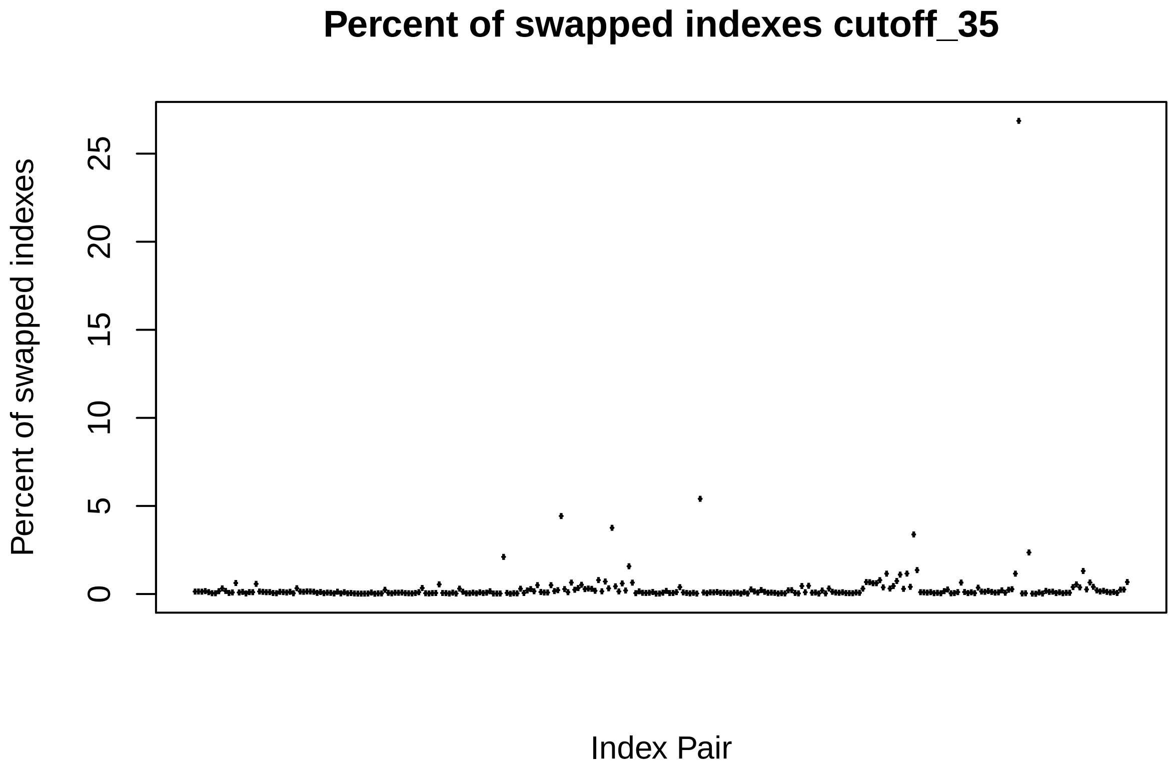
Swapping

Jake VanCampen

9/15/2017

```r
### Index hopping stats
library(stringr)

# enter directory with results from quality cutoff 30
setwd('/Users/JakeVanCampen/Documents/Bi622/Aug_7_2017/index_hop/Results_30')

# read in matched indexes from cutoff of 30
match_30 <- read.delim('match_out.tsv', sep='\t', header = 1)

# read in swapped indexes from cutoff of 30
swapped_30 <- read.delim('swapped_out.tsv', sep = '\t', header = 1)

# read in undetermined indexes from cutoff of 30
undet_30 <- read.delim('undetermined_out.tsv', sep = '\t', header = 1)


# determine the total number of reads for cutoff 30
total_30 <- sum(match_30$Counts)+sum(swapped_30$Counts)+sum(undet_30$Counts)


# determine the percent of swapped indexes for each pair
swapped_30$percent_swapped_30 <- swapped_30$Counts/sum(swapped_30$Counts) * 100


# plot the percent of swapped indexes vs. the index pairs
plot(swapped_30$percent_swapped_30 ~ swapped_30$Swapped.Index.Pair, xaxt='n',
     main = "Percent of swapped indexes cutoff_35",
     ylab = "Percent of swapped indexes",
     xlab = "Index Pair")
```

