诺禾致源
宏基因组交付目录说明手册
（V3.0）

诺禾致源微生物事业部

2015 年 07 月 31 日

目录

（注：单击即可跳转至相应文档的详细说明）

|-- 01.DataClean ——【测序数据预处理结果】

| |-- novototal.QCstat.info.xls ——【QC 结果基本信息表】

该文件即对应的是结题报告中的数据预处理统计表，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	列标题	说明
1	Sample	样品名称
2	InsertSize(bp)	建库时的插入片段大小，单位为 bp
3	SeqStrategy	测序的策略，若为 125:125 即代表采用的是 Pair-end 测序，测序 reads 长度为 125bp
4	RawData	RawData 的数据量，单位为 M,
5	CleanData	CleanData 数据量，单位为 M
6	Clean_Q20	CleanData 的 Q20
7	Clean_Q30	CleanData 的 Q30
8	Clean_GC(%)	CleanData 碱基的 GC 含量
9	Effective(%)	CleanData 占 RawData 的百分比

| |-- total.QCstat.info.xls ——【QC 结果详细信息表】

可以用 excel 打开该文件，其相比于 novototal.QCstat.info.xls 而言，多出了 RawReads、Low_Q、N_num、Adapter、Duplication、Poly 这几列，各列所代表的含义如下：

列数	列标题	说明
1	Sample	样品名称
2	InsertSize(bp)	建库时的插入片段大小，单位为 bp
3	SeqStrategy	测序的策略，若为 125:125 即代表采用的是 Pair-end 测序，测序 reads 长度为 125bp
4	RawData	RawData 的数据量，单位为 M,
5	RawReads(#)	原始下机的 Reads 数目
6	Low_Q	去除含低质量碱基超过一定比例的 reads 序列总长，单位为 M
7	N_num	去除含 N 碱基达到一定比例的 reads 序列总长，单位为 M
8	Adapter	去除与 Adapter 之间 overlap 超过一定阈值的 reads 序列总长，单位为 M
9	Duplication	去除的重复 reads 序列总长，单位为 M
10	Poly	代表去除的 Poly reads 的 reads 序列总长，单位为 M
11	CleanData	CleanData 数据量，单位为 M
12	Clean_Q20	CleanData 的 Q20
13	Clean_Q30	CleanData 的 Q30

14	Clean_GC(%)	CleanData 碱基的 GC 含量
15	Effective(%)	CleanData 占 RawData 的百分比

列数	列标题	说明
1	Sample	样品名称
2	InsertSize(bp)	建库时的插入片段大小，单位为 bp
3	SeqStrategy	测序的策略，若为 125:125 即代表采用的是 Pair-end 测序，测序 reads 长度为 125bp
4	RawData	RawData 的数据量，单位为 M,
5	RawReads(#)	原始下机的 Reads 数目
6	Low_Q	去除含低质量碱基超过一定比例的 reads 序列总长，单位为 M
7	N_num	去除含 N 碱基达到一定比例的 reads 序列总长，单位为 M
8	Adapter	去除与 Adapter 之间 overlap 超过一定阈值的 reads 序列总长，单位为 M
9	Duplication	去除的重复 reads 序列总长，单位为 M
10	Poly	代表去除的 Poly reads 的 reads 序列总长，单位为 M
11	CleanData	CleanData 数据量，单位为 M

	12	Clean_Q20	CleanData 的 Q20
	13	Clean_Q30	CleanData 的 Q30
	14	Clean_GC(%)	CleanData 碱基的 GC 含量
	15	Effective(%)	CleanData 占 RawData 的百分比
	16	NonHostData	去除宿主后，剩余数据量，单位为 M

total.*.NonHostQCstat.info.xls ——【去除宿主后的结果的详细信息表】

可以用 excel 打开该文件，其相比于 total.QCstat.info.xls 而言，多出了 NonHostData 这一列，各列所代表的含义如下：

| `-- Sample ——【各样品对应的质控结果，文件夹以样品名称来命名】

| |-- *.fq1.gz ——【质控后 read1 的 FASTQ 文件】

关于 FASTQ 文件格式介绍，请参考结题报告中的常见数据格式说明文档。

| |-- *.fq2.gz ——【质控后 read2 的 FASTQ 文件】

关于 FASTQ 文件格式介绍，请参考结题报告中的常见数据格式说明文档。

| |-- ***.nohost.fq1.gz ——【当存在宿主基因组序列时，去除宿主后 read1 的 FASTQ 文件】**

关于 FASTQ 文件格式介绍，请参考结题报告中的常见数据格式说明文档。

| |-- ***.nohost.fq2.gz ——【当存在宿主基因组序列时，去除宿主后 read2 的 FASTQ 文件】**

关于 FASTQ 文件格式介绍，请参考结题报告中的常见数据格式说明文档。

| |-- ***.qual.png ——【Clean Data 的碱基质量分布图】**

在该图中，横轴为 reads 上的碱基位置，从中间隔开,前后代表的是 PE reads，图中纵轴为该 reads 位置上碱基质量的分布。

坐标轴	标题	说明
X 轴	Position along reads	reads 上的碱基位置，从中间隔开，前后代表的是 PE reads
Y 轴	Quality Value	该 reads 位置上碱基质量的分布

| `-- ***.base.png ——【Clean Data 的碱基类型分布图】**

在该图中，右上角的图例中有各个碱基所代表的颜色说明，图中横轴为 reads 上的碱基位置，中间黄线前后则代表的是 PE reads，图中纵轴代表该 reads 位置上某个碱基的比例。

