

诺禾致源微生物单菌产品 FAQ

一、质控及组装	7
1.1 QC 部分	7
1.1.1 碱基含量分布图怎么解读？	7
1.1.2 Clean data 的 GC 含量与 Sequence GC 含量差别较大？	7
1.1.3 rawdata 到 cleandata 的过滤条件？	7
1.1.4 Read1 和 Read2 长度为什么不一致？	8
1.1.5 Q 值是什么，什么是 Q20 和 Q30	8
1.1.6 Error rate 图中的数据是怎么算出来的？	9
1.1.7 测序接头序列引物信息？	9
1.2 Survey 部分	9
1.2.1 什么是 K-mer，有什么作用？	9
1.2.2 为什么基因组大小会估计不准？	10
1.2.3 survey 中哪些情况显示有污染发生？	11
1.2.4 为什么真菌精细图要先做 survey？	11
1.2.5 Kmer 图中，有小峰的存在是否是样品污染？Kmer 两个图中，那些小峰是否是样品污染？还是正常范围内的误差？	11
1.2.6 all_Kmer.stat.xls 文件中 Heterozygous Rate 的含义	12
1.3 组装及评估	12
1.3.1 组装结果中的 N50 和 N90 是什么？	12

1.3.2 GC-depth 图是怎么做出来的？有什么意义？	12
1.3.3 GC 含量与测序深度（Depth）关联分析统计图中，出现多个中心的情况有哪些？原因是什么？	13
1.3.4 为什么有污染混杂的情况下得不到好的组装结果？	14
1.3.5 测序覆盖度与测序深度的区别	14
1.3.6 拼接中是否有 gap，如果有，采取了什么方式补？	16
1.3.7 对于框架图中有污染的情况怎么处理？	17
1.3.8 如何寻找测序菌株的质粒信息？	17
1.3.9 为什么完成图样品有的质粒可以成环，有的不成环呢？质粒是如何确认的呢？	17
1.3.10 细菌框架图组装过程	18
1.3.11 细菌完成图二代矫正三代过程	18
1.3.12 细菌完成图（PacBio）组装	18
1.3.13 细菌完成图（Nanopore）组装	20
1.3.14 真菌精细图组装过程（PacBio）组装	21
1.3.15 使用 MIT（MIT 软件：MITObim；版本：version 1.8）组装的线粒体小基因组，是否成环？	22
1.4 基因组分及功能注释	22
1.4.1 Genemark 不是用来做真核生物注释的么？能否用于细菌基因注释？	22
1.4.2 真菌基因预测的三种方法（从头预测、同源预测及基于转录组数据预测）？	22
1.4.3 如果老师关心的基因没有被注释出来，原因是什么？	23

1.4.4 关于 ncRNA 注释，为什么注释不到 5S/16S/23S 的序列？	24
1.4.5 为什么在 genbank 找不到基因？	24
1.4.6 如何在 KEGG 注释结果里查找某个特定的功能基因？	25
1.4.7 KEGG 数据库比对时用 EC 号和 KO 区别？	25
1.4.8 在功能注释结果中，Identity、Evalue 和 Score 的区别？	26
1.4.9 将基因组的 KO 号输入 KEGG 网址分析，为什么有的基因找不到？	27
1.4.10 KEGG 数据库的 map 图中框出来的代表什么，不同颜色表示什么意思？	27
1.4.11 KEGG 图中的代码知道酶的名字？	28
1.4.13 KEGG 代谢通路图，怎样筛选特定基因组，例如 mcr-1 基因相关通路	29
1.4.14 从哪里可以知道注释的每个基因的具体信息？	31
1.4.15 如何从结果文件中找到关注的基因的氨基酸序列和核苷酸序列？	31
1.4.16 完成图如何确定起始位点？	32
1.4.17 完成图甲基化修饰可以提供的信息？	32
1.4.18 次级代谢产物基因簇注释分析中为什么没有 PKS（聚酮合酶）和 NRPS（非核糖体肽合成酶）结构？	33
1.4.19 次级代谢产物分析流程	33
1.4.20 单菌产品各数据库版本号	33
1.4.21 注释的 pathway 中，KO ID 无法在 anno 文件中找到的原因	34
1.4.22 为什么蛋白序列中起始氨基酸都不是 Met，并且这些蛋白都截短了？	35
1.4.23 P450 基因及 CAZy 鉴定及注释的方法	35

1.4.24CAZy 功能基因丰度表中，不同 EC 编号 ID 的丰度是一样的，怎么解释：.....	35
1.4.25 CDS 文件详细解读.....	36
1.4.26 前噬菌体.....	36
1.4.27 Go.xls 文件中，为什么第一排也有几行是基因的名称？对应的应该是哪一类？.....	37
1.4.28 基因岛在线预测网站.....	37
1.4.29 真菌精细图圈图展示详细说明.....	38
1.5 重测序部分.....	38
1.5.1 选取参考序列的有什么要求？.....	38
1.5.2 重测序的测序深度及覆盖度统计中为什么只用 100x 的数据来分析，这样能代表整体水平吗？.....	39
1.5.3 如何在变异检测中用 gene ID 查找基因的具体信息？.....	39
1.5.4 为何在.SV_filt.ctx.xls 文件中 pos1 与 pos2 位点的差值与实际缺失的长度不一致？.....	40
1.5.5 重测序 SV 部分 Orientation1、Orientation2 和 num_Reads 有什么关系?.....	41
1.5.6 如何找到可信的 SV？.....	41
1.5.7 重测序产品中为什么能得到插入/缺失了碱基的数目，却得不到插入/缺失的具体位置与序列？有没有什么办法可以知道具体序列？.....	41
1.5.8 Coverage (%) 和 Genomics(%)的含义.....	42
1.5.9 某一基因片段没有 reads 的原因？.....	42

1.5.10 如何确定感兴趣基因上是否发生突变？	43
1.5.11 重测序中除了 Qs core , 还有什么参数可以 ascertain quequality of the data 呢？	43
1.5.12 重测序中 SNP 的分析步骤	44
1.5.13 出现在 OR 和 special 表格中的具体的基因怎么找？	44
1.5.14 重测序中相应名称解释	45
1.6 结果文件相关	45
1.6.1 数据应该如何下载？为什么下载不下来？	45
1.6.2 数据应该如何打开？	46
1.6.3 M8 文件怎么打开、怎么解读？	46
1.7 高级分析相关	47
1.7.1 系统进化树构建方法	47
1.7.2 系统进化树构建的三种方法	47
1.7.3 系统进化树解读	48
1.7.4 进化树分支可信度解释问题	49
1.7.5 进化树常见售后	50
1.7.6 共有和特有基因分析中花瓣图/韦恩图为何和与表中统计数字的不一致？	51
1.7.7 cgMLST 前期准备	51
二、运营增强版	57
2.1 标准分析	57
2.1.1 线粒体和叶绿体注释有什么特殊要求？为什么有时候无法进行注释？	57

2.1.2 真菌精细图测序中，61 个 contig 中，为什么只有 18 个 contig 得到注释？	58
2.1.3 重测序一代和二代有出入的回复模板	58
2.1.4 预测得到的基因岛的功能信息可否提供一下，或者告知通过何种途径能够知道基因岛的功能	59
2.1.5 细菌完成图中重复序列数据详细说明	61
2.1.6 覆盖率和覆盖深度的问题	62
2.1.7 针对于革兰氏阳性菌，TNSS 没有 T3SS，而在 T3SS 预测中却有很多，是怎么回事。	63
2.1.8 Pacbio 下机数据相关	63
2.2 高级分析	65
2.2.1 共线性分析的意义及如何解读	65
2.2.2 比较基因组中查找 SNP 的方法。	65
2.2.3 core.cog.anno.xls 文件、core.cog.class.catalog.xls 的共有基因及 Venn 图中共有的数目为什么不同？	65
2.2.4 （毒力、耐药、转座）、还有 core-pan 和进化树分析，分析过程和软件版本。	66
2.2.5 NCBI 上传问题	69
2.2.6 关于组装结果评价	71
2.2.7 BUSCO 评估	71
2.2.8 关于 L50 计算	72

一、质控及组装

1.1 QC 部分

1.1.1 碱基含量分布图怎么解读？

碱基含量分布图是展示 GC 随 reads 读长不同位置的分布比例变化，不同颜色曲线分别表示了 A, T, G, C 及 N 的比例。由于采用了随机 PCR 扩增和双端测序，因此，AT 比例之间和 GC 比例之间大体上应该是一致的，不过头端由于受到引物连接的一些偏好性影响，可能有一定的波动。

1.1.2 Clean data 的 GC 含量与 Sequence GC 含量差别较大？

首先 GC 含量指的是一种生物的基因组或特定 DNA、RNA 片段有特定的 GC 含量。

Clean data GC%是每个 reads 位置处 GC 的比例，是对所有 reads 计算得到的。

这种情况一般会发生在有污染的样品上。Clean data GC%是每个 reads 位置处 GC 的比例，是对所有 reads 计算得到的。Sequence GC%是组装好的序列中 GC 的比例。

如果样品没有污染，PE reads 测序随机的情况下，两者相等。但如果样品有污染，前者计算时包含污染 reads 在内，而组装后的序列就比较复杂，可能含有污染的序列，比例不确定，所以两者可能不等。

1.1.3 rawdata 到 cleandata 的过滤条件？

二代数据质控用的是 readfq.v10 软件。老师拿到的数据质量都是我们采用软件过滤过的，已经将低质量的碱基过滤掉。reads 中的 fq 文件中的第 4 行代表每个碱基的质量。过滤条件如下：

1：过滤含 N 比例高于 10%的 reads。

2：过滤掉碱基质量值 20 以下比例高于 40%的 reads。

3：过滤掉 duplication (完全一样的 PReads)。

4：过滤掉含测序接头的 PE reads (15 个碱基比对到接头序列，mismatch 为 3)

1.1.4 Read1 和 Read2 长度为什么不一致？

由于测序技术上的原因，尾端的部分碱基质量可能会有一定下滑，特别是对于部分 GC 较高的样品，下滑可能比较厉害，为保证组装的质量，我们会对平均质量较低的部分 Reads 尾端进行截取去除（一般会选择 Q 值 10 以下，即平均错误率 10%），而 Read1 质量一般好于 Read2，Read2 截取会更多一些，因此一般截取后 Read1 会长于 Read2。

1.1.5 Q 值是什么，什么是 Q20 和 Q30

Q 值是 Illumina 在做碱基测序过程中，从测序原始数据转换为碱基的过程中评估出的质量分数取整的结果，e 为错误率，则质量分数为 $-\lg e$ ，对于通常质量较高的碱基，可以近似看作，反推质量值的话则是，以 Q 值 20 为例，折合错误率约为 0.01，Q 值 30 时错误率则为 0.001。而 Q20 和 Q30 则是 Reads 中 Q 值高于 20 或 30 的碱基所占的比例，反映了整体测序的质量情况。

1.1.6 Error rate 图中的数据是怎么算出来的？

我们计算的 Error rate 是单个碱基位置错误率的期望值，采用的是对数平均，简单的计算方法可以对该碱基位点的 Q 值做平均，并折算回错误率。如果 Reads 某位置平均 Q 值是 20 的话，对应的错误率就是 0.01。

1.1.7 测序接头序列引物信息？

测序接头序列引物信息如下：5'通用接头全长引物

5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3'

(红色部分为与 flowcell 结合的 p5 部分，绿色为 read1 引物部分)

3' 接头全长引物 (index1)

5'-CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT-3'

(红色为与 flowcell 结合的 p7 部分，绿色为 read2 引物匹配部分 read2 引物部分，蓝色为 index1 的序列，其他 index 序列略有差异)

1.2 Survey 部分

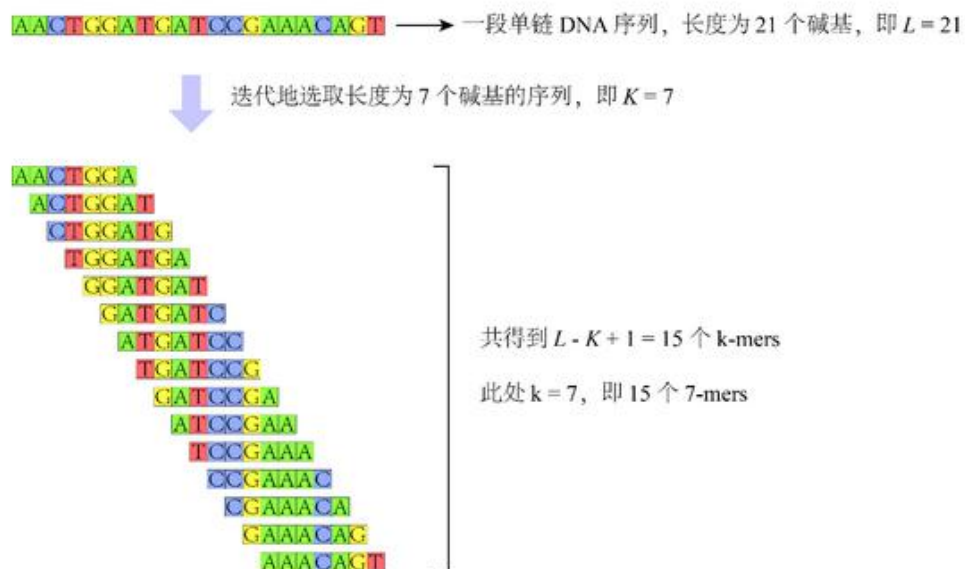
1.2.1 什么是 K-mer，有什么作用？

K-mer 就是一个长度为 K 的 DNA 序列，K 为正整数。如 K=17, 则称为 17-mer。K-mer 有多种用途，用于纠正测序错误、构建 contig 以及估计基因组大小、杂合率和重复序列含量等。假设测序 read 读长为 L, 则一条 read 上能取出 L-K+1 个 K-mer。用 K-mer 估算基因组大小时：基因组的大小 = K-mer 总数 / K-mer 期望深度。如下图所

示，假设这里存在某序列长度为 21，设定选取的 k-mer 长度为 7，则得到 $(21 - 7 + 1 = 15)$ 个 7-mers。

K-mer 分析的一个详细链接供老师参考 <http://blog.sciencenet.cn/blog-3406804-1162384.html>

对于微生物来说选 kmer 为 15 已经足够了，其原因是 ATCG 四种不通过的碱基组成的长度为 15 的核苷酸片段有 4 的 15 次方可能，足以覆盖微生物的大小。由于 Reads 上存在错误碱基，K-mer 并非越大越好，若 K-mer 选择的越大，则包含这个错误位点的 K-mer 的个数就会越多。



1.2.2 为什么基因组大小会估计不准？

由于我们估计基因组大小的数学模型，是基于鸟枪法测序随机抽样的模型，需要基因组大小较大（一般 1M 以上，最低 100k），测序覆盖深度足够深（一般要 50X 以上，最低 30X）才能近似符合数学模型近似假设的要求，对于一些较小的基因组，或

者覆盖深度不足の場合，由于模型近似假设已经不准确，估计的基因组大小会有偏差。另外，外源序列污染，也可能导致基因组大小估计发生偏差。

1.2.3 survey 中哪些情况显示有污染发生？

一般典型污染的特征有：kmer 图找不到峰，kmer 图中出现多个峰，估计的基因组大小明显高于预计大小等。出现这些情况，通常有很大可能是有外源 DNA 污染情况，需要结合后续初步组装及 GC-depth 和 NT 库比对来确定污染情况。

1.2.4 为什么真菌精细图要先做 survey？

真菌的基因组比较复杂，所以要先进行 survey，通过 survey 的结果，来判断是否加测大片段文库进行下一步组装。主要评估标准是杂合率，样品是否污染等，其中杂合率对基因组的组装影响很大。一般要求不超过 0.5%。

