R Notebook

Code

INTRODUCTION

Scientific Question: BRCA-1 is a tumor suppressor gene found in humans and many other species like dogs and cats. When they are mutated, breast cancer is very likely to develop. How similar is the normal gene across these different species and how does expression of this gene change after it is mutated during breast cancer?

Background: BRCA1 gene is a tumor suppressor gene found in many species including humans, dogs, and cats. This gene is in charge of making sure that tumor/cancer cells do not grow uncontrollably or rapidly. This gene is in charge for DNA repair. It interacts with other proteins to work together to restore any breakage in DNA. This allows the cell to maintain the stability of it's genetic information. When there is a mutation in this gene, cancer is more prone to develop. One cancer that is really prone to occur after a mutation in this gene is breast cancer. This not only poses a risk for the individual that develops the mutation, but also for future generations since this mutated gene can be passed along. This is why children of parents who developed breast cancer have a higher risk of developing it themselves.One other cancer that a mutation can cause is ovarian cancer along with breast cancer. Women ho have a mutation in this gene have about 30-60% chance of developing this cancer compared to an individual without the mutation. Prostate cancer and cholangiocarcinoma are also more likely to develop with a mutation in this gene. This gene is essential in many different species and whenever it is mutated, they all develop the risk of getting a type of cancer. I obtained this information from:
https://medlineplus.gov/genetics/gene/brca1/#conditions
(https://medlineplus.gov/genetics/gene/brca1/#conditions)

Scientific Hypothesis: If there was a mutation in the BRCA1 gene of these three different species, then all of them will have decreased levels of expression of this gene due to it being a tumor suppressor gene. Although the levels of expression will decrease for all three, they will vary in how much the expression decreases after mutated.

In this project, I will perform multiple sequence alignment, RNA sequencing, heatmap, and sequence logos. I will use multiple sequence alignment to compare the BRCA1 gene in humans, dogs, and cats and see how similar they are. I will download the sequences from NCBI. I will then use sequence logos to visualize the the sequence alignment across these species. I will also use RNA seq to look at the levels of expression of these genes after being mutated. I will obtain this data from the scientific articles that I found and look at how much this gene was down regulated during breast cancer. I will then use a heatmap to show the comparison between the levels of expression of this gene across all three species. This will allow me to see which species had the greater decrease in expression of this gene and which species had the least decrease.

I will list and define the packages needed to run below:

1. msa - provides a unified interface to the multiple sequence alignments algorithms ClustalW, ClustalOmega, and Muscle. The three algorithms are included in the package.
   https://bioconductor.org/packages/release/bioc/html/msa.html
   (https://bioconductor.org/packages/release/bioc/html/msa.html)
2. seqLogo - takes the position weight matrix of some DNA sequence motifs and plots the corresponding sequence logo. This will be used to show the visualization of the multiple sequence alignment.
   https://bioconductor.org/packages/release/bioc/html/seqLogo.html
   (https://bioconductor.org/packages/release/bioc/html/seqLogo.html)
3. DESeq2 - estimates variance-mean dependence in count data from sequencing assays and tests for any differential expression. This will be used to look at the levels of expression of the genes in the three species after being mutated. https://bioconductor.org/packages/release/bioc/html/DESeq2.html

(https://bioconductor.org/packages/release/bioc/html/DESeq2.html)

4. GEOquery - is a public repository of micro array data and the bridge between GEO and Bioconductor. https://bioconductor.org/packages/release/bioc/html/GEOquery.html (https://bioconductor.org/packages/release/bioc/html/GEOquery.html)
5. canvasXpress - enables creation of visualizations using the CanvasXpress framework in R. It is a Javascript library fo research with complete tracking of data and end-user modifications stored in a PNG image that can be played back. https://cran.rstudio.com/web/packages/canvasXpress/index.html (https://cran.rstudio.com/web/packages/canvasXpress/index.html)
6. ggplot2 - is a system for creating graphics, helps map the variables as long as we rovide the graphical primitives. https://cran.r-project.org/web/packages/ggplot2/index.html (https://cran.r-project.org/web/packages/ggplot2/index.html)
7. GGally - it is an extension to the ggplot2 package, it extends ggplot2 by adding several functions o reduce the complexity of combining geometric objects with transformed data. https://cran.r-project.org/web/packages/GGally/index.html (https://cran.r-project.org/web/packages/GGally/index.html)
8. factoextra - provides easy-to-use functions to extract and visualize output of multivariate data analyses including PCA, CA, MCA, FAMD, MFA, and HMFA functions from different packages in R. It also simplifies clustering analysis steps and provides ggplot2 based elegant data visualization. https://cran.r-project.org/web/packages/factoextra/index.html (https://cran.r-project.org/web/packages/factoextra/index.html)
9. pheatmap - implements heatmaps that offer control over dimensions and appearance. https://cran.r-project.org/web/packages/pheatmap/index.html (https://cran.r-project.org/web/packages/pheatmap/index.html)

