Initial Data Exploration

```
1294_S1_L008_R1_001.fastq.gz   1294_S1_L008_R3_001.fastq.gz   demul-
tiplexed  README.txt

1294_S1_L008_R2_001.fastq.gz          1294_S1_L008_R4_001.fastq.gz
indexes.txt
```

files in /projects/bgmp/shared/2017_sequencing

`ls - lah`
to look at the size of all the files

`zcat <file> | head`
confirmed the files were fastq

`zcat <file> | wc -l`
looking at number of lines for files
takes a long time!
1452986940 lines in 1294_S1_L008_R1_001.fastq.gz
all files have this many records
all reads are generated from the same piece of DNA

When comparing the first record from all four files, R2 and R3 are much smaller because they are indexes
remember that the order of files is as follows:
1 = forward read
2 = index 1
3 = index 2
4 = reverse read

`zcat 1294_S1_L008_R1_001.fastq.gz | head -2 | grep -v "^@" | wc -m`
to count the read length of the sequences in the file
subtract 1 from the output number to account for the newline character

`zcat 1294_S1_L008_R1_001.fastq.gz | head -4`
look at the first record and compare the fourth line (quality score line) to the ASCII table
Phred+33 for quality scores in these files

```
@NS500451:154:HWKTMBGXX:1:11101:10065:1121 1:N:0:TAGAACAC
AGGTTGCTATGAAATTTTAGTTGTCGTAGTAGGCAAACAATAAGGAATGTTGATCCAATAATTACATGGAGTCCATGGAA
+
AAAAAEEEA6EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEEEEEE
```

Quality scores

```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
.........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
.............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
.............................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
|                         |    |        |                              |                     |
33                        59   64       73                            104                   126

 S - Sanger        Phred+33,  raw reads typically (0, 40)
 X - Solexa        Solexa+64, raw reads typically (-5, 40)
 I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
 J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
 L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

ASCII values 33 through 73 correspond to phred scores 0 through 40 in Phred+33 encoding

'E' = 69 69 – 33 = **36** p_{error} = 0.025% Base call accuracy: 99.975%

http://en.wikipedia.org/wiki/FASTQ_format