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 and description of code within the scripts.

1. Using the clinical\_data.txt, a script named copyExomes.sh was created to meet the necessary conditions(20-30mm in diameter and were sequenced) will be outputed to the exomes directory within the script. Commands used in the script was sed, grep, awk and cp. Script was executed using ./copyExomes.sh and the will print out the results.

2. Another script was created named createCrisprReady.sh. This script will serve to find the top three highest occurrences of sequences for each motif outputting the files within the top\_exomes directory. To execute the script type ./createCrisprReady.sh. Loop, grep, sort, sed, and awk commands were used within the script to generate the (animal)\_topmotifs.fasta files. Sub directory and txt files were generated to locate the top 3 sequences of each motif. Here are the location of each file created : top\_exomes/sorted\_exomes, top\_exomes/exomes\_list, and top\_exomes/sequenced\_exomes. Print statements are created to show the script has been completed.

3. The next script called identifyCrisprSite.sh was created. This script will locate the suitable CRISPR site for each gene inside the animal\_topmotifs.fasta. The sequence must be 20 basepairs before the contained sequence which is "NGG" were N is any base. To execute the script, please type ./identifyCrisprSite.shWithin this script, variables were set, grep, and sed command was created 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.

4. editGenome.sh was created to add an A base pair before every NGG(N meaning any base) sequence of each gene. To execute the script please type ./editGenome.sh. Variables were set, loop and sed command was created to add A in front of every NGG basepair. Print statements were generated to inform the script is complete.

5. A python script named exomeReport.py was created to generate the results of the project. To execute the script, type python exomeReport.py. This will generate the results of each organism with the diameter, environment, discoverer, unquie genes and common coherts Within the script a dictionary was created with conditions of 20-30mm in diameter and all exomes were sequenced. Sets, loops, dictionaries and if statements were created to distingush common cohorts and unique genes. Print statements was scripted to show results after running the script.

Below is a list of scripts that can be executed to show the results of the project.

copy exomes: ./copyExomes.sh

create Crispr: ./createCrisprReady.sh

identify Crispr site: ./identifyCrisprSite.sh

edit gene: ./editGenome.sh

Report : python exomeReport.py

Results:

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

The following genes are unique : 'gene289', 'gene141', 'gene438', 'gene282', 'gene284', 'gene148', 'gene38', 'gene35', 'gene32', 'gene439', 'gene57', 'gene237', 'gene140', 'gene231', 'gene448', 'gene464', 'gene150', 'gene179', 'gene115', 'gene64', 'gene61', 'gene22', 'gene203', 'gene262', 'gene206', 'gene224', 'gene28', 'gene248', 'gene247'

there are 29 unique genes

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

The following genes are unique : 'gene162', 'gene389', 'gene102', 'gene106', 'gene431', 'gene10', 'gene19', 'gene371', 'gene193', 'gene374', 'gene406', 'gene136', 'gene85', 'gene135', 'gene335', 'gene332', 'gene390', 'gene5', 'gene410', 'gene361', 'gene185', 'gene45', 'gene44', 'gene493', 'gene84', 'gene125', 'gene37', 'gene323', 'gene450'

there are 29 unique genes

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

The following genes are unique : 'gene180', 'gene215', 'gene387', 'gene395', 'gene6', 'gene31', 'gene13', 'gene99', 'gene492', 'gene458', 'gene496', 'gene422', 'gene336', 'gene253', 'gene447', 'gene46'

there are 16 unique genes

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

The following genes are unique : 'gene380', 'gene145', 'gene165', 'gene413', 'gene337', 'gene498', 'gene385', 'gene370', 'gene191', 'gene359', 'gene312', 'gene196', 'gene311', 'gene198', 'gene73', 'gene272', 'gene333', 'gene119', 'gene317', 'gene67', 'gene481', 'gene360', 'gene187', 'gene341', 'gene352', 'gene455'

there are 26 unique genes

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

The following genes are unique : 'gene419', 'gene276', 'gene383', 'gene59', 'gene369', 'gene412', 'gene421', 'gene287', 'gene133', 'gene490', 'gene494', 'gene60', 'gene468', 'gene338', 'gene445', 'gene87', 'gene249'

there are 17 unique genes

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

The following genes are unique : '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'

there are 31 unique genes

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

The following genes are unique : 'gene160', 'gene388', 'gene283', 'gene104', 'gene373', 'gene238', 'gene377', 'gene219', 'gene52', 'gene137', 'gene466', 'gene68', 'gene409', 'gene429', 'gene110', 'gene292', 'gene25', 'gene375', 'gene183', 'gene348', 'gene134', 'gene121', 'gene122', 'gene244', 'gene246', 'gene242', 'gene86'

there are 27 unique genes