Part 1

1.

|  |  |  |
| --- | --- | --- |
| File name | Sequence/Index File? | Which end? |
| 1294\_S1\_L008\_R1\_001.fastq.gz | Sequence | 1 |
| 1294\_S1\_L008\_R2\_001.fastq.gz | Index | 1 |
| 1294\_S1\_L008\_R3\_001.fastq.gz | Index | 2 |
| 1294\_S1\_L008\_R4\_001.fastq.gz | Sequence | 2 |

2.

a)

see: GitHub

b)

30 is a good cutoff score, because it provides a good balance between accuracy and not throwing out too much data. With a Phred quality score of 30, there’s only a 1 in 1000 chance of a wrong base call (99.9% accuracy.) Despite this high accuracy, the average Phred score for all 101 bp positions are above 30, so it is almost certain that not too much data will be thrown out.

c)

INPUT:

zcat /projects/bgmp/shared/2017\_sequencing/1294\_S1\_L008\_R2\_001.fastq.gz /projects/bgmp/shared/2017\_sequencing/1294\_S1\_L008\_R3\_001.fastq.gz | awk 'NR%4 == 2' | grep -c 'N'

OUTPUT:

7304664

Part 2

All on Github (pseudocode.py)