Index Hopping

Questions & answers

David Ho

9/12/2017

1. Generate per base call distribution of quality scores for read1, read2, index1, and index2. Generate a per nucleotide distribution as you did in part 1 of PS4 (in Leslie's class).

Next, average the Quality scores for each read (for each of the four files) and plot frequency of the Quality Scores.

a. Turn in the 8 histograms.

See `/plots/` folder.

b. What is a good quality score cutoff for index reads and pairs to utilize for sample identification and downstream analysis, respectively?

We define cutoff as the following:

```
If ALL of the nucleotides of both index reads are above (or equal to) your minimum quali
ty score, that record should be retained.
```

If we look at just the index files (`R2` & `R3`), the lowest average quality score at a particular position between the two files is 30.49904 at the 2nd position in the `R3` file. Therefore, 30 should be the cutoff for sample identification and downstream analysis. If we choose anything higher than 30, many reads will be discarded because on average, the quality score at 2nd position of `R3` is just above 30.

c. 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)

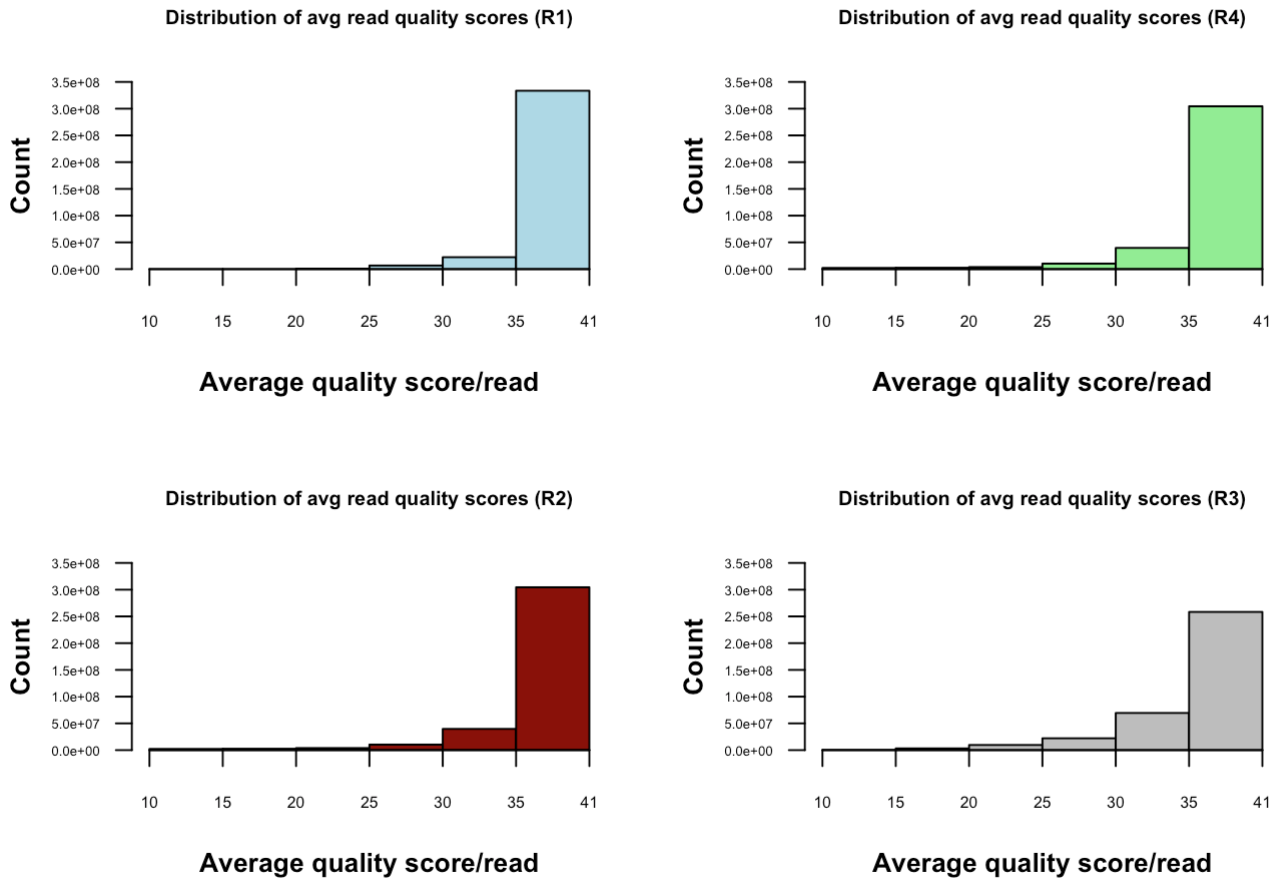
```
awk 'NR%4==2' /projects/bgmp/2017_sequencing/1294_S1_L008_R2_001.fastq | grep "N" | wc -
l
awk 'NR%4==2' /projects/bgmp/2017_sequencing/1294_S1_L008_R3_001.fastq | grep "N" | wc -
l

R2
3976613
R3
3328051
```

In `R2` , there are 3976613 indexes that have N base calls.

