Index Swapping Assignment

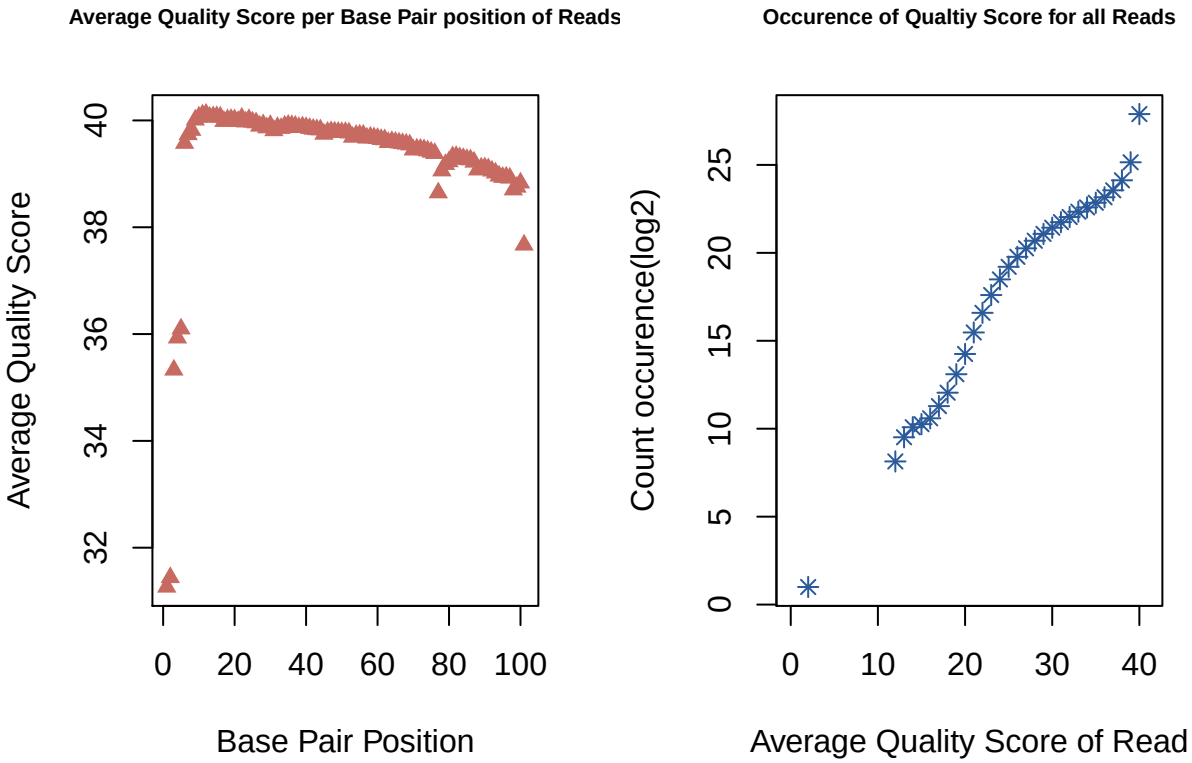
BI 622

Nicki Zavoshy

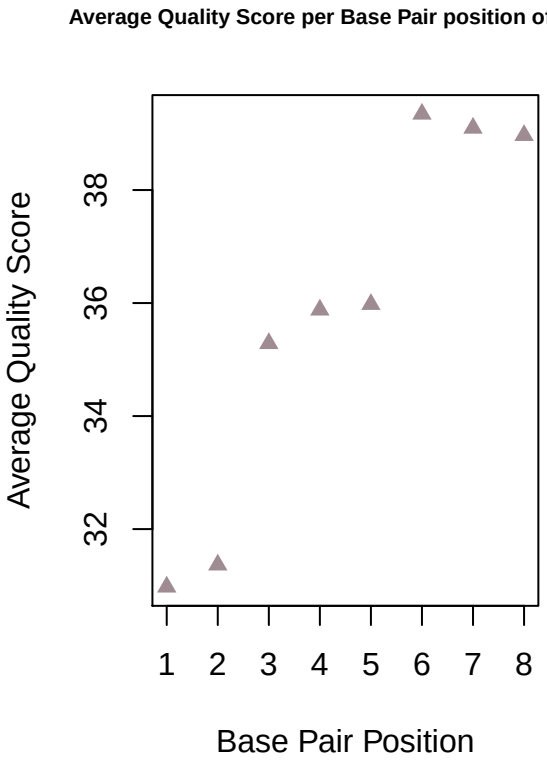
9/13/2017

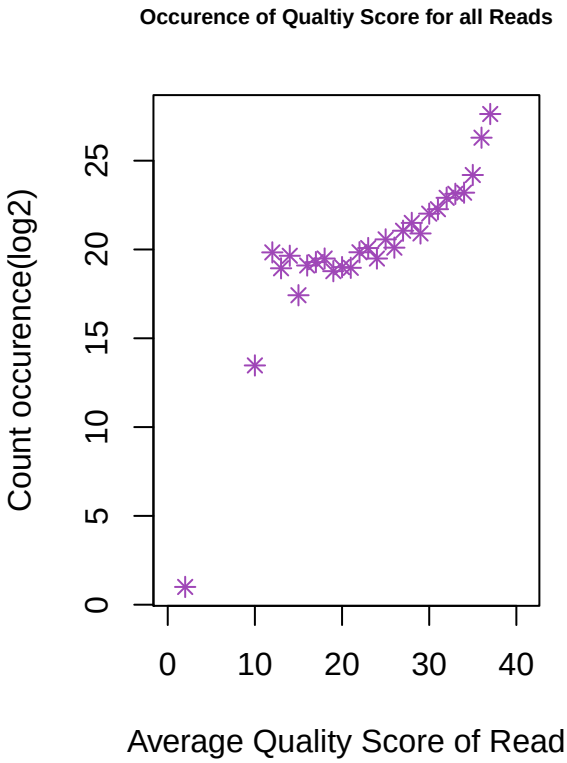
Part 1: Generate per base call distribution of quality scores for read1, read2, index1, and index2. Next, average the Quality scores for each read (for each of the four files) and plot frequency of the Quality Scores.

