诺禾致源
宏基因组交付目录说明手册
（V4.3）

诺禾致源微生物事业部

2017 年 11 月 13 日

目录

（注：单击即可跳转至相应文档的详细说明）

|-- 04.TaxAnnotation ——【物种注释及丰度分析结果】

| |-- MAT ——【物种注释统计矩阵】

| | |-- Absolute ——【绝对丰度矩阵：基于样品间基因数目均一化以后得到的绝对丰度矩阵】

| | | `-- Unigenes.absolute.{k,p,c,o,f,g,s}.xls——【各分类水平（界门纲目科属种）上物种绝对丰度矩阵】

这些表格都可以用 excel 打开，在这些表格中，第一行为样品名称，第一列为在某个水平上的物种信息，其余各列代表在相应物种上，各样品的相对丰度情况，最后一列为该物种的详细描述，包括其所属的详细分类层级的信息。值得注意的是，我们可以看到，在第一列中，有很多重复的 Unclassified 的物种，而从最后一列中我们可以看到，这些物种归属于不同的分类层级。因此，在这里我们会分开将其进行展示。

| | |-- Relative ——【相对丰度矩阵：基于绝对丰度矩阵得到的相对丰度矩阵】

| | `-- Unigenes.relative.{k,p,c,o,f,g,s}.xls ——【各分类水平（界门纲目科属种）上物种相对丰度矩阵】

这些表格都可以用 excel 打开，在这些表格中，第一行为样品名称，第一列为在某个水平上的物种信息，其余各列代表在相应物

种上，各样品的相对丰度情况，最后一列为该物种的详细描述，包括其所属的详细分类层级的信息。值得注意的是，我们可以看到，在第一列中，有很多重复的 Unclassified 的物种，而从最后一列中我们可以看到，这些物种归属于不同的分类层级。因此，在这里我们会分开将其进行展示。

| |-- GeneNums ——【物种注释基因数目统计】

| | |-- GeneNums.total ——【所有样品注释基因数目统计】

| | | `-- Unigenes. absolute.{k,p,c,o,f,g,s}.xls——【各分类水平（界门纲目科属种）所有样品注释基因数目矩阵】

这些表格都可以用 excel 打开，在这些表格中，第一列为在某个水平上的物种信息及其上一层级的信息，第三列为注释到该物种水平上的基因数目，第四列为这些基因的 id 号。

| | |-- GeneNums.BetweenSamples——【各样品间注释基因数目统计】

| | | `-- Unigenes.absolute.{k,p,c,o,f,g,s}.xls——【各分类水平（界门纲目科属种）注释基因数目矩阵】

这些表格都可以用 excel 打开，在这些表格中，第一行为样品名称，第一列为在某个水平上的物种信息及其上一层级的信息，其余各列代表在相应物种上个样品注释到的基因数目情况。

| |-- GeneNums.BetweenSamples.heatmap—— 【基于个样品间个层级上的基因数目的热图分析】

| | | `-- {k,p,c,o,f,g,s}.genenum.heatmap.txt.{png,pdf} ——【各样品间各分类水平基于基因数目的 heatmap 热图 pdf 和 png 格式】

储存了在界门纲目科属种水平上的基于基因数目的物种丰度聚类图，在每张图中,横向为样品信息,纵向为物种注释信息，图中左侧的聚类树为物种聚类树；上方的聚类树为样品聚类树;中间热图对应的值为每一行物种相对丰度经过标准化处理后得到的 Z 值，即一个样品在某个分类上的 Z 值为样品在该分类上的相对丰度和所有样品在该分类的平均相对丰度的差除以所有样品在该分类上的标准差所得到的值。

| | | `-- {k,p,c,o,f,g,s}.genenum.heatmap.txt——【各样品间各分类水平基于基因数目的 heatmap 热图分析所用文件】