坐标轴	标题	说明
X 轴	Position along reads	reads 上的碱基位置，中间黄线前后则代表的是 PE reads
Y 轴	Percent Value	该 reads 位置上某个碱基的比例，右上角的图例中有各个碱基所代表的颜色说明

|-- 02.Assembly —— 【宏基因组组装结果】

| |-- total.scaftigs.stat.info.xls —— 【所有样品 Scaftigs 信息表】

该文件即对应的是结题报告中的组装结果 Scaftigs 的统计表，可以用 excel 打开该文件，各列所代表的含义如下：

列数	列标题	说明
1	SampleID	样品名称
2	Total len.(bp)	组装得 到的 Scaftigs 的总长，单位为 bp
3	Num.	组装得到的 Scaftigs 总条数
4	Average len.(bp)	Scaftigs 的平均长度
5	N50 Len.(bp)	Scaftigs 的 N50

6	N90 Len.(bp)	Scaftigs 的 N90
7	Max len.(bp)	组装得到的最长 Scaftigs 的长度值

| |-- total.scafSeq.stat.info.xls ——【所有样品 Scaffold 信息表】

该文件为组装结果 Scaffold 的统计表，可以用 excel 打开该文件，各列所代表的含义如下：

列数	列标题	说明
1	SampleID	样品名称
2	Total len.(bp)	组装得 到的 Scaffold 的总长，单位为 bp
3	Num.	组装得到的 Scaffold 总条数
4	Average len.(bp)	Scaffold 的平均长度
5	N50 Len.(bp)	Scaffold 的 N50
6	N90 Len.(bp)	Scaffold 的 N90
7	Max len.(bp)	组装得到的最长 Scaffold 的长度值

| |-- **ReadsMapping ——【将各样品 Clean Data mapping 至组装 Scaftigs 上的结果】**

| | `-- **Sample ——【各样品 ReadsMapping 结果，文件夹以样品名称来命名】**

| | |-- **coverage_depth.{png|svg} ——【覆盖深度分布图，png 和 svg 格式】**

这两个文件为对应的样品的覆盖深度分布图，其横轴代表的是测序深度，纵轴代表的是属于该测序深度的序列数目。

| | |-- **coverage.depth.table.xls ——【各 Scaftigs 覆盖度总体情况统计,包含覆盖度，覆盖长度等信息】**

该文件是对 reads mapping 后的结果进行的统计，用 excel 打开该文件后，各列所代表的含义如下：

列数	列标题	说明
1	Reference_ID	Scaftigs 的编号
2	Reference_size(bp)	Scaftigs 长度
3	Covered_length(bp)	覆盖长度
4	Coverage(%)	覆盖度
5	Depth	深度
6	Depth_single	单碱基位点深度之和

| | |-- *.{PE|SE}.soap ——【soap 比对结果文件】

这两个文件是用 soapaligner 软件将对应样品的 Clean reads 比对至对应样品组装后的 Scaftigs，所获得的 soap 比对结果文件，可以用 excel 打开（文件过大时不推荐打开），在这些文件中，各列所代表的含义如下：

列数	说明
1	read 的编号，编号的有效字符有[a-zA-Z0-9.:^x!+_?-|]。
2	read 的序列，如果 read 比对上参考序列的负链，会被反向互补为正链。
3	质量值:序列的质量值，和序列顺序一致，如果 read 反向互补，质量值也会随着改变。
4	比对上的次数：最优比对的次数。没有比对上的 read 将被忽略。
5	a/b：pair-end 比对的标记，表示 read 属于来自哪个文件。
6	长度： read 长度,如果是容缺失的比对，长度将是加上缺失片断的长度。
7	+/-： 比对上参考序列的正链或负链

8	参考序列的名称。
9	位点：第一个碱基在参考序列上的位置，从 1 开始。
10	错配的个数。
11	错配的详细信息（"C->33G4" 意思是一个错配，在参考序列的位置是第 9 列+33（从 0 开始），在参考序列上是 C，read 上是 G，质量值是 4），如果错配数为 0，则无该列，即该行只有 12 列。
12	比对上的数目（"44M" 意思是 44 个碱基比对上了)。
13	对比的细节（"33C10"意思是前 33 个比对上了，第 34（参考序列上是第九列+34）个是错配，后面 10 个还是比对上了）

| | |-- ***.unmapping.{fq1|fg2}.gz** **——【各样品没有 map 上 Scaftigs 的 read1 和 read2 的 FASTQ 文件】**

