Assignment #2 - Normalizing Fly RNA-Seq Counts

The purpose of this assignment is to practice using functions from the tidyr and dplyr package. The data set is Drosophilia RNA-Seq count data (the number of sequencing reads mapped to a transcript/gene) (Brooks et al. Conservation of an RNA regulatory map between Drosophila and mammals. Genome Research, 2010) from the 'pasilla' package. tidyr and dplyr functions should be used to answer each of the following questions; please include a sentence stating your answer in addition to showing your code whenever possible:

a) Install the 'pasilla' package and the 'DESeq' package and load all libraries necessary for this assignment. Load the dataset with the following code:

```
data("pasillaGenes")
```

Of what class is this data set? (3 marks)

b) Look at the help documentation for this class of data. Find a function to retrieve the count data from this data class. Save the count data to an object called 'dat'. (1 mark)

c) What rules of tidy data is this count table currently defying? (2 marks)

d) Convert your count matrix to a data frame. Move you rownames to a column called 'gene'. (2 marks)

e) Transform the data from 'wide' to 'long' (AKA tidy) format. (1 mark)

f) Separate the sample names into 'treatment' type and 'group' number. (1 mark)

g) Some gene names appear to have an alternative transcript (2 gene names are present). Separate these 2 names into different columns. Name the 2nd column 'alternative_transcript'. (1 mark)

h) Are there any genes where all counts are 0? If so, how many of these genes are there? Filter them out of your data set. (3 marks)

i) Calculate the size of each sequence library (the total # of counts per sample). Calculate the mean library size and save it to a variable called 'mean_lib'. (2 marks)

j) To be able to compare the reads across experiments, we need to normalize our sample since there are different numbers of reads per library. For a simple example, we will do this by calculating a scaling factor for each sample and saving it to a new column. The scaling factor is the library size of a sample divided by the mean library size of all samples. (1 mark)

k) Mulitply the counts for each sample by its respective scaling factor and save the results to a new column called 'scaled_counts'. Round to nearest whole number. (2 marks)

l) Which top 5 genes have the greatest number of counts? Does this hold true after the data is scaled? (3 marks)

m) Now that we have normalized counts, we want to replace our original data matrix with our scaled counts. Use a tidyr function to recombine the treatment and group information (use an underscore to separate the 2 pieces of information). Get rid of the original counts, library sizes, alternative transcripts and scaling factors before converting your data back into wide format. (3 marks)

n) Convert your dataframe back to a numeric matrix. (2 marks)

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 marks: 30

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

Due date: 11:59pm October 9th, 2018