一般典型污染的特征有：kmer 图找不到峰，kmer 图中出现多个峰，估计的基因组大小明显高于预计大小等。出现这些情况，通常有很大可能是有外源 DNA 污染情况，需要结合后续初步组装及 GC-depth 和 NT 库比对来确定污染情况。

1.2.5 Kmer 图中，有小峰的存在是否是样品污染？Kmer 两个图中，那些小峰是否是样品污染？还是正常范围内的误差？

当基因组较小时，如果用抽样模型来解释鸟枪法测序的话，抽样的样本总体和抽样次数都明显低于理想值，结果会明显偏离正常的统计分布（这里是泊松分布），同时会

由于频率的波动导致峰的曲线不平滑，并不是小峰，属于正常现象，与样品污染无关。

那些小峰不是样品污染，都在正常范围内。敬请放心。那些小波动是由于基因组较小导致的，完全正常，与污染无关。

1.2.6 all_Kmer.stat.xls 文件中 Heterozygous Rate 的含义

all_Kmer.stat.xls 文件中，Heterozygous Rate：它指的是估计的基因组杂合比例。

杂合率高的话菌较难组装，杂合率一般要求不超过 0.5%，杂合率为“0”表明菌株无杂合情况。

1.3 组装及评估

1.3.1 组装结果中的 N50 和 N90 是什么？

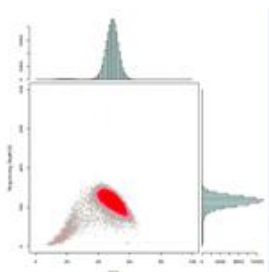
N50 和 N90 是基因组组装中的常用组装指标，其含义为，将序列按照长度从大到小排列，依次计算大于该序列长度的序列总长，找到序列总长度刚好大于基因组总长度的 50% (90%) 位置，该序列的长度即定义为 N50 (N90)。该数值反映了基因组 50% (90%) 以上的区域，都能被该数值以上长度的序列覆盖，体现了组装对于后续分析的质量贡献。

1.3.2 GC-depth 图是怎么做出来的？有什么意义？

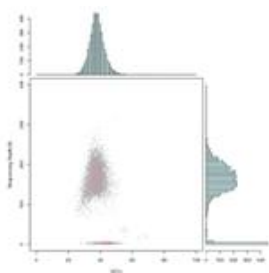
GC-depth 图是表征整个基因组 GC 和深度分布的关系，具体方法是用一定长度为单位对基因组进行切分，每个窗口都有特定的 GC 和 Reads 覆盖深度，对应图中的一个点。对于特定较纯的样品，其 GC 和深度会集中在某个区域，并向四周弥散，距离越远能找到的样本点越少。而如果 GC-depth 图分开成了多个集中区域，一般意味着该组装结果中包含来自不同来源的 DNA，特别是 GC 层面上如果分开的话，有外源污染可能性很大。而有时候，GC 不分离，仅深度分离的场合，也有可能是部分来自质粒的 DNA，需要结合其他信息，如 NT 比对结果来具体分析。

1.3.3 GC 含量与测序深度 (Depth) 关联分析统计图中，出现多个中心的情况有哪些？原因是什么？

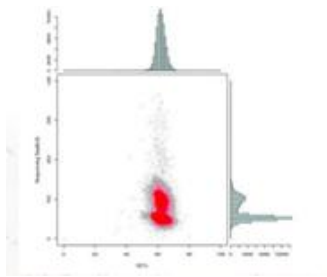
1) 真菌内可能在线粒体，正常。



2) 少量杂菌污染，可尝试分离。



3)近缘物种污染，较难分离干净。



1.3.4 为什么有污染混杂的情况下得不到好的组装结果？

由于组装软件在组装过程中是将测序数据看作来自同一个基因组的前提下进行的，而如果有外源 DNA 混杂，其中不同来源的 DNA 中会有不同程度的相似性序列和非相似性序列，这些复杂的关系会对组装软件产生干扰，而软件为保证组装的准确性，只能将可疑的部分切断成不同的碎片序列，而这也导致最终的组装只能拿到碎片化的序列，而失去了组装本身想要达到的效果。

如果能够找到足够近缘的参考基因组用于污染分离，是可以对上述的结果进行一定程度的改善的，不过受限于本身外源 DNA 可能带来的相似序列，及目标基因组和参考基因组间的潜在差异，分离是有一定的假阳性和假阴性的，因此无论如何，分离后的组装是不可能达到纯净 DNA 的标准的。由于受到污染情况和参考基因组的诸多限制，我们不对这样的样品组装做出结果承诺。

1.3.5 测序覆盖度与测序深度的区别

(1) 测序深度是指测序得到的总碱基数与待测基因组大小的比值。假设一个基因大小为 2M，测序深度为 10X，那么获得的总数据量为 20M。

覆盖度是指测序获得的序列占整个基因组的比例。由于基因组中的高 GC、重复序列等复杂结构的存在，测序最终拼接组装获得的序列往往无法覆盖所有的区域，这部分没有获得的区域就称为 Gap。例如一个细菌基因组测序，覆盖度是 98%，那么还有 2% 的序列区域是没有通过测序获得的。

(2) 测序覆盖率和测序深度是两个不同的概念。我们保证数据量足够，保证测序深度，也就是测序数据量与预测基因组大小的比值为 100 倍，在报告中展示为 100X。合同中没有 coverage 这一项。如果想知道的话，可以作为售后。

详细版介绍：

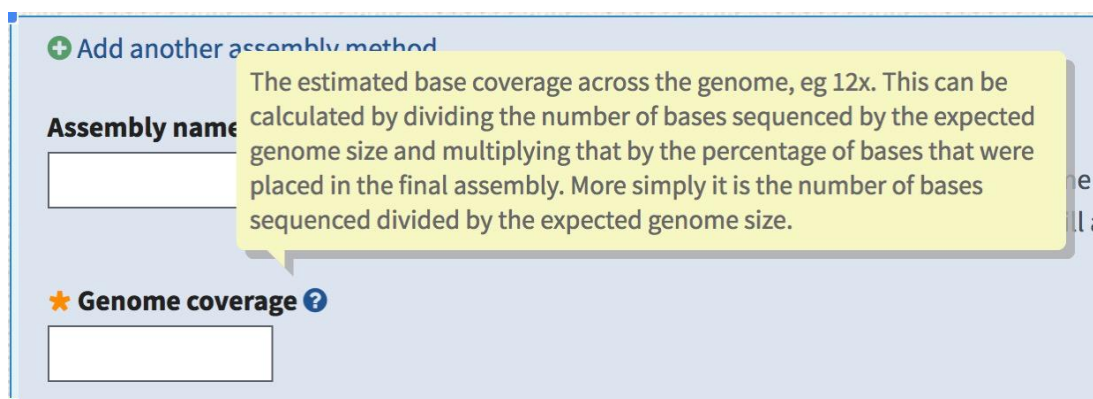
深度就是 reads 长度 * 多少条，得到的测序的数据量，也就是测序 1G / 2M 基因组 = 500X；

框架图的覆盖度不是指 scaffold 之间的 gap，而是指：组装好的 scaffold 序列，用组装前的原始 reads 去比对组装好的 scaffold 序列，如果 scaffold 上的某个碱基没有 reads 能比对上（可能是测序错误 / 组装拼接的过程中由于重复区等原因导致碱基错误 / 或是 N），那就认为这个碱基没有覆盖，然后用 scaffold 总碱基数减去没有覆盖的碱基数，然后去比上总碱基数，得到的覆盖度，因为错误的个数较少，一般情况在 98% 及以上，多数是 99.9%，所以可以默认为 100%，极个别特殊菌株除外；

完成图 3 代数据组装完成之后，会进行数据矫正，并且还会再使用 2 代测序数据进行矫正，正确率几乎在 99.999%，所以默认为是 100%；

以上是我们分析过程中的 coverage 含义，常规大家理解的 coverage 意思是组装出的序列与完成图参考基因组做 mapping，得到覆盖度，这个需要有参考基因组才能实现；

数据上传的部分有 Genome coverage，由于中英文的翻译问题导致多数认为是基因组覆盖度，实际是用测序深度值来填写，可以看 NCBI 中 Genome coverage 上的问号（黄色底文字部分）。



+ Add another assembly method

Assembly name

★ Genome coverage ?

The estimated base coverage across the genome, eg 12x. This can be calculated by dividing the number of bases sequenced by the expected genome size and multiplying that by the percentage of bases that were placed in the final assembly. More simply it is the number of bases sequenced divided by the expected genome size.

1.3.6 拼接中是否有 gap，如果有，采取了什么方式补？

对于非完成图的项目，目前的组装结果还没有组装到完成图，所以是有 gap 也是正常的。具体 gap 信息可以查看结果文件 02.Assembly/NO/NO.gapInfo.xls，02.Assembly/Assembly.readme.doc 这个文件有对 02.Assembly/NO/NO.gapInfo.xls 的说明。如果要做到无 gap，即完成图，可以通

过加测更大 insert size 的数据用于连接 scaffold，加以在洞两旁设计引物，Sanger 测序方法把洞中的序列测序出来用于补洞，但这个工作比较耗费时间和人力。

1.3.7 对于框架图中有污染的情况怎么处理？

一般我们会进行污染评估，同时反馈组装结果能否有效分离。对于可以有效分离的场合，经老师认可，可以作为售后，尝试分离并注释，但是我们无法保证序列分离的准确性和完备性。

1.3.8 如何寻找测序菌株的质粒信息？

老师可以在 NCBI 上使用 blast 工具进行 nt 库比对，

(http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome)，从比对结果可以得知质粒信息。

1.3.9 为什么完成图样品有的质粒可以成环，有的不成环呢？质粒是如何确认的呢？

我们分析样品基因组的测序深度发现：染色体的 reads 测序深度在 100x 左右，成环质粒的测序深度在 80x 左右，而不成环质粒的仅在在 20-40x 左右。所以，很可能是因为这些样品的质粒拷贝数少（这与质粒本身的稳定性有关，可能发生丢失），导致质粒的测序深度没有达到足够的乘数，因此质粒组装没有成环。

对于质粒的确认过程：首先将这些较短序列与所有基因组序列做比对，比对上的序列如果没有超过本身长度的 50%则保留，超过 50%的序列去除掉短的序列，保留下来的短序列，提交到 NCBI 上去看能否比对上质粒序列，如果能比对上质粒序列，认为是质粒序列。

1.3.10 细菌框架图组装过程

目前，我们采用最优 kmer，用三种软件（soap、spades、abyss）对样品进行组装，之后用 CISA 软件对组装软件组装生成的结果进行整合和优化，得到最后的组装结果。

1.3.11 细菌完成图二代矫正三代过程

1. 以初步组装的结果为参考基因组，利用软件 bwa-0.7.8 将下机二代 reads 与组装序列进行比对，生成 bam 文件；
2. 利用脚本以及 samtools-0.1.19 对 bam 文件进行排序；
3. 利用软件 bcftools 将 bam 文件转为 raw.vcf 文件
4. 利用脚本对 vcf 文件进行过滤获得 fiter_vcf，过滤的条件是，碱基的最低质量值为 20；最低 reads 深度 4；最大 reads 深度 1000
5. 脚本处理，根据 fiter_vcf 结果将初步组装中相应的位置进行矫正；

1.3.12 细菌完成图（PacBio）组装

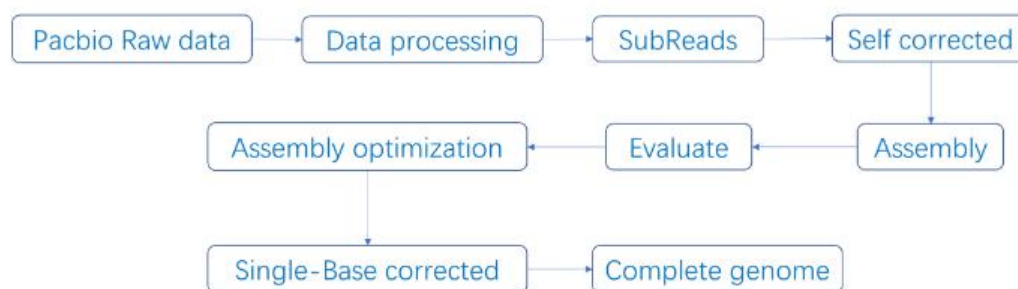


图 2 细菌完成图 (PacBio) 组装流程图

细菌完成图 (Pacbio) 具体组装步骤：

1、使用 SMRT Link v5.0.1 软件进行基因组组装：

1) 测序得到的原始数据存在一定比例的低质量 reads , 为确保后续信息分析结果的准确 , 首先要对原始数据进行过滤 (小于 500bp) , 得到 Clean data ;

2) 利用 SMRT portal 软件的自动纠错功能 , 挑选其中长的 reads (大于 6000bp) 作为 seed 序列 , 其他较短的 reads 通过 Blasr 与 seed 序列比对 , 进一步提高 seed 序列的准确度 , 组装得到一个能初步反应基因组情况的组装结果。

2、对初步组装的结果矫正：

使用 SMRT Link 软件的 variant Caller 模块 , 采用 arrow 算法矫正并统计初步组装结果中的变异位点。

3、二代数据矫正三代：

以矫正后的组装结果为参考基因组 , 利用软件 bwa 将下机二代 reads 与组装序列进行比对 , 过滤比对结果 , 过滤的条件 : 碱基的最低质量值为 20 , 最低 reads 深度 4 , 最大 reads 深度 1000 , 根据过滤 结果将初步组装中相应的位置进行二次矫正。

4、环化、起始位点矫正：

将二代矫正后的组装结果进行自身比对 , 查找首尾是否有 overlap , 基于 overlap 判断染色体序列是否组装成为一个环状基因组 , 然后与 DNAa 数据库比对 ,

进行起始位点矫正；之后进行比对分析（比对质粒数据库），筛分染色体与质粒序列，得到最终的完成图序列。

判定基于 overlap 判断染色体序列是否组装成为一个环状基因组的标准：

将组装好的序列，序列首部截取 2K 大小的片段与系列尾部 60k 比对，有大于 1.5k 的 overlap，就判定成环。

1.3.13 细菌完成图（Nanopore）组装

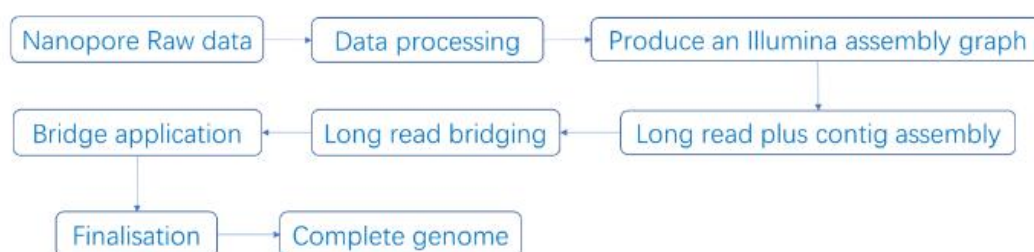


图 3 细菌完成图（Nanopore）组装流程图

细菌完成图（Nanopore）具体组装步骤：

从各样品质控后的有效数据出发，使用 Unicycler 软件对二代+三代数据进行基因组组装。

（1）初步组装：首先使用 SPAdes 将二代数据组装出一个框架图；

（2）创建桥接：然后使用 miniasm 和 Racon 将三代数据加到框架图上，使用三代的长读长数据创建桥接；

（3）桥接应用：出现冲突时，Unicycler 根据创建的桥接对应的质量分数选择高分数桥接；

（4）成环判断和最终确认：1.cotig 两端是否已经有 overlap；2.默认 dnaA 或 repA，通过 start_genes 进行起始位点矫正和最终确认，筛分染色体与质粒序列，并

