Part 1 Demultiplexing | BIO 622

Jared Galloway

Part 1

1.

Filename	Read
$1294_S1_L008_R1_001.fastq.gz$	Read 1
$1294_S1_L008_R2_001.fastq.gz$	Index 1
$1294_S1_L008_R3_001.fastq.gz$	Index 2
$1294_S1_L008_R4_001.fastq.gz$	Read 2

2.

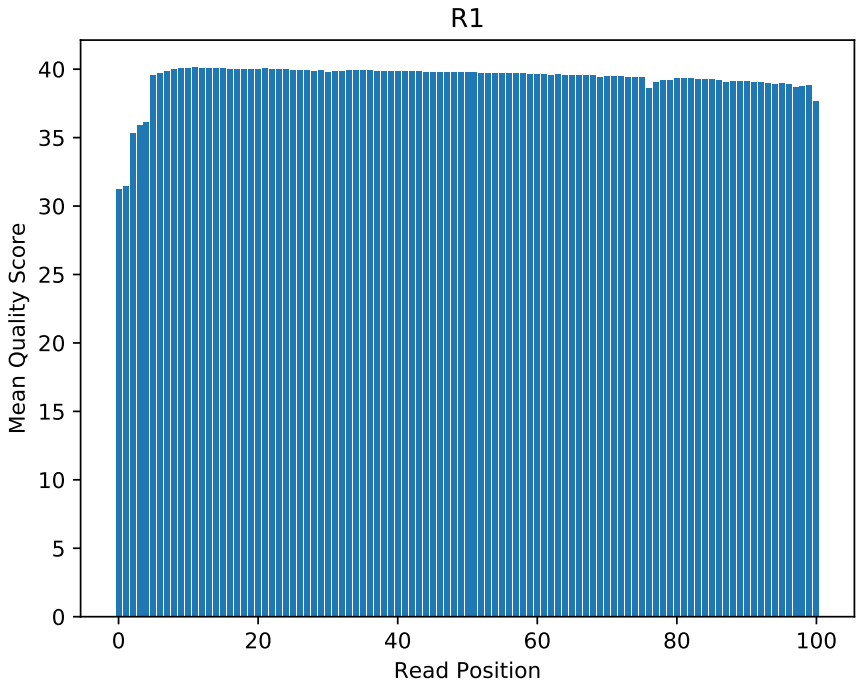

Figure 1: Distribution of average quality score R1

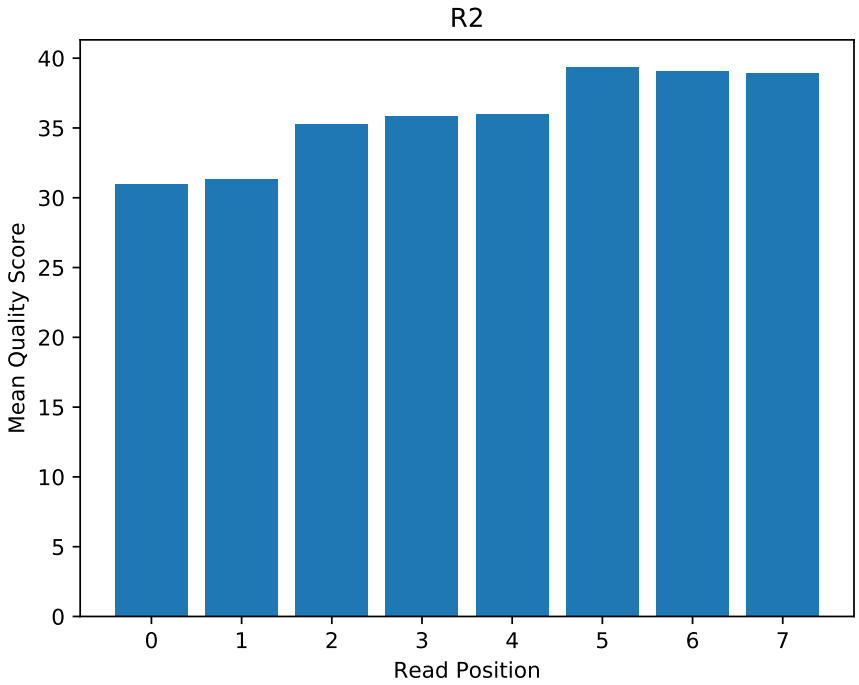

Figure 2: Distribution of average quality score R2

Due to the first two biologcal reads – for both R1 and R2 – are quite low $\approx<$ 36, I would set the cutoff for 37, to make sure we are getting high quality biolocial reads.

