Assignment #3 - Parsing a FASTA file from NCBI

The purpose of this assignment is to practice data cleaning. To do this, you will read in a fasta file of zebra fish rna sequences, extract information into columns, and save the data in spreadsheet format. Follow the steps below using functions from the stringr, tidyr or dplyr packages to accomplish your tasks whenever possible. Annotate your code, and explain any regular expressions used.

a) Install and load the 'seqinr' package. Use the 'read.fasta' function to load the fasta file into an object named 'fq'. The resulting list object holds the DNA sequence information as well as the annotation for each sequence. Use the following code to retrieve and store the annotation and sequence information into a data frame. (1 mark)

```r
fq <- data.frame(annotation = unlist(getAnnot(fq)),
                 sequence = unlist(getSequence(fq, as.string = TRUE)),
                 stringsAsFactors = FALSE)
```

b) The annotation column consists of a refseq identifier, the genus and species of our organism, a description of the gene product, the gene name or id, and the type of RNA represented. Split the annotation character string column into 4 columns: refseq, genus, species, and description. Do not retain the annotation column. (1 mark)

c) Abbreviate RNA types by changing 'long non-coding RNA' to 'lncRNA', 'antisense RNA' to 'asRNA' and 'non-coding RNA' to 'ncRNA'. (1 mark)

d) Extract this abbreviated form to a column called 'RNA_type' and remove the RNA type information from the description column (including the comma that separated the RNA type). (2 marks)

e) Extract the gene name or identifier (the last word in brackets) to a column called 'gene_name' and remove the gene name from the description column. (2 marks)

f) Remove the '>' symbol from the refseq column. Remove the brackets from the gene_name column. (2 marks)

g) Remove any extra spaces remaining in the description column. (1 mark)

h) Make an organism column that is a concatenation of genus and species. (1 mark)

i) Make a column called 'length' with the number of bases in each sequence. (1 mark)

j) Reorganize the data frame so that sequence is the last column. (1 mark)

k) Save the data frame to a .xls(x) file and submit the file with your assignment. (1 mark)

3 marks will be given according to the following rubric:

- 3.0 - code is well-documented and concise
- 1.5 - code is either well-documented or concise, but not both
- 0 - no attempt was made to document code, extra variables are created, code is difficult to read

Total: 17 marks

Submission: Each student will upload a .R file and a .xls file to Quercus. Please include your first and last name, the date of submission, and the assignment number.

Due date: 11:59pm February 6th, 2019