关于 FASTQ 文件格式介绍，请参考结题报告中的常见数据格式说明文档。

| | `-- **soap.coverage.depthsingle ——【单碱基位点覆盖深度文件】**

该文件和 FASTA 数据格式是一致的，每一个碱基位点上的数字代表了该碱基位点上的深度。

| |-- Sample/NOVO_MIX ——【各样品对应的组装结果，文件夹以样品名称来命名;NOVO_MIX 为 unmapped reads 混合组装的结果】

| | |-- *.scafSeq.fa ——【单样品 scaffold 序列，FASTA 格式】

关于 FASTA 文件格式介绍，请参考结题报告中的常见数据格式说明文档。

| | |-- *.scafSeq.500.ss.txt ——【按照长度 500 进行过滤后，单样品 scaffold 序列信息统计表】

在该文件中，储存的是相应样品组装所得到的 scaffold 的平均长度，N50，N90 等基本指标，可以用写字板或记事本打开该文件。该文件中，各列所代表的含义如下：

行数	行标题	说明
1	Statistical level	统计下方指标时的过滤阈值 ,例如括号中标明了 500 的即是过滤掉 500bp 以下的序列进行的统计

2	Total number	序列数目
3	Total length of (bp)	序列总长度
4	Gap number (bp)	Gap 的碱基长度
5	Average length (bp)	平均长度
6	N50 Length (bp)	序列 N50
7	N90 Length (bp)	序列 N90
8	Maximum length (bp)	最长序列长度
9	Minimum length (bp)	最短序列长度
10	GC content is (%)	序列 GC 含量

| | | |-- ***.scaftigs.fa** ——【单样品 Scaftigs 序列，FASTA 格式】

关于 FASTA 文件格式介绍，请参考结题报告中的常见数据格式说明文档。

| | | |-- ***.scaftigs.500.ss.txt** ——【单样品 Scaftigs 序列信息统计表】

在该文件中，储存的是相应样品组装所得到的 Scaftigs 的平均长度，N50，N90 等基本指标，可以用写字板或记事本打开该文件。在该文件中，各列所代表的含义如下：

行数	行标题	说明
1	Statistical level	统计下方指标时的过滤阈值 ,例如括号中标明了 500 的即是过滤掉 500bp 以下的序列进行的统计
2	Total number	序列数目
3	Total length of (bp)	序列总长度
4	Gap number (bp)	Gap 的碱基长度
5	Average length (bp)	平均长度
6	N50 Length (bp)	序列 N50
7	N90 Length (bp)	序列 N90
8	Maximum length (bp)	最长序列长度
9	Minimum length (bp)	最短序列长度
10	GC content is (%)	序列 GC 含量

| | `-- *.len.{png|svg} ——【Scaftigs 长度分布图，png 或 svg 格式】

这个图片展示的是某个样品中 Scaftigs 的长度分布，横轴表示 Scaftigs 的长度，第一纵轴（Frequence(#)）表示 Scaftigs 数目；第二纵轴（Percentage（%））表示 Scaftigs 数目的百分比，从这个图上我们可以看出，组装后得到的 Scaftigs 的长度分布情况。

坐标轴	标题	说明
横轴	Scaftig Length(bp)	Scaftigs 的长度
第一纵轴	Frequence	Scaftigs 数目
第二纵轴	Percentage(%)	Scaftigs 数目的百分比

|-- 03. GenePredict ——【基因预测结果及丰度分析结果】

| |-- GenePredict ——【基因预测结果】

| | `-- Sample /NOVO_MIX ——【各样品对应的基因预测结果，文件夹以样品名称来命名】

| | |-- *.CDS.fa ——【预测基因核苷酸 FASTA 文件】

关于 FASTA 文件格式介绍，请参考结题报告中的常见数据格式说明文档。

| | |-- ***.protein.fa ——【预测基因氨基酸 FASTA 文件】**

关于 FASTA 文件格式介绍，请参考结题报告中的常见数据格式说明文档。

| | |-- ***.CDS.fa.stat.xls ——【预测基因统计结果】**

预测基因基本信息表，各行所代表的含义如下所示：

行数	行标题	说明
1	ORFs NO.	预测得到 ORF(Open Reading Frame) 数目
2	integrity:end	只有终止密码子的基因数目
3	integrity:all	既有起始密码子也有终止密码子的基因数目
4	integrity:none	既无起始密码子也无终止密码子的基因数目
5	integrity:start	只含起始密码子的基因数目
6	Total Len.(Mbp)	预测得到的 ORF 的总长，单位是百万
7	Average Len.(bp)	ORF 的平均长度
8	GC percent	预测的 ORF 的整体 GC 含量值

