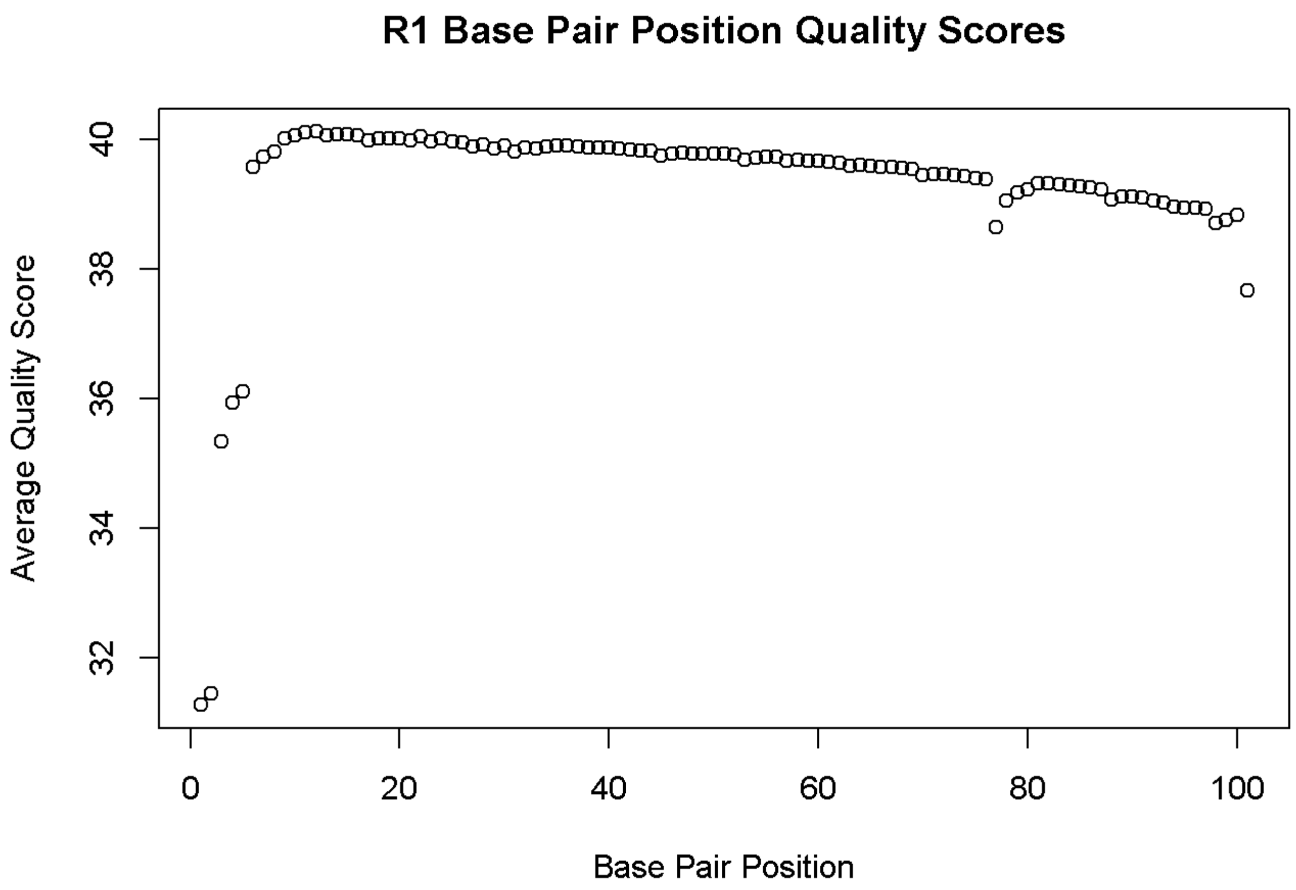
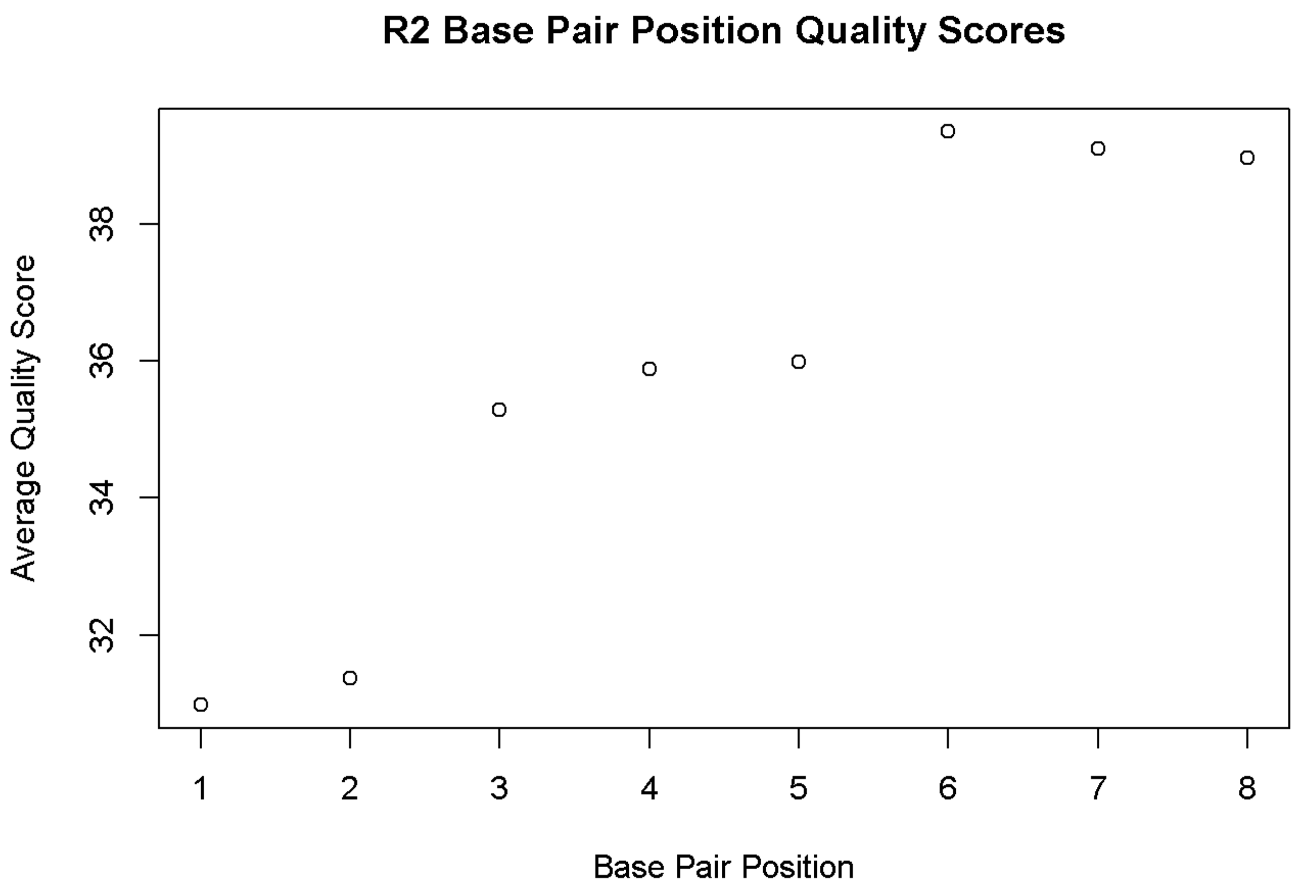
Index Swapping Revisited

