Class 31/03/2020
R Exercise

Needed libraries:
library(biomaRt)
library(MotifDb)
library(seqLogo)
library(PWMEnrich)
library(PWMEnrich.Hsapiens.background)

TASK 1

a) Select a chromosome.
b) Find basic annotations for all the human genes on the chromosome you selected.
c) Select only protein coding genes.
d) How many genes do you find?

TASK 2

a) Collect the promoter regions of 500 random genes.
    ○ consider a window of 30 nucleotides upstream and use the "sample" function to extract the ensembl_gene_id of the 500 random genes. Set “seqType” parameter to “gene_flank”.
b) Build the PFM from the sequences and compute the PPM
    ○ Use the function “consensusMatrix” applied on the vector of sequences.
c) Generate the logo
d) For the same genes, now generate the logo of the downstream 30 nucleotides.
e) Do you see differences between the upstream and the downstream region of your genes?

TASK 3

a) Find the top five TFs that are enriched in a collection of promoters.
    ○ generate the collection of promoters using the same code implemented in the previous task, but using a window of 300 nucleotides upstream and only 100 genes.

- ○ Use function “groupReport” to find TF enrichments across many sequences.

b) Compute empirical distributions of scores for all PWMs that you find in MotifDB for the previously selected top five TFs and determine for all of them the distribution (log2) threshold cutoff at 99.9%.

c) Calculate for each TF the fraction of the collected promoters having at least one binding score above the computed threshold for any of the TF related PWM matrices.

- ○ Use pattern matching function motifsScore with raw.score=FALSE and setting the parameter cutoff.

d) Calculate again the fractions using threshold cutoffs at 99%.