# 医口宏基因组-宏病毒组-代谢组关联分析内容介绍

1. **基本介绍**

无需在结题报告中体现

* 1. 本流程针对医口项目（最好是疾病-健康比较的这种），是之前医口宏病毒组项目的后续延伸
  2. 本流程可以进行细菌、病毒、功能基因和代谢物4者的关联分析
  3. 细菌和病毒的数据来自于医口宏病毒组的Pathseq注释丰度表
  4. 功能基因的数据来自于宏基因组测序的KEGG注释丰度表

医口宏病毒组也是使用宏基因组的测序方式，因此只需要在进行宏病毒组测序的同时增加一个宏基因组标准分析，或者在进行宏基因组测序的同时增加一个宏病毒组分析即可

* 1. 代谢组的数据来自于非靶向代谢组的代谢物识别丰度表
  2. 细菌与病毒的关联分析在医口宏病毒组项目中已经包含了
  3. 如果客户没有配套进行代谢组测试，本流程也可以单独完成细菌-病毒-功能基因的关联分析
  4. 所有分析均使用R v4.0.2完成

1. **内容设置**
   1. **单独分析内容**
      1. 分别对细菌、病毒、功能基因和代谢物数据先进行单独分析，结果分别在Bacteria、Virus、Gene和Metabolites文件夹中
      2. 4类分析共有分析内容包括beta多样性分析（Beta\_diversity）、差异分析(Diff\_analysis)和诊断分析(Diagnose)
      3. 针对细菌、病毒和功能基因还同时提供了组成分析(Community)

代谢组数据并不存在组成分析的内容，对于代谢组的分类、数量统计等等在非靶向代谢组的测序结果中有

* + 1. 对于细菌和病毒还同时提供了alpha多样性分析（Alpha\_diversity）

功能基因和代谢物没有alpha多样性的概念

* 1. **关联分析**
     1. 关联分析包括细菌、病毒和功能基因三者的关联分析（Association\_analysis\_virome）和细菌、病毒、功能基因与代谢物4者的关联分析（Association\_analysis\_metabolites）
     2. 分析内容包括组成结构关联（Mantel）、丰度关联网络（Network）和丰度相关性热图（Heatmap）

1. **结果解释**
   1. **单独分析内容**
      1. **Alpha多样性分析**

结果目录Alpha\_diversity，只在Bacteria和Virus结果中有

* + - 1. alpha\_diversity\_indices.txt：alpha多样性指数表格，包括Observed\_species、Chao1、ACE、Shannon、Simpson、Pielou\_J和Good coverage共7种alpha指数

vegan包计算

* + - 1. alpha\_div文件夹，alpha多样性指数组间差异检验结果，使用t-test检验两组差异，每个箱子上不同的小写字母代表组间差异显著，7个alpha指数分别进行计算，结果为指数名对应的pdf文件

ggplot2包完成

* + - 1. Venn文件夹，根据样品分组进行的共有和特有分析
         1. 使用VennDiagram包和ggplot2包完成
         2. Venn.png/pdf，组建种水平Venn图，在组内任意一个样品中有检出即认为该物种在该组中存在
         3. Shared\_species.txt，组间共有物种列表
         4. Shared\_sepecies\_abundance\_diff.pdf，共有物种总丰度组间差异检验结果，使用t-test检验两组差异，Tukey’s HSD test检验多组差异，每个箱子上不同的小写字母代表组间差异显著，7个alpha指数分别进行计算，结果为指数名对应的pdf文件
         5. Shared\_sepecies\_total\_abundance.pdf，共有物种总丰度样本内占比条形图
      2. Flower文件夹，根据单独样本进行的共有和特有分析
         1. ggplot2包完成
         2. Flower\_diagram.pdf，样本间共有物种花瓣图，共有物种表明在所有单个样本中均有检出

