诺禾致源
元基因组交付目录说明手册
（V4.2）

诺禾致源微生物事业部

2017 年 05 月 01 日

目录

（注：单击即可跳转至相应文档的详细说明）

|-- 02.Assembly —— 【元基因组组装结果】

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

该文件即对应的具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，Scaftigs 的编号

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

`-- 02.Assembly--ReadMe.pdf ——【 02.Assembly 交付结果目录说明】