QC方面

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。

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。

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。

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。

组装和评估

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，或者覆盖深度不足的场合，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。

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.

summary.txt文件中 Heterozygous 的含义

summary.txt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”表明菌株杂合情况。

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。

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。

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。

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

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有所的区域，这部分没 有获得的区域就称为 Gap。例如一细菌基因组测序，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.

细菌完成图组装的流程是什么?

在开发的流程中，我们按照以下流程：

1.使用二代测序数据，预估样本基因组大小和杂合度，评估是否为复杂基因组。

2.使用组装软件进行组装。

3.组装后的最长的序列是否达到预估的基因组的90%，如果达到90%阈值，则视为组装完成。若没有到达，自动调整组装参数或组装软件直至达到预估基因组大小90%。

4.达到预估基因组大小90%后，将最长的序列当作染色体序列，其他序列使用RFplmid软件鉴别序列是否为质粒。通过获得的质粒序列，再利用 BLAST软件和PLSDB数据库进行质粒注释。若其他序列鉴定不为质粒，则舍去。

5.若所有参数或软件都组装不成果，则进行人工手动调整。

为何NCBI上物种染色体有多条，而完成图组装结果只有1条?

可能的原因有以下情况：

1. 物种各条染色体差异特别巨大。大染色体和小染色体差异有10倍左右可能会遗漏小染色体。

> 在细菌世界中，拥有多条染色体的物种并不常见。在已知的细菌中，大多数都只有一条环状染色体。然而，仍然存在一些拥有多条染色体的细菌物种。以下是一些例子：
>
> 1. **Vibrio cholerae（霍乱弧菌）**：V. cholerae 是一种革兰氏阴性细菌，引起霍乱。这种细菌有两条染色体，分别为大染色体和小染色体。它们的大小分别为 2.96 Mb 和 1.07 Mb（兆碱基对）。
>
> 2. **Agrobacterium tumefaciens（冠状根癌土壤杆菌）**：A. tumefaciens 是一种革兰氏阴性细菌，侵染植物并导致冠状根癌。这种细菌有两条染色体，分别为大染色体和小染色体。它们的大小分别为 2.8 Mb 和 2.1 Mb。
>
> 3. **Burkholderia cepacia（博克霍尔德氏菌）**：B. cepacia 是一种革兰氏阴性细菌，与许多植物病害和人类感染有关。这种细菌有三条染色体，它们的大小分别为 3.9 Mb、3.2 Mb 和 0.9 Mb。
>
> 4. **Rhodobacter sphaeroides（球状红细菌）**：R. sphaeroides 是一种革兰氏阴性细菌，能通过光合作用产生能量。这种细菌有两条染色体，分别为大染色体和小染色体。它们的大小分别为 3.0 Mb 和 0.9 Mb。

> 除此之外，一些细菌物种如**Bradyrhizobium diazoefficiens**和**Leptospira interrogans**等也具有多条染色体。需要注意的是，多条染色体并不是细菌世界的普遍现象，因此这些例子相对较少。

2.组装出错

如有疑问，请联系售后。

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，从比对结果可以得知质粒信息。也可以根据02.Assembly/*/*.all.plassmid.blast.summary.top5文件内的ACC_NUCCORE号到ncbi上查找。

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，使用RFplmid软件鉴别序列是否为质粒。通过获得的质粒序列，再利用 BLAST软件和PLSDB数据库进行质粒注释。若其他序列鉴定不为质粒，则舍去。

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。

组装结果染色体不止1条?

- 1.该菌株确实有多条染色体（或巨质粒）；
- 2.该菌株是复杂基因组；

  基因组大小 >7M；染色体基因组个数 + 质粒个数 > 4； 转座子或其它重复序列在基因组中的比例异常高 (> 10%) 最长重复序列大于9K；GC含量异常高等可以称为复杂基因组
- 3.该菌株有污染等
- 详情请联系售后。

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%。

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。

基因预测

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.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。

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。

为何基因预测prokka文件夹预测的tRNA结果和rRNA结果与 ncRNA 注释有的结果不一致?

这是因为二者使用的软件是不同的导致的。一般来说，prokka结果适合大多数研究，而ncRNA结果对某些物种更精确。客户可以根据自身研究需要，选择合适的结果。

为什么在 genebank 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 genebank 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。

菌种鉴定

为什么16S鉴定结果和客户预期结果不同?