储存了在界门纲目科属种水平上的基于基因数目的物种丰度聚类图作图的数据，第一行为样品名称,第一列为物种名称，其余各列为各物种在各样品中的丰度信息。

| |-- MicroNR ——【MicroNR 注释结果统计】

| | |-- Unigenes.absolute.total.tax.xls——【代表性基因在各样品间的绝对丰度矩阵以及各代表性基因的 LCA 注释结果】

将代表性基因在各个样品中的绝对丰度信息和物种注释信息相结合。第一列为代表性基因的 id，第一行为样品名，表中数字为代表性基因在各个样品中的绝对丰度信息，最后一列为 LCA 注释结果。

| | |-- Unigenes.lca.tax.detail.xls ——【各代表性基因的对应的详细的 LCA 注释结果】

代表性基因的详细物种注释结果，包含了亚门、亚纲等更详细的层级注释信息。第一列为代表性基因的 id 第二列为对应 id 经 LCA 算法后的详细注释信息。

| | |-- Unigenes.lca.tax.xls ——【各代表性基因的对应的 LCA 注释结果】

代表性基因的 LCA 注释结果，第一列为代表性基因的 id 第二列为物种注释信息（这里不包含亚层级的信息，只有界门岗目科属种）。

| | |-- **Unigenes.m8.tax.xls ——【从 blast m8 结果出发添加了 reference 对应的 tax id 及物种信息】**

该文件为将代表性基因和 NR 库比对后，生成的 blast m8 格式的文件，与普通的 m8 格式的文件不同的是，在这个文件中，另外加入了两行数据，一行为该比对结果所对应的 taxonomy id 号，另外一行为该 taxonomy id 所对应的详细的物种信息，可以用 excel 打开该文件，在该文件中，各列所代表的含义如下：

列数	说明
1	目标核酸或氨基酸序列的 ID，编号的有效字符有[a-zA-Z0-9.:^x!+_?-|]。
2	数据库序列的 ID。
3	目标核酸或氨基酸序列与数据库序列比对的 Identity 值。
4	目标核酸或氨基酸序列与数据库序列比对的长度。
5	目标核酸或氨基酸序列与数据库序列比对区域的比对错配数。
6	目标核酸或氨基酸序列与数据库序列比对区域的比对空位数。
7	目标核酸或氨基酸序列的比对起始坐标。

8	目标核酸或氨基酸序列的比对终止坐标。
9	数据库序列的比对起始坐标。
10	数据库序列的比对终止坐标。
11	目标核酸或氨基酸序列与数据库序列比对的期望值。
12	目标核酸或氨基酸序列与数据库序列比对的比对得分。
13	比对结果所对应的 taxonomy id 号
14	该 taxonomy id 所对应的详细的物种信息

| | |-- **Unigenes.screening.m8.xls——【经过过滤后的 blast m8 结果】**

经过过滤后的代表性基因和 NR 库的比对结果。

| |-- Top10/Top10_group——【各样品（组）在各水平丰度排名前 10 柱状图】

| | |-- figure ——【样品（组）在各水平丰度排名前 10 柱状图，pdf 及 png 格式】

| | |-- table ——【样品（组）在各水平丰度排名前 10 柱状图分析所使用的文件】

| | | `-- Bar.tree.{k,p,c,o,f,g,s}10.png ——【样品在各水平上的聚类分析图，pdf 及 png 格式

| |-- Cluster_Tree ——【样品聚类分析结果】

| | |-- figure ——【样品在各水平上的聚类分析图，pdf 及 png 格式】

| | | `-- Bar.tree.{k,p,c,o,f,g,s}10.png ——【样品在各水平上的聚类分析图，pdf 及 png 格式】

即为结题报告中的物种聚类分析图，图中左侧是 Bray-Curtis 距离聚类树结构，右侧的是各样品在各个分类水平(界门纲目科属种)上的物种相对丰度分布图。

| | `-- table ——【各水平上样品聚类分析所使用的文件】

储存了在界门纲目科属种水平上用于计算 Bray-Curtis 距离的相对丰度矩阵，第一行为样品名称,第一列为物种名称，其余各列为各物种在各样品中的丰度信息。

| |-- heatmap ——【物种丰度聚类分析结果】

| | |-- figure ——【物种在各水平上的丰度聚类图，pdf 及 png 格式】

储存了在界门纲目科属种水平上的物种丰度聚类图，在每张图中,横向为样品信息,纵向为物种注释信息，图中左侧的聚类树为物种聚类树；上方的聚类树为样品聚类树；中间热图对应的值为每一行物种相对丰度经过标准化处理后得到的 Z 值,即一个样品在某个分类上的 Z 值为样品在该分类上的相对丰度和所有样品在该分类的平均相对丰度的差除以所有样品在该分类上的标准差所得到的值。

