# Demultiplexing, Assignment-the-first, Lab notebook

## Initial Data Exploration

I used zcat and less to open each file and look at the first few lines:

```
zcat 1294_S1_L008_R1_001.fastq.gz | less
zcat 1294_S1_L008_R2_001.fastq.gz | less
zcat 1294_S1_L008_R3_001.fastq.gz | less
zcat 1294_S1_L008_R4_001.fastq.gz | less
```

These are located:

```
/projects/bgmp/shared/2017_sequencing/
```

From this, I could tell that R1 and R4 contained our reads, as all sequences were 101 basepairs long, and R2 and R3 were out index files, as all sequences were 8 basepairs long.

To tell the phred encoding, we can see if on an initial exploration, there are any of the unique Phred 33+ letters in the qscore lines of any of the files. Heres just one read from zcat 1294_S1_L008_R2_001.fastq.gz.

```
@K00337:83:HJKJNBBXX:8:1101:1265:1191 2:N:0:1
NCTTCGAC
+
#AA<FJJJ
```

In this we can see # and < used in the phred score, both of these characters are only present in Phred 33 encoding.

This was verified with all of the different files, and its confirmed they all use Phred 33. It was helpful that the first reads are all of low quality so they have N's present in them, and the N is # in phred 33 encoding.

| File name | label | Read length | Phred encoding |
|---|---|---|---|
| 1294_S1_L008_R1_001.fastq.gz | Read1 | 101 | 33 |

| File name | label | Read length | Phred encoding |
|---|---|---|---|
| 1294_S1_L008_R2_001.fastq.gz | Index1 | 8 | 33 |
| 1294_S1_L008_R3_001.fastq.gz | Index2 | 8 | 33 |
| 1294_S1_L008_R4_001.fastq.gz | Read2 | 101 | 33 |
| | | | |

Using my python script perbase_dist.py, the input files, and 4 slurm bash scripts called run_R1.sh, run_R2.sh, run_R3.sh, run_R4.sh, I created the four graphs as outputs and they are now in github, as well as all the scripts.

https://github.com/Cameron-Grey-Kunstadt/Demultiplex/blob/master/Assignment-the-first/run_R1.sh