将染色体序列组装成为一个环状基因组（如果是线性基因组即为线性基因组序列），即最终的 0gap 完成图序列。

1.3.14 真菌精细图组装过程（PacBio）组装



图 1 真菌精细图（PacBio）组装流程图

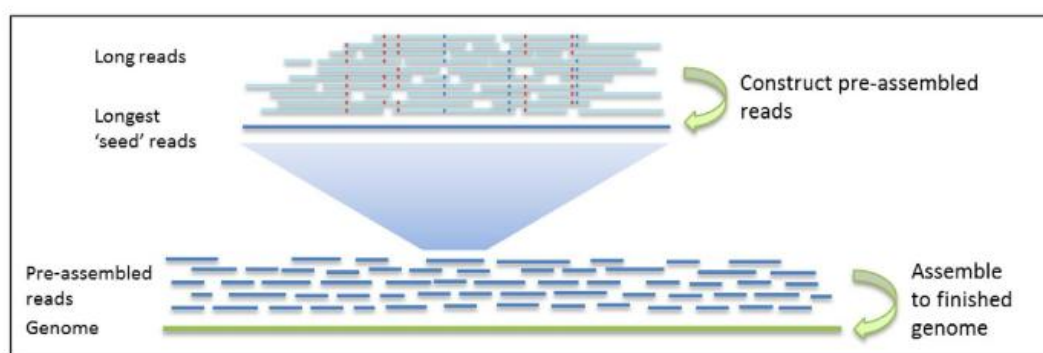


图 2 Self corrected 过程图

真菌精细图具体组装步骤：

1、初步组装：

使用 SMRT Link v5.0.1 软件进行基因组组装：

1) 测序得到的原始数据存在一定比例的低质量 reads，为确保后续信息分析结果的准确，首先要对原始数据进行过滤（小于 500bp），得到 Clean data；

2) 利用 SMRT portal 软件的自动纠错功能，挑选其中长的 reads（大于 6000bp）作为 seed 序列，其他较短的 reads 通过 Blasr 比对到 seed 序列，进一步提高 seed 序列的准确度，组装得到一个能初步反应基因组情况的组装结果。

2、对初步组装的结果矫正：

使用 SMRT Link 软件的 variant Caller 模块，采用 arrow 算法矫正并统计初步组装结果中的变异位点。

1.3.15 小基因组是否成环？

小基因组我们不承诺成环，因此我们这边没有判断是否成环的流程或标准，如果老师需要，可以将组装序列与 ref 进行 blast 比对的结果提供老师作为参考。

如果是序列完全一样，那么与参考是否成环的结果是一致的。

1.4 基因组分及功能注释

1.4.1 Genemark 不是用来做真核生物注释的么？能否用于细菌基因注释？

Genemark 包含了一系列软件，其中 GenemarkS 是含有 self-training 模块的版本，能够基于本身序列创建训练模型，其内置了包括细菌和病毒等不同的基因模型，而对应的真核生物由于含有 intron，模型比较复杂，对应的是另一个软件 GenemarkES，可以针对有 intron 的真核生物进行基因注释。

GenemarkS 是目前公认的注释细菌效果比较好的软件之一。

1.4.2 真菌基因预测的三种方法（从头预测、同源预测及基于转录组数据预测）？

从头预测使用 Augustus 软件预测，同源预测使用软件 Genewise。基于同源比对需要提供同物种的编码基因 cDNA 序列或者蛋白序列信息。而且越近缘对预测结果越好，非同属的物种信息几乎没有参考意义，反而会导致预测结果更差。所以真菌基因组项目客户最好能提供近缘物种的编码基因信息（cDNA 或 protein），或者组装好的转录本序列文件。如果老师不能提供或者没有同属物种编码基因信息或转录组信息，就只能还是基于之前的从头预测算法进行基因预测。

三种方法预测的结果将通过 EVM 进行整合，所以如果老师能提供近缘参考序列（已基因注释）和转录数据，结合三种方法的预测结果最好。

转录组数据预测 / 同源预测 / denovo 预测（基因预测）：提供哪个都可以做分析；

如果提供 2-3 种，3 者或 2 者的结果取交集，但结果中出现的比例是不一致的，转录 > 同源 > 基因预测。一般预测结果的蛋白序列都会以 Met 开头，但是如果比对结果出现不是 M 开头的情况，需要先查看老师提供的转录数据是否有降解情况或者参考基因组序列是否都是以 M 开头，二者均会影响比对结果（denovo 预测不会有影响），可尝试适当调节 3 者或 2 者之间的比例来改善。

1.4.3 如果老师关心的基因没有被注释出来，原因是什么？

- 1、这个基因没有被组装出来。
- 2、这个基因在目标基因组上不存在。
- 3、注释的数据库里没有这个基因，所以无法注释出来。
- 4、要测的这个菌株，根本就没有这个基因。

5、数据库版本低（咱们分析用的是本地数据库，会定期更新），请老师以最新版的为准。

1.4.4 关于 ncRNA 注释，为什么注释不到 5S/16S/23S 的序列？

上述情况在使用 denovo 方法预测 ncRNA 序列是出现的比较多，由于 denovo 预测 ncRNA，需要完整的 ncRNA 序列，才能确认 ncRNA 的结构，而由于 ncRNA，特别是 16S 和 23S 序列，往往本身就有一定的重复序列成分，在组装过程中很容易组装不完整，特别是框架图组装时，如果整条 rRNA 没有拼接成一条完整序列，是无法预测得到相应的 rRNA 序列的。

如组装较好，该样品对应的物种在数据库注释的少，所以虽然组装 scaffold 少但还是注释不到，不代表没有 rRNA。另外，如果精细图注释到 5S，没有注释到 16S，原因是：5S 较短，序列短比对到的可能性较大。

在一些真核新物种的项目中，会经常出现 18S 等数目为 0 的情况，这个是因为之前这个物种并没有进行过 18S 序列测序，所以数据库以及常用软件中没有收录该物种的 18S 序列，所以没有办法在组装结果中预测出 18s。

1.4.5 为什么在 genbank 找不到基因？

每一个测序菌株都会存在自己特异的基因，一般对于一个细菌来说，保守基因比率在 95%左右。所以有一部分基因在 genbank 数据库检索不到，这个属于正常现象。

其次编码基因在做同源比对的时候，往往选用蛋白与蛋白比对（blastp）来进行。因为不同菌株编码基因序列本身就会差异，密码子偏好性不同也会影响比对相似度，差异较多的话就会导致 blastn 比对不到。用基因氨基酸序列进行针对性的同源比对来进行检索。这样可以更准确的检索出相关基因。

1.4.6 如何在 KEGG 注释结果里查找某个特定的功能基因？

方法一：Result/04.Genome_Function/Anno Summary/***.AnnoSummary.xls 这个文件中去进行关键词检索，但是这个方法比较模糊。

方法二：通过 KEGG 数据库的注释结果，通过 EC 酶学编号来进行检索。比如找环己胺氧化酶，可以先在 KEGG 数据库网站 <http://www.kegg.jp/> 上检索 [cyclohexylamine](#) oxidase 关键词，然后它会告诉你对应的 EC 号，也就是 1.4.3.12，然后在 KEGG 的注释结果中，去检索这个 EC 编号，这样得到结果更加准确。

方法三：客户提供酶的基因的核酸或蛋白质序列，进行 blast 比对。

1.4.7 KEGG 数据库比对时用 EC 号和 KO 区别？

KEGG ORTHOLOGY 是同源基因数据库，表示方式为 K+5 位数字，一个 KO 表示一组同源的基因。如 K01623 (http://www.kegg.jp/dbget-bin/www_bget?ko:K01623)，如果一个 KO 是酶，往往也有对应的酶标号 EC:4.1.2.13；但一个 EC 号可能对应多个 KO 号，如下图。方框表示基因产物，经常为酶，故用 EC 号填充，当该基因产物没有对应的 EC 号时，用一个名称填充。所以 KO 统计注释到的基因个数更全面。

Entry	K01623	KO
Name	ALDO	
Definition	fructose-bisphosphate aldolase, class I [EC:4.1.2.13]	
Orthology	K01622 fructose 1,6-bisphosphate aldolase/phosphatase K01623 fructose-bisphosphate aldolase, class I K01624 fructose-bisphosphate aldolase, class II K11645 fructose-bisphosphate aldolase, class I K16305 fructose-bisphosphate aldolase / 6-deoxy-5-ketofructose 1-phosphate synthase K16306 fructose-bisphosphate aldolase / 2-amino-3,7-dideoxy-D-threo-hept-6-ulosonate synthase	

1.4.8 在功能注释结果中，Identity、Evalue 和 Score 的区别？

Identity 表示相似性，即序列的一致性。这个值越高，表示同源性越高，序列相似度越高，越有可能是行使相同功能的基因。这个值没有一个固定的范围，不过从经验来看，大于 80%可信度很高，小于 30%就几乎没有参考意义了。

Evalue 值是表示比对结果的可信度，是一个统计学的 P 值，用来进行判断这个比对结果是否可信。一般经验来看当 E 值小于 $1e-5$ 时，认为结果可信。E 值和 identity 没有关系，也不能换算。

比对的 Identity 是体现了比对区域相似性的高低，值越高相似性越高，但 Identity 的高低不能体现整体比对长度的大小，而 Evalue 值是结合序列长度计算出来的，所以在注释结果中，可以先根据 Evalue 的高低判断可靠性，Evalue 值越小可靠性越高。

用来挑选最佳比对结果的时候，往往是选用得分 score，得分 score 最高的那个比对结果是最相似最可信的。

有些 Identity 超过 80%，但 Evalue 是 0，这种应该怎么去理解？

答：这个 E-value 值实际上不是 0，是这个 E-value 极低，四舍五入为零，想知道哪个结果最可信，直接看 score 值就好，得分 score 最高的那个比对结果是最相似最可信的。这个 score 没有界定值，得分越高越好~

Very low e-values are rounded to zero. BLAST uses double float point representation, so e-values lower than approximately 5×10^{-324} are rounded to 0.

E-value 和 score 值该如何解读？

答：score 是打分矩阵计算出来的值，是搜索算法决定的，这个值越大说明你的序列跟目标序列匹配程度越大；

e 值是对比结果的期望值，解释是大概多少次随即比对才能出现一次这个

score, E-value 越小，表明这种情况从概率上越不可能发生，那么发生了即说明这更有可能是真实的相似序列。

1.4.9 将基因组的 KO 号输入 KEGG 网址分析，为什么有的基因找不到？

KEGG 库中注释到的基因，有一部分是参加代谢网络的或者有代谢通路图，可以在 KEGG 的 pathway 数据库中找到，但是有一部分基因是不参加代谢通路网络的，或者是 KEGG 的 pathway 数据库现有的代谢通路图中没有该基因参与的代谢通路图，这部分基因只能在 KEGG 的 gene 库中找到，不能在 pathway 数据库中找到。

1.4.10 KEGG 数据库的 map 图中框出来的代表什么，不同颜色表示什么意思？

有颜色的框表示在该样品中注释到了这个酶，不同颜色表示该样品注释到这个基因的个数，方框内为 EC 号，鼠标停留在框内可显示出对应的相关信息，点击可跳转到 KEGG 网站查看详细信息。也可在~.kegg.catalog.map.gene.xls 文件中对应各 map 编号查看注释到基因，对应 gene_ID 即可在 03.Genome_Component 文件夹的 Gene 中查找相应的序列及信息。相应图例会在结果文件中展示，color_direction.gif 示例如下，详细信息您可以查看 KEGG_KO.readme。框里面数字的颜色表示与疾病相关的基因。



1.4.11 KEGG 图中的代码知道酶的名字？

方法一：在 KEGG 主页 (<http://www.kegg.jp/>) 输入酶标号，如 5.1.1.3，进行搜索。

方法二：我们提供的 map 图的名称与 KEGG 数据库的 map 号是一致的，如 map00010.png。在 KEGG 主页 (<http://www.kegg.jp/>) 输入 “map00010 ”，可以搜索出这张 map 图，进而可以点击所有的方框查看对应酶的信息。（kegg 流程升级后，在我们提供的图上可以直接点击查看）。

1.4.12 map.result 表格里有一些通路在一株菌没有注释到基因，在另一株菌注释到了，但是在 map.ko 的表格里面没有看到这条通路。

这种情况出现的原因是因为 map351、472、901、5150，这些 map 中有和其它 map 共有的 KO，也就是说一个 KO 可以属于多个 pathway，所以差异的 KO 统计进了别的 map 中了，不过没关系，可以在结果文件 /Result/04.Genome_Function/General_Gene_Annotation/KEGG 这个文件夹中查看 *.kegg.catalog.map.gene.xls 这个文件，搜索相应的 map 号，然后看这个 map 中有哪些基因，哪些 KO（从第四列往后就是 KO 信息）。

1.4.13 KEGG 代谢通路图，怎样筛选特定基因组，例如 mcr-1 基因相关通路

首先，去以下目录 Result/04.Genome_Function/Anno Summary/-- *.gbk 中找到 [gbk 格式文件]；Gbk 格式文件是将预测基因的位置、ID、核酸序列和氨基酸序列以及注释到的功能信息整合的一种文件，格式类似于 NCBI 下载的 gbk 文件。文件部分展示如下：

```
LOCUS      Scaffold1          364339 bp    DNA      linear      07-JUL-2017
DEFINITION No definition line found.
ACCESSION
VERSION
KEYWORDS
SOURCE      Unknown.
ORGANISM     Unknown.
             Unclassified.
FEATURES             Location/Qualifiers
     source          1..364339
                     /organism="unknown"
                     /mol_type="genomic DNA"
     gene            3753..4079
                     /gene="14-128_GM000001"
                     /locus_tag="14-128_GM000001"
     CDS             3753..4079
                     /gene="14-128_GM000001"
                     /locus_tag="14-128_GM000001"
                     /codon_start=1
                     /product="conserved hypothetical protein; putative inner
                     membrane protein"
                     /translation="MKYIIIFLFRAIWLALSLILFFSMHRLSLLOSTRDVSSELISLMS
                     YGMMVICFPTGIVFFIALIFIGTVSDIIGVRIDSKYIMAIIWLYFLSGGYIQWFVLS
                     KRIINK"
```

图中 gene 对应基因 ID，product 对应基因名字，您可以先根据基因名字找到对应的基因 ID；

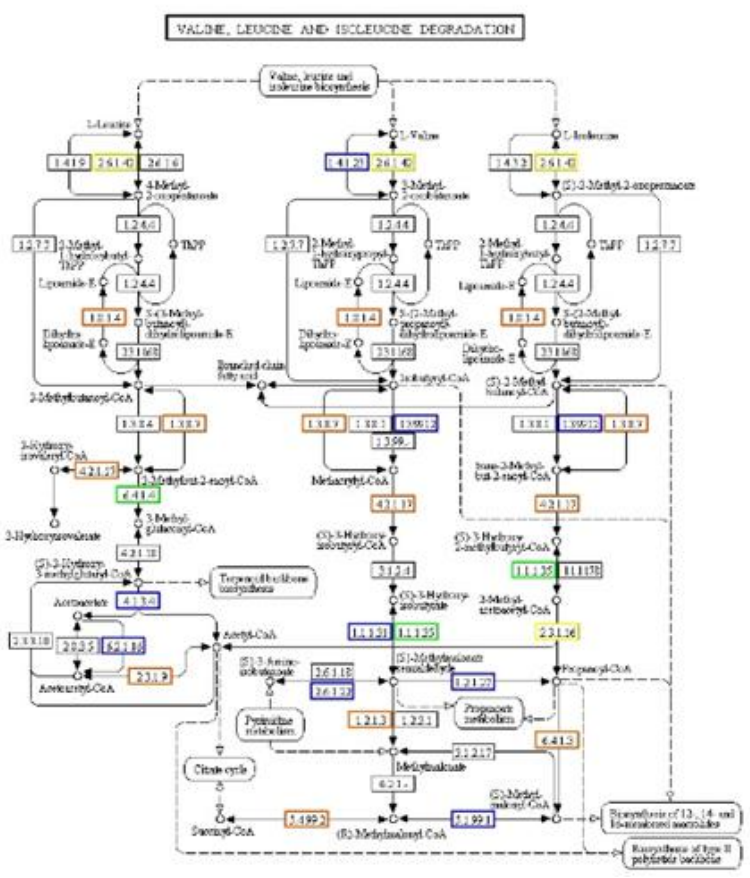
然后，Result/04.Genome_Function/KEGG/|-- *.kegg.anno.xls [KEGG 数据库注释的结果文件]，根据基因 ID 找到对应的 KO_id；