| | `-- table ——【各水平上物种丰度聚类所使用的文件】

储存了在界门纲目科属种水平上的基于物种丰度聚类图作图的数据，第一行为样品名称,第一列为物种名称,其余各列为各物种在各样品中的丰度信息。

| |-- **PCA ——【PCA 分析结果,下一级按照分类层级分为各个目录】**

| | |-- **{k,p,c,o,f,g}.PCA12_2.{pdf|png}——【没有标示样品名称的 PCA 分析结果，pdf 和 png 格式】**

在不同分类层级上的 PCA 图，图中没有标示样品名称,横坐标表示第一主成分，百分比则表示第一主成分对样品差异的贡献值；纵坐标表示第二主成分，百分比表示第二主成分对样品差异的贡献值；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。

| | |-- **{k,p,c,o,f,g}.PCA12.{pdf|png} ——【标示了样品名称的 PCA 分析结果，pdf 和 png 格式】**

在不同分类层级上的 PCA 图,图中标示了样品名称,横坐标表示第一主成分,百分比则表示第一主成分对样品差异的贡献值；纵坐标表示第二主成分,百分比表示第二主成分对样品差异的贡献值;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。

| | |-- **pca.csv ——【各个主成分分析结果】**

用于 PCA 作图的相关文件,第一列为样品名，第一行为主成分，表中数据分别对应相应样品在各个主成分上的坐标位置。

| | |-- **PCA_stat_correlation1.txt ——【第一主成分分析结果】**

存储着各物种与第一主成分相关性的文件，第一列为对应物种名，第二列为对应物种与第一主成分的相关性，第三列为 p 值。

| | `-- **PCA_stat_correlation2.txt ——【第二主成分分析结果】**

存储着各物种与第二主成分相关性的文件，第一列为对应物种名，第二列为对应物种与第二主成分的相关性，第三列为 p 值。

| |-- **MetaStats ——【各个分类层级下 MetaStats 及箱图结果】**

各个层级进行 MetaStats 分析的结果。

| | |-- **{ k,p,c,o,f,g,s} ——【各个分类层级下 MetaStats 及箱图结果】**

| | | `--**boxplot ——【具有显著性差异物种的箱图结果】**

存储着各个层级下具有显著想差异的的物种的箱图图片。

| | | `--**PCA——【具有显著性差异物种的 PCA 分析结果】**

存储着各个层级下基于显著性差异物种的进行 PCA 分析的图片和相关文件。

| | | ` **--cluster.species.diff.{png,pdf}——【具有显著性差异物种的 heatmap 热图分析结果】**

各个层级下基于具有显著性差异物种进行物种丰度聚类热图。

| | | ` **--A-vs-B.test.xls ——【MetaStats 分析计算结果】**

A 组与 B 组进行 MetaStats 分析的结果文件。第一列文物种名称，依次为对应物种在 group1 中的均值、方差、标准差、在 group2 中的均值方差标准差，p 值、q 值。

| | | `**--A-vs-B.qsig.xls ——【q 值小于 0.05 分析计算结果】**

基于 A-vs-B.test.xls 分析结果筛选出的 q 值小于 0.05 的物种信息,第一列文物种名称，依次为对应物种在 group1 中的均值、方差、标准差、在 group2 中的均值方差标准差，p 值、q 值。

| | | `**--A-vs-B.Psig.xls ——【p 值小于 0.05 分析计算结果】**

基于 A-vs-B.test.xls 分析结果筛选出的 p 值小于 0.05 的物种信息 ,第一列文物种名称 ,依次为对应物种在 group1 中的均值、方差、标准差、在 group2 中的均值方差标准差，p 值、q 值。

| | | `--{k,p,c,o,f,g,s}_qsig.xls ——【各层级 q 值小于 0.05 的差异物种结果】

根据 q 值的大小对具有显知性差异的物种进行“*” 或“**”标识。如果 0.01<=q 值<0.05 则为“*”,如果 q 值<0.01 则表为“**”。

| |-- Krona ——【Krona 网页展示相关文件】

存储了用 Krona 进行物种展示的网页等相关展示文件信息。

| |-- top/top_group ——【物种注释结果在各水平上丰度前 10 的物种统计及柱形图结果】

| |-- figure ——【物种注释结果各个层级排名前 10 的物种丰度柱形图,png 及 svg 格式】

从各水平上的相对丰度表出发,选取出在各样品中的最大相对丰度排名前 10 的物种,并将其余的物种设置为 Others,绘制出各 样品对应的物种注释结果在各水平的统计图，对应的含有结题报告中的门水平的相对丰度柱形图的分析结果。图中纵轴表示注释到某类型的物种的相对比例；横轴表示样品名称;各颜色区块对应的物种类别见右侧图例。

