Sequel下机数据为bam格式：

```
├── SAMPLE_B
│   ├── 1_A01
│   │   ├── m54000_666666_084229.adapters.fasta
│   │   ├── m54000_666666_084229.scraps.bam
│   │   ├── m54000_666666_084229.scraps.bam.pbi
│   │   ├── m54000_666666_084229.sts.xml
│   │   ├── m54000_666666_084229.subreads.bam
│   │   ├── m54000_666666_084229.subreads.bam.pbi
│   │   ├── m54000_666666_084229.subreadset.xml
│   │   └── m54000_666666_084229.transferdone
│   └── 2_B01
│   ├── m54000_666666_184156.adapters.fasta
│   ├── m54000_666666_184156.scraps.bam
│   ├── m54000_666666_184156.scraps.bam.pbi
│   ├── m54000_666666_184156.sts.xml
│   ├── m54000_666666_184156.subreads.bam
│   ├── m54000_666666_184156.subreads.bam.pbi
│   ├── m54000_666666_184156.subreadset.xml
│   └── m54000_666666_184156.transferdone
```

- Sequel下机数据主要存在bam格式的文件，其中subreads.bam保存为subreads的序列信息，scraps.bam保存过滤掉的接头、barcode、低质量序列等信息，通常比subreads.bam文件大小要大。bam.pbi为bam的index文件，记录了一些预先运算的统计信息以及一些辅助识别序列的信息。
- Sequel下机的bam文件，是高压缩格式的二进制文件，无法直接用文本方式查看。如果有需要查看的话，建议在Linux或者MacOS的终端命令行模式下，通过samtools工具来进行。
- 用于后续分析的文件一般是.subreads.bam，这等同于RS II 中的.subreads.fastq。

Sequel原始下机数据大小

- subreads.bam和scraps.bam都是高压缩的二进制文件，且包含的数据信息远多于fastq文件中的序列与碱基质量，所以文件大小与数据产量没有直接的关联。
- 一个cell数据下机的数据产量，一般会以subreads的数据总量为衡量标准

（polymerease的数据产量会比subreads多一些，但多的部分有限，偏差不大）。根据目前的经验，subreads.bam文件大小，约是subreads数据量的1.9倍。

- scraps.bam存放的是测序中的接头序列、barcode、低质量序列等信息。这一部分的数据量，与测序时的实际状况相关，无法人为干预与预估，大小也是不确定的。
- 一个完整的Sequel下机cell，数据同时包含scraps.bam和subreads.bam。subreads.bam的文件大小可以根据数据量推算出来，但是scraps.bam的大小无法预估，所以Sequel平台的原始数据大小，和数据量是没有直接关联的，文件大小也无法预估的。如果只是提供subreads.bam，则按照之前的说明评估即可。

---

bam文件中的内容格式说明：

> 详细参考：https://pacbiofileformats.readthedocs.io/en/5.1/BAM.html

1. 第一列：reads信息 {movieName}/{holeNumber}/{qStart}_{qEnd}
   [对于CCS：{movieName}/{holeNumber}/ccs]

   > 补充说明：
   > 1. 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. 这种未比对的bam必须是按第一列排序的。所有来自同一个ZMW hole的subreads放在一起，且按qStart排序，每个ZMW再按数字排序。特别是从subreads bam文件中随机取subreads分析时一定要用samtools sort -n处理一下
   > 3. qStart是从0开始的，qEnd – qStart = 第10列的碱基序列长度。

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。

   > 补充说明：
   > 如果是Subreads则全部为4表示read unmapped。在三代长reads比对结果中会出现两种：0或16。

3. 第三列 (RNAM)：参考序列 值为 ，代表无参考序列

   > 补充说明：
   > 如果是Subreads则全部为*如果是Subreads则全部为*

4. 第四列 (position) : 比对上的第一个碱基位置 0

> 补充说明：
> 如果是Subreads则全部为0如果是Subreads则全部为0

5. 第五列 (Mapping quality) : 比对质量分数 255

> 补充说明：
> 如果是Subreads则全部为255。
> Mapping qulity的计算方法是：Q=-10log10p，Q是一个非负值，p是这个序列来自这个位点的估计值。
> 假如说一条序列在某个参考序列上找到了两个位点，但是其中一个位点的Q明显大于另一个位点的Q值，这条序列来源于前一个位点的可能性就比较大。Q值的差距越大，这独特性越高。

6. 第六列 (CIGAR值) : 比对的具体情况

> 补充说明：
> 如果是Subreads则全部为*。如果是三代reads比对的结果与二代不同，不再用M字符（在二代的bam中它可以容许错配），三代中采用更为精确的X(它表示与ref不同)和=(它表示与ref相同)，如果CIGAR中检测到M字符Pacbio software将会中止运行。D表示deletion，I表示insertion。整个CIGAR字符串首尾的H和S表示硬切和软切（soft-clippedbase）。