- Read 1 Plots

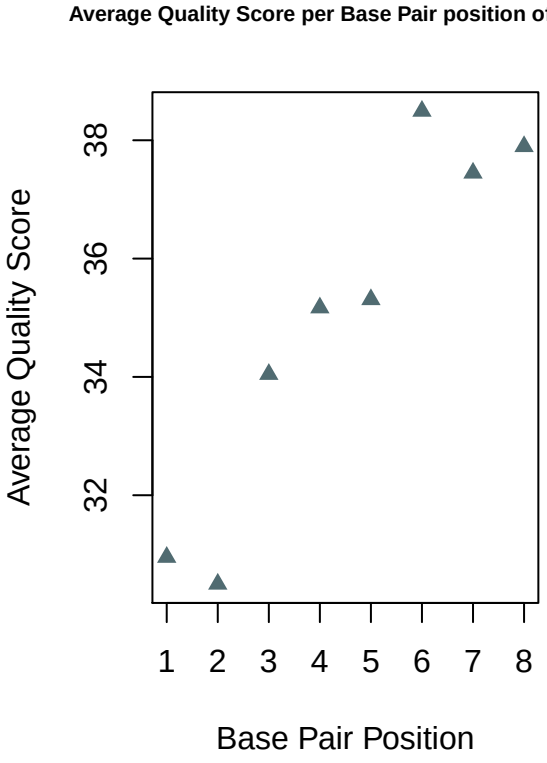


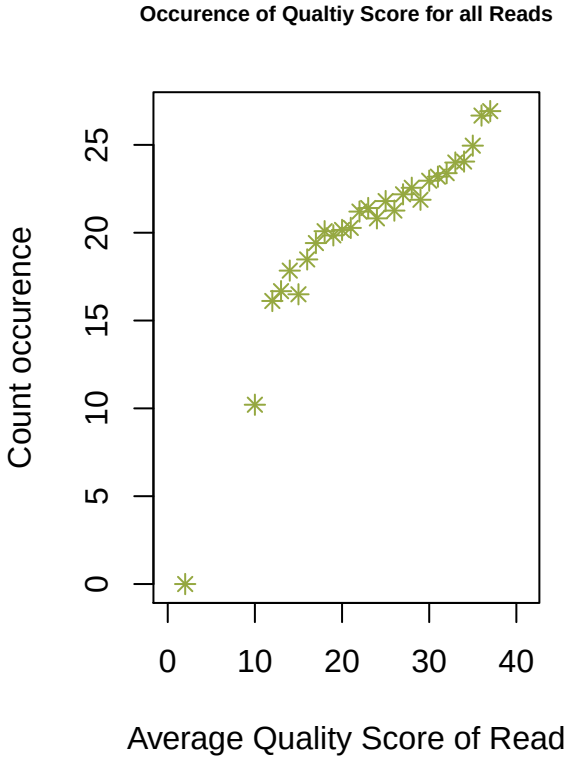
- Read 2 Plots



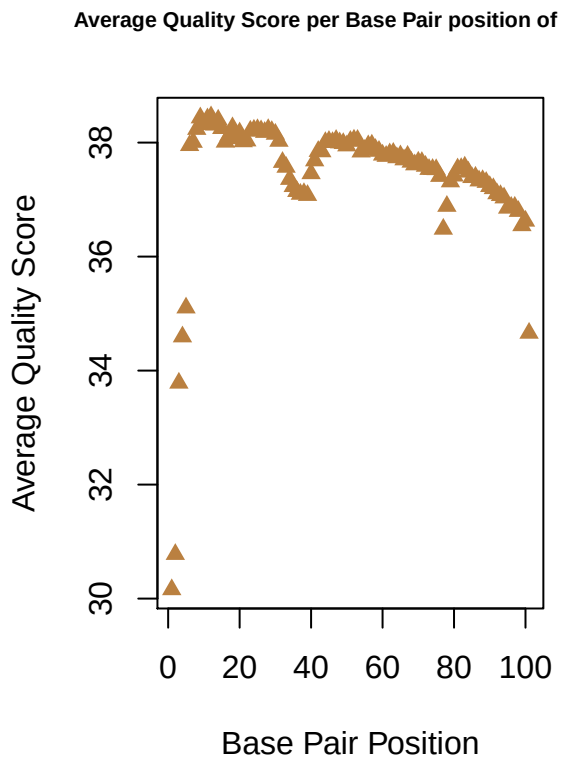


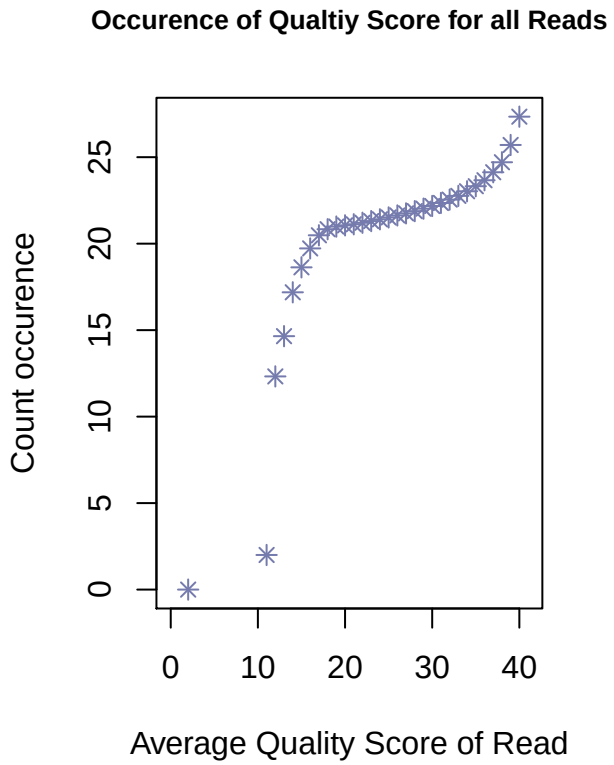
-Read 3 Plots





-Read 4 Plots





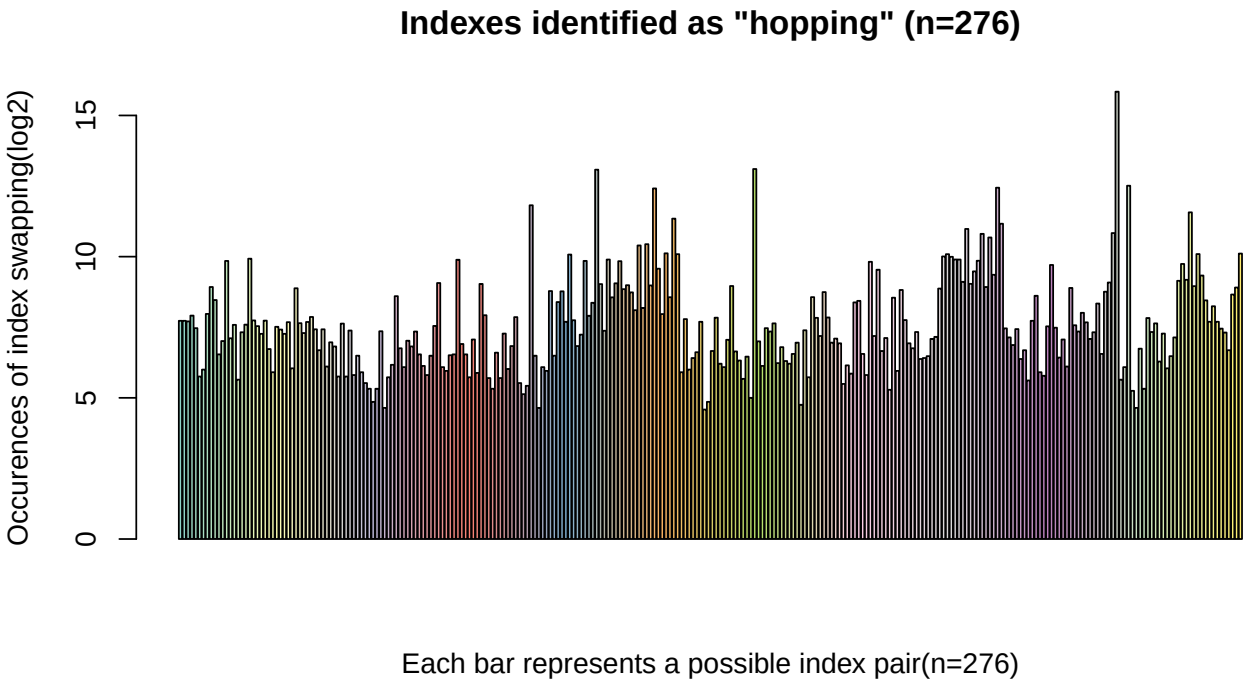
- What is a good quality score cutoff for index reads and pairs to utilize for sample identification and downstream analysis, respectively?
  - For the index reads, ~30 seems like it would be a good quality score cut off. There is a considerable spike in prevelance of the occurance of quality scores over 30, so any higher and we would be losing a significant amount of the data in both sample identification and downstream sample analysis. For R4, I think a cut off of around 20 