1.A Histogram Generation

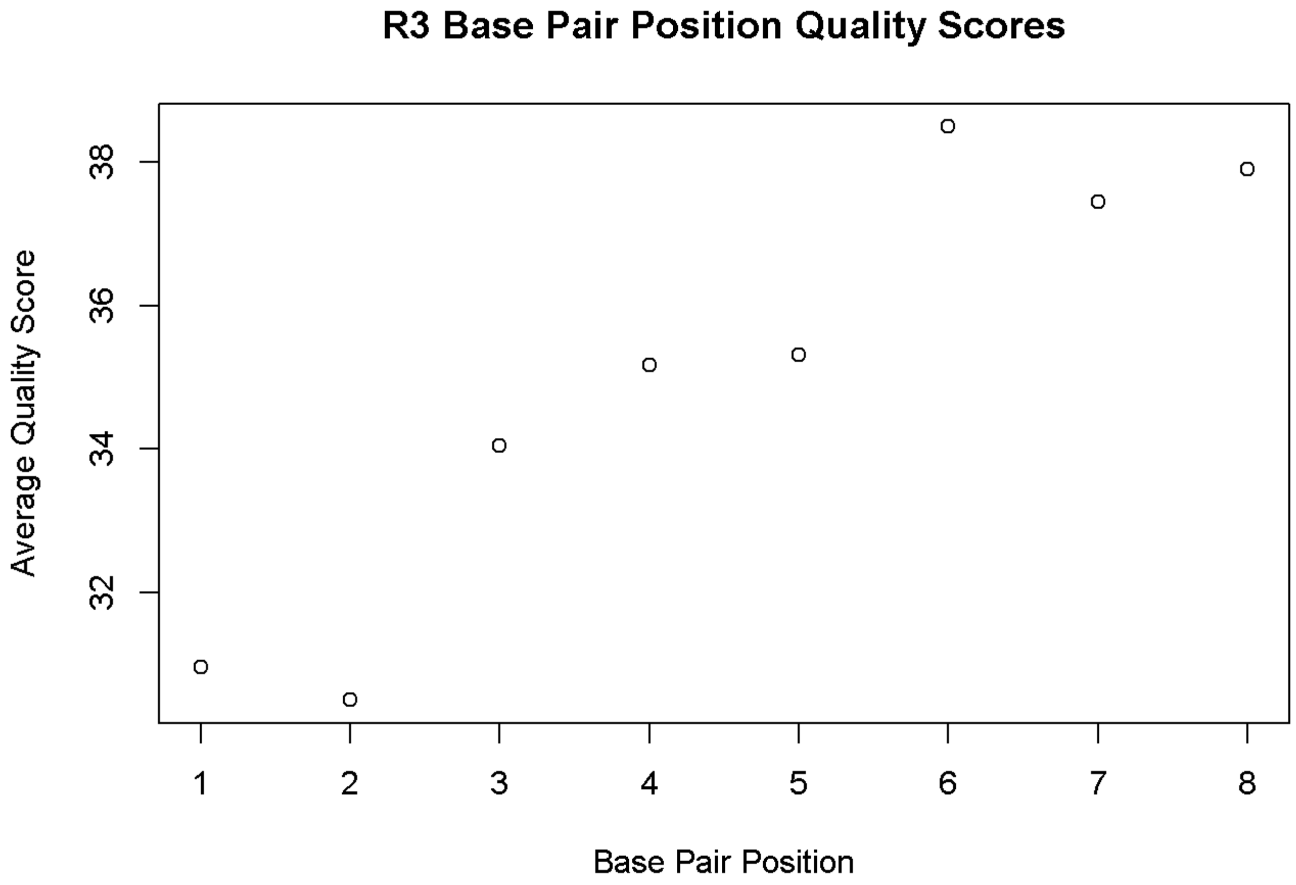
```r
setwd("C:/Users/Ryan1/Documents/FinalAssignment")
test1 <- read.table('test.txt')
test2 <- read.table('test1.txt')
test3 <- read.table('test2.txt')
test4 <- read.table('test3.txt')

test1 <- cbind(rownames(test1), test1)
rownames(test1) <- NULL
colnames(test1) <- c("V1","V2")

test2 <- cbind(rownames(test2), test2)
rownames(test2) <- NULL
colnames(test2) <- c("V1","V2")

test3 <- cbind(rownames(test3), test3)
rownames(test3) <- NULL
colnames(test3) <- c("V1","V2")

test4 <- cbind(rownames(test4), test4)
rownames(test4) <- NULL
colnames(test4) <- c("V1","V2")

test5 <- seq(1, 101, by=1)
test6 <- seq(1, 8, by=1)

plot(test1$V2~test5, xlab = 'Base Pair Position', ylab = 'Average Quality Score', main = 'R1 Base Pair Position Quality Scores')
```