| `-- table ——【物种注释结果各个层级排名前 10 的物种丰度数据】

存储的是物种注释结果各个层级上排名前 10 的物种的相对丰度情况,第一行为物种名称,第一列为样品名,反应了不同样品在不同物

种上的丰度情况。

| |-- NMDS ——【NMDS 分析结果,下一级按照分类层级分为各个目录】

| | |-- {p,c,o,f,g}.NMDS_2.{pdf|png}——【没有标示样品名称的 NMDS 分析结果，pdf 和 png 格式】

在不同分类层级上的 NMDS 降维图，图中没有标示样品名称，图中的每个点表示一个样品，点与点之间的距离表示差异程度，同一个组的样品使用同一种颜色表示；Stress 小于 0.2 时，表明 NMDS 分析具有一定的可靠性。

| | |-- {k,p,c,o,f,g}.PCA12.{pdf|png} ——【标示了样品名称的 PCA 分析结果，pdf 和 png 格式】

在不同分类层级上的 NMDS 降维图，图中标示了样品名称，图中的每个点表示一个样品，点与点之间的距离表示差异程度，同一个组的样品使用同一种颜色表示；Stress 小于 0.2 时，表明 NMDS 分析具有一定的可靠性。

| | |-- NMDS_scores.txt ——【NMDS 最终的降维结果】

用于 NMDS 作图的相关文件,第一列为样品名，第一行为 NMDS 的两个轴，表中数据分别对应相应样品在两个坐标轴上的位置。

| |-- Anosim ——【Anosim 分析结果,下一级按照分类层级分为各个目录】

| | |-- {k,p,c,o,f,g}.*.{pdf|png}——【各个分组间 Anosim 分析箱图，pdf 和 png 格式】

在不同分类层级上，绘制各个分组之间的组间和组内差异箱图，第 1 个箱子代表两组组间差异大小，第 2 个箱子代表分组 1 的组内差异，第 3 个箱子代表分组 2 的组内差异。

| | |-- {k,p,c,o,f,g}.stat_anosim.txt ——【Anosim 分析的统计结果】

第一列为两两分组的比较，第二列 R-value，介于（-1，1）之间，R-value 大于 0，说明组间差异显著，R-value 小于 0，说明组内差异大于组间差异；第三列为统计分析的可信度 P-value，P< 0.05 表示统计具有显著性。

| |-- LDA ——【基于物种丰度的 LEfSe 分析结果】

| | |-- LDA.*.{png,pdf} ——【LDA 柱状图结果，pdf 和 png 格式】

LDA 值分布柱状图中展示了 LDA Score 大于设定值（默认设置为 3）的物种，即组间具有统计学差异的 Biomarker。展示了不同组中丰度差异显著的物种，柱状图的长度代表差异物种的影响大小（即为 LDA Score）。

| | |-- **LDA.*.tree.{png,pdf} ——【进化分支图，pdf 和 png 格式】**

在进化分支图中，由内至外辐射的圆圈代表了由门至属（或种）的分类级别。在不同分类级别上的每一个小圆圈代表该水平下的一个分类，小圆圈直径大小与相对丰度大小呈正比。着色原则：无显著差异的物种统一着色为黄色，差异物种 Biomarker 跟随组进行着色，红色节点表示在红色组别中起到重要作用的微生物类群，绿色节点表示在绿色组别中起到重要作用的微生物类群。图中英文字母表示的物种名称在右侧图例中进行展示。

| | |-- **LDA.*.res ——【线性判别分析统计结果】**

| | | `--**heatmap——【具有显著性差异属水平物种的聚类热图结果】**

根据 LDA 分析结果检测出的属水平差异物种的丰度聚类热图。

| | | `--**ROC——【具有显著性差异物种的 ROC 曲线】**

为了检验通过 LEfSe 筛选出来的组间差异 Biomarker 的分类预测能力，绘制受试者工作特征曲线 （receiver operating characteristic curve，简称 ROC 曲线）。ROC 曲线常用来评价一个二值分类器的好坏，也是基于统计学上判断分组信息优劣的指标。AUC（Area Under Curve）被定义为 ROC 曲线下的面积。通常情况下，它的值在 1.0 和 0.5 之间。在 AUC>0.5 的情况下，AUC 越接近于 1，说明分类预测效果越好。

`-- 04.TaxAnnotation--ReadMe.pdf ——【 04.TaxAnnotation 交付结果目录说明】