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Gene_id	Identity	E_value	Kegg_geneID	Ko_id	Ko_name	Ko_defi	Ko_EC	Ko_class						
2	APGM000001	43.29	1.00E-50	rfa:Rsa133209_1377	K00599	82.1.1.-	—	—	Metabolism; Amino Acid Metabolism; Histidine metabolism [PATH:ko00340]						
3	APGM000002	75	0	hrh:HRH_02560	K01687	ilvD	dihydroxy-4.2.1.9		Metabolism; Amino Acid Metabolism; Valine, leucine and isoleucine biosynthesis						
4	APGM000003	57.97	0	ahr:Arch_0427	K07085	K07085	putative tr--		Unclassified; Cellular Processes and Signaling; Other ion-coupled transport						

接下来，Result/04.Genome_Function/KEGG/ *.kegg.catalog.map.gene / [KEGG 代谢通路图注释上的基因信息]，根据 KO_id 找到对应的 map 号（KEGG 信号通路图的编号），每一个 map 号对应的就是一个信号通路图，可以

	A	B	C	D	E
1	map00010	Glycolysis / Gluconeogenesis	28	APGM001727, K00016, 1.1.1.27	APGM001313, K00121, 1.1.1.1 1.1.1.284
2	map00020	Citrate cycle (TCA cycle)	18	APGM000786, K00024, 1.1.1.37	APGM000989, K00031, 1.1.1.42
3	map00030	Pentose phosphate pathway	20	APGM001330, K00033, 1.1.1.44	APGM001805, K00033, 1.1.1.44

最后，Result/04.Genome_Function/KEGG/`-- *.kegg.map/ [KEGG 功能代谢通路图存放目录]，根据 map 号查看信号通路图。*对应通路编号，每个通路对应一个文件，用不同颜色标示相应基因的数目。color_direction.gif 为对应图例。



1.4.14 从哪里可以知道注释的每个基因的具体信息？

CDS 序列：在 03.Genome_Component/Gene/XXX.gmhmp.cds 文件中，为 fasta 文件（可以用 EditPlus 之类的文本编辑器打开）。

每一个“>”后面跟的是基因名字，后面的格式为：locus = Scaffold1:102:971，表示这个基因对应的 Scaffold 名字：Scaffold 起始位置：Scaffold 终止位置。

gff 文件：03.Genome_Component/Gene/XXX.gmhmp.gff 为预测基因的 GFF3 格式文件，可用文本编辑器打开，或扩展名改为 xls 后用 excel 打开。

从左至右各列依次表示：①seqid 序列的编号②“source”注释信息的来源③“type”注释信息的类型④“start”起始位置⑤“end”终止位置⑥“score”得分（数字，是注释信息可能性的说明，可以是序列相似性比对时的 E-values 值或者基因预测时的 P-values 值。“.”表示为空）⑦“strand”序列的方向（+表示正义链，-反义链，?表示未知）⑧“phase”仅对注释类型为“CDS”有效，表示起始编码的位置，有效值为 0、1、2。⑨“attributes”以多个键值对组成的注释信息描述，键与值之间用“=”，不同的键值用“;”隔开，一个键可以有多个值，不同值用“,”分割。注意如果描述中包括 tab 键以及“,=;”，要用 URL 转义规则进行转义，如 tab 键用 %09 代替。键是区分大小写的，以大写字母开头的键是预先定义好的，在后面可能被其他注释信息所调用。

详见：03.Genome_Component/Gene/Gene.readme.doc

1.4.15 如何从结果文件中找到关注的基因的氨基酸序列和核苷酸序列？

例如从 “*.kegg.anno ” 文件找到相关功能的基因 ID (文件的第一列) , 就可以从 “*.gmhmmmp.pep 和 *.gmhmmmp.cds” 文件找到对应的基因的氨基酸序列和核苷酸序列。

1.4.16 完成图如何确定起始位点？

我们通过 GC_skew 以及比对 dnaA 基因的上游序列来预测起始位点。

1.4.17 完成图甲基化修饰可以提供的信息？

DNA 的双链结构，编码核心就是带有 ATGC 四种不同碱基的脱氧核糖核酸。而实际上在生物体内很多时候 DNA 的碱基形式不是单纯的 ATGC 碱基，其中 A 和 C 两种碱基经常会存在一些甲基化修饰的现象，甚至会影响到全基因组的甲基化水平，进而影响到整个菌株的毒力性状表现等等。

由于甲基化修饰过的碱基，其同配对碱基结合的化学动力学会有所差异，在底物碱基被甲基化修饰时，我们在测序的时候检测到的测序信号就会发生改变。3 代测序可以检测到 N6-methyladenine (6mA) 和 4-methylcytosine (4mC) ，这是组成细菌甲基化组的两种主要修饰。

细胞不同生长周期下，不同位置的甲基化修饰起到了重要的辅助功能。DNA 的甲基化修饰会影响到其与配对碱基之间结合的化学动力学过程，进而使基因的转录及表达受到影响。所以可以反映基因组在表观层面的变化。

我们提供的结果包括两部分，一个是甲基化修饰位点类型 (6mA 或 4mC) 和位置，另外一部分是甲基化位点周围的 motif 基序类型。

1.4.18 次级代谢产物基因簇注释分析中为什么没有 PKS (聚酮合酶) 和 NRPS (非核糖体肽合成酶) 结构？

在进行分析时，分两步进行：首先，我们先预测是否存在 PKS (聚酮合酶) 和 NRPS (非核糖体肽合成酶)。其次，根据目前软件训练集中的基因簇的结构进行预测的，如果训练集中的基因簇中有匹配的结构就会预测出来，当然也就有预测不到的情况。也就是说出现没有 PKS (聚酮合酶) 和 NRPS (非核糖体肽合成酶) 可能是由于样本本身就没有这两种酶；或者是这两种的酶的结果与训练集中的结果不匹配导致。

1.4.19 次级代谢产物分析流程

次级代谢产物分析用到的软件为 antiSMASH (2.0.2) ,

其在线网址：<http://antismash.secondarymetabolites.org/>

参数：--clusterblast (BLAST identified clusters against known clusters) ,

可提交 gbk 文件。

1.4.20 单菌产品各数据库版本号

数据库名称	数据库版本号	更新时间	数据库用途简答描述
GO_CLASS	201705	201705	GO数据库
NRB_DB	201704	201704	细菌NR数据库
NRF_DB	201704	201704	真菌NR数据库
SWISSPORT_DB	201807	201807	SWISSPORT数据库
KEGG_DB	201807	201807	KEGG数据库
PHI_DB	20180515	20180515	PHI数据库
CAZY_DB	201604	201604	CAZy数据库
TCDB_DB	201807	201807	TCDB数据库
COG_DB	20151214	20151214	COG数据库
VFDB_DB	20180726	20180726	VFDB数据库
ARDB_DB	201604	201604	ARDB数据库
CARD_DB	1.1.9	20170913	CARD_DB
KOG_DB	20140126	20140126	KOG数据库
P450	201808	201808	P450数据库
DFVF	2012	2012	DFVF数据库
POG_DB	20140914	20140914	POG数据库
miRNA	20130805	20160426	miRNA预测
snRNA	20130805	20160426	snRNA预测
sRNA	20130805	20160426	sRNA预测
SignalP	Version 4.1	Version 4.1	分泌蛋白预测
TMHMM	Version 2.0c	Version 2.0c	分泌蛋白预测
antiSMASH	version 4.0.2	version 4.0.2	次级代谢产物预测

1.4.21 注释的 pathway 中，KO ID 无法在 anno 文件中找到的原因

KO 指的是 KEGG ORTHOLOGY GROUP，也就是我们常说的同源基因簇，在一个 KO 中，包含若干个基因(gene)，这些基因是行使同一个功能的。因此，这些行使同一功能的 gene cluster，就称之为 orthology group。一个 KO ID 可以参与多个 pathway，如果 KO 目前不能界定属于某个 pathway，就不能注释到 pathway 中，这是正常的。

1.4.22 为什么蛋白序列中起始氨基酸都不是 Met，并且这些蛋白都截短了？

分泌蛋白预测采用的软件是 SignalP，成熟的分泌蛋白是将信号肽（分泌蛋白的 N 端是由 15 ~ 30 个氨基酸组成的信号肽）切下的蛋白序列，所以看到的起始氨基酸都不是 Met，且蛋白序列长度相比于原始蛋白要短。

1.4.23 CAZY 基因及 CAZY 鉴定及注释的方法

以 CAZY 为例，鉴定方法为使用 BLAST 软件，把目标物种的氨基酸序列，与 CAZY 数据库（Carbohydrate-Active enZYmes Database）进行比对，把目标物种的基因和其相对应的功能注释信息结合起来，得到注释结果。由于每一条序列比对结果可能超过一条，为保证其生物意义，注释时保留一条最优比对结果作为该基因的注释。

1.4.24 CAZY 功能基因丰度表中，不同 EC 编号 ID 的丰度是一样的，怎么解释：

结果中此类 EC 编号丰度相同，主要是由于预测出来的这些 EC 的序列是相同的，由于存在属于同一 family 下的多个 EC 编号下的序列相同，所以同一基因序列与数据库比对会得到多个 EC 编号的注释结果。即这些 EC 实际上是同一条序列在数据库中的比对结果，所以丰度是一样的。

EC ID	H12a	H55a	H12wp	H12wf	H12wp	H12wf	H55wp	H55wf	H55wp	H55wf
Alpha-L-arabinofuranosidase (EC 3.2.1.55)	0.001908	0.001738	0.000783	0.001545	0.001198	0.001172	0.001136	0.001613	0.000952	0.001101
lipopolysaccharide N-acetylglucosaminyltransferase (EC 2.4.1.30)	0.001288	0.001509	0.000556	0.001590	0.001387	0.001789	0.00045	0.001162	0.001283	0.001145
beta-1,3-glucan synthase (EC 2.4.1.34)	0.001302	0.001401	0.000407	0.001031	0.001362	0.00168	0.000427	0.000887	0.001037	0.000929
N-acetylglucosaminyltransferase (EC 2.4.1.-)	0.001302	0.001401	0.000407	0.001031	0.001362	0.00168	0.000427	0.000887	0.001037	0.000929
UDP-Gal4: galactofuranosyl-galactofuranosyl-rhamnosyl-N-acetylglucosaminyltransferase (EC 2.4.1.12)	0.001302	0.001401	0.000407	0.001031	0.001362	0.00168	0.000427	0.000887	0.001037	0.000929
cellulose synthase (EC 2.4.1.12)	0.001302	0.001401	0.000407	0.001031	0.001362	0.00168	0.000427	0.000887	0.001037	0.000929
chitin oligosaccharide synthase (EC 2.4.1.-)	0.001302	0.001401	0.000407	0.001031	0.001362	0.00168	0.000427	0.000887	0.001037	0.000929
chitin synthase (EC 2.4.1.16)	0.001302	0.001401	0.000407	0.001031	0.001362	0.00168	0.000427	0.000887	0.001037	0.000929
UDP-L-Rha: N-acetylglucosaminyl-PP-decaprenyl alpha-1,3-4	0.001302	0.001401	0.000407	0.001031	0.001362	0.00168	0.000427	0.000887	0.001037	0.000929
delichyl-phosphate beta-D-mannosyltransferase (EC 2.4.1.8)	0.001302	0.001401	0.000407	0.001031	0.001362	0.00168	0.000427	0.000887	0.001037	0.000929
delichyl-phosphate beta-glucosyltransferase (EC 2.4.1.117)	0.001302	0.001401	0.000407	0.001031	0.001362	0.00168	0.000427	0.000887	0.001037	0.000929
hyaluronan synthase (EC 2.4.1.212)	0.001302	0.001401	0.000407	0.001031	0.001362	0.00168	0.000427	0.000887	0.001037	0.000929
beta-1,4-mannan synthase (EC 2.4.1.-)	0.001302	0.001401	0.000407	0.001031	0.001362	0.00168	0.000427	0.000887	0.001037	0.000929
beta-mannosylphosphododecaprenyl-mannosyloligosaccharide alpha	0.001302	0.001401	0.000407	0.001031	0.001362	0.00168	0.000427	0.000887	0.001037	0.000929
N-acetylglucosaminyltransferase (EC 2.4.1.-)	0.001302	0.001401	0.000407	0.001031	0.001362	0.00168	0.000427	0.000887	0.001037	0.000929
beta-glucosidase (EC 3.2.1.21)	0.000735	0.000884	0.000973	0.001351	0.000934	0.000973	0.001341	0.00149	0.000831	0.000949
beta-mannosidase (EC 3.2.1.25)	0.001368	0.001019	0.000806	0.000864	0.000481	0.000426	0.000913	0.000754	0.000383	0.000459

1.4.25 CDS 文件详细解读

CDS 为编码序列，通过基因预测软件，对测序后的 seq 文件进行预测，得到的 CDS 序列。在结果给出的 CDS 序列文件中可以看到 CDS 的起始位点和终止位点，以及位于 seq 文件中是正向还是反向（如下图所示），后面的 “-” 表示其与 seq 序列成反向互补（即 seq 序列位于该片段的序列的反向互补序列位于编码区），其余的序列位于非编码区。以下面的例子为例，测序后的 seq 在 484-1659 位置的序列没有预测到蛋白，但是 484-1659 位置的反向互补序列可以预测到蛋白，即下面的核酸序列是反向互补链的序列。细菌是通过 GeneMarkS

（<http://topaz.gatech.edu/GeneMark/>）对 seq 进行的预测得到的 CDS 序列，结果是软件自动输出。下图为示例

```
>GT13_GM000001 locus=Scaffold1:484:1659:-
ATCACTGCTCTGACCACTGCCCCAAGCAATCTCCATCACTTCCACTCTAT
```

1.4.26 前噬菌体

Prophage.stat.xls 文件中，没有前噬菌体说明没有预测到。细菌中，不同样本会出现不同的情况，没有前噬菌体也是正常的现象。

1.4.27 Go.xls 文件中，为什么第一排也有几行是基因的名称？对应的应该是哪一类？

GO 的全称是 Gene Ontology，其分为三大类：1) 细胞学组件 (Cellular Component)：用于描述亚细胞结构、位置和大分子复合物，如核仁、端粒和识别起始的复合物等；2) 分子功能 (Molecular Function)：用于描述基因、基因产物个体的功能，如与碳水化合物结合或 ATP 水解酶活性等；3) 生物学途径 (Biological Process)：用于描述分子功能的有序组合，达成更广的生物功能，如有丝分裂或嘌呤代谢等。

Go.xls 文件是 GO 注释结果按分类汇总的文件。表格中对应的信息如下：

行数	行标题	说明
1	Ontology	GO 本体学分类。
2	Class	GO 二级分类。
3	number_of_AP	相应功能上的基因数目
4+	Genes_of_AP	相应功能上的基因列表，基因间用分号分割。

如下示例 “genes of NLN102” 这列显示的基因与 “Class” 中的条目所对应，另 “Ontology” 指的是 GO 的三大类别。

Ontology	Class	number_of_NLN102	genes_of_NLN102	
----------	-------	------------------	-----------------	--

