Comprehensive- mmc7

June 26, 2022

1 importing pandas , numpy and matplotlib.pyplot packages

[21]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

2 Reading the PDB databse

[22]:
```
ATG_PDB = pd.read_csv('PDB.txt')
ATG_PDB=ATG_PDB['symbol'].values.tolist()
# len(ATG_PDB)
```

3 Reading the Morishita & Mizushima's database

[23]:
```
ATG_Mizushima = pd.read_csv('Murshita.txt')
ATG_Mizushima=ATG_Mizushima['symbol'].values.tolist()
# len(ATG_Mizushima)
```

4 Reading the Tanpaku database

[24]:
```
ATG_tanpaku = pd.read_csv('Japaness(tanpaku).txt')
ATG_tanpaku=ATG_tanpaku['symbol'].values.tolist()
# len(ATG_tanpaku)
```

5 Reading the other papars

[25]:
```
ATG_Isaac = pd.read_excel('Pathogenic Single Nucleotide Polymorphisms on␣
↪Autophagy-Related Genes.xlsx')
ATG_Isaac=ATG_Isaac['Gene'].values.tolist()
ATG_Isaac =[x.strip(' ') for x in ATG_Isaac]
# len(ATG_Isaac)
```

6 Reading Tudor I. Oprea's dataset

[26]:
```
ATG_Tudor=pd.read_excel('Tudor Opera.
 ↪xlsx',sheet_name='Input_Output',engine='openpyxl')
ATG_Tudor=ATG_Tudor['symbol'].values.tolist()
# len(ATG_Tudor)
```

7 Reading Tudor I. Oprea's + dark genes dataset

[27]:
```
dark_genes = pd.read_csv('dark_genes.txt')
dark_genes=dark_genes['symbol'].values.tolist()
# len(dark_genes)
```

8 Making a set of the ATG genes (comprehensive list) and removing a NaN value from the list. The total number of comprehensive list is 9812

[28]:
```
comprehensive= ATG_PDB + ATG_Mizushima + ATG_tanpaku + ATG_Isaac + ATG_Tudor +␣
 ↪dark_genes
comprehensive= set(comprehensive)
comprehensive= [str(x) for x in comprehensive]
comprehensive = [x for x in comprehensive if x !='nan']
len(comprehensive)
```

[28]:
```
9812
```

9 Converting the list of comprehensive to dataframe

10 (https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.html

[29]:
```
df = pd.DataFrame (comprehensive,columns=['symbol'])
df
```

[29]:

	symbol
0	PGC1A
1	cathepsin E
2	DRAM2
3	G0A67_11560
4	SAPK1
…	…
9807	E5KLJ5
9808	O14617
9809	P17655

```
9810       EIF2B2
9811         1M17

[9812 rows x 1 columns]
```

11 Saving the comprehensive dataframe as a CSV file

[14]:
```
df.to_csv('comprehensive.csv')
```

12 Reading the MMC7 file with different sheet name.

[31]:
```
mmc7_virus2host=pd.read_excel('1-s2.0-S0896627318304215-mmc7.
 ↪xlsx',sheet_name='virus2host',engine='openpyxl')
mmc7_host2virus=pd.read_excel('1-s2.0-S0896627318304215-mmc7.
 ↪xlsx',sheet_name='host2virus',engine='openpyxl')
```

13 Filtering "MMC7" with sheetname "virus2host" in selecting rows with having the comprehensive value in a "hostGene" column.