| | `-- *.CDS.fa.len.{png|svg} ——【基因碱基序列长度分布统计图，png 格式和 svg 格式】

预测基因长度图，有 PNG 和 SVG 两种格式，SVG 为高清矢量图，可以无限放大而不失真，在该图中，第一纵轴 Frequence(#) 表示预测基因数目；第二纵轴 Percentage(%) 表示预测基因数目的百分比；横轴表示预测基因长度。

| |-- UniqGenes——【基因去冗余分析结果】

| | |-- Unigenes.CDS.cdhit.fa ——【去冗余后的基因核苷酸 FASTA 文件】

关于 FASTA 文件格式介绍，请参考结题报告中的常见数据格式说明文档。

| | |-- Unigenes.protein.fa ——【去冗余前的所有样品的预测基因氨基酸 FASTA 文件】

关于 FASTA 文件格式介绍，请参考结题报告中的常见数据格式说明文档。

| | |-- Unigenes.protein.cdhit.fa ——【去冗余后的基因氨基酸 FASTA 文件】

关于 FASTA 文件格式介绍，请参考结题报告中的常见数据格式说明文档。

| | |-- **Unigenes.CDS.cdhit.fa.len.{png|svg} ——【去冗余后的基因核苷酸 FASTA 序列的长度分布统计图，png 格式和 svg 格式】**

在该图中，第一纵轴 Frequence(#) 表示预测基因数目；第二纵轴 Percentage(%) 表示预测基因数目的百分比；横轴表示预测基因长度。

| | |-- **Unigenes.CDS.cdhit.fa.stat.xls ——【去冗余后的基因核苷酸 FASTA 序列基本信息统计表】**

文件格式与*.CDS.fa.stat.xls 一致。

| | `-- **Unigenes.protein.table.txt ——【去冗余后的代表基因 ,代表基因所属 cluster 数目及基因编号表】**

Uniq.Genes.protein.table.txt 是对去冗余结果进行的统计，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	列标题	说明
1	#Rep_id	代表性基因的 ID 号
2	Len(nt/aa)	该代表性基因的长度

3	Num	该代表性基因的 cluster 中的基因的数目
4	Seq_ID	该 cluster 中各基因的 ID 号

| |-- **GeneStat ——【基因特征的统计分析结果】**

| | |-- **core_pan ——【core 基因与 pan 基因分析结果】**

Core 基因与 Pan 基因相关分析的稀释度曲线图，横坐标为随机抽取的样本数目，纵坐标为样本组合的 core 基因与 pan 基因数目。

| | |-- **correlation ——【各样品基因丰度相关性分析结果】**

各样品之间基因丰度相关性热图，不同颜色对应不同的相关性系数。

| | |-- **genebox ——【样品组间基因数目箱图】**

各样品组之间基因数目箱图，横坐标为样品的分组情况，纵坐标为基因数目。

| | `-- venn ——【基因韦恩图结果】

指定样品间共有基因分布情况韦恩图。

| `-- GeneTable ——【基因丰度分析结果】

| |-- Unigenes.readsNum.xls ——【基因在各样品中的覆盖 reads 数】

通过 readsmapping 结果得到的，非冗余基因在各样品中覆盖 reads 数目统计表

| |-- Unigenes.readsNum.even.tree ——【从基因在各样品中均一化后的绝对丰度表出发，获得的 BC 距离聚类树】

从 Unigenes.readsNum.even.xls 结果出发，所获得的样品 BC 聚类分析结果，为 tree 格式的文件，可以使用 treeviewer 等可以查看树文件结构的软件打开。

| |-- Unigenes.readsNum.relative.xls ——【基因在各样品中的相对丰度表】

从 Unigenes.readsNum.xls 结果出发，采用基因长度进行均一化后，得到的基因在各样品中的相对丰度表。

| `-- **Unigenes.readsNum.even.xls ——【基因在各样品中的相对丰度表进行均一化后的结果】**

将 Unigenes.readsNum.relative.xls 按照在各样品中，比对上的 reads 数目之和的最大值进行均一化后，得到的结果。

|-- 04.TaxAnnotation ——【物种注释分析结果】

| |-- MAT ——【物种注释丰度统计矩阵】

| | |-- Absolute ——【均一化后的绝对丰度矩阵：基于 Unigenes.readsNum.even.xls，所获得不同分类层级的绝对丰度均一化矩阵】

| | | `-- **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，各样品的绝对丰度情况，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。