In `R3` , there are 3328051 indexes that have N base calls.

d. What do the averaged Quality Scores across the reads tell you? Interpret your data specifically.



Overall, the majority of the reads have an average read quality score of 35 more higher.

`R1` has the highest proportion of reads that have average quality reads of 35 or higher.

`R3` has the lowest proportion of reads that have average quality reads of 35 or higher.

It was surprising to me to see that in all four files, there were reads that had average quality scores less than 15, even though they make up a very low proportion.

Considering the number of reads there are and the majority of them have average qualities of 35+, it seems like this was a good sequencing run.

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

See the file `demultiplex_counts.py`

SLURM script:

```
/usr/bin/time -v python demultiplex_counts.py -R1 /projects/bgmp/2017_sequencing/1294_S1
_L008_R1_001.fastq.gz \
-R2 /projects/bgmp/2017_sequencing/1294_S1_L008_R2_001.fastq.gz \
-R3 /projects/bgmp/2017_sequencing/1294_S1_L008_R3_001.fastq.gz \
-R4 /projects/bgmp/2017_sequencing/1294_S1_L008_R4_001.fastq.gz \
-i indexes_edited.txt \
-c 30
```

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

Retained means that the read pass cutoff.

There were 363246735 in total, in which 2.322281410^{8} reads passed the cutoff of 30. Which means that the quality score of each base position in the index reads were at or above 30.

Out of the retained reads, 226715602 had expected index pairs.

Therefore, out of the retained reads, 97.6262386% had expcted index pairs. Out of all reads, 62.41% of them passed cutoff and had expected index pairs.

The following is part of the output from `demultiplex_counts.py` :

```
Cutoff:          30
Number of reads:         363246735
Number of retained, EXPECTED index pairs:        226715602
Percentage of retained, EXPECTED index pairs out of all reads:    62.41366546625671 %
Number of retained, BAD index pairs (AKA Index-swapping):        330975
Number of retained reads with Ns in the barcode:          0
Number of retained reads that have index not matching list (sequencing errors):   518156
7
Number of retained reads with Ns or sequencing errors:    5181567
Number of reads below cutoff:    131018591
```

index	counts	percent_of_above_cutoff	percent_of_all_reads
GTAGCGTA_GTAGCGTA	5774439	2.487	1.59
CGATCGAT_CGATCGAT	4237854	1.825	1.167
GATCAAGG_GATCAAGG	4628196	1.993	1.274
AACAGCGA_AACAGCGA	6368144	2.742	1.753
TAGCCATG_TAGCCATG	7148153	3.078	1.968
CGGTAATC_CGGTAATC	2393021	1.03	0.6588
CTCTGGAT_CTCTGGAT	24515042	10.56	6.749
TACCGGAT_TACCGGAT	49686878	21.4	13.68
CTAGCTCA_CTAGCTCA	13034311	5.613	3.588
CACTTCAC_CACTTCAC	2577666	1.11	0.7096