[39]:
```
intresect_virus2host_comprehensive =␣
 ↪mmc7_virus2host[mmc7_virus2host['hostGene'].isin(comprehensive)]
intresect_virus2host_comprehensive
```

[39]:

	tissue	marker	chr	pos	\
74	BM_10	rs7987664	chr13	77566158	
76	BM_10	rs7987664	chr13	77566158	
78	BM_10	rs7987664	chr13	77566158	
85	BM_10	rs7987664	chr13	77566158	
95	BM_10	rs7987664	chr13	77566158	
…	…	…	…	…	
91368	BM_44	rs1540000	chr21	19666901	
91372	BM_44	rs1540000	chr21	19666901	
91397	BM_44	rs1540000	chr21	19666901	
91437	BM_44	rs1540000	chr21	19666901	
91441	BM_44	rs1540000	chr21	19666901	

	viral_feature	\
74	NC_000898.1_repeat_region_153322_162114__ID.id63	
76	NC_000898.1_repeat_region_153322_162114__ID.id63	
78	NC_000898.1_repeat_region_153322_162114__ID.id63	
85	NC_000898.1_repeat_region_153322_162114__ID.id63	
95	NC_000898.1_repeat_region_153322_162114__ID.id63	

```
…                                                  …
91368               NC_020810.1_region_1_11976__ID.id0
91372               NC_020810.1_region_1_11976__ID.id0
91397   NC_020810.1_three_prime_UTR_11907_11976__ID.id2
91437   NC_020810.1_three_prime_UTR_11907_11976__ID.id2
91441   NC_020810.1_three_prime_UTR_11907_11976__ID.id2

                                          virus_name hostGene  \
74               Human herpesvirus 6B, complete genome     KAT8
76               Human herpesvirus 6B, complete genome    DAPK3
78               Human herpesvirus 6B, complete genome     SIM2
85               Human herpesvirus 6B, complete genome    ESRRA
95               Human herpesvirus 6B, complete genome     CDK1
…                                                  …       …
91368  Duvenhage virus isolate 86132SA, complete genome   FBXL17
91372  Duvenhage virus isolate 86132SA, complete genome   HSPA1B
91397  Duvenhage virus isolate 86132SA, complete genome     KAT8
91437  Duvenhage virus isolate 86132SA, complete genome   FBXL17
91441  Duvenhage virus isolate 86132SA, complete genome   HSPA1B

       p_cit_permuted  p_TassocL_permuted  p_TassocGgvnL_permuted  \
74           0.005994            0.173826                0.026973
76           0.027972            0.344655                0.046953
78           0.016983            0.610390                0.023976
85           0.018981            0.434565                0.083916
95           0.027972            0.192807                0.231768
…               …                   …                       …
91368        0.018981            0.243756                0.027972
91372        0.019980            0.395604                0.047952
91397        0.027972            0.854146                0.036963
91437        0.018981            0.243756                0.027972
91441        0.019980            0.395604                0.047952

       p_GassocLgvnT_permuted  p_LindTgvnG_permuted  p_cit_reactive_permuted  \
74                   0.987013              0.007992                 0.190809
76                   0.973027              0.014985                 0.073926
78                   0.984016              0.017982                 0.280719
85                   0.950050              0.049950                 0.224775
95                   0.919081              0.066933                 0.368631
…                       …                     …                        …
91368                0.973027              0.035964                 0.474525
91372                0.981019              0.017982                 0.138861
91397                0.968032              0.013986                 0.290709
91437                0.973027              0.035964                 0.474525
91441                0.981019              0.018981                 0.138861

       p_TassocL_reactive_permuted  p_TassocGgvnL_reactive_permuted  \
```

```
74                        0.000999                        0.026973
76                        0.000999                        0.046953
78                        0.000999                        0.023976
85                        0.000999                        0.083916
95                        0.000999                        0.231768
…                            …                               …
91368                     0.000999                        0.027972
91372                     0.000999                        0.047952
91397                     0.000999                        0.036963
91437                     0.000999                        0.027972
91441                     0.000999                        0.047952

       p_GassocLgvnT_reactive_permuted  p_LindTgvnG_reactive_permuted  \
74                            0.987013                        0.013986
76                            0.973027                        0.017982
78                            0.984016                        0.017982
85                            0.950050                        0.013986
95                            0.919081                        0.090909
…                                …                               …
91368                         0.973027                        0.019980
91372                         0.981019                        0.040959
91397                         0.968032                        0.019980
91437                         0.973027                        0.004995
91441                         0.981019                        0.035964

       transgene_cor_with_virus  transgene_cor_with_virus_pvalue
74                     -0.237486                         0.001603
76                     -0.236257                         0.001698
78                      0.209307                         0.005574
85                     -0.186659                         0.013657
95                     -0.117642                         0.122107
…                          …                               …
91368                   0.225220                         0.006094
91372                  -0.197256                         0.016631
91397                   0.290420                         0.000359
91437                   0.225220                         0.006094
91441                  -0.197256                         0.016631

[16624 rows x 19 columns]
```