1.4.28 基因岛在线预测网站

基因岛在线预测网站：<http://www.pathogenomics.sfu.ca/islandviewer/upload/>

教程：<http://blog.sciencenet.cn/blog-3406804-1192278.html>

1.4.29 真菌精细图圈图展示详细说明

第一圈：最外圈是基因组序列位置坐标

第二圈：基因组 GC 含量：以窗口 200000 bp，步长 (200000) bp 来统计 GC 含量，向内的蓝色部分表示该区域 GC 含量低于全基因组平均 GC 含量，向外的紫色部分与之相反，且峰值越高表示与平均 GC 含量差值越大；

第三圈：基因组 GC skew 值：窗口 (200000) bp，步长 (200000) bp，具体算法为 $G-C/G+C$ ，向内的绿色部分表示该区域 G 的含量低于 C 的含量，向外的粉色部分与之相反；

第四圈：基因密度 (编码基因)

第五圈：基因密度 (rRNA)

第六圈：基因密度 (snRNA)

第七圈：基因密度 (tRNA)

第四圈-七圈 基因密度 (以窗口 200000 bp，步长 200000bp 分别统计编码基因、rRNA snRNA tRNA 的基因密度，颜色越深，代表窗口内的基因密度越大)

最里圈：染色体 duplication

真菌精细图是用 circos 软件来做的

1.5 重测序部分

1.5.1 选取参考序列的有什么要求？

参考基因组最好有核酸序列 (seq) 和基因注释信息 (cds 序列)，提供 NCBI 链接 (信息搜集表里有详细要求) 或直接提供序列。

用于 SNP，InDel 和 SV，共线性分析的参考序列，必须要有基因组的核酸序列和基因的注释信息。如果没有基因的信息，只有基因组序列，细菌的 core-pan 分析我们可

以帮助基因预测，但这部分分析老师最好提供基因信息。如果实在找不到有基因注释的序列，有另一种策略：因为我们这个样品的分析内容包括组装和注释，因此可以把我们的样品作为参考序列，其他两个都和我们的样品作比较。这样，基于同一个参考序列的分析结果还可以进行比较。

最低要求：SNP、InDel 和 SV 查找：必须要有基因组序列。SNP、InDel 和 SV 查找和注释，构建进化树，进化选择压力：必须要有基因组序列和基因注释信息。

1.5.2 重测序的测序深度及覆盖度统计中为什么只用 100x 的数据来分析，这样能代表整体水平吗？

我们全部分析都是基于 100x 的数据，但是这 100x 的数据是完全随机的，并没有任何偏好性。病毒基因组大小只有不到 20KB，按照 20K 计算,最低上机量 1G 的数据相当于平均 50000X 的覆盖深度了,这样会使测序深度无限增大，从而无法分辨得到的结果是低频突变还是测序错误。

1.5.3 如何在变异检测中用 gene ID 查找基因的具体信息？

Gene ID 是指 indel 或 SNP 所在的参考序列基因的 ID 信息，该 ID 在参考序列的 gbk 文件里（在 NCBI 进行下载）都有对应注释，老师您可以下载下来查看的。

例如，打开参考序列 gbk 文件，查找关键字“XXX”定位至该 ID 所对应的基因，在 CDS 中/product 一行便可知道该基因控制的功能。

1.5.4 为何在.SV_filt.ctx.xls 文件中 pos1 与 pos2 位点的差值与实际缺失的长度不一致？

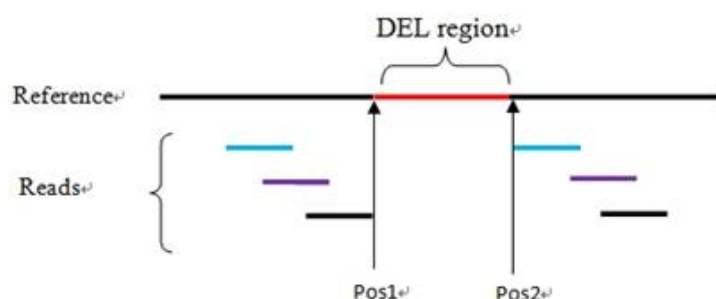
我们采用的是 breakdancer 软件检测 SV 的，一般的检测 DEL 序列是按照下图的方式检测的，两位 pos 的位置是最内侧的两个位置，具体的 size 大小是根据软件的内嵌模型和相应的数学模型预估出来的大小，只是一个参考值（来自 breakdance 软件的发表文献），并不是起止位置的一个简单加和。另外在 readsmapping 过程中 reads 的 mapping 不是 reads 全部 mapping 上，有的 reads 是部分匹配上，还有 mismatch 数，这些都是计算 size 时数学模型需要考虑的。下图是 breakdance 软件作者对此问题的解释。

```
The deletion size is estimated as the mean ISIZE of supporting read pairs for this event minus the mean ISIZE of the library. If the supporting read pairs are from multiple libraries, this estimation is done per library and then weighted by the number of read pairs per library.
```

-Ken

Thomas Wieland wrote:

```
> Hi,  
>  
> I'm using the latest version and I'm wondering how the size of an SV is  
> calculated? See, for example, this deletion:  
>  
> chr12 2880569 7+0- chr12 2880673 0+7- DEL 98 99  
> 7 NA|7 0.62  
>  
> I thought that the size would be the area between the two breakpoints,  
> i.e. 2880673-2880569 = 104 but instead the outputfile says it's 98.  
> What's the reason for this?  
>  
> Best Regards,  
>  
> Thomas
```



1.5.5 重测序 SV 部分 Orientation1、Orientation2 和 num_Reads 有什么关系？

Orientation1 为前端 reads 比对中，比对到染色体正链（+）和反向互补链（-）的 reads 个数。Orientation2 为后端 reads 比对中，比对到染色体正链（+）和反向互补链（-）的 reads 个数。num_Reads 为 reads 比对中，支持 SV 发生的 Paired-end reads 对数。

这个结果是 BreakDancer 软件进行直接给出的，Orientation1、2 分别是支持 SV 前段和后端的 reads 数目，+ 代表正链，- 代表反向互补链，num_Reads 是支持的 PReads 对数。前两列包含了最后这一列的 reads，但因为不是所有的都成对，最后一列可能少一些。

1.5.6 如何找到可信的 SV ？

您还可以参考*SV_filt.ctx.xls 文件来判断比较可信的 SV 结果，主要有以下三条：

1. Size：介于[-10000,-200]或[200,10000]；
2. num_Reads：>=20；
3. Orientation1 和 Orientation2 的正负链个数接近（如 100+100-）。

另外，测序深度与 reads 支持的条数相接近比较可信。

1.5.7 重测序产品中为什么能得到插入/缺失了碱基的数目，却得不到插入/缺失的具体位置与序列？有没有什么办法可以知道具体序列？

重测序产品中 SV 检测分析，是可以提供样本相对于参考基因组的一个大概的 DEL 序列，INS 的具体序列重测序是检测不到的，由于测序只测文库片段的两端，因此中间 INS 序列暂时是不能知道的。

理论上说这种插入序列是可以通过在插入位置的附近设计引物来 PCR 扩增出具体的序列。另外，也可以通过局部组装附近的 reads 来获取中间的序列（这个也很依赖局部组装的效果）。

1.5.8 Coverage (%) 和 Genomics(%)的含义

$Coverage (\%) = Cover_Len(bp) / Scaffolds_Len(bp)$,

$Genomics(\%) = Scaffolds_Len(bp) / \text{组装 scaffold 总长度}$ 。

Cover_Len(bp)为比对到 ref 的长度；Scaffolds_Len(bp)为比对到 refscaffold，这些 scaffold 的长度（可能有少数碱基没有比对到），Coverage (%)

$= Cover_Len(bp) / Scaffolds_Len(bp)$ 。反映了能比对到当前 ref 的 scaffold，与参考序列的相似度。Genomics(%)=Scaffolds_Len(bp)/组装 scaffold 总长度

反映了所有 scaffold 中能比对到 ref 的比例。

1.5.9 某一基因片段没有 reads 的原因？

对于参考基因组上一段基因片段没有 reads 覆盖的原因：

（1）可能是 sample 中没有该基因片段，也就是 DEL 缺失；

（2）该基因片段出现了突变，与参考基因组的该基因片段同源性很差，因此 reads 没有比对上。

1.5.10 如何确定感兴趣基因上是否发生突变？

首先要知道这条序列在参考基因组的哪个位置，因为我们所做的分析，突变位点都是给定的相对于参考基因组的位置信息，下面用截图进行说明：

RCLC 以 GS115 为参考基因组的 SNP 分析：

ref_name	SNP_site	ref_base	sample_base	type	SNP_reads	total
FN392319.1	1900	C	G	hete	112	
FN392319.1	1989	A	T	hete	143	
FN392319.1	2138	C	A	hete	146	
FN392319.1	2317	A	G	hete	141	
FN392319.1	2428	G	C	hete	126	
FN392319.1	2484	T	C	hete	121	
FN392319.1	2795	C	G	hete	62	
FN392319.1	2821	A	C	hete	60	
FN392319.1	2864	T	C	hete	66	

RCLC 以 GS115 为参考基因组的 Indel 分析：

ref_name	InDel_site	InDel_type	InDel_seq	reads_strain	mutation_t	InDel_read	total_read	mapping_quality
FN392319.1	3796	I1	T	*	hete	86	249	60
FN392319.1	4413	I1	A	*	homo	123	137	60
FN392319.1	4842	I1	T	*	homo	125	141	60
FN392319.1	7131	I1	A	*	homo	121	136	60

第一列的信息就是相对于参考基因组的位置，如果您想要的基因确定在参考基因组的位置是 FN392319.1 上的 1000-3000 的染色体上，那么根据第一个截图就可以看出，中间有 9 个 SNP 位点的，根据第二个截图发现该基因没有 Indel。

参考序列染色体位置是所提供的参考基因组 NCBI 自带的信息，老师可以自己进行比对并找到相应的染色体，然后再结合我们的结果进行查找即可。

1.5.11 重测序中除了 Qs core，还有什么参数可以 ascertain quequality of the data 呢？

“que quality of the data” 是指的是 “mapping query quality of the data”，与 SNP 等结果的可信度有关。

首先数据质控方面，我们会去除所含低质量碱基（质量值 ≤ 38 ）超过一定比例（默认 40bp）、N 碱基达到一定比例（默认 10bp）、去除与 Adapter 之间 overlap 超过一定阈值（默认 15bp）的 reads，并去除 duplication 的 reads，以此来保证后续分析结果的准确性。

在比对方面，会利用 samtools 软件过滤掉质量值低于 20、测序深度低于 4x 的的 SNP 及 InDel 的结果，即保证最后交付的结果都至少有 4 条以上且质量值均高于 20 的 reads 支持，进一步保证分析结果的准确性。

1.5.12 重测序中 SNP 的分析步骤

- 1、利用 bwa 软件将样本的测序数据与 reference 比对；
- 2、利用 samtools 软件对比对结果（bam 文件）进行筛选过滤，过滤参数为：SNP 的 reads 质量最低阈值（=20），最低测序深度（=4），其他参数为软件默认值。

1.5.13 出现在 OR 和 special 表格中的具体的基因怎么找？

由于 OR 和 special 中并未展示具体的基因信息，因此我们还是需要去前边基因注释信息的部分通过基因 ID 来寻找具体的基因。详细位置在：

```
|-- *.SNP.cds_info.xls  -[位于编码序列区 SNP 注释文件]
|-- *.SNP.gene_info.xls    [位于基因区 SNP 注释文件]
|-- *.SNP.exon_info.xls    [位于外显子区 SNP 注释文件]
|-- *.SNP.mRNA_info.xls    [位于基因转录区 SNP 注释文件]
```

1.5.14 重测序中相应名称解释

mismatch 指的是可以比对到参考基因组的碱基，但是中间有比对不上的碱基或者片段缺失，也就是 SNP/Indel。

mismatch rate 就是 $\text{mismatch} / \text{total_map_base}$ 的值。

后续我们分析的 SNP/SV/Indel 的数据主要来自于 total_mismatch 经过过滤后的结果，我们默认支持的 SNP 的 reads 数要多余 4 条，如果 mismatch 位置的 reads 数少于 4 条则会被过滤到。

1.6 结果文件相关

1.6.1 数据应该如何下载？为什么下载不下来？

我们释放数据采用的是 ftp 平台，该链接可以在 IE 等浏览器打开，不过由于网页浏览器对于 ftp 协议支持不好，下载会带来一系列问题，我们推荐老师使用专业的 ftp 工具进行数据下载，这些工具通常还支持断点续传等功能，方便老师在网络状况不佳的场合保证下载数据的完整性。

推荐一个免费开源的 ftp 工具 filezilla，下载地址：

<http://rj.baidu.com/soft/detail/13432.html?ald>

另外，现在我们的数据释放时会提供文件的 md5 值，可以使用 md5 计算工具来检查文件下载是否完整。

1.6.2 数据应该如何打开？

我们在对应文件夹内一般会有较为详细的文件说明。生物信息的结果文件通常分两类，一类是普通文本文件（或压缩的文本文件），可以用文本编辑器打开（或解压后用文本编辑器打开），推荐的文本编辑器：Notepad2，<http://www.flos-freeware.ch/notepad2.html>

另外一类是图片文件，可以用图片浏览器打开，如 windows 中的图片预览工具。其中需要注意的一类是 svg 文件，这类也是图片文件，不过是用于网络的矢量图格式，可以用 8.0 以上的 IE，或者 firefox，chrome 之类的第三方浏览器打开。更老版本的 IE 需要安装插件才能显示，推荐使用火狐 firefox，使用起来更方便一些。

另外 pdf 之类的常用文件，一般可以用对应类型的工具打开，不再一一详述。

此外需要提到的一点是超大文件的打开，如 cleandata 的 fastq 文件，这些文件也是纯文本文件，可以用文本编辑器打开，不过由于 windows 下读取文本文件的内存机制，会将整个文件读入内存，瞬间将内存占满，因此通常读取会产生障碍。实际上这些文件是测序的原始数据，具体的结果在我们提供的结果文件中都已经涵盖的比较好，建议老师不要在 windows 下尝试打开类似文件。如果老师有需求，我们可以提供截取的部分文件内容作为示例，老师可以了解一下文件的内容和结构。

1.6.3 M8 文件怎么打开、怎么解读？

m8 文件可以直接用文本编辑器打开（例如 EditPlus）也可以用 excel 打开（将文件的后缀名改为 .xls）。m8 文件格式：列表格式的比对结果，从左到右各列的意义依次是：query 名、subject 名、identity、比对长度、错配数、空位数、query 比对起始

坐标、query 比对终止坐标 subject 比对起始坐标、subject 比对终止坐标、期望值、比对得分。在这里 query 是我们预测到的基因，subject 是数据库中的基因。