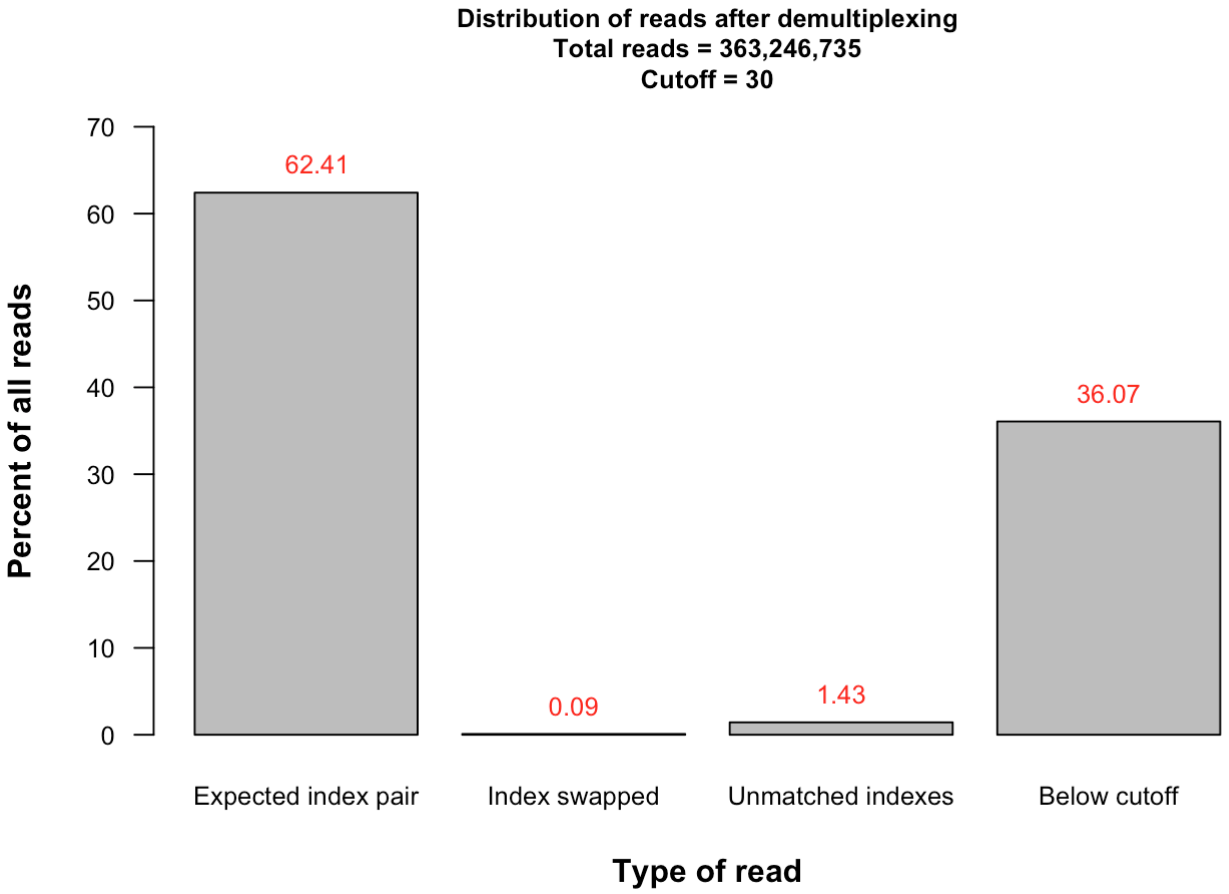
index	counts	percent_of_above_cutoff	percent_of_all_reads
GCTACTCT_GCTACTCT	4301318	1.852	1.184
ACGATCAG_ACGATCAG	5933528	2.555	1.633
TATGGCAC_TATGGCAC	7651472	3.295	2.106
TGTTCCGT_TGTTCCGT	11450554	4.931	3.152
GTCCTAAG_GTCCTAAG	6200133	2.67	1.707
TCGACAAG_TCGACAAG	2644260	1.139	0.728
TCTTCGAC_TCTTCGAC	30089661	12.96	8.284
ATCATGCG_ATCATGCG	6927867	2.983	1.907
ATCGTGGT_ATCGTGGT	4730009	2.037	1.302
TCGAGAGT_TCGAGAGT	7448072	3.207	2.05
TCGGATTC_TCGGATTC	2874320	1.238	0.7913
GATCTTGC_GATCTTGC	2636332	1.135	0.7258
AGAGTCCA_AGAGTCCA	7602663	3.274	2.093
AGGATAGC_AGGATAGC	5861709	2.524	1.614

b. How many reads are indicative of index swapping?

After discarding reads that did not meet a cutoff of 30, **330975** reads had index pairs that suggested index swapping occurred.

See files in `/index_pair_counts/` for counts of each expected & swapped index pair.