| | |-- **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。

| |-- **GeneNums ——【物种注释基因数目统计】**

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

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）上的注释基因数目矩阵】**

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，对应样品注释到的基因数目情况。

|　　|-- GeneNums.BetweenSamples.heatmap——【基于 GeneNums.BetweenSamples，进行的基因数目热图分析】

|　　|　　|　　`-- {k,p,c,o,f,g,s}.genenum.heatmap.txt.{png,pdf}　——【各分类水平上，各样品间基于基因数目的 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热图，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；中间热图对应的值为每一行物种在对应样品中的基因数目；各颜色对应的基因数目见右侧图例。

|　　|　　|　　`-- {k,p,c,o,f,g,s}.genenum.heatmap.txt——【各分类水平上，各样品间基于基因数目的 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聚类热图的作图数据，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基因数目信息。

| |-- **MicroNR_Anno ——【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、亚纲等更详细层级的物种注释信息。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，只有界门纲目科属种）。

| | |-- **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。

| |-- 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分析的相对丰度矩阵，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；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。

|　　|　　|-- { kingdom,phylum,class,order,family,genus,species} ——【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性差异的的物种的丰度箱图图片。

|　　|　　|　　`--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,第一列为物种名称，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 ,第一列为物种名称 ,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 值的大小对具有显著性差异的物种进行“*” 或“**”标识。如果 0.01<=q 值<0.05 则为“*”,如果 q 值<0.01 则标为“**”。

| |-- 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 ——【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。

|-- 05.FunctionAnnotation ——【功能注释分析结果】

| |-- CAZy ——【CAZy 数据库分析结果】

| | |-- CAZy_Anno ——【CAZy 注释结果统计】

| | | |-- cazy.unigenes.num.{pdf|png} ——【注释到 CAZy 第一层级的基因数目统计图，pdf 及 png 格式】

对应的是结题报告中的 CAZy 注释结果统计图，图中，横轴是数据库中各功能类型的代码，代码的解释见对应的图例说明，纵轴代表注释为相应功能类的基因数目。

| | | |-- cazy.unigenes.num.txt ——【注释到 CAZy 第一层级的基因数目统计结果】

储存了六大功能类对应的总体注释结果，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	列标题	说明
1	CAZy_Class	六大功能类的缩写

2	Discription	各功能类对应的描述
3	Num	该功能类注释上的基因数目

| | | |-- **Unigenes.blast.m8.filter.anno.xls ——【过滤后的 blast 结果的注释信息】**

为基于 Unigenes.blast.m8.filter.xls 进行的 CAZy 注释结果，可以用 excel 打开。该文件共分为 4 列，各列所代表的含义如下：

列数	列标题	说明
1	Gene_id	基因的 ID 号
2	Subject_id	比对上的序列在 genebank 中的 GI 号
3	CAZy_Family	比对上的序列所属的子功能类
4	Family_Description	该功能类的描述

| | | |-- **Unigenes.blast.m8.filter.xls ——【过滤后的 blast 结果文件，Blast 软件的 m8 格式】**

即为 blast 后的 m8 格式的文件，关于 m8 格式的详细解释可以参看结题报告中的常用数据格式说明。

| | | |-- **Unigenes.level1.bar.{png|svg} ——【CAZy 第一层级上的相对丰度统计图】**

对应的是结题报告中的 CAZy 基因注释结果在第一层级上的统计图，纵轴表示注释到某类型的功能的相对比例；横轴表示样品名称；各颜色区块对应的功能类别见右侧图例。

| | | |-- **Unigenes.level1.bar.tree.{png|svg} ——【CAZy 功能聚类分析结果】**

对应的是结题报告中的 CAZy 功能聚类分析图，图中心是 BC 距离聚类树结构，外圈各层是各样品在第一层级上的功能相对丰度分布，各颜色区块对应的功能类别见左上角图例。

| | |-- **CAZy_MAT ——【CAZy 相对丰度和绝对丰度分析结果：ec 为酶，level1 为六大功能类，level2 为子功能】**

| | | |-- **Absolute ——【各样品在 ec，level1 和 level2 不同水平的均一化后的绝对丰度矩阵】**

在该文件夹中，一共有三个文件，可以用 excel 打开这些文件，其中 Unigenes.absolute.level1.xls 是对第一层级的六大功能类在各样品中的绝对丰度进行的统计，在该文件中，第一行为样品信息，第一列为第一层级六大功能类的代号，最后一列为代号的详细说明，其余的数字则代表某个功能类在某个样品中的绝对丰度。

在 Absolute 文件夹中，存在的另外两个文件 Unigenes.absolute.level2.xls 和 Unigenes.absolute.ec.xls 的展示形式和 Unigenes.absolute.level1.xls 文件的展示形式是一样的，不同的是，Unigenes.absolute.level2.xls 是对子功能类在各样品中的绝对丰度进行的统计，而 Unigenes.absolute.ec.xls 则是对所有能够注释上 CAZy 数据库的基因所属的酶，在各样品中的绝对丰度进行的统计。

| | | `-- **Relative ——【各样品在 ec，level1 和 level2 不同水平的相对丰度矩阵】**

在该文件夹中，一共有三个文件，可以用 excel 打开这些文件，其中 Unigenes.relative.level1.xls 是对第一层级的六大功能类在各样品中的相对丰度进行的统计，在该文件中，第一行为样品信息，第一列为第一层级六大功能类的代号，最后一列为代号的详细说明，其余的数字则代表某个功能类在某个样品中的相对丰度。