Hide

```
library(BiocManager)
library(msa)
#if (!require("BiocManager", quietly = TRUE))
    #install.packages("BiocManager")
#BiocManager::install("ggseqLogo")
library(seqLogo)

#BiocManager::install(c("DESeq2", "GEOquery", "canvasXpress", "ggplot2", "clinfun", "GGa
lly", "factoextra"))
library(DESeq2)
library(GEOquery)
library(canvasXpress)
library(ggplot2)
library(ggseqlogo)
library(clinfun)
library(GGally)
library(factoextra)

library(pheatmap)
```

PERFORMING BIONFORMATICS ANALYSES

Below, this code is responsible for loading in the 3 different sequences of the BRCA1 gene of the human, dog, and cat and comparing them using multiple sequence alignment. I obtained the Fasta sequence files of this gene in all three species from the NCBI database. This function (msa) will scan the sequences and find spot where there are differences and similarities.

Hide

```
#Multiple Sequence Alignment code below
mySequences1 <- readAAStringSet("sequence (1).txt")
mySequences2 <- readAAStringSet("sequence (2).txt")
mySequences3 <- readAAStringSet("sequence (3).txt")

mySequences1
```

```
AAStringSet object of length 5:
    width seq                                                   names
[1]  2280 ATGGATTTATCTGCTCTTCGCGTTGAAGA...TACCCCAGATCCCCCACAGCCACTACTGA lcl|NC_000017.11
_...
[2]  5592 ATGGATTTATCTGCTCTTCGCGTTGAAGA...TACCCCAGATCCCCCACAGCCACTACTGA lcl|NC_000017.11
_...
[3]  5655 ATGGATTTATCTGCTCTTCGCGTTGAAGA...TACCCCAGATCCCCCACAGCCACTACTGA lcl|NC_000017.11
_...
[4]  5451 ATGCTGAAACTTCTCAACCAGAAGAAAGG...TACCCCAGATCCCCCACAGCCACTACTGA lcl|NC_000017.11
_...
[5]  2100 ATGGATTTATCTGCTCTTCGCGTTGAAGA...GGCTTCCATGCAATTGGGCAGATGTGTGA lcl|NC_000017.11
_...
```

Hide

```
mySequences2
```

```
AAStringSet object of length 12:
     width seq                                                 names
 [1]  5805 ATGGTGAGGGAACAGGGCCCCTTTTTGCG...TGCAGACTCCAGCCAGCCATGCGTGTAA lcl|NC_051813.1_
c...
 [2]  2475 ATGGTGAGGGAACAGGGCCCCTTTTTGCG...TGCAGACTCCAGCCAGCCATGCGTGTAA lcl|NC_051813.1_
c...
 [3]  5766 ATGGTGAGGGAACAGGGCCCCTTTTTGCG...TGCAGACTCCAGCCAGCCATGCGTGTAA lcl|NC_051813.1_
c...
 [4]  2436 ATGGTGAGGGAACAGGGCCCCTTTTTGCG...TGCAGACTCCAGCCAGCCATGCGTGTAA lcl|NC_051813.1_
c...
 [5]  5541 ATGGTGAGGGAACAGGGCCCCTTTTTGCG...GAGCAAAGACGGAGACTCTCCTACTTAG lcl|NC_051813.1_
c...
 ...   ... ...
 [8]  5676 ATGGATTTATCTGCGGATCGTGTTGAAGA...TGCAGACTCCAGCCAGCCATGCGTGTAA lcl|NC_051813.1_
c...
 [9]  5637 ATGGATTTATCTGCGGATCGTGTTGAAGA...TGCAGACTCCAGCCAGCCATGCGTGTAA lcl|NC_051813.1_
c...
[10]  5637 ATGGATTTATCTGCGGATCGTGTTGAAGA...TGCAGACTCCAGCCAGCCATGCGTGTAA lcl|NC_051813.1_
c...
[11]  5637 ATGGATTTATCTGCGGATCGTGTTGAAGA...TGCAGACTCCAGCCAGCCATGCGTGTAA lcl|NC_051813.1_
c...
[12]  5637 ATGGATTTATCTGCGGATCGTGTTGAAGA...TGCAGACTCCAGCCAGCCATGCGTGTAA lcl|NC_051813.1_
c...
```