⚠️有可能存在共有物种结果为0的情况

* + - * 1. Shared\_species.txt，样本间共有物种列表
        2. Shared\_sepecies\_abundance\_diff.pdf，共有物种总丰度组间差异检验结果，使用t-test检验两组差异，Tukey’s HSD test检验多组差异，每个箱子上不同的小写字母代表组间差异显著，7个alpha指数分别进行计算，结果为指数名对应的pdf文件
        3. Shared\_sepecies\_total\_abundance.pdf，共有物种总丰度样本内占比条形图
    1. **Beta多样性分析**

结果目录Beta\_diversity

* + - 1. Distance文件夹内为根据样本种水平物种丰度计算得到的样本间bray-curtis距离结果
      2. PCA、PCoA和NMDS文件夹为样本基于bray-curtis距离的PCA、PCoA和NMDS分析结果

结果的解释可以参考16S扩增子结题报告和宏基因组抗性基因结题报告相关内容
PCA为ape包，PCoA和NMDS为vegan包

* + - 1. Beta\_tests文件夹内为样本基于bray-curtis距离的Adonis、ANOSIM和MRPP检验结果

结果的解释可以参考16S扩增子结题报告相关内容
vegan包

* + - 1. Hcluster文件夹为样本基于bray-curtis距离的UPGMA聚类树

结果的解释可以参考16S扩增子结题报告相关内容
vegan包

* + - 1. Hcluster\_bar文件夹为样本基于bray-curtis距离的UPGMA聚类树与物种丰度组合图

结果分为门、纲、目、科、属，结果的解释可以参考16S扩增子结题报告相关内容
Gene和Metabolites不适合此分析，因此在Gene和Metabolites文件夹中无此结果

* + - 1. PLS-DA

代谢组研究常见分析方法，只在Metabolites文件夹中提供
​mixOmics包

* + 1. **物种丰度分析**

结果目录Community，只在Bacteria、Virus和Gene结果中有

* + - 1. tax\_summary\_a文件夹为门至属水平每个样本注释得到的物种tpm丰度表格

Gene没有此文件，因为宏基因组的数据直接就给出了不同层级的tpm丰度表格

* + - 1. tax\_summary\_a\_Group文件夹为门至属水平每组样本注释得到的物种tpm丰度平均值的表格

Gene没有此文件，因为宏基因组的数据直接就给出了不同层级的tpm丰度表格

* + - 1. tax\_summary\_r文件夹为门至属水平每个样本注释得到的物种reads数目计算的百分比相对丰度表格

Gene结果为KEGG的L1、L2、L3和ko4个层级

* + - 1. tax\_summary\_r\_Group文件夹为门至属水平每组样本注释得到的物种reads数目计算的百分比相对丰度平均值的表格

Gene结果为KEGG的L1、L2、L3和ko4个层级

* + - 1. Alluvial\_chart，样本间物种丰度冲积图

分别在门、纲、目、科、属水平进行分析，选择丰度排名前10的物种进行绘图，其它物种归类为Others
Gene文件夹无此结果
ggalluvial包

* + - 1. Alluvial\_chart\_Group，组间物种丰度冲积图

分别在门、纲、目、科、属水平进行分析，选择丰度排名前10的物种进行绘图，其它物种归类为Others
Gene文件夹此结果合并到Barplot中
ggalluvial包

* + - 1. Bubble\_chart，样本间物种丰度气泡图

分别在门、纲、目、科、属水平进行分析，选择丰度排名前30的物种进行绘图
由于L1水平基因条目过少，不适合top30形式展示，Gene文件夹中只包含L2、L3和ko水平结果
ggplot2包