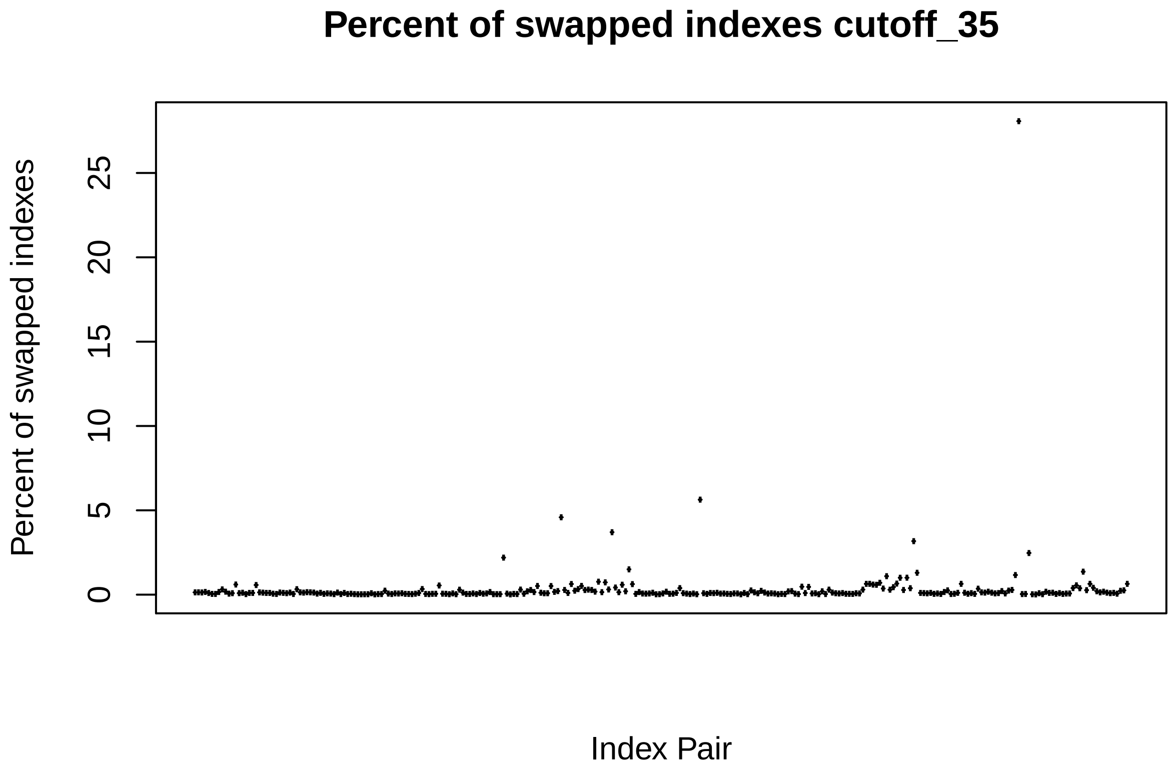


```
# This plot shows most swapping was at low percent, though some swapping oocured
# more frequently, which pairs were swapped the most?

highly_swapped_30 <- swapped_30[which(swapped_30$percent_swapped_30 > 1), ]

print(highly_swapped_30)
```

```
##     Swapped.Index.Pair Counts percent_swapped_30
## 59   GATCAAGG_TCTTCGAC  33103           5.406372
## 107  CGGTAATC_TACCGGAT  12917           2.109601
## 124  CTCTGGAT_TACCGGAT  23014           3.758640
## 133  CTCTGGAT_TCTTCGAC   9650           1.576035
## 141  TACCGGAT_CTAGCTCA   7085           1.157120
## 145  TACCGGAT_TATGGCAC   6747           1.101918
## 146  TACCGGAT_TGTTCCGT   8312           1.357513
## 149  TACCGGAT_TCTTCGAC  20721           3.384148
## 152  TACCGGAT_TCGAGAGT   7115           1.162020
## 163  CTAGCTCA_TCGACAAG  27104           4.426617
## 211  TATGGCAC_TGTTCCGT 164471          26.861355
## 214  TATGGCAC_TCTTCGAC   7081           1.156467
## 242  TCGACAAG_ATCATGCG  14458           2.361276
## 250  TCTTCGAC_ATCGTGGT   8023           1.310314
```

```
# enter directory with results from quality cutoff 35
setwd('/Users/JakeVanCampen/Documents/Bi622/Aug_7_2017/index_hop/Results_35')

# read in matched indexes from cutoff of 35
match_35 <- read.delim('match_out.tsv', sep = '\t', header = 1)

# read in swapped indexes from cutoff of 35
```

```r
swapped_35 <- read.delim('swapped_out.tsv', sep = '\t', header = 1)

# read in undetermined indexes from cutoff of 35
undet_35 <- read.delim('undetermined_out.tsv', sep = '\t', header = 1)

# determine the total number of reads for cutoff 35
total_35 <- sum(match_35$Counts)+sum(swapped_35$Counts)+sum(undet_35$Counts)

swapped_35$percent_swapped_35 <- swapped_35$Counts/sum(swapped_35$Counts) * 100

plot(swapped_35$percent_swapped_35 ~ swapped_35$Swapped.Index.Pair, xaxt='n',
     main = "Percent of swapped indexes cutoff_35",
     ylab = "Percent of swapped indexes",
     xlab = "Index Pair")
```



```r
# This plot shows most swapping was at low percent, though some swapping oocured
# more frequently, which pairs were swapped the most?

highly_swapped_35 <- swapped_35[which(swapped_35$percent_swapped_35 > 1), ]

print(highly_swapped_35)
```

```
##     Swapped.Index.Pair Counts percent_swapped_35
## 59   GATCAAGG_TCTTCGAC  15743           5.633788
## 107  CGGTAATC_TACCGGAT   6139           2.196902
## 124  CTCTGGAT_TACCGGAT  10346           3.702418
## 133  CTCTGGAT_TCTTCGAC   4194           1.500864
## 141  TACCGGAT_CTAGCTCA   3063           1.096125
## 145  TACCGGAT_TATGGCAC   2806           1.004155
## 146  TACCGGAT_TGTTCCGT   3631           1.299389
## 149  TACCGGAT_TCTTCGAC   8865           3.172428
## 152  TACCGGAT_TCGAGAGT   2815           1.007375
## 163  CTAGCTCA_TCGACAAG  12820           4.587763
## 211  TATGGCAC_TGTTCCGT  78436          28.069096
## 214  TATGGCAC_TCTTCGAC   3253           1.164118
## 242  TCGACAAG_ATCATGCG   6898           2.468517
## 250  TCTTCGAC_ATCGTGGT   3814           1.364877
```

It is interesting to note that the indexes with the highest percent of swapping were `TATGGCAC_TGTTCCGT`, accounting for 26 and 28 percent of the swapped indexed. Further research will look at the differences in indices, and if some indices should be discontinued because of their high degree of swapping.