更详细 m8 文件格式见：

04.Genome_Function/General_Gen_Annotation/KEGG/KEGG.readme.doc

1.7 高级分析相关

1.7.1 系统进化树构建方法

基因 SNP 数据，采用 TreeBeST 的 PHYML（最大似然法）算法构建系统进化树。

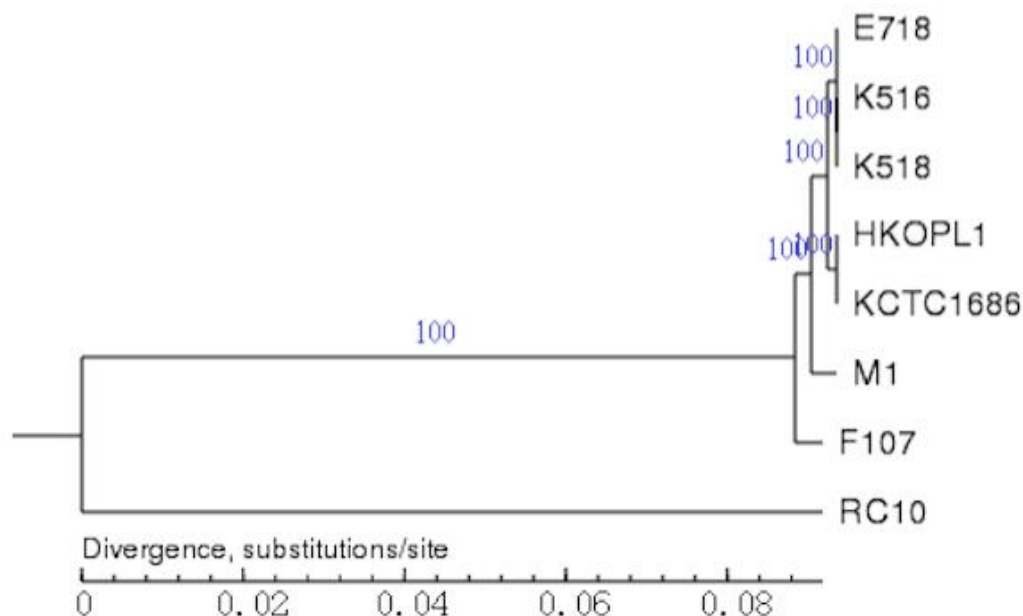
1.7.2 系统进化树构建的三种方法

1. 基于 SNP 建树，用样品和参考菌株群体的 SNP 矩阵构建系统进化树。对于每一株菌，按照相同顺序将 所有 SNP 相连，获得相同长度的 fasta 格式的序列(其中一个为参 考序列)，作为输入文件。用 TreeBeST[1]的 PHYML(最大似然法)算法构建系统进化树。
2. 基于 core-pan 分析建树，用 core-pan 分析鉴定出样本的单拷贝 core 基因，利用了 MUSCLE 软件进行蛋白多序列的 比对，并转化为 CDS 结果，并用 Treebest 的 maximum likelihood 模型进行进化树构建。
3. 基于基因家族建树，用基因家族聚类鉴定出的单拷贝直系同源基因结果，利用了 MUSCLE 软件进行蛋白多 序列的比对，并转化为 CDS 结果，并用 Treebest 的 maximum likelihood 模型进行进化树 构建。

1.7.3 系统进化树解读

1、tree 格式文件中紧跟在每个样品后面的数字代表分歧度还是括号括起来之后括号外面的数字代表分歧度？

```
(((((E718:2.62924e-05,(K516:1.29518e-06,K518:1.33878e-09):2.54713e-05):0.00103384,(HKOPL1:7.57712e-05,KCTC1686:3.94036e-05):0.000924158):0.000191608,M1:0.0013661):0.00205763,F107:0.00334939):0.0884277,RC
```



可以根据图形来看，K516 和 K518 是同一个节点上的两个分支(K516:1.29518e-06,K518:1.33878e-09)，每个冒号后面的数字表示分歧度（平均每个位点碱基的替换次数），然后 K516 和 K518 作为一个整体和 E718 再次作为一个节点（E718:2.62924e-05,(K516:1.29518e-06,K518:1.33878e-09):2.54713e-05），然后 E718\K516\K518 作为一个整体和 HKOPL1\KCTC1686 再次作为一个节点（E718:2.62924e-05,(K516:1.29518e-06,K518:1.33878e-09):2.54713e-

05):0.00103384,(HKOPL1:7.57712e-05,KCTC1686:3.94036e-05):0.000924158),

以此类推，您可以根据图形进行判定。

2、如何知道 F107 和哪株菌的距离最近？是计算分歧度的差值吗？

您可以直接从进化关系来看，目前的进化树可以看出 F107 是和 M1 较近的。分歧度只是计算进化距离，具体的进化关系还是要看整体的进化关系的。

1.7.4 进化树分支可信度解释问题

分支上的数字其实是 bootstrap，即自展值，是用来检验你所计算的进化树分支可信度的。简单地讲就是把序列的位点都重排，重排后的序列再用相同的办法构树，如果原来树的分枝在重排后构的树中也出现了，就给这个分枝打上一分，如果没出现就给 0 分，这样经过你给定的 repetitions 次（例如：1000 次）重排构树打分后，每个分枝就都得出分值，计算机会给你换算成 bootstrap 值。

如果是同源物种构树时，软件在计算时的精度会高，而加入外参及同源性很远的时候，则软件计算时要考虑很多因素，所以精度会相应降低，则 bootstrap 值就会降低。不加外参的进化树和加入外参的进化树的 bootstrap 值也恰好验证了这一点。

从整体的拓扑结构看，加入外参前后是一致的，所以目前的结果是没有问题的。

Bootstrap 值说明:

进化树中有的数值为 Bootstrap 值。在这里没有任何阈值的限定。

Bootstrap 值不等于序列同源度水平，不是同源度水平高就 Bootstrap value 就很高，他们之间没有任何必然的联系，Bootstrap 方法是自举法，或者也叫自助法，是一种有放回的抽样方法，树枝中间的数值是 Bootstrap 的统计数值，也就是两次计算

他们的聚类都聚在一块这就是 2，三次都聚在一块都是 3，以此类推，如果 1000 次 bootstrap 有 1000 次在二元分法情况下 A 和 B 都聚在一块他们就是 1000，换算成百分制就是 100，如果是 1000 次只有 200 次聚在一块，那就是 200 换算成百分制就是 20，这也就是 bootstrap 支持率的原理。同源性高，但是 bootstrap value 很低，是因为 A、B 和 C 序列的关系太近了，差异性太小了，A 和 B 聚在一起或者跟 C 聚在一起，软件都会默认成有效的，有统计价值的，无法分辨出他们真正的系统发育关系。这样在抽样放回，在抽样的时候，尤其是很多个序列差异度都非常小的情况下，每一次 A 和 B 真正能聚在机会其实是很小的。

分支可信度，表明拓扑结构的可靠程度，不说明亲缘关系。亲缘关系看分支长度

<http://www.bacteria.cn/html/2014/43.html>

1.7.5 进化树常见售后

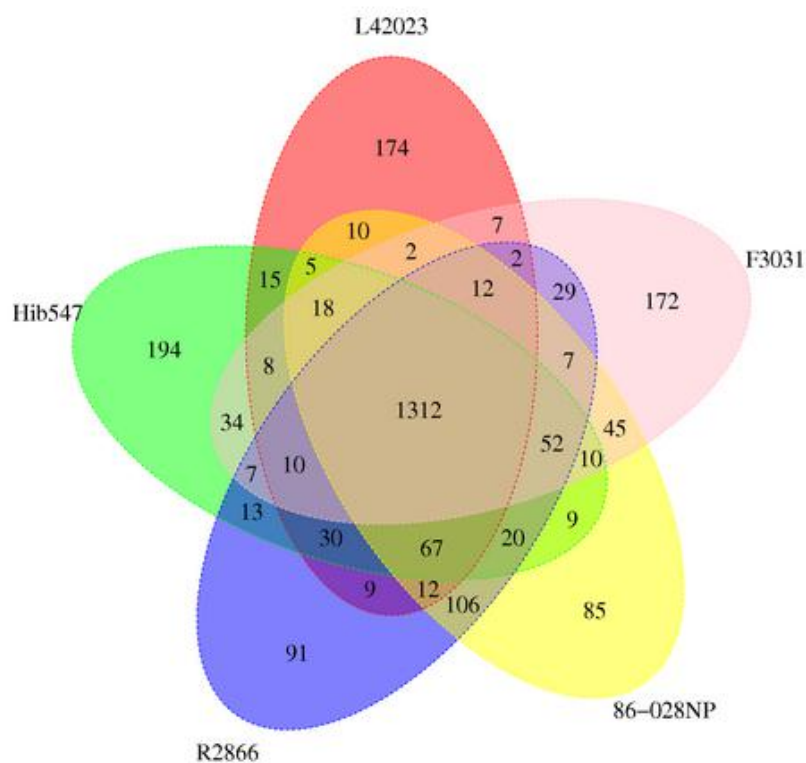
I would like more detail in figure legend detailing tree method, i.e MP tree.

Also gives no details on how bootstrap values were generated i.e how many replicates were used? This is also something that should be added to the M&M section.意思就是构建这个树状图采用的方法？bootstrap values 怎么产生的？还有 replicates 的次数？

答：基于 SNP 做进化树分析方法：基于样品和参考菌株群体的 SNP 矩阵构建系统进化树。对于每一株菌，按照相同顺序将所有 SNP 相连，获得相同长度的 fasta 格式的序列（其中一个为参考序列），作为输入文件。用 TreeBeST 的 PHYML（最大似然法）算法构建系统进化树。

分支上的数字表示分支可信度，值越接近 100 表示可信度越高；分支长度表示进化距离的大小，进化距离以平均每个核苷酸的替换次数来计算。

1.7.6 共有和特有基因分析中花瓣图/韦恩图为何和与表中统计数字的不一致？



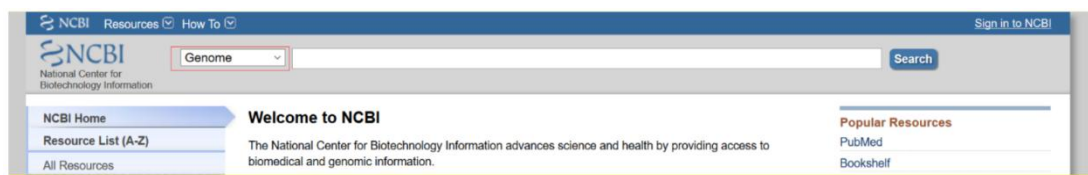
图中每个椭圆表示一个样品，每个区域上的数据表示在且仅在此区域的样品中出现的 group 的个数，一个 group 表示一组具有大于 50%相似性、序列长度差异低于 0.7 的基因集。表格中统计的是基因的个数，图说明的是基因集的个数~

1.7.7 cgMLST 前期准备

1. NCBI 上查询菌株种级别所发表的菌株个数，后面用于建立 cgMLST 核心基因集，若是样本个数少，例如 200 株，后期若做客户需要增加自己的样本个数(总样本个数约 500 株以上)。

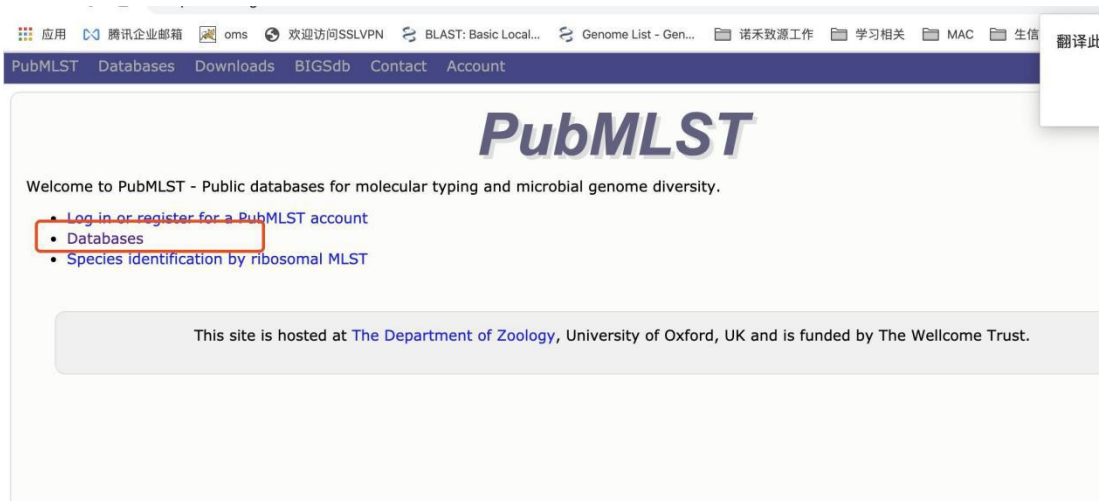
注意△ 这里的样本必须是同种的，实际上每个分组可以几个样本，也可以几十个样本，都可以，具体几个分组也可以做，但菌株越多越好

网址:<https://www.ncbi.nlm.nih.gov/>



2. 进入网址，查询此菌是否做过 MLST(管家基因，用于核心基因集印证)或 cgMLST (做过可不用构建核心基因集，直接拿客户样本做最小生成树) 网

址:<https://pubmlst.org/databases/>;



例 1 *Achromobacter*

a. 查询要找的菌(按照字母顺序)，点击进入。

Databases hosted on PubMLST

These databases host MLST schemes and isolate data, increasingly including whole genome sequences.

Bacteria

- [Achromobacter](#)
- [Acinetobacter baumannii](#)
- [Aeromonas spp.](#)
- [Anaplasma phagocytophilum](#)
- [Arcobacter spp.](#)
- [Bacillus cereus](#)
- [Bacillus licheniformis](#)
- [Bacillus subtilis](#)
- [Bordetella spp.](#)
- [Borrelia spp.](#)
- [Bartonella bacilliformis](#)
- [Bartonella henselae](#)
- [Brachyspira spp.](#)
- [Brucella spp.](#)

b. 出现这株菌的信息，点击 [Sequence/ profile definitions database](#)

Achromobacter MLST Databases

This site uses two linked databases powered by the [BIGSdb genomics platform](#). The sequence database contains sequence data, whereas the isolate database contains provenance and epidemiological information. Further details are available in [Bioinformatics 11:595](#).



- Information
 - [Primers used for amplification and sequencing](#)
- Access main databases
 - [Sequence/profile definitions database](#)
 - [Isolates database](#)
- [Policy document](#)
- [Submission of data](#)
- [BIGSdb software](#)
- [Recent publications using MLST in Achromobacter research](#)

This MLST scheme was developed by [Theodore Spilker](#)

c. 选择等位基因序列 ([Allele sequences](#)), 点击进入

Achromobacter locus/sequence definitions database



Query database

- [Sequence query](#) - query an allele sequence or genome.
- [Batch sequence query](#) - query multiple sequences in FASTA format.
- [Sequence attribute search](#) - find alleles by matching criteria (all loci together)
- [Locus-specific sequence attribute search](#) - select, analyse and download specific alleles.
- [Search, browse or enter list of MLST profiles](#)
- [Search by combinations of MLST alleles](#) - including partial matching.
- [Batch profile query](#) - lookup MLST profiles copied from a spreadsheet.



Downloads

- [Allele sequences](#)
- [MLST profiles](#)



Export

- [Profiles](#)
- [Sequences](#) - XMFA / concatenated FASTA formats



Analysis

- [Sequence similarity](#) - find sequences most similar to select
- [Sequence comparison](#) - display a comparison between two
- [Locus Explorer](#) - tool for analysing allele sequences stored

d. 如下图所示，这种情况就只有 MLST

Download allele sequences

Select loci by scheme | [Alphabetical list](#) | All loci by scheme

MLST

Locus	Download	Type	Alleles	Length (setting)	Min length	Max length	Curator(s)	Last updated
nusA		DNA	148	Fixed: 355 bp	355	355	T. Spilker	2018-01-04
rpoB		DNA	148	Fixed: 413 bp	413	413	T. Spilker	2018-01-04
eno		DNA	126	Fixed: 214 bp	214	214	T. Spilker	2018-01-04
gltB		DNA	117	Fixed: 241 bp	241	241	T. Spilker	2018-01-04
lepA		DNA	156	Fixed: 347 bp	347	347	T. Spilker	2018-01-04
nuoL		DNA	130	Fixed: 230 bp	230	230	T. Spilker	2018-01-04
nrdA		DNA	134	Fixed: 449 bp	449	449	T. Spilker	2018-01-04

Other loci

Locus	Download	Type	Alleles	Length (setting)	Min length	Max length	Curator(s)	Last updated
nrdA 765		DNA	306	Fixed: 765 bp	765	765	T. Spilker	2017-12-12

Download table: [tab-delimited text](#) | [Excel format](#)

例2 *Campylobacter* spp.

a. 查询要找的菌(*Campylobacter* spp.), 点击进入

Databases hosted on PubMLST

These databases host MLST schemes and isolate data, increasingly including whole genome sequences.

