Part 1. We will go over scripts from the challenges assigned last week.

Part 2. Work as a group to develop a strategy to remove unusual characters from the Gulf pipefish cds file.

1. Copy “ssc\_2016\_12\_20\_cds.fa” from the shared folder to a new week04 folder in your CSB folder.
2. Check to make sure none of the sequences have lower-case letters.
3. Write a script that scans the file for every capital letter in the alphabet and returns the number of lines with at least one occurrence of each letter. For example, thousands of lines will contain an A but no lines should contain a Z. This list will tell you which non-standard symbols are present in the file (and will have to be changed to Ns).
   1. See here for a list of what these different letters mean: [IUPAC ambiguity codes. Nucleotide ambiguity code. Nomenclature for Incompletely Specified Bases in Nucleic Acid Sequences. (dnabaser.com)](http://www.dnabaser.com/articles/IUPAC%20ambiguity%20codes.html).
   2. Be sure to ignore the header line somehow.
4. Write a script that replaces any letter other than A, C, G, T, or N with N in the sequence.
   1. Again, leave the header line intact.
   2. This process should not alter the total length of the file. Use the command wc to ensure that the file is the same length before and after the changes.
   3. Use your script from step 2 to make sure the atypical base pairs are removed from the sequences. You should find that only A, C, G, T, and N have non-zero numbers of lines.
   4. The most important utilities to use here are probably grep and sed.

Part 3. Work as a group to verify that everyone can analyze the same fastq paired-read files with fastqc and trimmomatic and get the same results.

1. Copy the sample paired-end read files from the shared folder to the week04 folder in your CSB directory.
2. Read the following website to understand how fastqc works: <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
3. Run fastqc on the provided files (see <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/INSTALL.txt>).
4. Run trimmomatic on the paired read files. See <http://www.usadellab.org/cms/?page=trimmomatic>.
   1. You will want this command to be in a script.
   2. Use the default settings for paired-end data provided by Trimmomatic (i.e., the first set of parameters listed under Paired End in the Quick start section).
   3. On our servers, the command is trimmomatic followed by the arguments. You don’t need to type the java -jar part.
   4. Use the Illumina adapter sequences (TruSeq3-PE-2.fa).
5. Rerun fastqc on the files.
   1. Do you see any improvement?

Part 4. Take the Canvas quiz based on this assignment (Thursday). Keep your scripts from above and your output from fastqc (both before and after) handy so you can access them during the quiz.