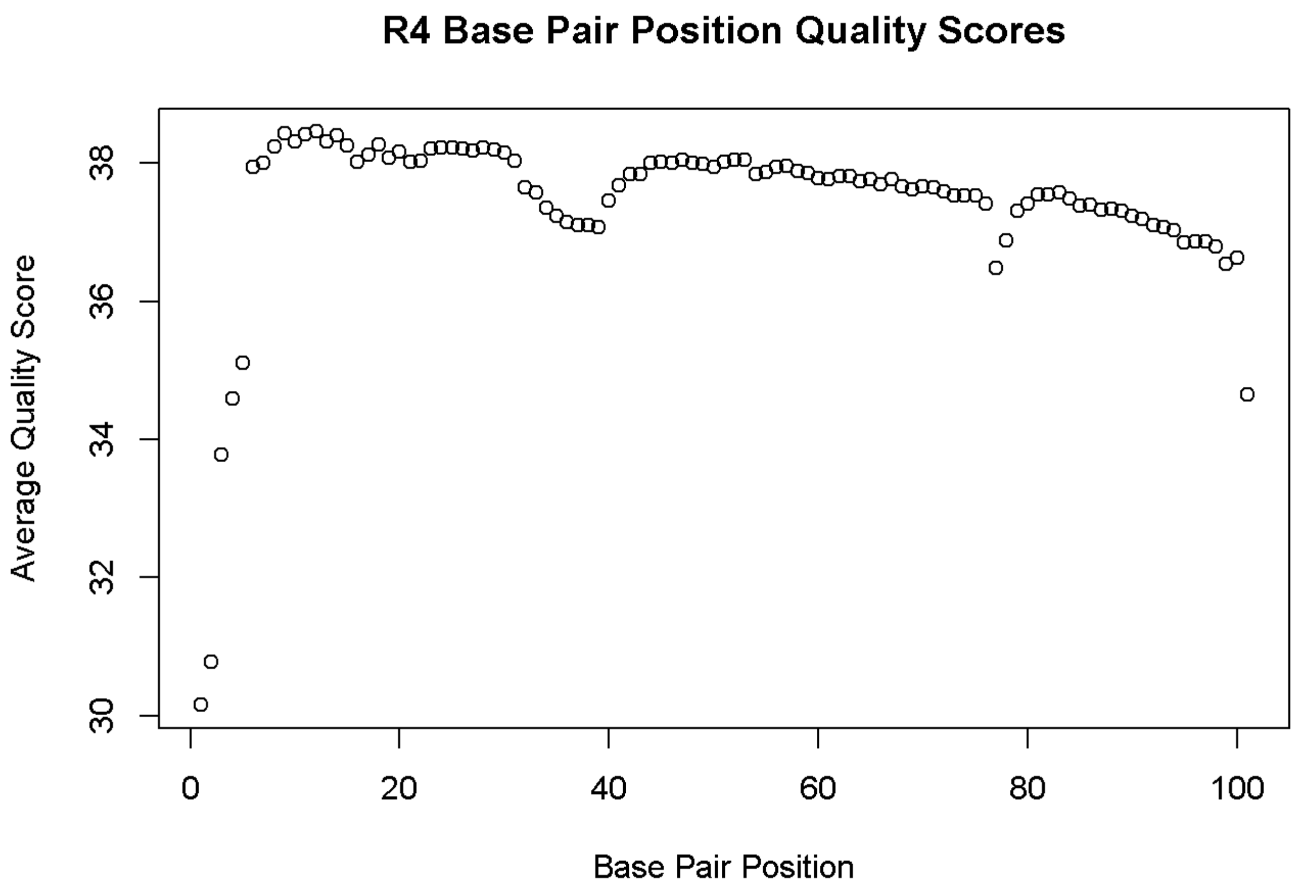
```
plot(test2[1:8,]$V2~test6, xlab = 'Base Pair Position', ylab = 'Average Quality Sco
re', main = 'R2 Base Pair Position Quality Scores')
```