在 Relative 文件夹中，存在的另外两个文件 Unigenes.relative.level2.xls 和 Unigenes.relative.ec.xls 的展示形式和

Unigenes.relative.level1.xls 文件的展示形式是一样的，不同的是，Unigenes.relative.level2.xls 是对子功能类在各样品中的相对丰度进行的统计，而 Unigenes.relative.ec.xls 则是对所有能够注释上 CAZy 数据库的基因所属的酶，在各样品中的相对丰度进行的统计。

|　　|-- GeneNums ——【注释到的基因数目、基因 id 统计，ec 为酶，level1 为六大功能类，level2 为子功能】

在该文件夹中有三个文件，文件展示形式类似。文件 Unigenes.absolute.level1.xls，共三列，第一列为第一层级六大功能类的代号，第二列为注释上的基因数，第三列为注释上该功能的所有基因 id。文件 Unigenes.absolute.level2.xls，共四列，第一列为子功能名称，第二列为子功能的详细描述，第三列为注释上的基因数，第四列为注释上该功能的所有基因 id。文件 Unigenes.absolute.ec.xls，共三列，第一列为酶的名称，第二列为注释上该酶的基因数，第三列为注释上该酶的所有基因 id。

|　　|-- GeneNums.BetweenSamples ——【各样品中基因数目统计】

在该文件夹中有三个文件，文件展示形式相同。文件 Unigenes.absolute.level1.xls，Unigenes.absolute.level2.xls，Unigenes.absolute.ec.xls， 内容为不同功能水平上，在各样品中注释上的基因数目统计结果。

| | |-- **GeneNums.BetweenSamples.heatmap ——【各样品注释到的基因数目的聚类热图】**

| | |-- **heatmap ——【不同层级在各样品中的相对丰度聚类热图以及作图数据】**

| | |-- **MetaStats ——【不同层级的 MetaStats 统计结果，EC 为酶，level1 为六大功能类，level2 为子功能】**

| |-- **eggNOG ——【eggNOG 数据库分析结果】**

| | |-- **eggNOG_Anno ——【eggNOG 注释结果统计】**

| | | |-- **eggNOG.unigenes.num.{pdf|png} ——【注释到 eggNOG 第一层级的基因数目统计图，pdf 及 png 格式】**

结果与 CAZy 数据库的结果类似。

| | | |-- **eggNOG.unigenes.num.txt ——【注释到 eggNOG 第一层级的基因数目统计结果】**

结果与 CAZy 数据库的结果类似。

| | | |-- **Unigenes.blast.m8.filter.anno.xls ——【过滤后的 blast 结果的注释信息】**

储存的是基于 blast m8 文件进行的 eggNOG 注释结果，该文件可以用 excel 打开，该文件共分为 5 列，各列所代表的含义如下：

列数	列标题	说明
1	Query_id	基因的 ID 号
4	Subject_id	比对上的序列号
5	Ortholog_Group	该序列所属的 Ortholog Group ID
6	Functional_Category	该 Ortholog Group ID 所属的功能类代号
7	OG_Description	该 Ortholog Group ID 所对应的描述

| | | |-- **Unigenes.blast.m8.filter.xls ——【过滤后的 blast 结果文件，Blast 软件的 m8 格式】**

即为 blast 后的 m8 格式的文件，关于 m8 格式的详细解释可以参看结题报告中的常用数据格式说明。

| | | |-- **Unigenes.level1.bar.{png|svg} ——【eggNOG 第一层级上的相对丰度统计图】**

| | | |-- **Unigenes.level1.bar.tree.{png|svg} ——【eggNOG 功能聚类分析结果，pdf 与 png 格式】**

| | |-- **eggNOG_MAT ——【eggNOG 相对丰度和绝对丰度分析结果：level1 为第一层级，level2 为第二层级，og 为直系同源簇】**

| | | |-- **Absolute ——【各样品在 level1，level2 和 og 水平注释到的均一化后的绝对丰度矩阵】**

在该文件夹中，一共有三个文件，可以用 excel 打开这些文件，其中 Unigenes.absolute.level1.xls 是对第一层级的 25 大功能类在各样品中的绝对丰度进行的统计，在该文件中，第一行为样品信息，第一列为第一层级 25 大功能类的代号，最后一列为代号的详细说明，其余的数字则代表某个功能类在某个样品中的绝对丰度。

