Demultiplexing and Index Swapping

Goals: Our goal is to look through a lane of sequencing generated from the 2017 BGMP cohort's library preps and determine the level of index swapping and undetermined index-pairs, before and after quality filtering of index reads. In order to do this, we must first demultiplex the data. Develop a strategy to de-multiplex samples to create 48 FASTQ files that contain acceptable index pairs (read1 and read2) and two files of undetermined files that contain unacceptable index pairs, low quality, or undetermined (read1 and read2).

De-multiplexing is necessary for downstream analysis.

We submitted 24 indexed (dual matched) libraries. The indexes are:

B1	GTAGCGTA	A11	CTAGCTCA	C10	TCTTCGAC
A5	CGATCGAT	C7	CACTTCAC	A2	ATCATGCG
C1	GATCAAGG	B2	GCTACTCT	C2	ATCGTGGT
B9	AACAGCGA	A1	ACGATCAG	A10	TCGAGAGT
C9	TAGCCATG	B7	TATGGCAC	B8	TCGGATTC
C3	CGGTAATC	A3	TGTTCCGT	A7	GATCTTGC
B3	CTCTGGAT	B4	GTCCTAAG	B10	AGAGTCCA
C4	TACCGGAT	A12	TCGACAAG	A8	AGGATAGC

4 FASTQ files are: 1294_S1_L008_R1_001.fastq.gz, 1294_S1_L008_R2_001.fastq.gz, 1294_S1_L008_R3_001.fastq.gz, and 1294_S1_L008_R4_001.fastq.gz, in /projects/bgmp/shared/2017_sequencing/.

Part 1 – Quality Score Distribution per-nucleotide

1. Determine which files contain the indexes, and which contain the paired end reads containing the biological data of interest. Create a table and label each file with either read1, read2, index1, or index2.
2. Generate per base call distribution of quality scores for read1, read2, index1, and index2. Average the QSs for each read and generate a per nucleotide distribution as you did in part 1 of PS4 (in Leslie's class).
    a. Turn in the 4 histograms.
    b. What is a good quality score cutoff for index reads and pairs to utilize for sample identification and downstream analysis, respectively?
    c. How many indexes have Undetermined (N) base calls? (Utilize your command line tool knowledge. Submit the command you used. CHALLENGE: use a one-line command)

Part 2 – Develop an algorithm to de-multiplex the samples

Write up a strategy for writing an algorithm for de-multiplexing files and reporting index-hopping. That is, given four files (2 with biological reads, 2 with index reads) and known indexes, sort reads by index, outputting one forward file and one reverse file per index, plus a pair of files for unknown indexes. Additionally, your algorithm should report the number of properly matched indexes (per index) and the level of index hopping observed. You should not write any code for this portion of the assignment. Be sure to:

- Define the problem
- Determine/describe what output would be informative
- Write examples (unit tests!):
  - Include four properly formatted input fastq files
  - Include the appropriate number of properly formatted output fastq files
- Develop your algorithm using pseudocode
- Determine high level functions
  - Description
  - Function headers
  - Test examples for individual functions
  - Return statement

Turn in:
Answers to questions, Python script for part 1, plots, and anything outlined in part 2.