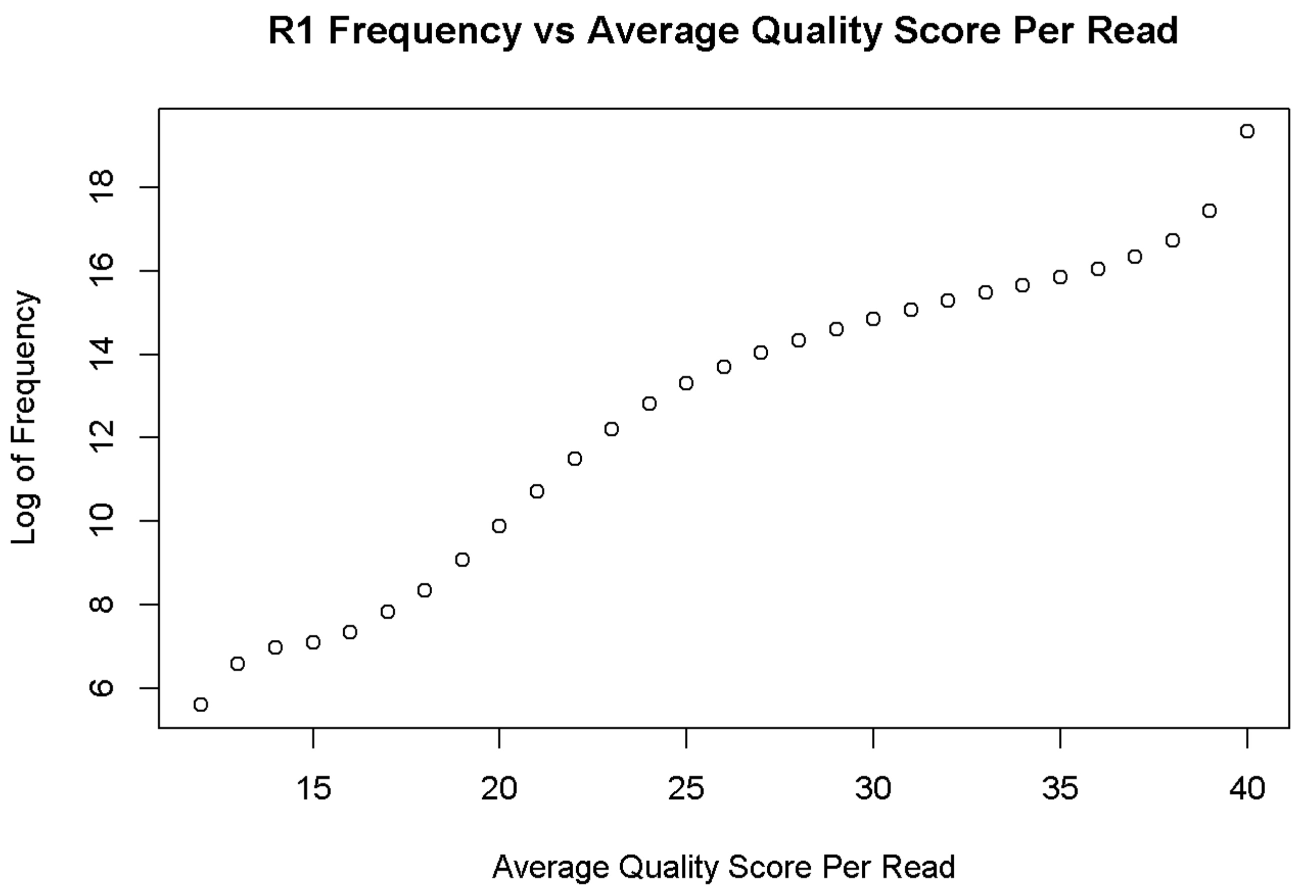
```
plot(test3[1:8,]$V2~test6, xlab = 'Base Pair Position', ylab = 'Average Quality Score', main = 'R3 Base Pair Position Quality Scores')
```