would probably be better because there are a significant amount of reads that have a score lower than 30. Discarding too many reads could lose a significant amount of samples that might not have sequenced as well.For R1, I would also probably use a cutoff of 20 because that seems to be the inflection point of the data.
- How many indexes have Undetermined (N) base calls? (Utilize your command line tool knowledge. Summit the command you used. CHALLENGE: use a one line command)
- grep tool used `cat 1294_S1_L008_R2_001.fastq | grep -A 1 "^@K00337" | grep -v "^@" | grep -v "^--" | grep 'N' | wc -l`
  - For R2: 3,976,613 indexes have an undetermine base call. For R3: 3,328,051 indexes have an undetermine base call.
- What do the averaged Quality Scores across the reads tell you? Interpret your data specifically
  - The average quality score across all the reads gives us a relatively good indication of where the median quality score is for all of the reads. While it doesn't give us a good understanding of the temporal aspect of read qualtiy increasing, it does show that the reads get better over time and end up being, on average, of pretty good quality. For my data, we can see that it peaks somewhere around 38 for both R2 and R3, which are the index pairs. For R4, the peak is at around 40 which is a good sign that the reads are of high quality. The same pattern holds true for R1. Of note in R4 is the sharp drop off of counts below 20, which is very intersting to me, since that pattern isn't mimicked in R1.