To find the total number of N's within the two index reads, I used the following bash command

zcat emp_files/1294_S1_L008_R[2-3]* | awk 'NR%4==2' | grep N | wc -l > count$N_{I}ndex.txt$

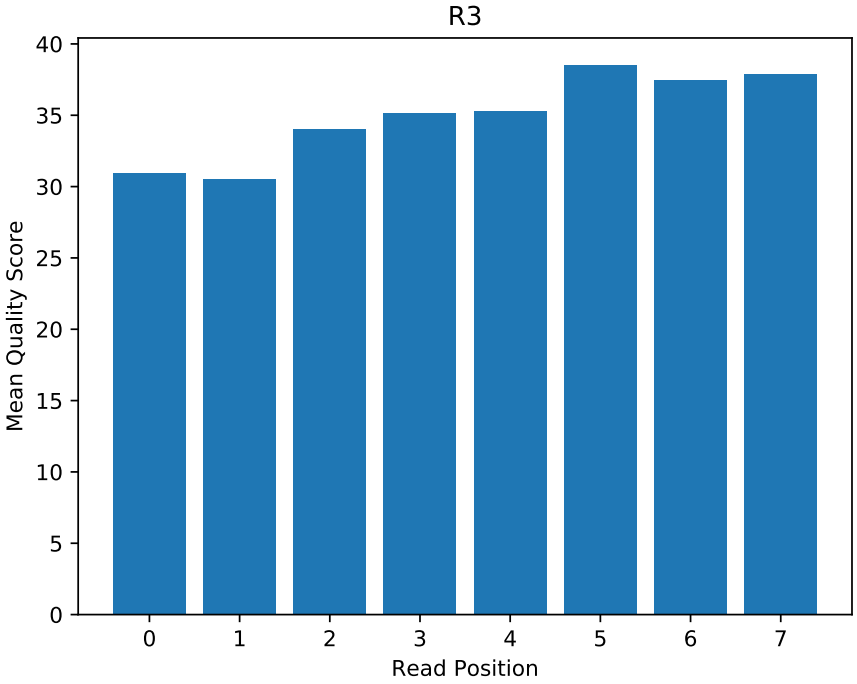

Figure 3: Distribution of average quality score R3

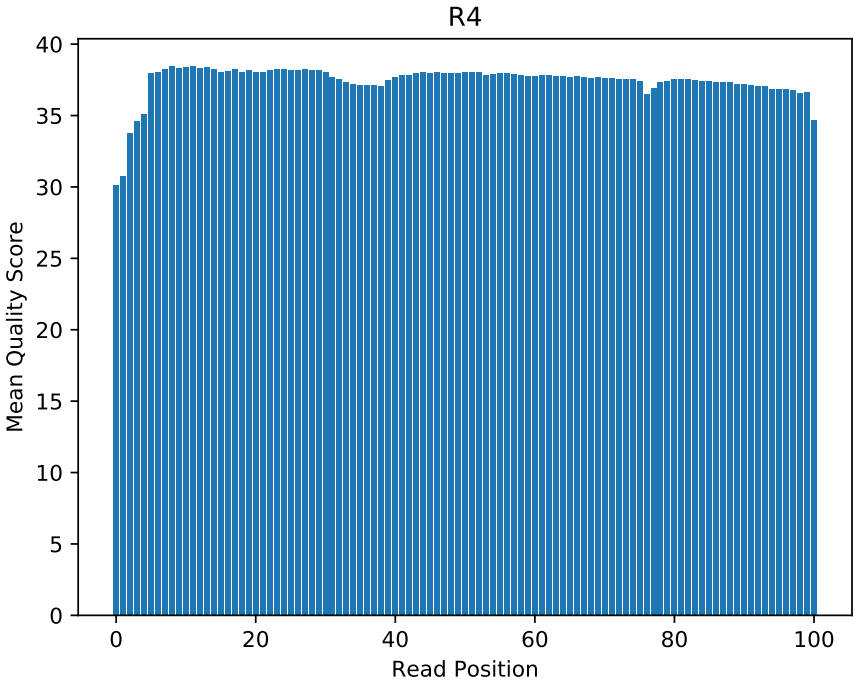

Figure 4: Distribution of average quality score R4