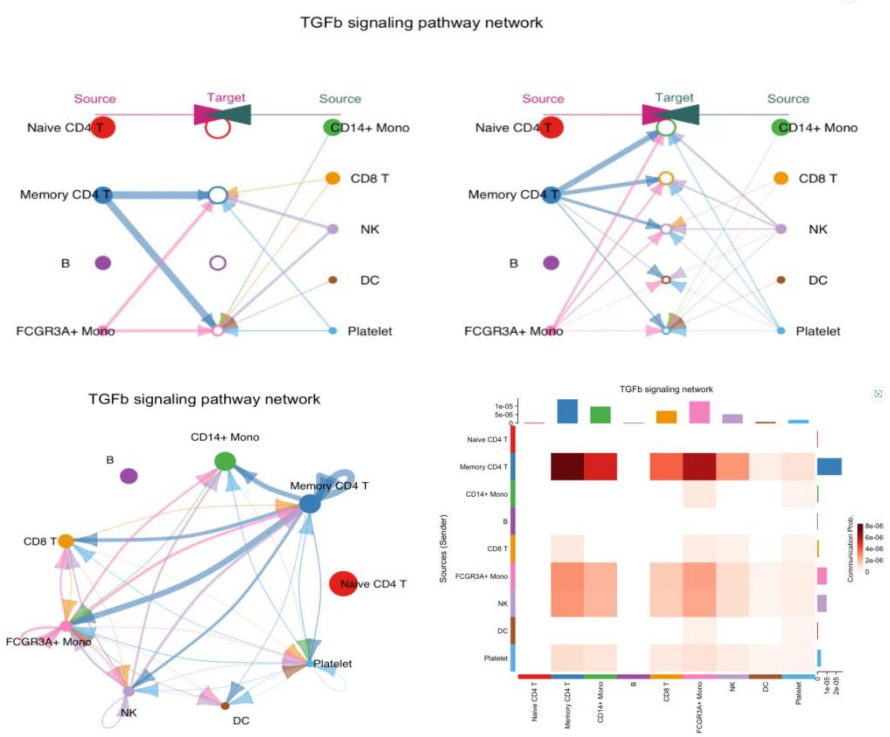


图 3-11-3 层次图和热图

注：层次图：在层次图中，实体圆和空心圆分别表示源和目标。圆的大小与每个细胞组的细胞数成比例。线越粗，互作信号越强。左侧中间的 target 是我们选定的靶细胞，右图是选中的靶细胞之外的另外一组放在中间看互作（vertex.receiver = c(1,2,4,6) 定义一个数字向量（淋系细胞），将细胞类型的索引作为目标）。两图的实心为我们定义的细胞类型（即定义的 vertex.receiver）的自分泌和旁分泌信号。

热图：单个信号通路的细胞互作热图。纵轴是发出信号的细胞，横轴是接收信号的细胞，热图颜色深浅代表信号强度。上侧和右侧的柱子是纵轴和横轴强度的累积。

3.12.4 配体-受体层级的可视化（计算各个 ligand-receptor pair 对信号通路的贡献）

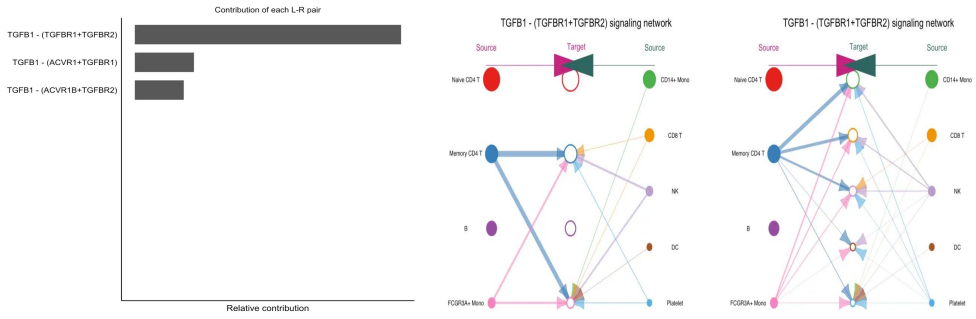
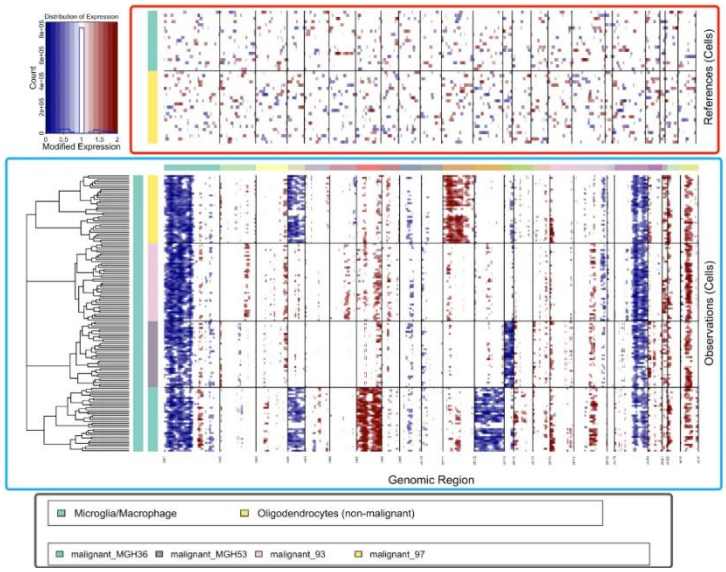