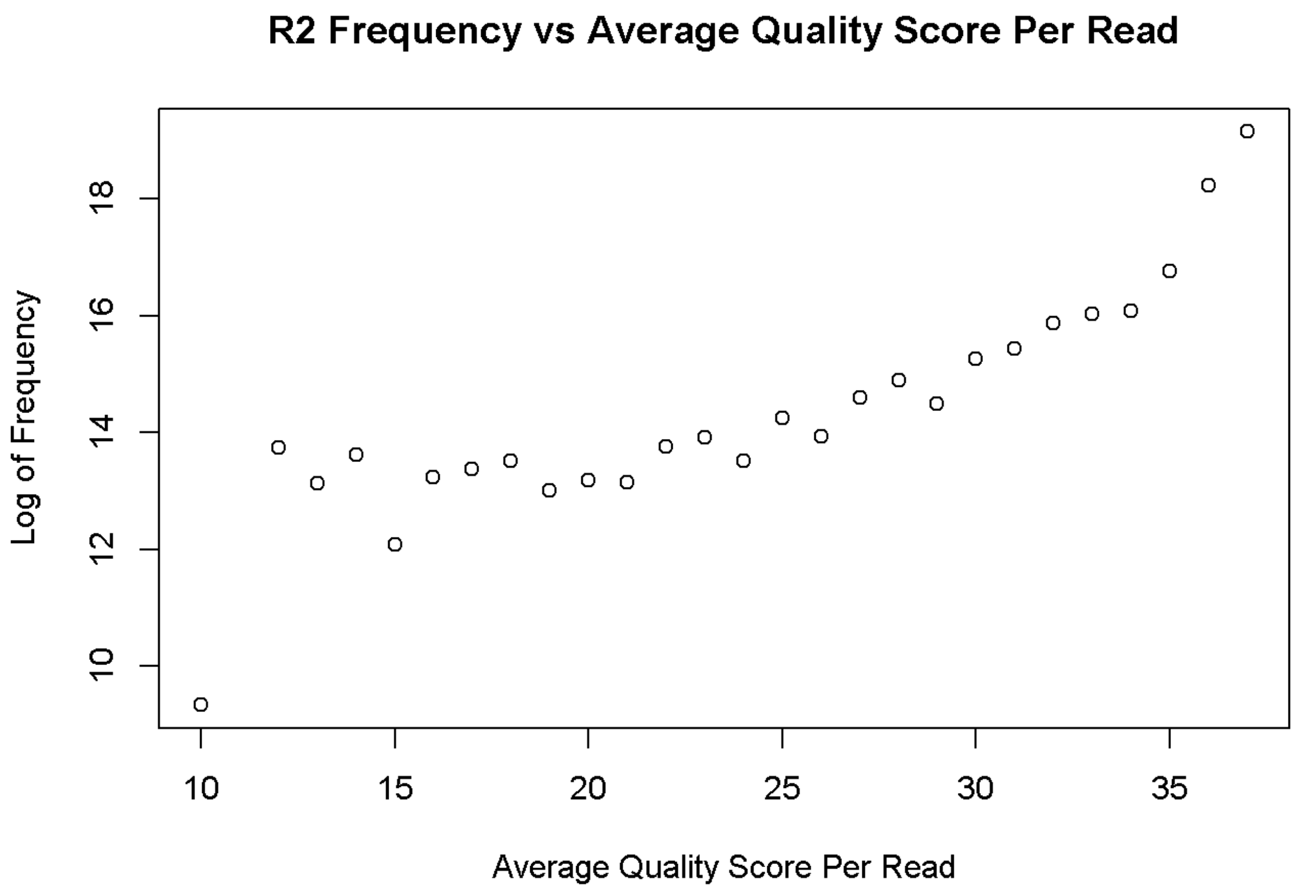
```
plot(test4$V2~test5, xlab = 'Base Pair Position', ylab = 'Average Quality Score', m
ain = 'R4 Base Pair Position Quality Scores')
```