Bacteria

- [Achromobacter](#)
- [Acinetobacter baumannii](#)
- [Aeromonas spp.](#)
- [Anaplasma phagocytophilum](#)
- [Arcobacter spp.](#)
- [Bacillus cereus](#)
- [Bacillus licheniformis](#)
- [Bacillus subtilis](#)
- [Bordetella spp.](#)
- [Borrelia spp.](#)
- [Bartonella bacilliformis](#)
- [Bartonella henselae](#)
- [Brachyspira spp.](#)
- [Brucella spp.](#)
- [Burkholderia cepacia complex](#)
- [Burkholderia pseudomallei](#)
- [Campylobacter spp.](#)
- [Carnobacterium maltaromaticum](#)
- [Chlamydiales spp.](#)
- [Citrobacter freundii](#)

b. 出现这株菌的信息，点击 [Sequence and profile definitions](#)

Campylobacter MLST Home Page

This site uses two linked databases powered by the [BIGSdb genomics platform](#). The sequence definitions database contains sequence definitions and allele frequencies, whereas the isolate database contains provenance and epidemiological information. Further details [Bioinformatics 11:595](#).



- Information
 - [Sequencing primers and experimental conditions](#)
 - [Campylobacteriosis sentinel surveillance in Oxfordshire, UK](#)
 - [C. jejuni/C. coli core genome MLST \(cgMLST\)](#)
- Access main databases
 - [Campylobacter jejuni/coli](#)
 - [Sequence and profile definitions](#) (MLST, fla, porA)
 - [PubMLST Isolate Database](#)
 - [Non jejuni/coli Campylobacter](#)
 - [Sequence and profile definitions](#)
 - [PubMLST Isolate Database](#)
- [Policy document](#)
- [Submission of data](#)
- [Submission history](#)
- [News and updates](#)
- [BIGSdb software](#)
- [Recent publications using MLST in Campylobacter research](#)

c. 选择等位基因序列 ([Allele sequence](#)), 点击进入

Campylobacter locus/sequence definitions database

The Campylobacter PubMLST sequence definition database contains allele and profile data representing the total known diversity of *C. jejuni* and *C. coli*. Every

Query database

- Sequence query - query an allele sequence or genome.
- Batch sequence query - query multiple sequences in FASTA format.
- Sequence attribute search - find alleles by matching criteria (all loci together)
- Locus-specific sequence attribute search - select, analyse and download specific alleles.
- Search, browse or enter list of profiles
- Search by combinations of alleles - including partial matching.
- Batch profile query - lookup profiles copied from a spreadsheet.

Downloads

- Allele sequences
- MLST**
C. jejuni / C. coli cgMLST v1.0

Profiles

Export

- Profiles
- Sequences - XMFA / concatenated FASTA formats

Analysis

- Sequence similarity - find sequences most similar to selected allele.
- Sequence comparison - display a comparison between two sequences.
- Locus Explorer - tool for analysing allele sequences stored for particular locus.

从下图可看到，这种情况是有 cgMLST 靶基因的

Download allele sequences

Select loci by scheme | [Alphabetical list](#) | [All loci by scheme](#)

Click within the tree to display details of loci belonging to schemes or groups of schemes - clicking a group folder will display the loci for all schemes

All loci

- Typing
 - MLST
 - C. jejuni / C. coli **cgMLST v1.0**
 - eMLST (21 partial genes)
- Other schemes
 - 1643 Gundogdu loci
 - C. coli 15-537360
 - ED pathway
 - Ribosomal Protein Genes
- Loci not in schemes

MLST

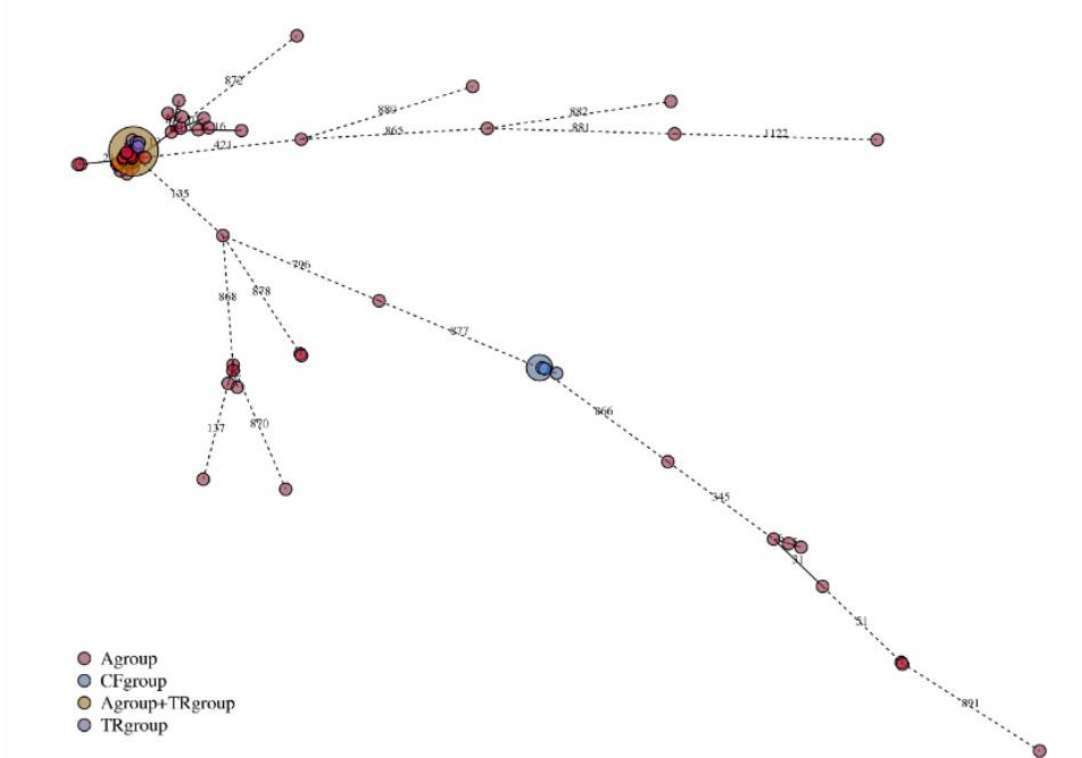
Locus	Download	Type	Alleles	Length (setting)	Min length	Max length	Full name/product	Curator(s)	Last updated
aspA	Download	DNA	449	Fixed: 477 bp	477	477		A. Cody, F. Colles	2017-11-06
ginA	Download	DNA	618	Fixed: 477 bp	477	477		A. Cody, F. Colles	2018-01-11

cgMLST 靶基因: 从下表可得到 cgMLST 的位点、等位基因、长度、全名称、别名等详细信息

C. jejuni / C. coli cgMLST v1.0

Locus	Download	Type	Alleles	Length (setting)	Min length	Max length	Full name/product	Aliases	Curator(s)	Last updated
CAMP0001 (dnaA)	Download	DNA	1491	Variable: No limits set	1320	1329	Chromosomal replication initiator protein (K02313)	Cj0001; N149_0001	A. Cody, F. Colles	2018-01-28
CAMP0002 (dnaN)	Download	DNA	1238	Variable: No limits set	1065	1068	DNA polymerase III, beta chain (K02338)	Cj0002; N149_0002	A. Cody, F. Colles	2017-11-26
CAMP0003 (gyrB)	Download	DNA	1948	Variable: No limits set	2310	2322	DNA gyrase subunit B (K02470)	Cj0003; N149_0003	A. Cody, F. Colles	2017-11-26
CAMP0006	Download	DNA	875	Variable: No limits set	1284	1320	Putative Na ⁺ /H ⁺ antiporter family protein (K07084)	Cj0006; N149_0010	A. Cody, F. Colles	2018-01-12
CAMP0007	Download	DNA	2014	Variable: No	4428	4491	Glutamate synthase (NADPH) large subunit (K00265)	Cj0007;	A. Cody, F.	2017-11-

注意: 像例 2 这种情况，数据库中有 cgMLST 靶基因的，可直接采用，不用建立核心基因集，交付结果只有最小生成树(如下图所示)



二、运营增强版

2.1 标准分析

2.1.1 线粒体和叶绿体注释有什么特殊要求？为什么有时候无法进行注释？

线粒体和叶绿体与细菌和真核生物都有所不同，一方面，其编码规则和其宿主的真核生物非常接近，另一方面，又不像真核生物那样存在大量的 intron，而是几乎没有 intron 的存在。目前没有好的本地注释软件用于对这些基因组进行注释，公认的在线注释工具中，由于其注释的原理是依靠同源蛋白比对为基础，结合 DNA 序列结构做出预测，因此如果不能做到 single scaffold 的程度，注释效果会大打折扣，甚至无法

做出准确注释。我们的经验是，最多 5 条 scaffold 可以尝试人工进行注释，更多 scaffold 的场合，注释结果是无法得到保证的。为保证结果的可靠性，我们更推荐对于这样的数据，转做基于 Reads 或 scaffold 的变异检测，即便信息量会打折扣，至少能保证结果的可靠性。

2.1.2 真菌精细图测序中，61 个 contig 中，为什么只有 18 个 contig 得到注释？

这些基因是全部基因组的预测结果，个数相对较少主要有以下两个可能：

- 1、该物种基因个数本身相对较少，因此结果偏少；
- 2、我们预测基因的软件为 augustus，是根据现有研究的真菌基因模型及序列结构对编码区进行预测，而该种在 ncbi 中没有找到已公布的信息，表明研究结果相对较少，对基因结构的研究不是很成熟，因此导致预测结果偏少。

2.1.3 重测序一代和二代有出入的回复模板

- 1、S13 样本扩增一代测序结果与二代组装序列有出入。

基于二代测序组装的结果，即使测序深度再高，与一代测序相比时难免出现一些区别，属于正常情况，gap 区（即 N）为根据测序建库的片段长度进行的预估，可能与实际情况有一定差值，但理论来说不会很大。

- 2、FA、PQ 重测序检测结果与一代测序结果有出入。

FA 样本检测到 29 个 SNP，PQ 样本检测到 20 个 SNP，其中分别有 2、7 个位点存疑，687、41929 为共有存疑位点，可能是 S13 样本组装导致的碱基差异。

687 位点检测均为 hete (杂合) , 即部分 reads 检测到碱基突变, 部分 reads 未突变, 在保证一代测序质量较高的情况下可以以一代测序结果为准。其他位点均检测为 homo (纯和) , 这几个位点 85% 以上的 reads 均检测到碱基突变, 这些位点中除 PQ 的 14805 的测序深度低于 100x , 其他均高于 110x , 其中 41929 位点超过 500x。

二代检测与一代测序有出入主要有两个原因: 二代变异检测结果具有一定假阳性比率, 迄今为止没有任何软件可以保证其检测到的变异结果具有 100% 的正确率, 但假阳性比率并不是很高 (属于小概率事件) , 尤其当测序深度较高时假阳性概率更低; 老师使用一代 PCR 测序验证, 由于 PCR 过程中存在引入突变的可能性, 也是可能导致二者结果有出入的原因。

对于 InDel 结果, 检测到的几个位点测序到的 reads 均超过 100x , 支持的 reads (InDel_reads_cov) 相对较低, 即检测到部分 reads 确实存在 InDel 的情况, 但比率较低, 因此在确定一代测序质量的情况下可以一代测序为准。

对于 SNP 变异, 建议老师在深入分析时以二者共有的变异结果为主, 对于存疑位点, 其中 hete 位点在确定一代测序质量的情况下可以一代测序为准; 对于 homo 位点, 当测序深度较高时, 出现假阳性概率理论来说较低, 还请老师核查下 PCR 及一代测序结果。

如有其他问题, 欢迎与我沟通。

2.1.4 预测得到的基因岛的功能信息可否提供一下, 或者告知通过何种途径能够知道基因岛的功能

首先，您可在文件 Result\03.Genome_Component*\Genomic_Island*.GI.xls 中找到某一基因岛对应的具体基因 ID，示例如下：

[illegible]

之后在\Result\04.Genome_Function*\AnnoSummary*.AnnoSummary.xls 中查找其对应的不同数据库中的功能注释信息，示例如下：

Gene_id	Locus	[nr]	[SwissProt]	[KEGG]	[COG]	[TCDB]	[GO]	[PHI]	[VFDB]	[ARDB]	[CARD]
Annotation_num		5,310	3,774	5,240	4,446	1,011	3,740	298	150	36	
K21_GM000001	[Chr1:70: [AER7948E [B5XT51; [kpa:KPN] [YP_00223	[NA]				[NA]	[GO:00037	[NA]	[NA]	[NA]	[NA]
K21_GM000002	[Chr1:140: [WP_00414 [P26464; [kpb:FH42 [YP_00223	[NA]				[NA]	[GO:00038	[NA]	[NA]	[NA]	[NA]
K21_GM000003	[Chr1:265: [WP_00418 [A6TG02; [kpb:FH42 [YP_00223	[NA]				[NA]	[GO:00055	[NA]	[NA]	[NA]	[NA]
K21_GM000004	[Chr1:375: [WP_00417 [POAES7; [kpa:KPN] [YP_00223	[NA]				[NA]	[GO:00038	[NA]	[NA]	[NA]	[NA]
K21_GM000005	[Chr1:637: [WP_00418 [POA8Y6; [kpb:FH42 [YP_00223	[NA]				[NA]	[NA]	[NA]	[NA]	[NA]	[NA]

进而综合判断基因岛整体的功能情况（其中*代表结题报告中的物种名称）。

2.1.5 细菌完成图中重复序列数据详细说明

1、TR 为串联重复序列，根据串联重复单元的长度进行细分，包括 Minisatellite DNA（小卫星 DNA、一般为 15-65 bp 的串联重复单位）、Microsatellite DNA（微卫星 DNA、一般为 2-10 bp 的串联重复单位），另外可能还包括其他片段区间的类型，但是由于小卫星和微卫星 DNA 具有极高多态性，是科研研究和应用的重点，因此我们将这两部分单独拉出来，分别进行了统计，其他类型并未深入研究。因此您看到的结果中，tr.gff 包含了 Minisatellite DNA.gff 或者 Microsatellite DNA.gff 中的信息，您可理解为 tr.gff 为总的的结果，其余两个只是单独摘出来了部分内容，目的是便于您分类查看；

2、如下图所示，两种注释结果起始、终止位点相同，但是 score 得分值、ID 号、重复序列长度及数目等均不同，表征在该片段内，可能存在多种重复序列组合，举例如下：

当重复序列长度为 15 时，该类重复序列数目为 3.6，相邻两个重复序列序列匹配度为 89%，相邻两个重复序列之间 indel 的比例为 0；

```
4;Consensus=TTTCATCATATATAAGAAAAGC;
Chr1 TRF TandemRepeat 5139619 5139672 90 + .
ID=S-R-52_TR299;PeriodSize=15;CopyNumber=3.6;PercentMatches=89;PercentIndels=
0;Consensus=CTTCTTTTTTTAGGCT;
Chr1 TRF TandemRepeat 5139612 5139672 104 + .
ID=S-R-52_TR300;PeriodSize=30;CopyNumber=2.0;PercentMatches=93;PercentIndels=
0;Consensus=TTAGGTGCATGCTTTTTTAGGATCTTCTTTT;
Chr1 TRF TandemRepeat 5158909 5158970 70 + .
ID=S-R-52_TR303;PeriodSize=18;CopyNumber=3.4;PercentMatches=84;PercentIndels=
```

我们在进行重复序列分析时，一般会设置 7 个以上的参数，其中能够比对上的概率设置为 80，插入的概率设置为 10，被匹配上的串联重复序列的最小分值是 50，最大的重复单元是 500bp，以及匹配上，没匹配上，插入的权重，分别是 2，7，7 等等。

这些参数的设置不仅来自于大量的文献研究，更基于众多的项目经验，可以最大限度的帮助老师分析数据。

参考文献可见：

Saha S, Bridges S, Magbanua Z V, et al. Empirical comparison of ab initio repeat finding programs[J]. Nucleic acids research, 2008, 36(7): 2284-2294.