Hide

```
mySequences3
```

```
AAStringSet object of length 9:
    width seq                                                 names
[1]  5616 ATGGATTTATCTGCAGATCGTGTTGAAGA...CTGCTGACCTCAGCCAGCCATGTGTGTAG lcl|NC_058381.1_
c...
[2]  5619 ATGGATTTATCTGCAGATCGTGTTGAAGA...CTGCTGACCTCAGCCAGCCATGTGTGTAG lcl|NC_058381.1_
c...
[3]  2298 ATGGATTTATCTGCAGATCGTGTTGAAGA...CTGCTGACCTCAGCCAGCCATGTGTGTAG lcl|NC_058381.1_
c...
[4]  2301 ATGGATTTATCTGCAGATCGTGTTGAAGA...CTGCTGACCTCAGCCAGCCATGTGTGTAG lcl|NC_058381.1_
c...
[5]  5619 ATGGATTTATCTGCAGATCGTGTTGAAGA...CTGCTGACCTCAGCCAGCCATGTGTGTAG lcl|NC_058381.1_
c...
[6]  5622 ATGGATTTATCTGCAGATCGTGTTGAAGA...CTGCTGACCTCAGCCAGCCATGTGTGTAG lcl|NC_058381.1_
c...
[7]  5622 ATGGATTTATCTGCAGATCGTGTTGAAGA...CTGCTGACCTCAGCCAGCCATGTGTGTAG lcl|NC_058381.1_
c...
[8]  5622 ATGGATTTATCTGCAGATCGTGTTGAAGA...CTGCTGACCTCAGCCAGCCATGTGTGTAG lcl|NC_058381.1_
c...
[9]  5622 ATGGATTTATCTGCAGATCGTGTTGAAGA...CTGCTGACCTCAGCCAGCCATGTGTGTAG lcl|NC_058381.1_
c...
```

Hide

```
Combined_Sequences <- c(mySequences1, mySequences2, mySequences3)
Combined_Sequences
```

```
AAStringSet object of length 26:
     width seq                                                 names
 [1]  2280 ATGGATTTATCTGCTCTTCGCGTTGAAGA...ACCCCAGATCCCCCACAGCCACTACTGA lcl|NC_000017.11
_...
 [2]  5592 ATGGATTTATCTGCTCTTCGCGTTGAAGA...ACCCCAGATCCCCCACAGCCACTACTGA lcl|NC_000017.11
_...
 [3]  5655 ATGGATTTATCTGCTCTTCGCGTTGAAGA...ACCCCAGATCCCCCACAGCCACTACTGA lcl|NC_000017.11
_...
 [4]  5451 ATGCTGAAACTTCTCAACCAGAAGAAAGG...ACCCCAGATCCCCCACAGCCACTACTGA lcl|NC_000017.11
_...
 [5]  2100 ATGGATTTATCTGCTCTTCGCGTTGAAGA...GCTTCCATGCAATTGGGCAGATGTGTGA lcl|NC_000017.11
_...
 ...   ... ...
[22]  5619 ATGGATTTATCTGCAGATCGTGTTGAAGA...TGCTGACCTCAGCCAGCCATGTGTGTAG lcl|NC_058381.1_
c...
[23]  5622 ATGGATTTATCTGCAGATCGTGTTGAAGA...TGCTGACCTCAGCCAGCCATGTGTGTAG lcl|NC_058381.1_
c...
[24]  5622 ATGGATTTATCTGCAGATCGTGTTGAAGA...TGCTGACCTCAGCCAGCCATGTGTGTAG lcl|NC_058381.1_
c...
[25]  5622 ATGGATTTATCTGCAGATCGTGTTGAAGA...TGCTGACCTCAGCCAGCCATGTGTGTAG lcl|NC_058381.1_
c...
[26]  5622 ATGGATTTATCTGCAGATCGTGTTGAAGA...TGCTGACCTCAGCCAGCCATGTGTGTAG lcl|NC_058381.1_
c...
```

Hide

```
myFirstAlignment <- msa(Combined_Sequences)
```

```
use default substitution matrix
```