图 3-11-4 信号通路条形图和层次图

注：左图计算配体受体对选定信号通路的贡献值（在这里就是查看哪条信号通路对 TGFb 贡献最大）；右图为层次图：提取对这个通路贡献最大的配体受体对来展示（也可以选择其他的配体受体对）。

3.13 inferCNV 分析

InferCNV 用于分析肿瘤单细胞 RNA 测序数据，通过比较肿瘤细胞与参考"正常"细胞的基因表达强度，识别体细胞大规模染色体拷贝数变异的证据（如整条染色体或大片段染色体的扩增或缺失）。该方法可生成展示肿瘤基因组各染色体区域相对表达强度的热图，相较于正常细胞，肿瘤基因组中过度富集或缺失的区域通常会直观显现。



注：图中的红色部分是 References(Cells)，也就是我们自行定义的对照(正常)细胞。这些对照细胞可以是相对于肿瘤细胞的正常细胞(癌与非癌)，也可选用免疫细胞进行对照。蓝色部分是 Observations(Cells),也就是我们想要重点分析的细胞。灰色部分是细胞的图注，上边的色块代表了对照细胞的情况，该图研究者纳入了 Microglia/Macrophage 和 Oligodendrocytes (non-malignant) 作为对照细胞，恶性细胞作为观察细胞。

网络药理学

一、项目基本信息

样品与数据信息	内容选项
药物名称	中文名称：_____ 拉丁名/英文名：_____
药物来源	请注明：_____
疾病基本信息	疾病名称：中英文名称
蛋白名称	请注明：_____
分析需求	参考以下内容，选择需要的分析内容

二、生物信息分析流程

2.1 分析流程

网络药理学（Network Pharmacology）是一种基于系统生物学和计算生物学的方法，旨在理解药物在复杂生物系统中的作用机制。与传统药理学通常关注单一靶点不同，网络药理学将药物、靶点、疾病和生物网络中的其他元素视为一个整体。其核心思想是“多靶点、多通路”协同作用，从而研究它们之间的互作以及相互依赖关系。

2.2 分析内容

分析	结果	备注
药物成分收集	药物或中药的活性成分信息，包括化学成分、结构式等	TCMSP
靶点预测	获取已知药物-靶点相互作用数据	SwissTargetPrediction
疾病靶点筛选	收集与目标疾病相关的靶点基因	GeneCards
靶点可视化	Venn 图	Rstudio
PPI 网络互作分析	PPI 网络图	STRING+Cytoscape
KEGG 富集分析	柱状图，气泡图	clusterProfiler
分子对接与模拟	验证药物成分与核心靶点的结合能力，预测结合模式和亲和力	Autodock
对接可视化	对接全局图+细节对接图	Pymol