14 Filtering "MMC7" with sheetname "host2virus" in selecting rows with having the comprehensive value in a "hostGene" column.

[40]:
```
intresect_host2virus_comprehensive =␣
↪mmc7_host2virus[mmc7_host2virus['hostGene'].isin(comprehensive)]
intresect_host2virus_comprehensive
```

[40]:

	tissue	marker	chr	pos	viral_feature	virus_name	hostGene
2	BM_10	rs506721:193241400:T:G	chr1	193241400	NC_000898.1_gene_138591_140054_Name.U91__ID.ge…	Human herpesvirus 6B, complete genome	GCK
6	BM_10	rs74130935:193191646:T:C	chr1	193191646	NC_000898.1_gene_138591_140054_Name.U91__ID.ge…	Human herpesvirus 6B, complete genome	GCK
10	BM_10	rs74130941:193216226:G:T	chr1	193216226	NC_000898.1_gene_138591_140054_Name.U91__ID.ge…	Human herpesvirus 6B, complete genome	GCK
14	BM_10	rs74130942:193216300:A:T	chr1	193216300	NC_000898.1_gene_138591_140054_Name.U91__ID.ge…	Human herpesvirus 6B, complete genome	GCK
18	BM_10	rs74909976:193216895:G:A	chr1	193216895	NC_000898.1_gene_138591_140054_Name.U91__ID.ge…	Human herpesvirus 6B, complete genome	GCK
…	…	…	…	…	…	…	…
19263	BM_22	rs7525340:52910974:A:C	chr1	52910974	NC_020808.1_region_1_11918__ID.id0	Aravan virus, complete genome	CDK1
19273	BM_22	rs76993084:53029552:A:T	chr1	53029552	NC_020808.1_region_1_11918__ID.id0	Aravan virus, complete genome	CDK1
19283	BM_22	rs77436971:53036467:A:T	chr1	53036467	NC_020808.1_region_1_11918__ID.id0	Aravan virus, complete genome	CDK1
19292	BM_22	rs72903653:52948750:T:C	chr1	52948750	NC_020808.1_region_1_11918__ID.id0	Aravan virus, complete genome	CDK1
19304	BM_44	rs1540000	chr21	19666901	NC_020810.1_region_1_11976__ID.id0	Duvenhage virus isolate 86132SA, complete genome	SMG9

	p_cit_permuted	p_TassocL_permuted	p_TassocGgvnL_permuted	\
2	0.072927	0.000999	0.280719	
6	0.074925	0.000999	0.285714	
10	0.074925	0.000999	0.285714	
14	0.074925	0.000999	0.285714	
18	0.072927	0.000999	0.280719	
…	…	…	…	
19263	0.117882	0.273726	0.232767	
19273	0.099900	0.272727	0.201798	
19283	0.113886	0.249750	0.219780	
19292	0.106893	0.328671	0.199800	
19304	0.080919	0.010989	0.099900	