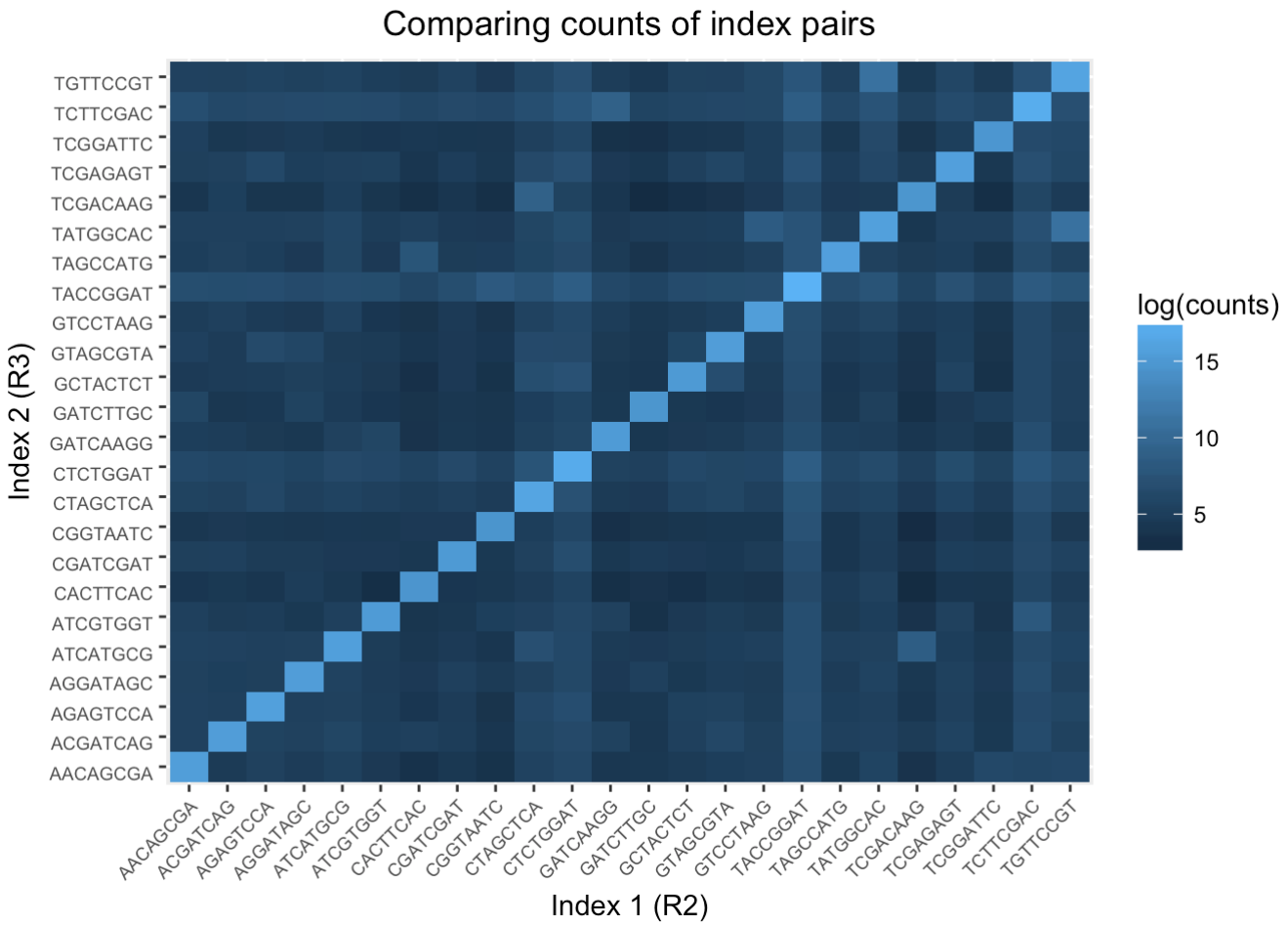
c. Create a distribution of swapped indexes. What does this histogram tell you/what is your interpretation of this data?



Even after filtering the reads by a cutoff of 30, there were still 330975 reads that were indicative of index swapping. With Illumina technology, especially with patterned flow cells, it is important to use unique dual indices when multiplexing samples because that would be the only way to know if there was index swapping.

Overall, having actually help create the samples that were sequenced, utilizing only 62% of the data is disheartening. But even so, we still have 226,715,602 potentially good reads to do downstream analysis.

Comparing the actually index pairs…



Lighter colors indicate greater abundance of an index pair (log-scale). Overall, there are way more expected index pairs from the sequencing but every swapped index pair was sequenced as well. Reads with TACCGGAT_TACCGGAT were the most abundant, and the number of swapped indexes with TACCGGAT as one of the pair was the most abundant as well.

There was uneven coverage as certain samples had more reads after sequencing, perhaps due to initial library pooling. TACCGGAT_TACCGGAT was the most abundant with 49,686,878 reads, while CGGTAATC_CGGTAATC had only 2,393,021 reads. Because the number of swapped indexes with TACCGGAT was the most abundant, this could suggest that swapping of a particular index correlates with the initial amount of that index pair in the pooled library.

~~3. List your filenames and the number of reads contained within each one.~~