

1. 准备输入文件

1.1 新建数据目录

```
mkdir /demo/data
```

```
cd /demo/data
```

1.2 准备参考基因组及注释文件

如果是人的数据，可通过如下命令直接下载我们提供的参考基因组文件

```
wget -c -O demo3k.tar "http://seekgene-public.oss-cn-beijing-internal.aliyuncs.com/seekonetools%2Fdemo3k.tar?Expires=100001662537732&OSSAccessKeyId=LTAI5t7eyRCz8dGaEhwFaMSV&Signature=4SmjSKp0zd8%2FZpkSGpII6oz52JI%3D"
```

```
tar xf demo3k.tar
```

如果是其他物种，请自行从NCBI UCSC Ensembl等数据库下载需要版本的fasta文件以及gtf文件。

之后请用seekonetools中的STAR软件构建索引，示例如下：

```
/demo/seekonetools.1.2.0/bin/STAR \
  --runMode genomeGenerate \
  --runThreadN 16 \ #线程数目
  --genomeDir /yourpath/refdata/star \ #存放构建好索引的目录，请先mkdir
  --genomeFastaFiles /yourpath/refdata/genome.fa \ #下载好的参考基因组的fa文件
  --sjdbGTFfile /yourpath/refdata/genome.gtf \ #下载好的注释文件，gtf格式
  --sjdbOverhang 149 \ #根据测序方式选择，比如PE150时此参数选择149
  --limitGenomeGenerateRAM 17179869184 #加上此参数防止参考基因组太大报错
```

其他参数可参考STARmanual。

2. 运行seekonetools

2.1 新建分析目录

```
mkdir /demo/myproject/demo3k
```

```
cd /demo/myproject/demo3k
```

2.2 运行分析命令

```
seekonetools rna run \  
  --fq1 /demo/data/demo3k_R1_001.fastq.gz \  
  --fq2 /demo/data/demo3k_R2_001.fastq.gz \  
  --samplename demo3k \  
  --outdir /demo/myproject/demo3k \  
  --genomeDir /demo/refdata/GRCh38-3.0.0/star \  
  --gtf /demo/refdata/GRCh38-3.0.0/genes.gtf \  
  --chemistry S001V3 \  
  --core 4
```

3. 其他使用说明

3.1 指定STAR版本和索引

分析时默认使用环境变量下的STAR, 也可以使用--star_path指定分析时的STAR。

注意使用的STAR版本需要和构建的索引匹配。

```
seekonetools rna run \  
  --fq1 /demo/data/demo3k_R1_001.fastq.gz \  
  --fq2 /demo/data/demo3k_R2_001.fastq.gz \  
  --samplename demo3k \  
  --outdir /demo/myproject/demo3k \  
  --star_path /yourpath/STAR \  
  --genomeDir /yourpath/refdata/star \ # 与STAR版本匹配的 索引路径  
  --gtf /yourpath/refdata/genes/genes.gtf \  
  --chemistry S001V3 \  
  --core 4
```

3.2 处理自定义结构的数据

--structure用于描述read1的接头结构, B表示cellbarcode, L表示linker, U表示UMI, 后面数字表示碱基长度, 比如B9L12B9L13B9U8; 如果是P3CBGB类的结构, 只有barcode和UMI, 没有linker, structure参数就是B17U12。

使用--barcode依次指定3段barcode, --linker依次指定2段linker, 没有linker的结构只需指定barcode文件即可。

使用--misB和--misL设置是否允许错配和缺失, 第一个数字描述错配, 第二个数字描述缺失, '1 0'表示允许错配, 不允许缺失。

```
seekonetools rna run \  
  --fq1 /demo/data/demo_R1_001.fastq.gz \  
  --fq2 /demo/data/demo_R2_001.fastq.gz \  
  --samplename demo \ --outdir /demo/myproject/demo \  
  --genomeDir /demo/refdata/GRCh38-3.0.0/star \  
  --gtf /demo/refdata/GRCh38-3.0.0/genes.gtf \
```

```
--barcode /demo/utils/CLS1.txt \  
--barcode /demo/utils/CLS2.txt \  
--barcode /demo/utils/CLS3.txt \  
--linker /demo/utils/Linker1.txt \  
--linker /demo/utils/Linker2.txt \  
--structure B9L12B9L13B9U8 \  
--misB 1 0 \  
--misL 1 0 \  
--core 4
```

3.3 其他参数

--sc5p 用于分析5'转录组数据；

--forceCell 当正常分析得到的细胞数目不理想时，选用此参数，后面加一个预期的数值N，seekonetools软件会按照UMI从高到低取前n个细胞；

--region 此参数默认为exon，此时只会选择exon reads用于定量，当此参数选择transcript时，intron reads也会用于定量。