Benson G. Tandem repeats finder: a program to analyze DNA sequences[J]. Nucleic acids research, 1999, 27(2): 573.

鉴于不同的菌种、不同的序列结构具有其特异性，受限于对具体物种的专业知识，我们只能将所有分析、比对到的结果向您展示，您可根据自己的切身需要筛选具体信息，如有问题，您可再与我联系~

2.1.6 覆盖率和覆盖深度的问题

覆盖率的计算方法是 covered length/referance size，其中，referance 是指组装得到的 scaffolds，而 covered 是指原始的 clean reads 序列通过 mapping 能覆盖到 scaffold 上面的长度，二者的比值即为覆盖率。

因为完成图是 0gap，因此是 100%，框架图可以按照售后来，截图如下：

ChrID	Referance_size(bp)		Covered_length(bp)		Coverage(%)	Depth
Scaffold1	1595606	1595600	100	110		
Scaffold10	106640	106620	99.98	110		
Scaffold11	98713	98711	100	110		
Scaffold12	92180	92175	99.99	100		
Scaffold13	86761	86759	100	110		
Scaffold14	71732	71727	99.99	93		
Scaffold15	64768	64765	100	89		
Scaffold16	48292	48289	99.99	92		
Scaffold17	40517	40513	99.99	110		

Scaffold32	1096	1095	99.91	150
Scaffold4	470129	470125	100	110
Scaffold5	372951	372945	100	110
Scaffold6	364185	364183	100	110
Scaffold7	255571	255560	100	110
Scaffold8	220621	220620	100	110
Scaffold9	181863	181862	100	110
Total	5820821	5820708	100	111

还有一个概念是覆盖深度，覆盖深度是说每个位点支持的 reads 数目，截图中的

depth 代表的是覆盖深度，这个覆盖深度其实是跟测序深度差不多。

测序深度=测序数据量/基因组大小

2.1.7 针对于革兰氏阳性菌，TNSS 没有 T3SS，而在 T3SS 预测中却有很多，是怎么回事。

关于 TNSS 与 T3SS 的预测问题，TNSS 是对蛋白功能数据库注释结果中找到分泌蛋白，再对其分型，而 T3SS 是一种软件，对 pep 序列文件直接进行预测获得的，二者方法不同，但是结果都有罗列

III 型分泌系统 (Type III secretion system, T3SS) 主要是革兰氏阴性菌的分泌蛋白分泌到细胞外的运输途径，因此 III 型分泌系统效应蛋白 (Type III secretion system Effector protein) 与革兰氏阴性致病菌致病机理有关。^[4]

使用软件 EffectiveT3 对输入的氨基酸序列进行预测，通过其内部特定的计算模型对每条氨基酸序列进行评分，分值越高，可信度越高，选出评分高于阈值的序列，认为这些序列为 III 型分泌系统效应蛋白。^[4]

2.1.8 Pacbio 下机数据相关

目前 PB 是 sequel 平台 (之前是 RSII)，在下机文件中，主要有三类文件，bam 文件，bam.pbi 文件，以及 xml 文件。在 sequel 平台中 bam 文件成为了 fastq 格式文

件的替代者，因为其更节约储存空间。用于后续分析的文件一般是.subreads.bam，
等同于 RS II 中的.subreads.fastq

咱们整个大库下机的所有文件如下：

```
.adapters.fasta  
.baz2bam_1.log  
.scraps.bam.pbi  
.sts.xml  
.subreads.bam  
.subreads.bam.pbi  
.subreadset.xml  
.transferdone
```

由于我们三代是混库上机，这些文件中包含其他客户的数据，所以我们最终不交付给客户整个
大库的下机结果（上图），只会交付给客户拆分过的以下数据文件

```
bam.pbi  
subreads.bam  
subreads.bam.pbi  
subreadset.xml
```

1.bam 文件

主要分为两个部分，第一行是 Header，储存测序的相关信息，剩余部分也即是文件的主要部分 records，存储序列信息。主要作用就是储存序列。其中，scraps.bam 格式保存的是获取 subreads 时废弃的序列，包括 adapter，以及一些低质量的序列，不用于分析，因此最终交付我们也不会给客户。

2.是 bam 文件的索引文件(PacBio BAM index)

主要用于随机访问和在无需完全访问 BAM 文件的情况下，获取信息。

3.XML 文件

MetaData, 储存数据描述。可用于 filter 或者 subset 等功能。

sts.xml 储存数据的统计信息。

2.2 高级分析

2.2.1 共线性分析的意义及如何解读

共线性分析是衡量他们之间进化距离的尺度，可以知道物种间的亲缘关系。我们可以直观地观察到两个基因组在进化过程中发生的易位、倒置以及比对的相似度等情况。如需要详细了解基因位置的变化请您查看数据释放中的 05.Comparative_Genomics 中 Synteny” 文件。

2.2.2 比较基因组中查找 SNP 的方法。

利用 MUMmer 比对软件，将每个样品与参考序列进行全局比对，找出样品序列与参考序列之间有差异的位点并进行了初步的过滤，检测出潜在 SNP 位点；提取参考序列 SNP 位点两边各 100 bp 的序列，然后使用 BLAT 软件将提取的序列和组装结果进行比对，验证 SNP 位点。如果比对的长度小于 101 bp，则认为是不可信的 SNP，将去除；如比对上多次，认为是重复区域的 SNP，也将被去除；最后用 BLAST、TRF、Repeatmask 软件预测参考序列的重复序列区，过滤位于重复区的 SNP。最后得到可靠的 SNP。

2.2.3 core.cog.anno.xls 文件、core.cog.class.catalog.xls 的共有基因及 Venn 图中共有的数目为什么不同？

基因在各数据库注释的时候，并不是所有基因都会注释到具体的 COG 结果，因此 core.cog.anno.xls 中基因个数会比 Venn 图中三菌共有的数目少；有的基因注释到的 COG_num 是归属与两个或以上的 Functional_class，因此在统计 core.cog.class.catalog.xls 时会对这些基因进行多次统计，使 core.cog.class.catalog.xls 比 core.cog.anno.xls 中基因个数多一些，如 core 基因中的 “E407-8GM000140_E407-8” ，同时注释到 E、K 两个 class，所以在这两个 class 统计时出现了两次。

2.2.4 （毒力、耐药、转座）、还有 core-pan 和进化树分析，分析过程和软件版本。

1. 毒力/耐药：

使用 BLAST 软件，把目标物种的氨基酸序列，与 VFDB/ARDB 数据库进行比对，把目标物种的基因和其相对应的功能注释信息结合起来，得到注释结果。由于每一条序列比对结果可能超过一条，为保证其生物意义，注释时保留一条最优比对结果作为该基因的注释。最后提供的 BLAST 结果为 M8 格式。

参数：BLAST 比对参数：blastp， $\text{evaluate} \leq 1\text{e-}5$

注释时保留一条最优比对结果作为该基因的注释，这个是自己写的脚本筛选最优比对

结果：在 $\text{identity} \geq 40\%$ ， $\text{coverage} \geq 40\%$ 的基础上选取 score 最高的；

2. 转座：

IS 预测，是采用的 IS finder 网站(<https://www-is.biotoul.fr/blast.php>) 进行的在线预测；

TN 预测，我们是采用的 IS finder 网站和 TransposonPSI 预测结合的方法；type 为 TransposonPSI 的是 TransposonPSI 预测的结果，用的是 TransposonPSI 软件预测，参数：-p psitblastn -F F -M BLOSUM62 -t -1 -e 1e-5 -v 10000 -b 10000

参数解释：-p Program Name [String]

-F Filter query sequence (DUST with blastn, SEG with others)

[String] default = T

-M Matrix [String] default = BLOSUM62

-t Length of the largest intron allowed in a translated nucleotide sequence when linking multiple distinct alignments. (0 invokes default behavior; a negative value disables linking.) [Integer] default = 0

-l Restrict search of database to list of GI's [String] Optional

-e Expectation value (E) [Real]

-v Number of database sequences to show one-line descriptions for (V) [Integer]

-b Number of database sequence to show alignments for (B) [Integer]

type 为 TN 的为 IS finder 网站预测的结果。

3. core-pan :

使用 cd-hit 软件(v4.6.1 版本)对需要分析的多个样品的蛋白序列进行聚类，参数-c 0.5 -n 3 -p 1 -T 4 -g 1 -d 0 -s 0.7 -aL 0.7 -aS 0.7

参数解释：

-c sequence identity threshold, default 0.9

this is the default cd-hit's "global sequence identity" calculated as: number of identical amino acids in alignment divided by the full length of the shorter sequence

-T number of threads, default 1; with 0, all CPUs will be used

-n word_length, default 5, see user's guide for choosing it

-d length of description in .clstr file, default 20

if set to 0, it takes the fasta defline and stops at first space

-s length difference cutoff, default 0.0

if set to 0.9, the shorter sequences need to be at least 90% length of the representative of the cluster

-aL alignment coverage for the longer sequence, default 0.0

if set to 0.9, the alignment must covers 90% of the sequence

-aS alignment coverage for the shorter sequence, default 0.0

if set to 0.9, the alignment must covers 90% of the sequence

-p 1 or 0, default 0

if set to 1, print alignment overlap in .clstr file

-g 1 or 0, default 0

by cd-hit's default algorithm, a sequence is clustered to the first cluster that meet the threshold (fast cluster). If set to 1, the program will cluster it into the most similar cluster that meet the threshold (accurate but slow mode) but either 1 or 0 won't change the representatives of final clusters

4. 基于 core-pan 进化树分析：

基于 corepan 分析鉴定出的单拷贝 core 基因，利用了 MUSCLE 软件(3.8.31 版本)进行蛋白多序列的比对，并转化为 CDS 结果，并用 Treebest (Neighbor-Joining,NJ) 构建进化树，参数 treebest nj -t mm -b 100。

参数解释：

-m FILE tree to be compared [null]
-t TYPE codon NT: ntmm, dn, ds, dm; AA: mm, jtt, kimura [mm]
ntmm p-distance (codon alignment)
dn non-synonymous distance
ds synonymous distance
dm dn-ds merge (tree merge)
mm p-distance (amino acid alignment)
-b NUM bootstrapping times [100]

2.2.5 NCBI 上传问题

【A】 Did the assembly program combine the sequences into scaffolds using runs of N's to represent gaps between ordered and oriented contiguous sequences? Alternatively, did you randomly merge the sequences into a single sequence (for example, may be you just linked the sequences together by size without using an assembly program)?

答：对于组装过程中的 gaps 我们是用 N 来进行连接的，在组装过程中首先选取不同的 K-mer（默认选取 35、47、59、71、83、95、107、119）进行组装，再选择最优的 kmer（细菌项目选取最少 scaffold）并调节其他参数（-d -u -R -F 等）再次筛选得到初步组装结果，然后采用 krskgf（Version: 1.2）、gapclose（Version: 1.12）等软件对初步组装结果进行优化和补洞，从而得到最终的组装结果。

[B] Does every N in your sequence represent a gap? Alternatively, does your sequence include single or short runs of N's that represent ambiguous base calls? If not every N is a gap, what is the minimum number of N's that represent a gap? In order for us to add the assembly_gap features for you, we need to input a minimum gap size to tell our software which N's to convert.

[C] In each gap, does the number of N's represent the estimated gap size? Alternatively, are all or some of the gap sizes unknown? If there are unknown gaps, please specify which ones. For example, are all of the gaps of 100 N's unknown length?

答：N 的数目代表的就是 gaps 的大小。

[D] What type of evidence was used to assert linkage across assembly gaps? The type of linkage evidence must be one of the following:

paired-ends

align-genus

align-xgenus

align-trnscpt

within-clone

clone-contig

map

strobe

答：paired-ends

目前公司不提供序列上传服务，关于数据问题，可以帮忙核实，但是和 NCBI 格式不匹配等相关要求，需要老师按照 NCBI 官网要求格式自己修改，NCBI 系统一直在不断升级，但是我们不会由于 NCBI 升级而更改我们交付的结果形式，请知悉。

2.2.6 关于组装结果评价

框架图我们不做组装结果好坏的评价，由于不同的菌株是不一样的，这个无法统一评价，可以让老师根据基因组大小 N50 scaffold 等指标可以自行判定；

三代组装结果评价除了细菌基因组成环与否，真菌的 N50 指标外，还可以进一步做 BUSCO 评估(version 3.0.2)评估基因组的质量、完整性。

2.2.7 BUSCO 评估

BUSCO 利用 OrthoDB 直系同源数据库构建了六种主要的系统进化分枝 (Bacteria、Eukaryota、Protists、Metazoa、Fungi、Plants) 的基因集，通过同源基因数据库从基因完整度层面上评估基因组的组装质量。虽然每个物种的基因组各不相同，但对于亲缘关系较近的物种来讲，它们之间总存在一些保守的序列。基于这个特征，BUSCO 根据 OrthoDB 数据库，针对几个大的进化分支分别构建了单拷贝基因集。在得到某物种组装后的基因组或者转录本序列后，可以将组装结果与该物种所属进化分

支的基因集中的保守序列进行比对，鉴定组装的结果是否包含这些序列，包含单条、多条还是部分或者不包含等情况给出结果。其中，对于基因组，BUSCO 首先调用 Augustus 软件进行基因结构预测，再使用 HMMER3 比对参考基因集。最终根据比对上的序列比例、完整性等，评估组装结果的准确性和完整性。

交付结果如下：

```
# BUSCO version is: 3.0.2
# The lineage dataset is: fungi_odb9 (Creation date: 2016-02-13, number of species: 85,
number of BUSCOs: 290)
# To reproduce this run: python run_BUSCO.py -i
TN3.1.assembly.fasta.review.assembly.FINAL.fasta -o BUSCO -l BUSCO/fungi_odb9/ -m genome
-c 30 -sp aspergillus nidulans
#
# Summarized benchmarking in BUSCO notation for file
TN3.1.assembly.fasta.review.assembly.FINAL.fasta
# BUSCO was run in mode: genome

C:99.3%[S:3.4%,D:95.9%],F:0.3%,M:0.4%,n:290

288   Complete BUSCOs (C)
10    Complete and single-copy BUSCOs (S)
278   Complete and duplicated BUSCOs (D)
1     Fragmented BUSCOs (F)
1     Missing BUSCOs (M)
290   Total BUSCO groups searched
```

Complete BUSCOs (C)：多少个 BUSCO 测试基因被覆盖；

Complete and single-copy BUSCOs (S)：多少个基因经过比对发现是单拷贝；

Complete and duplicated BUSCOs (D)：多少个基因经过比对发现是多拷贝；

Fragmented BUSCOs (F)：多少个基因经过比对覆盖不完全，只是部分比对上；

Missing BUSCOs (M)：没有得到比对结果的基因数；

Total BUSCO groups searched (n)：测试基因的总条目。

2.2.8 关于 L50 计算

L50-产生 N50 长度的 contig 最小数量

统计 scaffold L50 或是 contig 的 L50 的话，老师可以把需要统计的序列发给我们，不需要 reads。

更具体的 L50 解释说明：

http://www.360doc.com/content/19/1224/14/68068867_881789567.shtml

