Demultiplexing Assignment

1. The R1 file is the first read, and R2 is its corresponding index. R4 is the second read, with R3 as its corresponding index.
2. A
   1. D
   2. A quality score of 30 or above is usually considered “high”. If we use that score for our insert read sequences (downstream analysis) we should preserve most of our data. For the index reads (sample identification), the mean scores fall in about the same range, with lower values towards the start of the read. The index reads are short and at the start of the read, so if you were assessing quality based on the average quality score value, a larger percentage of the reads might be thrown out. However, I chose to sort by minimum value so this shouldn’t be an issue.
   3. command line: zcat 1294\_S1\_L008\_R2\_001.fastq.gz | sed -n '2~4p'| grep -c "N"

I submitted the command in a bash script and got 3976613 + 3328051 = 7,304,664 lines with N.