Project-1

Points: 100

Due Date: 10/15

A bed file: The minimum structure of a bed file is three columns, each column separated by tab. First column the chromosome name, second column the starting location and the third column the end location. There can be other parameters too but this is the bare minimum. If you pass a bed file with the minimum three column structure to a bedtools it is going to give us the nucleotide (ATGCs) present in the range.

For example: bed file-
chr1 345 452
…….
…….
chr2 1000 1007

Then passing this file to a bedtool will give:
>chr1 345 452
ATTTGGCC
…..
….
>chr2 1000 1007
ATGGTTT

This output file has a structure. Two lines for each line in the bed file. It starts with a **>** symbol then followed by the range and in the next line the nucleotides.

Creating the negative bed file

From the original (downloaded) bed file find the interval of negative sequence. E.g:

chr1 10124 **10134** **……..**
chr1 **10298** 10345 ……..
chr1 11234 11289 ……..

Then you have to create the negative bed file, which will look something like this

chr1 **10135** **10297**
chr1 10346 11233
Etc…

The above file is a **tab separated file**, this is important. The range of the new bed file should be consistent with the number you have chosen to crop the positive file with. The above numbers are just random.

Now pass this file to the bedtools to create the nucleotide file.

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 (containing ATGCs) file, using the bedtools. Let's say the file name is **atgc.txt**.
3. The length of each nucleotide 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. Output all the nucleotide sequences having the same length of let's say X to a different file.
4. 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. **So this file only contains strings of nucleotides of a fixed length**.
5. Now we need a negative file Or a file containing all the regions that do not contain accessible regions.
6. To create this file use the **bed narrowpeak** file (or the **atgc.txt file**, it depends on your programming logic) and subtract the distance of two consecutive accessible regions. This is going to give the negative regions or non

accessible regions. Assuming all the numbers are ascending, if not apply some programming logic. **There will be other inconsistencies while creating this file, think about it and address it.**

7. While creating a negative bed file, with a minimum of three columns. Remember that the range should be X. The number that you have chosen in point-4 above. Pass this newly created bed file to bedtools.
8. Extract only the nucleotides from the above file and put them into a new file. Let's say there will be N_2 number of sequences each of length X (same as the positive file) in this file.
9. Now you have two files, each having a different number of DNA sequences, but all having the same sequence length. The dimensions will be positive file $N_1 * X$ and negative file $N_2 * X$.
10. Now repeat this for the 10 different narrowPeak files from 10 different experiments.

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.

There will be an in person interview. A working code does not guarantee any points, the interview and working code guarantees points. So please present during the interview. Zip the python codes and upload in Moodle on the day of the deadline.