

Shell scripting - 0704/2020

Programming for Bioinformatics - Module 3

Laurea Magistrale in Bioinformatics

A.Y. 2019-2020

Consider the following data, saved into the space-separated file `fastq.txt`.

```
1 @SRR5109807.811 811/1
2 GGCCTTGGAGGATGCATGGTTCTGGCATCTGCTTCTGGTGAGAGCCTCAGGAAACATTTATTCA
3 +
4 AAAA?C11BC?AB1B1B3F1BAG1EHH1AA13DADDDG1AGHHFHHGEBAA0FA0FGHFGFGFGB
5 @SRR5109807.812 812/1
6 GGCCTTGTAGGATGCATGGTTCTGGCATCTGCTTCTGGTGAGAGCCTCAGGAAACATTTATTCA
7 +
8 @BBABFFFFFFFFFGGFGGGGGGGGHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHGHHGGHHHHHHFGHHF
9 @SRR5109807.813 813/1
10 GGCCTTGTAGGATGCATGGTTCTGGCATCTGCTTCTGGTGAGAGCCTCAGGAAACATTTATTCA
11 +
12 BB@?AF4FFFBBGGGGGDCGGGFGHBFEHFHHHHHCFHHCF25BGHGGEC3FHGHHHGHFFGEG
13 @SRR5109807.814 814/1
14 GGCCTTGTAGGATGCATGGTTCTGGCATCTGCTTCTGGTGAGAGCCTCAGGAAACATTTATTCA
15 +
16 AAABBBFFFFFFFFFGGGGGGGGGGHHHHHHHHHHGHHHHHGBGHHGHHHHHHHHHHHHHHHHHHHHGHHH
17 @SRR5109807.815 815/1
18 GGCCTTGTAGGATGCATGGTTCTGGCATCTGCTTCTGGTGAGAGCCTCAGGAAACATTTATTCA
19 +
20 ABBCCFFFFFFFFFFGGGGGGGGGGHHHHHHHHHHHHHHHHHHHGFHHHHHHHHHHHHHHHHHHHHGHIHHHHHH
```

1. (2 points) What is the output of the following two commands?

- (a) `cat fastq.txt | sort | cut -c 1-11 | tail -4 | uniq | wc -l`
- (b) `cat fastq.txt | cut -c 1-11 | sort | tail -4 | uniq | wc -l`

2. (2 points) What is the output of the following command?

```
cat fastq.txt | grep "@.*6" | wc -l
```

3. (3 points) What is the output of the following command?

```
cat fastq.txt | awk '$1 ~ /^@/ {if(!a[$1]) {a[$1]=1; n++;}}END{print n}'
```

4. (3 points) What is the output of the following command?

```
cat fastq.txt | sed '/^@/d' | wc -l
```