Hide

```
myFirstAlignment
```

```
CLUSTAL 2.1

Call:
   msa(Combined_Sequences)

MsaAAMultipleAlignment with 26 rows and 5887 columns
     aln                                                                names
 [1] --------------------------------...TACTGCTGACCTCAGCCAGCCATGTGTGTAG lcl|NC_058381.1_
c...
 [2] --------------------------------...TACTGCTGACCTCAGCCAGCCATGTGTGTAG lcl|NC_058381.1_
c...
 [3] --------------------------------...TACTGCTGACCTCAGCCAGCCATGTGTGTAG lcl|NC_058381.1_
c...
 [4] --------------------------------...TACTGCTGACCTCAGCCAGCCATGTGTGTAG lcl|NC_058381.1_
c...
 [5] --------------------------------...TACTGCTGACCTCAGCCAGCCATGTGTGTAG lcl|NC_058381.1_
c...
 [6] --------------------------------...TACTGCTGACCTCAGCCAGCCATGTGTGTAG lcl|NC_058381.1_
c...
 [7] --------------------------------...TACTGCTGACCTCAGCCAGCCATGTGTGTAG lcl|NC_058381.1_
c...
 [8] --------------------------------...TACTGCTGACCTCAGCCAGCCATGTGTGTAG lcl|NC_058381.1_
c...
 [9] --------------------------------...TACTGCTGACCTCAGCCAGCCATGTGTGTAG lcl|NC_058381.1_
c...
 ... ...
[19] --------------------------------...TGCTGCAGACTCCAGCCAGCCATGCGTGTAA lcl|NC_051813.1_
c...
[20] --------------------------------...TGCTGCAGACTCCAGCCAGCCATGCGTGTAA lcl|NC_051813.1_
c...
[21] --------------------------------...TGCTGCAGACTCCAGCCAGCCATGCGTGTAA lcl|NC_051813.1_
c...
[22] --------------------------------...CACTACTGA---------------------- lcl|NC_000017.11
_...
[23] --------------------------------...CACTACTGA---------------------- lcl|NC_000017.11
_...
[24] --------------------------------...CACTACTGA---------------------- lcl|NC_000017.11
_...
[25] --------------------------------...CACTACTGA---------------------- lcl|NC_000017.11
_...
[26] --------------------------------...------------------------------- lcl|NC_000017.11
_...
 Con --------------------------------...TACTGCTGAC??CAGCCAGCCATG?GTGTA? Consensus
```

Below, this code is responsible for sequence. This allows me to visualize the differences in the sequences of the three species. It will give me a graphical representation of the sequence conservation of nucleotides and will show me the diversity that exists.

Hide

```
#Sequence Logo below
require(ggplot2)
require(ggseqlogo)

#data(myFirstAlignment)
#ggseqlogo(mySequences1$MA0001.1 )
# kept getting "Error in geom_logo(data = data, ...) : Missing "data" parameter!"

#data(myFirstAlignment)
#p1 = ggseqlogo( mySequences1[[1]] )
# kept getting error: "Error in if (seq_type == "auto") stop("\"col_scheme\" and \"seq_t
ype\" cannot both be \"auto\"") : argument is of length zero"

#ggseqlogo(mySequences1)
#could not figure out a code to run the sequence logo for my data. Kept getting error "E
rror in if (seq_type == "auto") stop("\"col_scheme\" and \"seq_type\" cannot both be \"a
uto\"") :argument is of length zero" which prevented me from being able to visualize the
difference in nucleotides across all three species.
```

Below, this code is responsible for finding the difference in expression of the BRCA1 gene after being mutated for all three species. I will obtain this data from the articles that I have found and plug these into the code to run the RNA sequence.

Hide

```
#RNA Seq code below
```

Below, this code is responsible for showing a visual representation of the difference in expression of the BRCA1 gene obtained by RNA sequencing. It involves a heatmap that, depending on the color, will show whether the gene was upregulated or downregulated after mutated and will be compared between the three species.

Hide

```
#Heatmap code below
```

ANALYSIS OF RESULTS

Based on the output of the multiple sequence alignment, the cat BRCA1 gene seems to be more similar to the dog BRCA1 gene. This can be seen through the level of relatedness in the sequence when aligned next to each other. The cat has more nucleotides that coincide with the sequence of the dog's BRCA1 gene. The human has more blank spaces when compared to the other two species showing that it is more different. There are some spot in the gene that coincide between all three species given that it is the same gene but there is some variety. I would have wanted to show these results through the Sequence Logo method, but I could not get the code to run.

I was only able to find the code for the multiple sequence alignment. I was having a really hard time looking for a function that could run my sequence logo. I kept obtaining error after error and was not able to get anywhere for that portion. In regards to the RNA Seq, I really had no idea where to even start. I was able to find some info from the articles I used but did not know how to use it or implement it into my coding to find the differential expression analysis.Since I could not do the RNA sequencing portion, I was not able to create the heat map neither and get my results; therefore, there could be no analysis of the results.