Index Hopping

Questions & answers

David Ho

9/12/2017

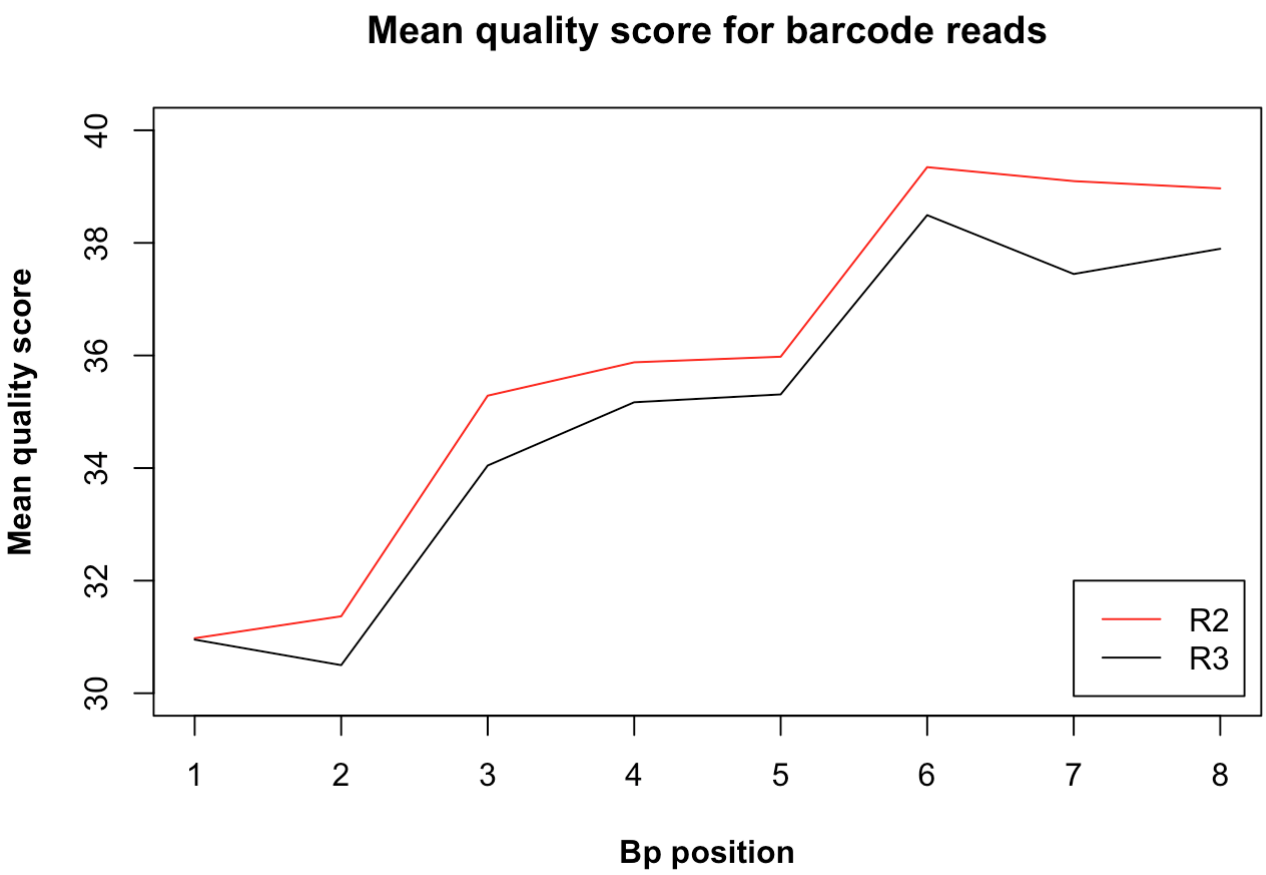
1. Generate per base call distribution of quality scores for read1, read2, index1, and index2. Generate a per nucleotide distribution as you did in part 1 of PS4 (in Leslie's class).

Next, average the Quality scores for each read (for each of the four files) and plot frequency of the Quality Scores.

a. Turn in the 8 histograms.

See `/plots/` folder.

b. What is a good quality score cutoff for index reads and pairs to utilize for sample identification and downstream analysis, respectively?



We define cutoff as the following:

```
If ALL of the nucleotides of both index reads are above (or equal to) your minimum quali
ty score, that record should be retained.
```

If we look at just the index files (`R2` & `R3`), the lowest average quality score at a particular position between the two files is 30.49904 at the 2nd position in the `R3` file. Therefore, 30 should be the cutoff for sample identification and downstream analysis. If we choose anything higher than 30, many reads will be discarded because on average, the quality score at 2nd position of `R3` is just above 30.

c. How many indexes have Undetermined (N) base calls? (Utilize your command line tool knowledge. Summit the command you used. CHALLENGE: use a one line command)

```
awk 'NR%4==2' /projects/bgmp/2017_sequencing/1294_S1_L008_R2_001.fastq | grep "N" | wc -
l
awk 'NR%4==2' /projects/bgmp/2017_sequencing/1294_S1_L008_R3_001.fastq | grep "N" | wc -
l

R2
3976613
R3
3328051
```