在 Absolute 文件夹中，存在的另外两个文件 Unigenes.absolute.level2.xls 和 Unigenes.absolute.og.xls 的展示形式和 Unigenes.absolute.level1.xls 文件的展示形式是一样的，不同的是，Unigenes.absolute.level2.xls 是对子功能类的描述在各样品中的绝对丰度进行的统计，而 Unigenes.absolute.og.xls 则是对所有能够注释上 eggNOG 数据库的基因所属的 Ortholog Group ID，在各样品中的绝对丰度进行的统计。

| | | `-- **Relative ——【各样品在 level1，level2 和 og 水平注释到的相对丰度矩阵】**

在该文件夹中，一共有三个文件，可以用 excel 打开这些文件，其中 Unigenes.relative.level1.xls 是对第一层级的 25 大功能类在各样品中的相对丰度进行的统计，在该文件中，第一行为样品信息，第一列为第一层级 25 大功能类的代号，最后一列为代号的详细说明，其余的数字则代表某个功能类在某个样品中的相对丰度。

在 Relative 文件夹中，存在的另外两个文件 Unigenes.relative.level2.xls 和 Unigenes.relative.og.xls 的展示形式和 Unigenes.relative.level1.xls 文件的展示形式是一样的，不同的是，Unigenes.relative.level2.xls 是对子功能类的描述在各样品中的相对丰度进行的统计，而 Unigenes. relative.og.xls 则是对所有能够注释上 eggNOG 数据库的基因所属的 Ortholog Group ID，在各样品中的相对丰度进行的统计。

| | |-- **GeneNums ——【各层级注释到的基因数目、基因 id 统计】**

在该文件夹中有三个文件，文件展示形式类似。文件 Unigenes.absolute.level1.xls，共三列，第一列为第一层级 25 大功能类的代号，第二列为注释上的基因数，第三列为注释上该功能的所有基因 id。文件 Unigenes.absolute.level2.xls，共四列，第一列为子功能描述，第二列为子功能的详细描述，第三列为注释上的基因数，第四列为注释上该功能的所有基因 id。文件 Unigenes.absolute.og.xls，共三列，第一列为 Ortholog Group ID，第二列为注释上该 Ortholog Group ID 的基因数，第三列为注释上该 Ortholog Group ID 的所有基因 id。

| | |-- **GeneNums.BetweenSamples ——【各样品中基因数目统计结果】**

在该文件夹中有三个文件，文件展示形式相同。文件 Unigenes.absolute.level1.xls ， Unigenes.absolute.level2.xls ， Unigenes.absolute.og.xls， 内容为不同功能水平上在各样品中注释上的基因数目统计结果。

| | |-- GeneNums.BetweenSamples.heatmap ——【各样品注释到的基因数目的聚类热图】

| | |-- heatmap ——【各样品中的相对丰度聚类热图以及作图数据】

| | |-- NOG.Tax ——【ortholog group 物种归属分析结果】

| | |-- MetaStats ——【各样品的 MetaStats 统计结果，level1 为第一层级，level2 为第二层级，og 为直系同源簇】

| | `-- PCA ——【各样品的 PCA 分析结果】

| `-- KEGG ——【KEGG 数据库分析结果】

| |-- GeneNums ——【注释到的基因数目、基因 id 统计，ec 为酶，ko 为直系同源, level1 为六大代谢通路，level2 为子代谢通路，level3 为代谢通路图】

在该文件夹中，一共有五个文件，可以用 excel 打开这些文件，其中 Unigenes.absolute.level1.xls，共三列，第一列为第一层级六大

代谢通路的名称，第二列为注释上的基因数，第三列为注释上该功能的所有基因 id。Unigenes.absolute.level2.xls、Unigenes.absolute.level3.xls、Unigenes.absolute.ko.xls 和 Unigenes.absolute.ec.xls 也是三列，第一列分别为 level2、level3、Orthologous groups 的 id 和酶的名称，第二列为注释上的基因数，第三列为注释上该功能的所有基因 id。

| |-- **GeneNums.BetweenSamples ——【各样品中基因数目统计】**

在该文件夹中有三个文件，文件展示形式相同。文件 Unigenes.absolute.level1.xls，Unigenes.absolute.level2.xls，Unigenes.absolute.level3.xls，Unigenes.absolute.ec.xls，Unigenes.absolute.ko.xls 内容为不同功能水平上在各样品中注释上的基因数目统计结果。

| |-- **GeneNums.BetweenSamples.heatmap ——【各样品注释到的基因数目的聚类热图】**

| |-- **heatmap ——【各样品中的酶的相对丰度聚类图以及作图数据】**

| |-- **KEGG_Anno ——【KEGG 注释结果统计】**

| | |-- **kegg.unigenes.num.{pdf|png} ——【注释到 KEGG 第一层级的基因数目统计图】**

对应的是结题报告中的 KEGG 注释结果统计图，图中，左侧为代谢通路的描述，图中条形图上的数字为注释为该通路的基因数目。

| | |-- **kegg.unigenes.num.txt ——【注释到 KEGG 第一层级的基因数目统计图】**

该文件一共分为三列，各列所代表的含义如下：

列数	列标题	说明
1	First Level	第一层六大代谢通路的描述
2	Second Level	注释出来的属于该代谢通路的第二层子通路的描述
3	Gene of Unique.Genes.KEGG.catalog	该子通路注释上的基因数目