	p_GassocLgvnT_permuted	p_LindTgvnG_permuted	p_cit_reactive_permuted	\
2	0.944056	0.133866	0.001998	
6	0.950050	0.122877	0.000999	
10	0.950050	0.120879	0.000999	
14	0.950050	0.144855	0.000999	
18	0.944056	0.128871	0.001998	
…	…	…	…	
19263	0.883117	0.101898	0.031968	
19273	0.901099	0.093906	0.029970	
19283	0.887113	0.111888	0.039960	
19292	0.894106	0.101898	0.038961	
19304	0.920080	0.072927	0.049950	

	p_TassocL_reactive_permuted	p_TassocGgvnL_reactive_permuted	\
2	0.000999	0.280719	
6	0.000999	0.285714	
10	0.000999	0.285714	
14	0.000999	0.285714	
18	0.000999	0.280719	
…	…	…	
19263	0.000999	0.232767	
19273	0.000999	0.201798	
19283	0.000999	0.219780	
19292	0.000999	0.199800	
19304	0.000999	0.099900	

	p_GassocLgvnT_reactive_permuted	p_LindTgvnG_reactive_permuted	\
2	0.610390	0.068931	
6	0.627373	0.046953	
10	0.627373	0.083916	
14	0.627373	0.071928	
18	0.610390	0.074925	
…	…	…	

```
19263                          0.883117                          0.088911
19273                          0.901099                          0.096903
19283                          0.887113                          0.103896
19292                          0.894106                          0.155844
19304                          0.920080                          0.088911

       transgene_cor_with_virus  transgene_cor_with_virus_pvalue
2                      0.161049                         0.033758
6                      0.161049                         0.033758
10                     0.161049                         0.033758
14                     0.161049                         0.033758
18                     0.161049                         0.033758
…                           …                                …
19263                 -0.207446                         0.056777
19273                 -0.207446                         0.056777
19283                 -0.207446                         0.056777
19292                 -0.207446                         0.056777
19304                 -0.217410                         0.008165

[1925 rows x 19 columns]
```

15 Length of ATG genes in MMC7-virus2host

[45]:
```
len(intresect_virus2host_comprehensive['hostGene'])
```

[45]:
```
16624
```

16 Length of ATG genes in MMC7-host2virus

[48]:
```
len(intresect_host2virus_comprehensive['hostGene'])
```

[48]:
```
1925
```

17 Length of unique ATG genes in MMC7-virus2host

[49]:
```
len(set(intresect_virus2host_comprehensive['hostGene']))
```

[49]:
```
410
```

18 Length of unique ATG genes in MMC7-host2virus

[50]:
```
len(set(intresect_host2virus_comprehensive['hostGene']))
```

[50]: 223

19 Filtering based on "Herpesvirus"

[58]:
```
list_herpesvirus_virus2host=intresect_virus2host_comprehensive[intresect_virus2host_comprehens
 ↪str.contains('herpesvirus')]
list_herpesvirus_host2virus=intresect_host2virus_comprehensive[intresect_host2virus_comprehens
 ↪str.contains('herpesvirus')]
```

20 Length of ATG genes in MMC7-virus2host based on "Herpesvirus"

[69]:
```
len(list_herpesvirus_virus2host['hostGene'])
```

[69]: 4875

21 Length of ATG genes in MMC7-host2virus based on "Herpesvirus"

[70]:
```
len(list_herpesvirus_host2virus['hostGene'])
```

[70]: 814

22 Length of unique ATG genes in MMC7-virus2host based on "Herpesvirus"

[72]:
```
len(set(list_herpesvirus_virus2host['hostGene']))
```

[72]: 306

23 Length of unique ATG genes in MMC7-host2virus based on "Herpesvirus"

[73]:
```
len(set(list_herpesvirus_host2virus['hostGene']))
```

[73]: 165

24 Saving the filtering MMC7 as a CSV file

[16]:
```
writer = pd.ExcelWriter('mmc7_Herpesvirus.xlsx', engine='xlsxwriter')
list_herpesvirus_virus2host.to_excel(writer, 'virus2host')
list_herpesvirus_host2virus.to_excel(writer, 'host2virus')

writer.save()
```