Project-1

Due Date: 03/17

As discussed in the class you project has the following description:

1. Download the data from the encode website. Only use DNase-seq data. Search for DNase and there will be thousands of experiments. On the search result page, on the left side there is a filter- choose the following parameters:

   Assay type ->DNase-seq,
   Biosample -> Homo sapiens,
   Organ-> blood, brain, bodily fluid (any one),
   Cell -> choose with the highest number of experiments associated with it.
   Analysis -> GRCh38.
   Read length (nt) ->101 OR 151 (choose any one).

   That's enough filter. If you have to, you can change the cell/organ to find out a read length that has a higher number of experiments associated with it. Download at least 10 different files **bed narrowPeak** file type from 10 different experiments.
2. Convert the narrowPeak file into a nucleotide (ATGCs) file, using the bedtools. Let's say these files are atgc_file just an example.
3. Extract all the nucleotides from this file and store the data into a different file containing only the nucleotide sequences and nothing else. So the file will contain many many rows of nucleotides each of them having different lengths.
4. The length of each sequence in the above file is of variable length. Come up with a number that optimally crops the sequences in such a manner that all the sequences are of the same length. If you have to discard sequences that are smaller than the number that you come up with, discard them. And trim the sequence which is longer than the number that you come up with.
5. Now you should have a file containing N_1 number of sequences each having length of X. For example 500 sequences, each of length 120. This is the positive file, Or the file containing all the accessible regions in the DNA.
6. Now we need a negative file Or a file containing all the regions that do not contain accessible regions.
7. To create this file use the **narrowpeak** file and subtract the distance of two consecutive accessible regions. That is to find the middle region between two accessible regions.
8. Go through the same process as described above and finally you are going to have a file containing N_1 number of sequences each having length of X. This

X and the X from point-5 should be similar. N_1 and N_2 may differ, since this is the total number of sequences.

9. This file contains all the negative values or the sequences (ATGC) that do not represent accessible regions in the DNA.
10. Now you have two files, each having a different number (row) of DNA sequences but all having the same sequence length.
11. Now repeat this for the 10 different narrowPeak files from 10 different experiments.

Please follow the guidelines given in the assignment on submitting the solution to the TAs. Upload to Github and send it to the TA and also zip the files, all of them, and submit it in Moodle.

If you have any questions, happy to discuss during class, since this will help other students too. Most of the time I have to repeat the same solution to each individual which is not optimum. So I encourage you to ask those questions during the class time, I am more than happy to explain.