* + - 1. Bubble\_chart\_Group，组间物种丰度气泡图

分别在门、纲、目、科、属水平进行分析，选择丰度排名前30的物种进行绘图
Gene文件夹此结果合并到Bubble\_chart中
ggplot2包

* + - 1. Heatmap，样本间物种丰度聚类热图

分别在门、纲、目、科、属水平进行分析，选择丰度排名前30的物种进行绘图，对每个物种的丰度按行进行了标准化，结果的解释可以参考16S扩增子结题报告相关内容
由于L1水平基因条目过少，不适合top30形式展示，Gene文件夹中只包含L2、L3和ko水平结果
pheatmap包

* + - 1. Heatmap\_Group，组间物种丰度聚类热图

分别在门、纲、目、科、属水平进行分析，选择丰度排名前30的物种进行绘图，对每个物种的丰度按行进行了标准化，结果的解释可以参考16S扩增子结题报告相关内容
Gene文件夹此结果合并到Heatmap中
pheatmap包

* + - 1. Barplot，样本间物种丰度条形图，按照物种的丰度大小对样本进行重新排序并且表明了样本的分组信息，便于更直观的查看组间物种变化

分别在门、纲、目、科、属水平进行分析，选择丰度排名前10的物种进行绘图，其它物种归类为Others
由于L2、L3和ko水平基因条目数量非常多，并不适合以top10形式展示，因此，Gene文件夹中此结果只有L1水平
ggplot2包

* + 1. **物种丰度差异分析**

结果目录Diff\_analysis

* + - 1. 分别使用了wilcox rank-sum test、lefse和随机森林分析识别组间区分关键物种，结果文件夹分别为Wilcox、LEfSe和RandomForest

wilcox和lefse使用自编写代码，随机森林分析使用RandomForest包

* + - 1. 在分析之前由于很多物种、基因和代谢物只在个别样本中有检出，首先去除了在研究样本中检出率低于60%的物种、基因和代谢物，并以0.1%为阈值过滤到稀有物种和基因

代谢物丰度数据与物种和基因不同，不进行丰度过滤

* + - 1. 结果的解释参考新版的多样性+代谢组关联分析的结题报告
      2. ⚠️wilcox rank-sum test和随机森林分析分别在门、纲、目、科、属水平进行分析，结果文件夹分别为Phylum、Class、Order、Family、Genus
      3. ⚠️对于Gene，wilcox rank-sum test和随机森林分析分别在L1、L2、L3和ko水平进行分析
      4. ⚠️对于细菌和病毒，lefse分析是使用所有分类学水平的数据进行分析，因此只有一个结果
      5. ⚠️对于Gene，lefse只在ko水平进行分析
      6. ⚠️对于Metabolites，只有一个丰度水平，因此所有差异检验只有一个结果
      7. ⚠️可能有一些分类学水平结果是空白的文件，这是因为在该水平的数据可能组间并未发现任何显著差异的结果
    1. **疾病诊断分析**

结果目录Diagnose

* + - 1. 与物种丰度差异分析结果相对应，疾病诊断分析同样分别基于3种差异检验的结果进行，结果文件夹分别为Wilcox、lefse和RandomForest
      2. ⚠️wilcox rank-sum test和随机森林分析分别在门、纲、目、科、属水平进行分析，结果文件夹分别为Phylum、Class、Order、Family、Genus
      3. ⚠️lefse分析是使用所有分类学水平的数据进行分析，因此只有一个结果，对细菌和病毒的lefse分析结果，只使用属水平差异物种进行诊断分析
      4. ⚠️对于在差异检验部分并未找到任何差异结果的情况，此部分分析将不会有对应的结果文件夹
      5. 每个文件夹内的结果文件包括3个：
         1. index.txt为根据差异物种计算的疾病诊断指数，计算方法为疾病组富集指示物种丰度减去健康组富集指示物种丰度
         2. index.pdf为组间疾病诊断指数差异检验，使用t-test检验组间差异，每个箱子上不同的小写字母代表组间差异显著

ggplot2包

* + - * 1. ROC.pdf为基于疾病诊断指数绘制的ROC曲线用于评估指数的诊断效果

pROC包

ROC 曲线指受试者工作特征曲线(receiver operating characteristic curve), 是反映敏感性和特异性连续变量的综合指标，通过构图法揭示敏感性和特异性的相互关系。ROC 曲线将连续变量设定出多个不同的临界值，从而计算出一系列敏感性和特异性，再以敏感性为纵坐标、（1-特异性）为横坐标绘制成曲线，曲线下面积越大，诊断准确性越高。