三、项目分析结果

3.1 药物成分收集

药物成分收集：从中药数据库（如 TCMSP、TCMID）或化合物库中获取药物或中药的活性成分信息，包括化学成分、结构式等。

3.2 靶点预测

利用工具（如 SwissTargetPrediction）预测药物成分可能作用的靶点蛋白，或从数据库（如 PubChem）获取已知药物-靶点相互作用数据。

Target	Common name	Uniprot ID	ChEMBL ID	Target Class	Probability*	Known actives (3D/2D)
Niemann-Pick C1-like protein 1	NPC1L1	Q9UHC9	CHEMBL2027	Other membrane protein		13 / 13
Cytochrome P450 2C19	CYP2C19	P33261	CHEMBL3622	Cytochrome P450		0 / 3
Androgen Receptor	AR	P10275	CHEMBL1871	Nuclear receptor		5 / 106
Norepinephrine transporter	SLC6A2	P23975	CHEMBL222	Electrochemical transporter		1 / 2
Nuclear receptor ROR-gamma	RORC	P51449	CHEMBL1741186	Nuclear receptor		12 / 9
LXR-alpha	NR1H3	Q13133	CHEMBL2808	Nuclear receptor		17 / 20
Acetylcholinesterase	ACHE	P22303	CHEMBL220	Hydrolase		2 / 1
HMG-CoA reductase	HMGCR	P04035	CHEMBL402	Oxidoreductase		18 / 22
Cytochrome P450 51 *(by homology)*	CYP51A1	Q16850	CHEMBL3849	Cytochrome P450		3 / 2
Butyrylcholinesterase	BCHE	P06276	CHEMBL1914	Hydrolase		4 / 2
Protein-tyrosine phosphatase 1B	PTPN1	P18031	CHEMBL335	Phosphatase		9 / 52
Testis-specific androgen-binding protein	SHBG	P04278	CHEMBL3305	Secreted protein		0 / 48
Muscarinic acetylcholine receptor M2	CHRM2	P08172	CHEMBL211	Family A G protein-coupled receptor		1 / 2
Serotonin transporter	SLC6A4	P31645	CHEMBL228	Electrochemical transporter		4 / 5
Cytochrome P450 17A1	CYP17A1	P05093	CHEMBL3522	Cytochrome P450		4 / 51

图 3-1 获取药物靶点结果

3.3 疾病靶点筛选

通过疾病数据库（GeneCards）收集与目标疾病相关的靶点基因，筛选出与疾病发生发展密切相关的基因集合。

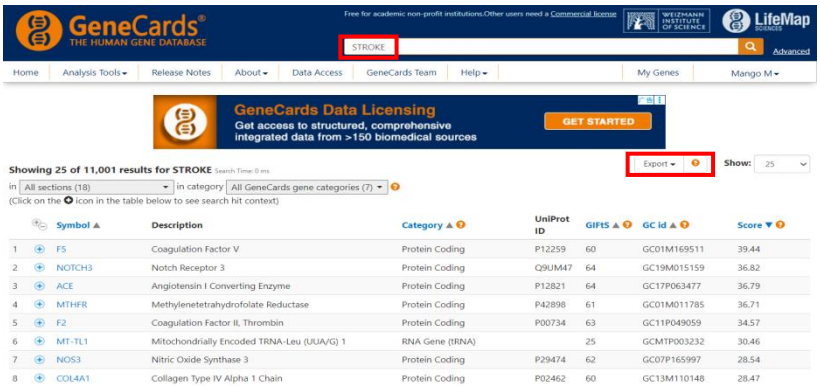


图 3-2 获取疾病靶点结果

3.4 靶点可视化

将药物靶点与疾病靶点取交集，确定药物可能作用于疾病的关键靶点。

图 3-3 Venn 图

3.5 PPI 网络互作分析

利用 STRING 等数据库构建交集靶点的 PPI 网络，分析靶点之间的相互作用关系，识别网络中的关键节点（如 hub 基因）。使用 Cytoscape 软件及插件（如 CytoHubba、MCODE）对网络进行拓扑分析，计算节点的度中心性、介数中心性、接近中心性等指标，筛选出核心靶点。

