This common genes python script takes in ANNOVAR formatted csv or excel files with variants info for two individual traits and compares them to find the number of common variants and the variant locations. By utilizing the bash script to automate this process for all traits, the outputs are stored in an output txt file.

Starting with

import sys, os

import pandas as pd

inFile1=sys.argv[1]

inFile2=sys.argv[2]

inFile1t=os.path.splitext(inFile1)[1]

inFile2t=os.path.splitext(inFile2)[1]

outFile1 = os.path.splitext(inFile1)[0]

outFile2 = os.path.splitext(inFile2)[0]

Sys.argv[1] and Sys.argv[2] are connected with the bash script which can automate the process of all traits by replacing sys.argv[1] and sys.argv[2] with the two variants tables.

#saving the excels as pandas tables

if inFile1t=='.xls' or inFile1t=='.xlsx':

tb1 = pd.read\_excel(inFile1)

if inFile1t=='.csv':

tb1 = pd.read\_csv(inFile1)

if inFile2t=='.xls' or inFile2t=='.xlsx':

tb2 = pd.read\_excel(inFile2)

if inFile2t=='.csv':

tb2 = pd.read\_csv(inFile2)

The inputted excel or csv tables are saved in pandas dataframe for processing

#creating a new column with Chr + Start info

df1 = pd.DataFrame(tb1)

df1['chrstart'] = df1[['Chr', 'Start']].astype(str).apply(lambda x: ','.join(x), axis=1)

df2 = pd.DataFrame(tb2)

df2['chrstart'] = df2[['Chr', 'Start']].astype(str).apply(lambda x: ','.join(x), axis=1)

#Saving all the rows in the chrstart to a list

list1 = df1["chrstart"].tolist()

list2 = df2["chrstart"].tolist()

#comparing the two lists

#another way of doing is using set if you prefer that

#set1 = set(list1)

#commons = set1.intersection(list2)

#num\_commons = set1.len()

The locations of the variants for each trait (Chr number plus start location) are stored in two separate lists

commons = [element for element in list1 if element in list2]

num\_commons = len(commons)

print(num\_commons )

print(commons)

Num\_commons prints the number of common variants, while ‘commons’ prints the variants locations.