Part 2: Write a program to de-multiplex the samples and document index swapping and number of reads retained per sample

- How many reads are retained for each expected index pair? What is the percentage?

Expected Index Pair	Occurences	Percentage
GTAGCGTA_GTAGCGTA	5774439	1.5896740
CGATCGAT_CGATCGAT	4237854	1.1666599
GATCAAGG_GATCAAGG	4628196	1.2741191
AACAGCGA_AACAGCGA	6368144	1.7531180
TAGCCATG_TAGCCATG	7148153	1.9678506
CGGTAATC_CGGTAATC	2393021	0.6587867
CTCTGGAT_CTCTGGAT	24515042	6.7488678
TACCGGAT_TACCGGAT	49686878	13.6785477
CTAGCTCA_CTAGCTCA	13034311	3.5882803
CACTTCAC_CACTTCAC	2577666	0.7096185
GCTACTCT_GCTACTCT	4301318	1.1841312
ACGATCAG_ACGATCAG	5933528	1.6334704
TATGGCAC_TATGGCAC	7651472	2.1064118
TGTTCCGT_TGTTCCGT	11450554	3.1522800
GTCCTAAG_GTCCTAAG	6200133	1.7068654
TCGACAAG_TCGACAAG	2644260	0.7279515
TCTTCGAC_TCTTCGAC	30089661	8.2835324
ATCATGCG_ATCATGCG	6927867	1.9072070
ATCGTGGT_ATCGTGGT	4730009	1.3021477
TCGAGAGT_TCGAGAGT	7448072	2.0504168
TCGGATTC_TCGGATTC	2874320	0.7912858
GATCTTGC_GATCTTGC	2636332	0.7257689
AGAGTCCA_AGAGTCCA	7602663	2.0929749
AGGATAGC_AGGATAGC	5861709	1.6136990

- Approximately 62% of all of the reads generated from this sample, after being quality filtered so that every nucleotide of both indexes had a quality score >= 30, were retained from known index pairs. Approximately 1.5% of the reads that passed the quality filter did not align to a known and intended index pairs, and could be the consequence of index swapping or sequencing error.
- How many reads are indicative of index swapping?
  - There are 179,624 reads that have been indicated as potential index swaps.
- Create a distribution of swapped indexes. What does this histogram tell you/what is your interpretation of this data?



- The histogram tells us that every single one of 276 possible swapped index combinations were found in our generated data. I am shocked by this information, because I didn't think that we would see all of the index combinations possible. I thought maybe there would be one or two index pairs that swapped but not all of them. Even though this only represents ~0.08% of our data, this is still an issue that should be considered during downstream analysis.There are some spikes in the data which indicate sequences that more likely to incur index swapping. There are a couple index pairs that swapped more frequently than the others, with the highest occurence being of `TATGGCAC_TGTTCCGT` despite the fact that `TATGGCAC` index pair was only represented in ~2% of the reads and `TGTTCCGT` was only represented in ~3% of reads. The levenshtein distance is 6 between these two sequences, and so that indicates to me that it cannot be blamed on sequencing error alone and likely swapping has occurred.