说明：ROC 曲线下的面积值在 1.0 和0.5 之间。在 AUC>0.5 的情况下，AUC 越接近于 1，说明诊断效果越好。AUC在0.5~0.7 时有较低准确性，AUC 在 0.7~0.9 时有一定准确性，AUC 在 0.9 以上时有较高准确性。AUC=0.5时，说明诊断方法完全不起作用，无诊断价值。AUC<0.5 不符合真实情况，在实际中极少出现。在 ROC 曲线上，最靠近坐标图左上方的点为敏感性和特异性均较高的临界值。

* 1. **关联分析**
     1. **细菌、病毒、功能基因关联分析**

结果目录Association\_analysis\_virome

* + - 1. **组成相关性分析**

结果目录Mantel

* + - * 1. 分别针对细菌、病毒和功能基因，使用Mantel test和普氏分析评估这3者两两之间的相关性
        2. 结果解释参考宏基因组抗性基因Mantel分析部分内容
        3. vegan和ggplot2包
      1. 差异物种、基因相关性分析

结果目录Heatmap

* + - * 1. 分别针对细菌、病毒和功能基因分析两两之间差异条目的相关性，分析使用Spearman相关性分析，结果使用热图展示

psych和pheatmap包

* + - * 1. 与差异分析相对应，相关性分析同样分别基于3种差异分析方法的结果进行，结果文件夹分别为Wilcox、lefse和RandomForest，如果缺少某个结果文件夹，代表用于相关性分析的两个数据至少有一个在该差异分析种没有结果
        2. 对于wilcox rank-sum test和随机森林分析，针对细菌和病毒关联分析，分别在门、纲、目、科、属水平进行分析，结果文件夹分别为Phylum、Class、Order、Family、Genus和Species
        3. 对于wilcox rank-sum test和随机森林分析，针对细菌与功能基因或病毒与功能基因，在门、纲、目、科、属水平分别与L1、L2、L3和ko水平的结果进行分析，结果文件夹分别为L1、L2、L3和ko
        4. 对于lefse分析，只有属水平物种和ko水平功能基因的关联分析
        5. 同时还针对3种差异分析方法的共有差异条目进行了相关性分析，结果目录为Shared，由于lefse只针对属水平物种和ko水平基因，因此，Shared目录结果种只有属水平物种和ko水平基因的差异条目相关性
        6. 每个结果文件夹包含3个文件：

correlation.xls：相关性分析R值结果表

pvalue.xls：相关性分析p值结果表

heatmap.pdf：相关性分析热图，颜色代表相关性R值，红色正相关，蓝色负相关，+代表p< 0.05，++代表p<0.01

* + - 1. **关联网络分析**

结果目录Network

* + - * 1. 首先对物种和功能基因进行过滤，去除检出率低于60%，丰度低于0.1%的条目
        2. 使用Spearman相关性在种水平分析物种与功能基因两两之间的丰度相关性，保留相关系数大于0.8，并且p小于0.05的结果，使用ggraph包绘制网络图