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. 第十二列及以后 : 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。

---

bam的格式转化

- bam转为fastq、fasta文件，可以用Pacbio官方提供的bam2fasta、bam2fastq、bamtools处理，也可以用第三方工具来处理，
  如：https://gsl.hudsonalpha.org/information/software/bam2fastq。

1. 用bam2fastq/bam2fasta转化：
   需要安装smrtlink，并配置环境变量，准备工作较复杂，程序运行较简单：

```
$ bam2fastq m54000_666666_084229.subreads.bam -o
m54000_666666_084229.subreads.bam
```

说明：a、命令行输入文件是subreads.bam(绝对或者相对路径)，输出的名字可以自己指定，上述示例中，输出文件名设置的是与输入文件名一致的，但是程序会自动在输出文件名加上.fastq文件，并进行gzip压缩，所以实际产生的文件名会是m54000_666666_084229.subreads.bam.fastq.gz

b、bam2fastq转格式的时候，对文件目录格式有要求，需要在subreads.bam存在的目录中，同时存在与之对应的.pbi文件，即m54000_666666_084229.subreads.bam.pbi，否则会报错。

c、bam2fasta的使用方式及需求与bam2fastq一致，输出文件为gzip后的fasta文件。

2. 用bamtools转化：

   需要安装smrtlink，并配置环境变量，准备工作较复杂，程序运行较简单：

```
$ bamtools convert -in m54000_666666_084229.subreads.bam -out
m54000_666666_084229.subreads.bam.fastq -format fastq
```

说明：a、bamtools有很多功能，可以对bam文件进行统计、过滤、转化、切分、合并等操作，这里只用它来进行格式转化。

b、-in后面是需要转化的bam文件的绝对或者相对路径，-out后面是输出文件的名字（没有gzip压缩的），-format是需要转化的格式，一般为fastq或者fasta。

c、bamtools不要求读取.pbi文件，可以直接对单独的bam进行转化。

3. 用bam2fastq转化：

   参考：https://gsl.hudsonalpha.org/information/software/bam2fastq

---

关于sequel质量值

- Sequel的下机数据中没有质量值。由subreads转成的fastq中的质量值部分全部为“!”。
- 官方回复说明：

```
To clarify, there is no base QV on Sequel generated sequencing data.
although we did have to put something there in order to conform to the bam
spec, so we elected to use "!".
Sequel uses high quality region finder HQRF to identify longest
contiguous region of singly-loaded enzymatic activity. The current read
quality is assigned based on a SNR threshold cutoff. If the SNR is above
the threshold, then we assign a default score of 0.8 as the predicted read
quality, otherwise it's 0. Note that 0.8 is just a placeholder and should
not be used as a filter parameter.
Now, if you run a consensus algorithm on the data, either CCS2 or Arrow,
you will get meaningful per-base and overall read quality values in the
consensus output.
To conclude, there is no need to do any filtering on instrument generated
data since its already processed and high quality.
```

- 参考PPT：http://www.pacb.com/wp-content/uploads/4May2017_JimDrake_WhatsNewSMRTLink5.pdf

---

PACBIO 术语说明:

- Zero-Mode Waveguide（ZMW）孔：位于SMRT Cell上的数百纳米的小孔，测序反应在其中进行。由于孔径小于荧光基团激发所需激发光的波长，导致激发光在孔中发生了零模式波导效应，只能在孔底部几十纳米的范围传播，有效地降低了背景噪音；
- insert size: SMRTbell template中不含有发夹接头的双链核苷酸片段
- SMRT®Cells: 含有零模波导纳米结构（ZMWs）的上机耗材
- Reads：高通量测序中产生的序列标签称为Reads；
- Polymerase Reads：Pacbio产生的原始read称为Polymerase Reads；
- Subreads：去掉Polymerase Read中所包含的测序接头序列后的Polymerase Read称为subreads；
- Adapter：测序过程中添加在待测DNA片段两端的颈环状接头序列为Adapter，序列信息已知，主要用来结合测序引物；
- Circular consensus（CCS）Reads：环形一致性序列，是指使用同一条Polymerase Reads（通读待测DNA片段两圈以上）所产生的多个Subread进行一致性校正所产生的高准确度序列称为CCS Reads；
- Full-length reads：在全长转录组测序中，包含有5'端测序引物结合区、3'端

PolyA区域、3'端测序引物结合区的Polymerase Reads称为Full-length reads，通常代表的了一条完整的转录本。