Author: Christopher Tran

Goal: Identify suitable CRISPR genes with the given motifs for the experiment by creating multiple scripts and files to locate the suitable genes. A report of the genes and codenames of the motifs will be provided.

Usage: clinical\_data.txt copyExomes.sh createCrisprReady.sh editGenome.sh exomeReport.py identifyCrisprSite.sh motif\_list.txt chicken.fasta dromedary.fasta fox.fasta goat.fasta gopher.fasta gorilla.fasta lamb.fasta

Output files: exomes exomes\_list postcrispr precrispr top\_exomes readme.txt week4\_copy\_files

Instructions: Obtained files in home/rbif/week4 by coping files into home/tranc/week4/week4\_copy\_files. Created "readme.txt" in home/tranc/week4/ to give step by step instructions. Using the clinical\_data.txt, a script named copyExomes.sh was created to align the columns of the file and pull data that was in between 20-30 mm in diameter that were sequenced. Commands used in the script was sed, grep, and awk. The results showed 7 motifs that qualified for the requirement. The 7 motifs were chicken, dromedary, fox, goat, gopher, gorilla, and lamb. An exomes directory was created and a copy of these motifs were taken from week4\_copy\_files. Print and cat command were created to show results of the script.

A new script was created named createCrisprReady.sh. This scripts objective was to find the top three highest occurrences of sequences for each motif. Inside the script,a set varibale was created to full and label the exome files correctly. Loop, grep and sort commands were also created to generate the top 3 sequences of each motif. After generating the top 3 sequences, the 3 seqeunces was set as variables in the same script.An awk command was also used on the top 3 sequences list of occurences to just have the sequences in a (animal).txt file. Another grep command was created to pull all corresponding genes and sequences from the exome/(animal).fasta files and appending it to a new file named (animal)\_topmotifs.fasta. Within the script, top\_exomes directory was created and rearranged in sub directories. The location of each directories are top\_exomes/sorted\_exomes, top\_exomes/exomes\_list, and top\_exomes/sequenced\_exomes. Print statements are created to show the script has been completed.

The next script called identifyCrisprSite.sh was created. This script was to locate the suitable CRISPR site for each gene inside the animal\_topmotifs.fasta file. The sequence has to be 20 basepairs before the contained sequence which is "NGG" were N is any base. Within this script, variable were set and sed command was created to set and label each exome files correctly. A grep command was created again to pull 20 base pairs before the intended sequence site. (animal)\_precrispr.fasta files were created and left in the precrispr directory. Print statements were generated to inform the script is complete.

editGenome.sh was created next. This script is meant to add an A base pair before every NGG sequence of each gene. Within the script, variables were set and sed command was created to set labels for each exome files correctly. A looping sed command was created to add A in front of every NGG basepair. Print statements were generated to inform the script is complete.

A python script named exomeReport.py was created to generate the results of the project. Within the script a dictionary was created with conditions of 20-30mm in diameter and all exomes were sequenced. Sets, loops, if statements and keys were created to distingush common cohorts and unique genes. The codename, diameter, environment, and discoverer. Multiple dictionaries, keys, and loops were generated to generate the the uniquen genes and common coherts. Print statements was scripted to show results after running the script. Please see below for the results of the project.

Results:

Results have shown that organism chicken was discovered by Rivet, has a diameter of 26, and in the Swamp enviroment.

Results have shown that organism dromedary was discovered by Forness, has a diameter of 30, and in the Sewage enviroment.

Results have shown that organism fox was discovered by Vanderburg, has a diameter of 26, and in the River enviroment.

Results have shown that organism goat was discovered by Gorham, has a diameter of 24, and in the Sewage enviroment.

Results have shown that organism gopher was discovered by Mccafferty, has a diameter of 21, and in the Sewage enviroment.

Results have shown that organism gorilla was discovered by Landey, has a diameter of 30, and in the River enviroment.

Results have shown that organism lamb was discovered by Kesner, has a diameter of 28, and in the Sewage enviroment.

7 coherts are in common with all exomes

gorilla has the following genes unique to itself: set(['gene8', 'gene146', 'gene109', 'gene382', 'gene143', 'gene411', 'gene101', 'gene100', 'gene286', 'gene499', 'gene147', 'gene471', 'gene442', 'gene378', 'gene212', 'gene239', 'gene195', 'gene56', 'gene74', 'gene404', 'gene76', 'gene460', 'gene461', 'gene130', 'gene405', 'gene425', 'gene257', 'gene138', 'gene462', 'gene299', 'gene398', 'gene153', 'gene293', 'gene297', 'gene394', 'gene114', 'gene271', 'gene463', 'gene347', 'gene20', 'gene266', 'gene349', 'gene305', 'gene220', 'gene70', 'gene127', 'gene453', 'gene474', 'gene245', 'gene457', 'gene302', 'gene240', 'gene304', 'gene321']) set(['gene147', 'gene146', 'gene109', 'gene382', 'gene143', 'gene411', 'gene101', 'gene499', 'gene212', 'gene239', 'gene195', 'gene56', 'gene460', 'gene462', 'gene405', 'gene70', 'gene257', 'gene138', 'gene293', 'gene425', 'gene442', 'gene349', 'gene220', 'gene378', 'gene321', 'gene474', 'gene463', 'gene457', 'gene302', 'gene304', 'gene453'])

dromedary has the following genes unique to itself: set(['gene162', 'gene389', 'gene102', 'gene100', 'gene107', 'gene5', 'gene431', 'gene416', 'gene10', 'gene323', 'gene19', 'gene309', 'gene193', 'gene374', 'gene235', 'gene234', 'gene218', 'gene406', 'gene404', 'gene136', 'gene135', 'gene335', 'gene332', 'gene428', 'gene371', 'gene394', 'gene390', 'gene106', 'gene410', 'gene361', 'gene456', 'gene185', 'gene45', 'gene44', 'gene493', 'gene450', 'gene125', 'gene37', 'gene82', 'gene85', 'gene84', 'gene307']) set(['gene162', 'gene389', 'gene102', 'gene5', 'gene431', 'gene10', 'gene19', 'gene371', 'gene193', 'gene374', 'gene406', 'gene136', 'gene85', 'gene410', 'gene335', 'gene332', 'gene135', 'gene390', 'gene106', 'gene361', 'gene185', 'gene45', 'gene44', 'gene493', 'gene84', 'gene125', 'gene37', 'gene323', 'gene450'])

lamb has the following genes unique to itself: set(['gene8', 'gene160', 'gene388', 'gene283', 'gene107', 'gene7', 'gene466', 'gene104', 'gene190', 'gene373', 'gene238', 'gene377', 'gene235', 'gene199', 'gene219', 'gene52', 'gene137', 'gene134', 'gene271', 'gene409', 'gene375', 'gene156', 'gene428', 'gene429', 'gene153', 'gene110', 'gene292', 'gene348', 'gene20', 'gene25', 'gene68', 'gene183', 'gene186', 'gene184', 'gene189', 'gene41', 'gene268', 'gene121', 'gene122', 'gene244', 'gene246', 'gene82', 'gene347', 'gene242', 'gene86']) set(['gene160', 'gene388', 'gene283', 'gene104', 'gene373', 'gene375', 'gene377', 'gene219', 'gene52', 'gene137', 'gene466', 'gene68', 'gene409', 'gene429', 'gene110', 'gene292', 'gene25', 'gene238', 'gene183', 'gene348', 'gene134', 'gene121', 'gene122', 'gene244', 'gene246', 'gene242', 'gene86'])

"readme.txt" 51L, 10487C