# Demultiplexing Lab Notebook

## First Assignment

**Dual matched index**
All reads contain an index on each end and each read's indices are unique to their read but contain the same index. i.e. read 1 has index A on each end, read 2 has index B on both ends etc.

## Part 1

1. Do data exploration
    1. To find the length of each sequence line:
        1. `zcat 1294_S1_L008_R1_001.fastq.gz | head -12 | awk 'NR%4 == 2 { print length(), $0}'`
        2. `zcat 1294_S1_L008_R2_001.fastq.gz | head -12 | awk 'NR%4 == 2 { print length(), $0}'`
        3. `zcat 1294_S1_L008_R4_001.fastq.gz | head -12 | awk 'NR%4 == 2 { print length(), $0}'`
        4. `zcat 1294_S1_L008_R4_001.fastq.gz | head -12 | awk 'NR%4 == 2 { print length(), $0}'`
    2. To find the phred encoding for each file
        1. `zcat 1294_S1_L008_R4_001.fastq.gz | head -8 | awk 'NR%4 == 0 { print($0)}' | grep ".<"`

| R1 | R2 | R3 | R4 |
|---|---|---|---|
| Forward read (read 1) | Index 1 | Index 2 | Reverse read (Read 2) |
| 101 | 8 | 8 | 101 |
| PHRED +33 | PHRED +33 | PHRED +33 | PHRED +33 |

2. Per base nucleotide distribution
    1. Ran the following sbatch scripts:
        1. `sbatch R1_demultiplex_qual_score_dist_sbatch.sh` | Job ID = 7761171
        2. `sbatch R2_demultiplex_qual_score_dist_sbatch.sh` | Job ID = 7761172
        3. `sbatch R3_demultiplex_qual_score_dist_sbatch.sh` | Job ID = 7761174
        4. `sbatch R4_demultiplex_qual_score_dist_sbatch.sh` | Job ID = 7761175

2. To determine a good quality score cutoff, I added a standard deviation code function to get a better understanding of the lower bound for the average quality scores for the indexes. The quality score threshold will be one standard deviation below the mean of the lowest nucleotide distribution.
    1. If I make the threshold at 2 standard deviations below the mean quality score for all nucleotide positions (35.875), then (if normally distributed) I should be pulling out ~2.5% of my reads due to low quality score indexes.
3. To count how many indexes contain undetermined base calls (contains an N) I used the following bash command:
    1. `zcat /projects/bgmp/shared/2017_sequencing/1294_S1_L008_R2_001.fastq.gz | head -12 | awk 'NR%4 == 2 {print($0)} | grep -c "N"`
    2. 3,976,613 indexes in file 2 contain undetermined base calls.
    3. 3,328,051 indexes in file 3 contain undetermined base calls.

Make a quality score distribution per nucleotide position.
y-axis: Mean QS
x-axis: NT position

## Part 2

Make pseudocode for the script
Will have 52 files for the 24 known indexes.
.gitignore will allow FASTQ files in the test folders only.
don't forget to add the functions.

# Part 3

Command to test: (Should put all files into test folder)

`./demultiplex.py -b indexes.txt -r1 R1_unit_test.fq -r2 R2_unit_test.fq -r3 R3_unit_test.fq -r4 R4_unit_test.fq -l 101 -p test_folder/test_plot.png -q 27`

sbatch was giving a permission denied error so I had to run an interactive node. Command ran with quality score threshold:

```
./demultiplex.py -b indexes.txt -r1 /projects/bgmp/shared/2017_sequencing/1294_S1_L008_R1_001.fastq.gz -r2 /projects/bgmp/shared/2017_sequencing/1294_S1_L008_R2_001.fastq.gz -r3
```

```
/projects/bgmp/shared/2017_sequencing/1294_S1_L008_R3_001.fastq.gz -r4
/projects/bgmp/shared/2017_sequencing/1294_S1_L008_R4_001.fastq.gz -l 101 -p
Final_dir20_real/Demultiplex_dist_plot.png -q 20
```

I had to run this on my hotspot which kept cutting out and stopping the run because I had to use an interactive node.

Distribution plot did not work correctly but the stats file appears correct.