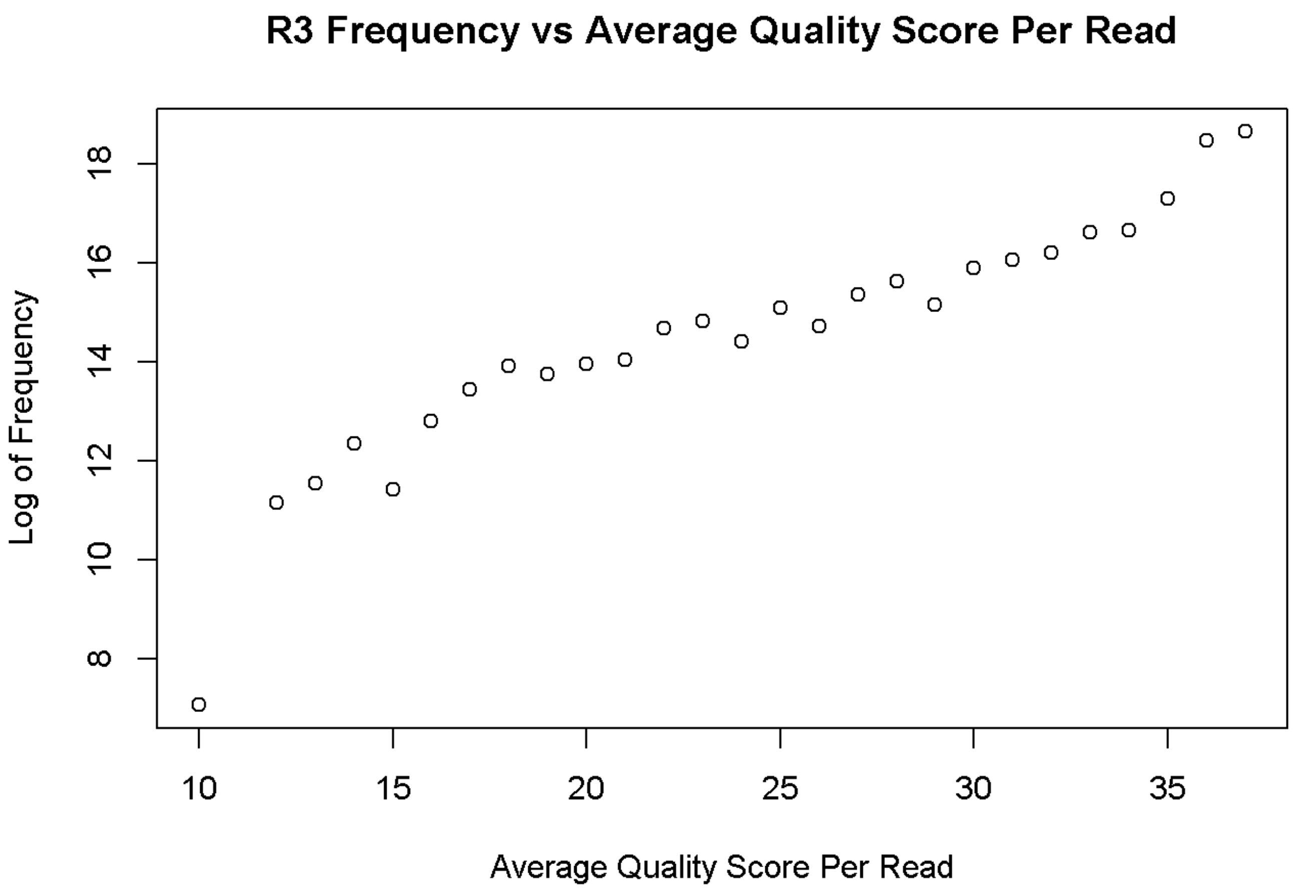
```
test20 <- read.table('test20.txt', sep = ':')
test21 <- read.table('test21.txt', sep = ':')
test22 <- read.table('test22.txt', sep = ':')
test23 <- read.table('test23.txt', sep = ':')

plot(log(test20$V2)~test20$V1, xlab = 'Average Quality Score Per Read', ylab = 'Log
of Frequency', main = 'R1 Frequency vs Average Quality Score Per Read')
```