相关性分析使用linkET包

* + - * 1. 分别在3个层次进行关联网络分析，物种门水平与L2水平功能基因（Phylum-L2），物种属水平与L3水平功能基因（Genus-L3）和物种属水平与ko水平功能基因（Genus-ko）
        2. 每个文件夹包含5个结果文件：

network\_abun.pdf：节点大小为物种或基因丰度的网络图，并且包含节点名称标签

network\_abun\_nolabel.pdf：节点大小为物种或基因丰度的网络图，不包含节点名称标签

network\_degree.pdf：节点大小为物种或基因degree的网络图，并且包含节点名称标签

network\_degree\_nolabel.pdf：节点大小为物种或基因degree的网络图，不包含节点名称标签

edge.csv：网络分析边文件，可以输入Gephi或Cytoscape进行后续分析

* + 1. **细菌、病毒、功能基因、代谢物关联分析**

结果目录：Association\_analysis\_metabolites

* + - 1. **组成相关性分析**

结果目录Mantel

* + - * 1. 分别针对细菌、病毒和功能基因，使用Mantel test和普氏分析评估这3者与代谢组数据的相关性
        2. 结果解释参考宏基因组抗性基因Mantel分析部分内容
        3. vegan和ggolot2包
      1. **差异物种、基因与代谢物丰度相关性分析**

结果目录Heatmap

* + - * 1. 使用Spearman相关性分析，分别评估差异细菌、病毒和功能基因与差异代谢物丰度的相关性，结果使用热图展示，结果目录分别为Bacteria\_DAM、Virus\_DAM和Function\_DAM

psych和pheatmap包

* + - * 1. 与差异分析相对应，相关性分析同样分别基于3种差异分析方法的结果进行，结果文件夹分别为Wilcox、lefse和RandomForest，如果缺少某个结果文件夹，代表用于相关性分析的两个数据至少有一个在该差异分析种没有结果
        2. 对于wilcox rank-sum test和随机森林分析，针对细菌和病毒与差异代谢物的关联分析，分别在门、纲、目、科、属水平进行分析，结果文件夹分别为Phylum、Class、Order、Family、Genus和Species
        3. 对于wilcox rank-sum test和随机森林分析，针对功能基因与差异代谢物的关联分析，分别在L1、L2、L3和ko水平的结果进行分析，结果文件夹分别为L1、L2、L3和ko
        4. 对于lefse分析，只有属水平物种和ko水平功能基因与差异代谢物的关联分析
        5. 同时还针对3种差异分析方法的共有差异条目进行了相关性分析，结果目录为Shared，由于lefse只针对属水平物种和ko水平基因，因此，Shared目录结果种只有属水平物种和ko水平基因与差异代谢物的相关性分析
        6. 每个结果文件夹包含3个文件：

correlation.xls：相关性分析R值结果表

pvalue.xls：相关性分析p值结果表

heatmap.pdf：相关性分析热图，颜色代表相关性R值，红色正相关，蓝色负相关，+代表p< 0.05，++代表p<0.01

* + - 1. **关联网络分析**

结果目录Network

* + - * 1. 首先对物种和功能基因进行过滤，去除检出率低于60%，丰度低于0.1%的条目
        2. 使用Spearman相关性在种水平分析物种、功能基因与差异代谢物两两之间的丰度相关性，保留相关系数大于0.8，并且p小于0.05的结果，使用ggraph包绘制网络图

相关性分析使用linkET包

* + - * 1. 分别在3个层次进行关联网络分析，物种门水平与L2水平功能基因（Phylum-L2-DAM），物种属水平与L3水平功能基因（Genus-L3-DAM）和物种属水平与ko水平功能基因（Genus-ko-DAM）
        2. 由于差异分析有3种方法，因此，每个网络分析分别针对3种差异分析识别到的差异代谢物进行，结果目录分别为Wilcox、lefse和RandomForest
        3. 每个文件夹包含3个结果文件：

由于代谢物与物种和功能基因丰度不在一个尺度范围内，因此无法提供节点大小为条目丰度的结果图

network\_degree.pdf：节点大小为物种或基因degree的网络图，并且包含节点名称标签

network\_degree\_nolabel.pdf：节点大小为物种或基因degree的网络图，不包含节点名称标签

edge.csv：网络分析边文件，可以输入Gephi或Cytoscape进行后续分析