This 3UTR python script takes in ANNOVAR formatted.csv files and counts the number of unique regions (intergenic, intronic, ncRNA\_intronic, UTR3, exonic, downstream, upstream, ncRNA\_exonic, UTR5, upstream;downstream, splicing, ncRNA\_splicing, and UTR5;UTR3) in each ANNOVAR formatted trait file. By utilizing the bash script to automate this process for all traits, the outputs are stored in a created outputfile.

Starting with

cand = open(sys.argv[1], 'r')

Sys.argv[1] is connected with the bash script which can automate the process of all traits by replacing sys.argv[1] with each trait.

try:

f = open('ZTraitsOutputfile.csv')

f.close()

cand1 = open('ZTraitsOutputfile.csv', 'a')

except OSError:

cand1 = open('ZTraitsOutputfile.csv', 'w')

Opens and allows the writing of the output file for all outputs generated by the python script.

words= []

for line in cand:

line = line.strip('\n').split(',')

words.append(line[6])

cand1.write("\*\*\*\*\*\* start read " + sys.argv[1] + " \*\*\*\*\*\*\*\*\*" + '\n')

words\_counted = []

for word in words:

if word not in [row[0] for row in words\_counted]:

wordcount = words.count(word)

words\_counted.append((word,wordcount))

print(word,wordcount)

lineresult = word + ':' + str(wordcount)

cand1.write(lineresult + '\n')

This part of the script isolates the region column and counts the number of each region which will be printed on your screen and documented in the output file. .