Demultiplexing part 2

Define the problem

Our read and index files have read counts measured in billions. Part of the problem is that any method that stores information across the whole file can’t be used. Each of our files numbered respectively are the reverse compliment read, the first index, the second index, and the read that mirrors what the RNA should be. A noted confusion with illumina nomenclature is that the R1 file is the reverse compliment of the RNA and the R4 file contains the reads of the template strand itself. Our problem is that we need to develop a method to read each of the 4 files at once, and then go 4 lines at a time sorting each read into an appropriate file based off on if it fails any of our standards for an acceptable read. These standards arise from potential errors that could occur during the sequencing process. Index hopping has occurred during sequencing, meaning that the 2 index reads do not match up. Low quality scores where the indexes are degraded to the point where it is hard to tell if index hopping has occurred. Finally, if the indexes that have matched to the reads is one that is not within our set of known indexes. Output files must be made for these cases and the reads or indexes need to be put into these files. The rest of the “good” files should be given their own outputs and then have the index values added to the end of their header.

• Determine/describe what output would be informative

In addition to sorting the entire reads into each file it would be good to generate an output file

• Write examples (unit tests!):

o Include four properly formatted input FASTQ files with read pairs that cover all three categories (dual matched, index-hopped, unknown index)

o Include the appropriate number of properly formatted output FASTQ files given your input files

• Develop your algorithm using pseudocode

• Determine high level functions

o Description/doc string – What does this function do?

o Function headers (name and parameters)

o Test examples for individual functions

o Return statement