| | |-- **Unigenes.blast.m8.filter.anno.xls ——【过滤后的 blast 结果的注释信息】**

储存的是基于 blast m8 文件进行的 KEGG 注释结果，该文件也可以用 excel 打开，共分为 7 列，在最后一列中，不同的 pathway 之间用相应的 pathway 编号隔开，在每一个 pathway 中，都分为三个层级，第一层级为六大代谢通路，第二层级为 43 种子通路，第三层级则为具体的通路信息。在该文件中，各列所代表的含义如下：

列数	列标题	说明
1	Query_id	基因的 ID 号
2	Kegg_geneID	比对上的序列 ID 号也即 KEGG 中的 geneID 号
3	KO_ID	该序列所对应的，在 KEGG 数据库中的 Orthology ID 号
4	KO_NAME	该 Orthology ID 所对应的名称
5	KO_DEFINITION	该 Orthology ID 所对应的描述
6	KO_EC	该 Orthology ID 所对应的 EC 编号
7	KO_PATHWAY	该 Orthologous id 所对应的 pathway 信息

| | |-- **Unigenes.blast.m8.filter.xls ——【过滤后的 blast 结果文件，Blast 软件的 m8 格式】**

| | |-- **Unigenes.level1.bar.{png|svg} ——【KEGG 第一层级上的相对丰度统计图】**

| | |-- **Unigenes.level1.bar.tree.{png|svg} ——【KEGG 样品聚类分析结果】**

| |-- **KEGG_MAT ——【KEGG 相对丰度和绝对丰度分析结果：ec 为酶，ko 为直系同源】**

| | |-- **Absolute ——【各样品在 ec，ko,level1，level2 和 level3 均一化后的绝对丰度矩阵】**

在该文件夹中，一共有五个文件，可以用 excel 打开这些文件，其中 Unigenes.absolute.level1.xls 是对第一层级的六大生物代谢通路在各样品中的绝对丰度进行的统计，在该文件中，第一行为样品信息，第一列为第一层级六大生物代谢通路的描述，其余的数字则代表某个代谢通路在某个样品中的绝对丰度。

在 Absolute 文件夹中，存在的另外三个文件 Unigenes.absolute.level2.xls、Unigenes.absolute.level3.xls 、Unigenes.absolute.ko.xls 和 Unigenes.absolute.ec.xls 的展示形式和 Unigenes.absolute.level1.xls 文件的展示形式是一样的，不同的是，Unigenes.absolute.level2.xls 是对子通路在各样品中的绝对丰度进行的统计，Unigenes.absolute.level3.xls 是对代谢通路图在各样品中的绝对丰度进行的统计，Unigenes.absolute.ko.xls 是对 Orthologous groups 在各样品中的绝对丰度进行的统计，而 Unigenes.absolute.ec.xls 则是对所有能够注释上 KEGG 数据库的基因所属的酶，在各样品中的绝对丰度进行的统计。

| | `-- **Relative ——【各样品在 ec，ko，level1，level2 和 level3 相对丰度矩阵】**

在该文件夹中，一共有五个文件，可以用 excel 打开这些文件，其中 Unigenes.relative.level1.xls 是对第一层级的六大生物代谢通路在各样品中的相对丰度进行的统计，在该文件中，第一行为样品信息，第一列为第一层级六大生物代谢通路的描述，其余的数字则代表某个代谢通路在某个样品中的相对丰度。

在 Relative 文件夹中，存在的另外三个文件 Unigenes.relative.level2.xls、Unigenes.relative.level3.xls 、Unigenes.relative.ko.xls 和 Unigenes.relative.ec.xls 的展示形式和 Unigenes.relative.level1.xls 文件的展示形式是一样的，不同的是，Unigenes.relative.level2.xls 是对子通路在各样品中的相对丰度进行的统计，Unigenes.relative.level3.xls 是对代谢通路图在各样品中的相对丰度进行的统计，Unigenes.relative.ko.xls 是对 Orthologous groups 在各样品中的相对丰度进行的统计，而 Unigenes.relative.ec.xls 则是对所有能够注释上 KEGG 数据库的基因所属的酶，在各样品中的相对丰度进行的统计。

| |-- **MetaStats ——【各样品的 MetaStats 统计结果】**

| |-- **pathwaymaps ——【代谢通路比较结果】**

| `-- **PCA ——【各样品的 PCA 分析结果】**

`-- **ReadMe.pdf ——【交付结果目录说明】**