诺禾致源
元基因组交付目录说明手册
（V5.0）

诺禾致源微生物事业部

2018 年 2 月 25 日

目录

（注：单击即可跳转至相应文档的详细说明）

|-- 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：

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

| | `-- *.CDS.fa.len.{png|svg|txt} ——【基因碱基序列长度分布统计表及图，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。

| | `-- *.CDS.fa.integrity.stat.xls ——【预测基因的完整性统计信息】

| | `-- *.mgm.gff ——【基因预测结果文件，gff 格式】

| |-- **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|xls} ——【去冗余后的基因核苷酸 FASTA 序列的长度分布统计表及图，**

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.CDS.cdhit.fa.integrity.stat.xls ——【去冗余后的基因完整性统计表】**

文件格式与*.CDS.fa.integrity.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】**

Unigenes.proteinin.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

| | `-- **Unigenes.protein.cdhit.fa.len.xls ——【去冗余后的基因氨基酸序列的长度分布统计表】**

| |-- **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_flower ——【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】

| | |-- Sample ——【各样品的基因 Reads Mapping 结果，文件夹以样品名称来命名】

| | |------ coverage_depth.png——【覆盖深度分布图，png 格式】

| | |------ coverage_depth.svg——【覆盖深度分布图，svg 格式】

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

| | |-- Total ——【基因丰度表文件夹】

| | |-- 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。

| | |-- Unigenes.readsNum.screening.fa ——【按照 reads 数目进行过滤后的基因序列文件，FASTA 格式】

`-- 03.GenePredict--ReadMe.pdf ——【 03.GenePredict 交付结果目录说明】