这很有可能是16s物种鉴定限制导致的。

16S rRNA序列是一种广泛用于微生物分类和进化研究的分子标记。虽然16S rRNA序列鉴定可以用于大多数微生物的物种鉴定，但也存在一些限制。以下是一些常见的16S序列物种鉴定的限制：

1. 限制于特定分类级别：16S rRNA序列鉴定通常用于物种或属级别的鉴定，但对于更低的分类级别，如亚种或生物型，其鉴定效果可能会受到限制。
2. 低分辨率：16S rRNA序列鉴定的分辨率相对较低，可能无法区分高度相似的菌株或菌种。
3. 基因组重复：一些微生物的基因组中可能存在多个16S rRNA基因副本，这些副本之间可能存在差异，导致鉴定结果的不确定性。
4. 基因水平的进化：16S rRNA序列的进化速率相对较慢，可能无法区分一些演化分化较近的菌株或菌种。
5. 缺乏参考序列：一些微生物的16S rRNA序列可能尚未被记录在已知的数据库中，这可能会限制其鉴定的准确性和可靠性。

综上所述，16S rRNA序列鉴定对于微生物物种鉴定有一定的限制。为了获得更准确和可靠的鉴定结果，建议采用多种分子标记和/或生物学特征进行鉴定，以及结合传统的生物学方法进行确认。

为什么ANI结果中有些超过阈值，有些没有?

本流程是根据16S得到的物种分类号后从NCBI上随机下载10条基因组序列进行ANI分析，其下载的基因组序列质量不能保证（尽管尽可能按照组装完整程度进行下载）。ANI的结果可能会受到数据质量的影响。低质量的序列或有缺失的序列可能会影响ANI的计算结果。

组分预测

为何在基因组内没有找到基因岛/噬菌体序列/crispris序列/插入序列?

可能原因有以下几点:

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。细菌中，不同样本会出现不同的情况，没有基因岛/噬菌体序列/crispris序列/插入序列也是正常的现象。

3.软件参数阈值设置太严格。调整参数阈值或者换软件进行查找。

客户可以按照3，2，1的顺序进行排查。

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 SingnalP，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。

功能预测

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。

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。

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。

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。E 值和 identity 没有关系，也不能换算。E值适合于有一定长度，而且复杂度不能太低的序列。一般经验来看当E值小于10-5时，表明两序列有较高的同源性，而不是因为计算错误。当E值小于10-6时，表明两序列的同源性非常高，几乎没有必要再做确认。

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?

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.

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

杂项

数据应该如何打开?

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）。

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 了，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。

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 47 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。

Pacbio 下机数据相关格式说明?

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。

1.bam 文件 主要分为两个部分， Bam文件主要分为两个部分，头一部分是Header，储存测序的相关信息，另一部分也即是文件的主要部分是records，这里头保存了我们的序列信息。我们这里就以subreads.bam文件为例，分析下bam文件的具体格式。可以用 `samtools view` 命令查看bam文件.

> 1. 第一列：reads信息{movieName}/{holeNumber}/{qStart}_{qEnd}.MovieName 是cell的名字，holeNumer是ZMW孔的编号，qStart和qEnd是subreads相对于ZMW reads的位置。
> 2. 第二列 (sum of flags)：比对信息 均为4 代表没有比对上，也表明了bam文件只储存了序列信息，而没有比对信息。
> 3. 第三列 (RNAM)：参考序列 值为 ，代表无参考序列
> 4. 第四列 (position) : 比对上的第一个碱基位置 0
> 5. 第五列 (Mapping quality) : 比对质量分数 255
> 6. 第六列 (CIGAR值) : 比对的具体情况
> 7. 第七列 (MRNM,) : mate 对应的染色体
> 8. 第八列 (mate position) : mate对应的位置 0
> 9. 第九列 (ISIZE, Inferred fragment size) : 推断的插入片段大小 0*
> 10. 第十列(Sequence) : 序列信息 具体的ATCG
> 11. 第十一列 (ASCII码) : 碱基质量分数 ASCII+33
> 12. 第十二列 : 可选区域 记录Reads 的总体属性包括信号长度，信号强度等信息。

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