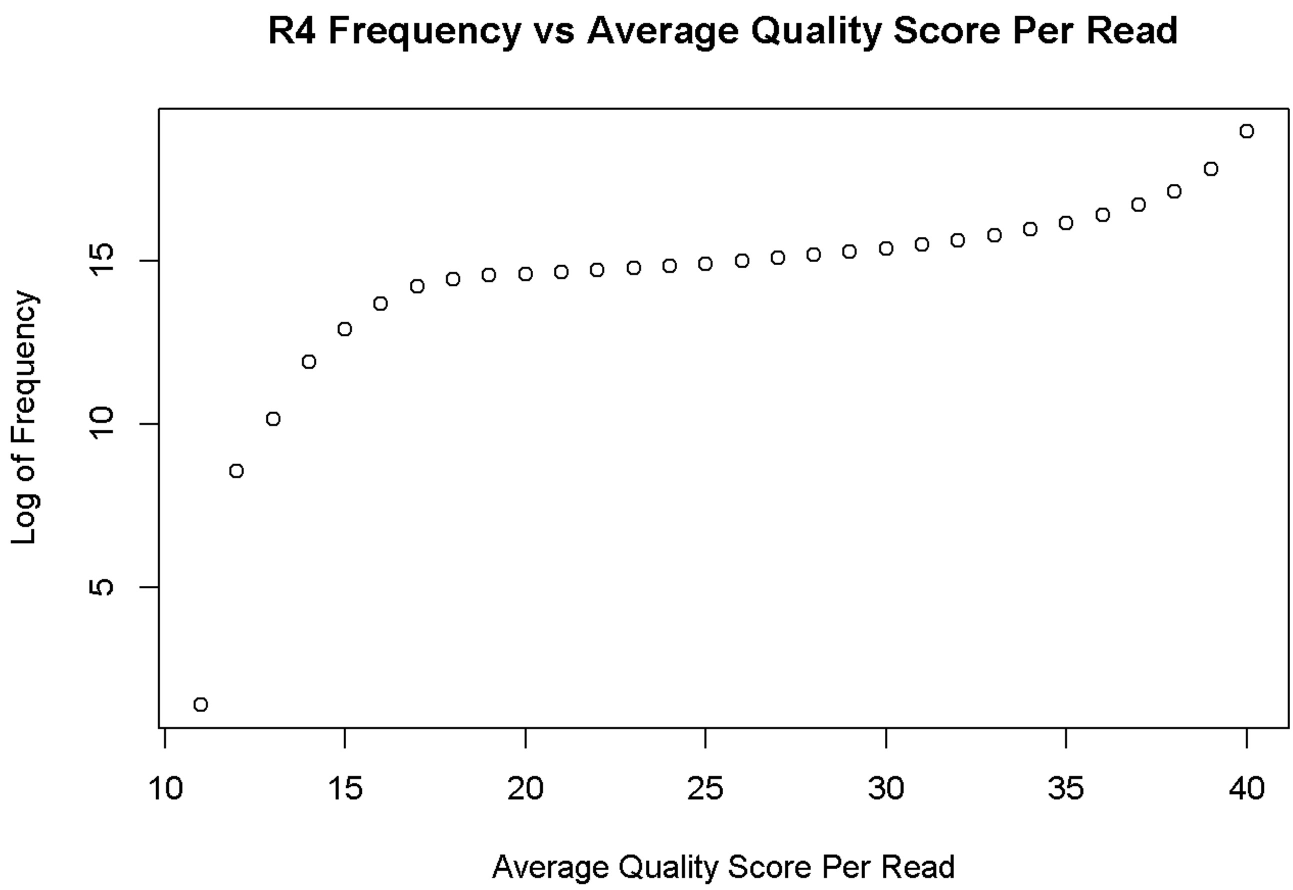
```
plot(log(test21$V2)~test21$V1, xlab = 'Average Quality Score Per Read', ylab = 'Log of Frequency', main = 'R2 Frequency vs Average Quality Score Per Read')
```



```
plot(log(test22$V2)~test22$V1, xlab = 'Average Quality Score Per Read', ylab = 'Log
of Frequency', main = 'R3 Frequency vs Average Quality Score Per Read')
```



```
plot(log(test23$V2)~test23$V1, xlab = 'Average Quality Score Per Read', ylab = 'Log
of Frequency', main = 'R4 Frequency vs Average Quality Score Per Read')
```



1.B

I believe a good quality score cutoff would be 30 for both index reads and pairs based on the plots generated.

1.C awk “NR%4==2” /projects/bgmp/2017_sequencing/1294_S1_L008_R2_001.fastq | grep “N” | wc -l

3976613 reads for 1294_S1_L008_R2_001.fastq

awk “NR%4==2” /projects/bgmp/2017_sequencing/1294_S1_L008_R3_001.fastq | grep “N” | wc -l

3328051 reads for 1294_S1_L008_R3_001.fastq

1.D

Based on the averaged quality scores across our reads, it tells us that the quality of our data is very high given that the majority of our reads are above a cutoff of 30.