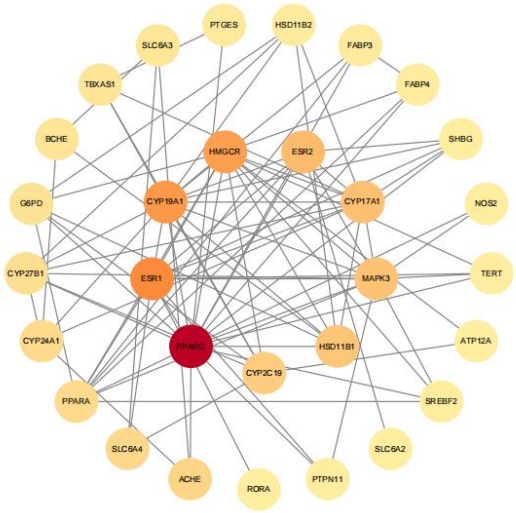


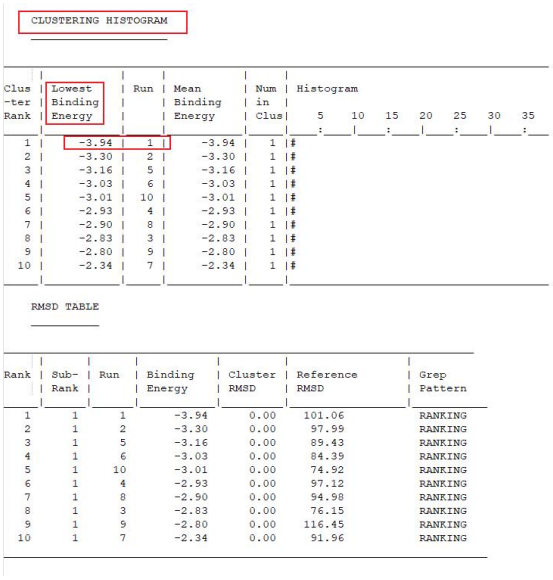
图 3-4 PPI 网络互作图

3.6 KEGG 富集分析

通过 clusterProfiler 等工具对核心靶点进行基因本体（GO）和京都基因与基因组百科全书（KEGG）富集分析，揭示靶点参与的生物过程、细胞组分、分子功能及信号通路，解释药物作用机制（图同转录组）。

3.7 分子对接与模拟

利用 Autodock Vina 等分子对接软件，验证药物成分与核心靶点的结合能力，预测结合模式和亲和力；通过分子动力学模拟进一步分析结合稳定性。



CLUSTERING HISTOGRAM

Cluster Rank	Lowest Binding Energy	Run	Mean Binding Energy	Num in Clus	Histogram 5 10 15 20 25 30 35
1	-3.94	1	-3.94	1	#
2	-3.30	2	-3.30	1	#
3	-3.16	5	-3.16	1	#
4	-3.03	6	-3.03	1	#
5	-3.01	10	-3.01	1	#
6	-2.93	4	-2.93	1	#
7	-2.90	8	-2.90	1	#
8	-2.83	3	-2.83	1	#
9	-2.80	9	-2.80	1	#
10	-2.34	7	-2.34	1	#

RMSD TABLE

Rank	Sub-Rank	Run	Binding Energy	Cluster RMSD	Reference RMSD	Grep Pattern
1	1	1	-3.94	0.00	101.06	RANKING
2	1	2	-3.30	0.00	97.99	RANKING
3	1	5	-3.16	0.00	89.43	RANKING
4	1	6	-3.03	0.00	84.39	RANKING
5	1	10	-3.01	0.00	74.92	RANKING
6	1	4	-2.93	0.00	97.12	RANKING
7	1	8	-2.90	0.00	94.98	RANKING
8	1	3	-2.83	0.00	76.15	RANKING
9	1	9	-2.80	0.00	116.45	RANKING
10	1	7	-2.34	0.00	91.96	RANKING

图 3-5 对接结果

3.8 对接可视化

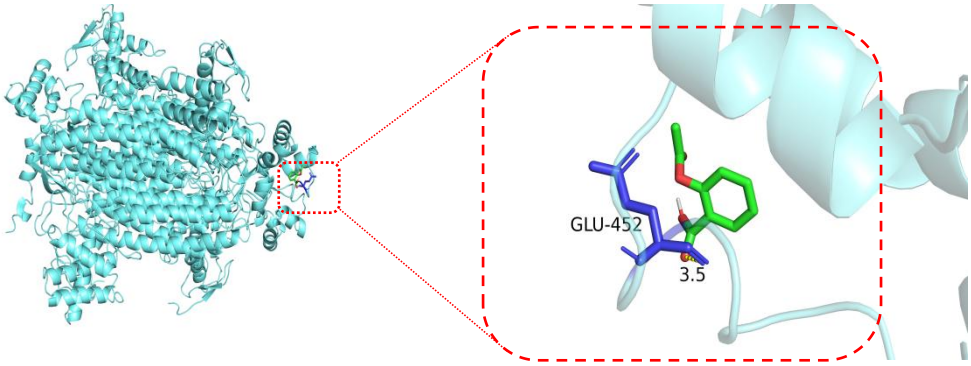


图 3-6 对接美化结果

注：左侧区域展示了分子对接的整体结构，右侧区域为蛋白的结合口袋被放大以显示细节，重点突出了配体与蛋白质局部氨基酸残基的相互作用。蓝色：氨基酸残基（名称 GLU-452）；黄色：氢键（长度 3.5）。