2.A

GTAGCGTA_GTAGCGTA:8119243 2.235%

CGATCGAT_CGATCGAT:5604966 1.543%

GATCAAGG_GATCAAGG:6587100 1.813%

TAGCCATG_TAGCCATG:10629633 2.926%

CGGTAATC_CGGTAATC:5064906 1.394%

CTCTGGAT_CTCTGGAT:34976387 9.628%

TACCGGAT_TACCGGAT:76363857 21.022%

CTAGCTCA_CTAGCTCA:17332036 4.7714%

CACTTCAC_CACTTCAC:4191388 1.154%

GCTACTCT_GCTACTCT:7416557 2.017%

ACGATCAG_ACGATCAG:7942853 2.187%

TATGGCAC_TATGGCAC:11184304 3.079%

TGTTCCGT_TGTTCCGT:15733007 4.331%

GTCCTAAG_GTCCTAAG:8830276 2.431%

TCGACAAG_TCGACAAG:3853350 1.061%

TCTTCGAC_TCTTCGAC:42094112 11.588%

ATCATGCG_ATCATGCG:10087503 2.777%

ATCGTGGT_ATCGTGGT:6887592 1.896%

TCGAGAGT_TCGAGAGT:11741547 3.232%

TCGGATTC_TCGGATTC:4611350 1.269%

GATCTTGC_GATCTTGC:3641072 1.002%

AGAGTCCA_AGAGTCCA:11316780 3.115%

AGGATAGC_AGGATAGC:8673180 2.388%

(322882999 / 363246735) *100 = 88.888% of reads had the desired index

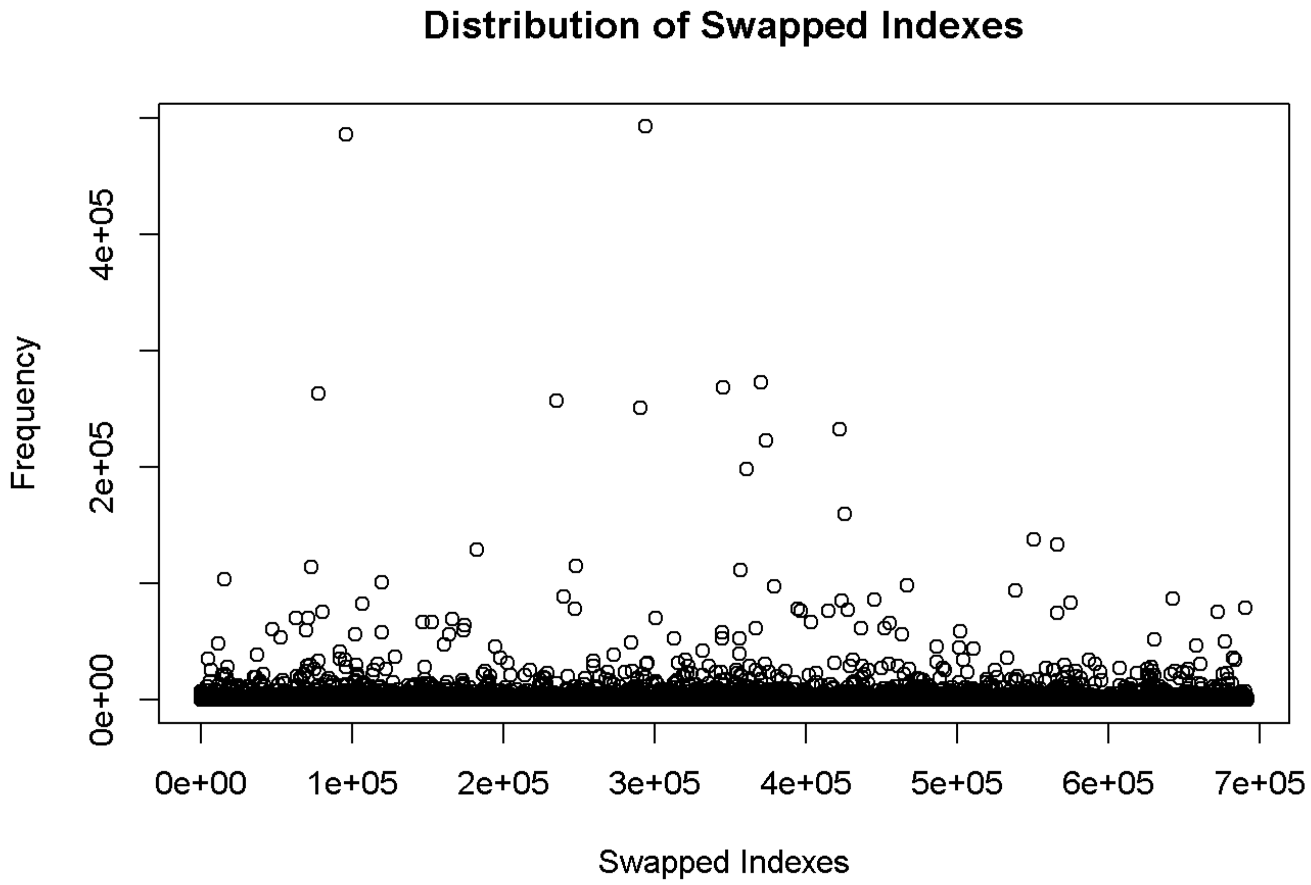
(4205183 / 363246735) *100 = 1.158% of reads had an N in the index

(36158553/ 363246735) *100 = 9.954% of reads had swapped indexes.

3.B. 36158553 reads were indicative of index swapping.

3.C

```
test24 <- read.table('FINALLYDONEWITHTHESUMMERTERM2.TXT', sep=':')
#test25 <- test24[order(test24$V2),c(1,2)]
test26 <- subset(test24, V2 > 250000, select=c(V1, V2))
plot(test24$V2, xlab = "Swapped Indexes", ylab = "Frequency", main = "Distribution
of Swapped Indexes")
```



```
test26
```

```
##                           V1     V2
## 77782  TACCGGAT_CACCGGAT 262619
## 96225  TACCGGAT_TCCCGGAT 485660
## 234997 TACCGGAT_TACCGTAT 256968
## 290772 TACCGGAT_TACCGGTT 250683
## 294044 TACCGGAT_TGCCGGAT 492220
## 345190 TACCGGAT_GACCGGAT 268356
## 370285 TACCGGAT_TTCCGGAT 272693
```

I was unable to plot the labels in R and thus just plotted all of the indexes that were swapped and the frequency in which they were swapped at. I also displayed the table of the top 7 indexes that were frequently swapped. Overall the data tells us that there was a high number of indexes swapped (almost 10% of the reads) and the majority of the swapped indexes had a very similar sequence indicating that it was likely one library that contributed to the majority of the index hopping.