main

March 7, 2022

1 Comparison with public TWAS associations for SZ

[1]:
```
import urllib
import numpy as np
import pandas as pd
from html.parser import HTMLParser
from scipy.stats import fisher_exact
from html.entities import name2codepoint
```

[2]:
```
class MyHTMLParser(HTMLParser):
    def __init__(self):
        super().__init__()
        self.genes = []

    def handle_starttag(self, tag, attrs):
        for attr in attrs:
            if attr[0] == 'href' and attr[1].startswith('/genes/'):
                gene = [attr[1][len('/genes/'):]]
                self.genes += gene
```

1.1 MHC genes

[3]:
```
mhc_genes = pd.read_csv('/ceph/projects/v4_phase3_paper/analysis/twas_ea/'+\
                        "gene_weights/fusion_pgc2/_m/PGC2.SCZ.Caudate.6.dat.
 ↪MHC", sep='\t').ID
len(set(mhc_genes))
```

[3]:
```
105
```

1.2 Caudate specific of BrainSeq brain regions

[4]:
```
fn = "../../libd_twas_comparison/_m/caudate_only_twasList_genes.txt"
caudate = pd.read_csv(fn, sep='\t')
caudate.shape
```

[4]:
```
(309, 47)
```

```
[5]: len(set(caudate.ID) - set(mhc_genes))
```

```
[5]: 241
```

```
[6]: caudate_noMHC = caudate[(caudate["ID"].isin(list(set(caudate.ID) -␣
      ↪set(mhc_genes))))].copy()
     caudate_noMHC.iloc[0:2, 0:5]
```

```
[6]:               FILE        ID  CHR_TWAS        P0          P1
     0  ENSG00000242866      STRC        15  43599398  43618800.0
     1  ENSG00000183379  SYNDIG1L        14  74405893  74426102.0
```

1.3 TWAS Hub comparison

1.3.1 Schizophrenia 2014

```
[7]: parser = MyHTMLParser()
     html_str = urllib.request.urlopen('http://twas-hub.org/traits/Schizophrenia/').
      ↪read().decode()
     parser.feed(html_str)
     scz_2014 = np.unique(parser.genes)
```

```
[8]: scz_2014.shape
```

```
[8]: (49,)
```

1.3.2 Schizophrenia 2018

```
[9]: parser = MyHTMLParser()
     html_str = urllib.request.urlopen('http://twas-hub.org/traits/SCZ_2018/').
      ↪read().decode()
     parser.feed(html_str)
     scz_2018 = np.unique(parser.genes)
```

1.3.3 TWAS hub comparison

```
[10]: twas_hub = np.unique(np.append(scz_2014, scz_2018))
      len(twas_hub)
```

```
[10]: 67
```

```
[11]: twas_hub
```

```
[11]: array(['AC011816.1', 'AC103965.1', 'ACTR5', 'AKT3', 'ALMS1P', 'ANKRD44',
              'AS3MT', 'ATG13', 'BAI1', 'C12orf65', 'C17orf39', 'C2orf47',
              'CACNA2D4', 'CEP170', 'CHRNA5', 'CLCN3', 'CNTN4', 'CPEB1', 'CPNE7',
              'CRELD2', 'CUL3', 'CYSTM1', 'DCP1B', 'ELAC2', 'ERCC8', 'FAM53C',
              'FAM83H', 'FES', 'FLJ10661', 'GIGYF1', 'GMIP', 'HSPD1', 'IK',
              'IMMP2L', 'ITIH4-AS1', 'KIAA0319', 'KLC1', 'LGSN', 'LRRC48',
              'MAD1L1', 'MAP7D1', 'MAPK3', 'MGAT3', 'NAGA', 'NGEF', 'PCCB',
              'PCNX', 'PITPNM2', 'PLEKHO1', 'PPP1R13B', 'PPP1R14B', 'RERE',
              'RP11-981G7.6', 'SDCCAG8', 'SF3B1', 'SLCO4C1', 'SMG6', 'SNAP91',
              'SUGP2', 'THOC7', 'TRIM38', 'TTC14', 'U91328.21', 'VPS29',
              'VPS37A', 'XPNPEP3', 'ZNF318'], dtype='<U12')
```

```
[12]: print("There are %d caudate only genes present in the TWAS hub" %
            len(set(caudate.ID) & set(twas_hub)))
      set(caudate.ID) & set(twas_hub)
```

```
There are 6 caudate only genes present in the TWAS hub
```

```
[12]: {'ELAC2', 'MAD1L1', 'PPP1R13B', 'TRIM38', 'VPS29', 'VPS37A'}
```

```
[13]: print("There are %d caudate only genes (no MHC) present in the TWAS hub" %
            len(set(caudate_noMHC.ID) & set(twas_hub)))
      set(caudate.ID) & set(twas_hub)
```

```
There are 6 caudate only genes (no MHC) present in the TWAS hub
```

```
[13]: {'ELAC2', 'MAD1L1', 'PPP1R13B', 'TRIM38', 'VPS29', 'VPS37A'}
```

1.4 Gandal comparison

```
[14]: gandal = pd.read_excel("../_h/aat8127_Table_S4.xlsx", sheet_name="SCZ.TWAS")
      gandal_twas = gandal[(gandal['TWAS.Bonferroni'] <= 0.05)].copy()
      np.sum(gandal.loc[:, 'TWAS.Bonferroni'] < 0.05)
```

```
[14]: 193
```

```
[15]: caudate_bonferroni = caudate[(caudate['Bonferroni'] <= 0.05)].copy()
      np.sum(caudate.Bonferroni <= 0.05)
```

```
[15]: 85
```

```
[16]: caudate_noMHC_bonferroni = caudate_noMHC[(caudate_noMHC['Bonferroni'] <= 0.05)].
       ↪copy()
      np.sum(caudate_noMHC.Bonferroni <= 0.05)
```

```
[16]: 34
```

```
[17]: print("There are %d caudate only genes present in the Gandal at Bonferroni < 0.
       ↪05." %
            len(set(caudate_bonferroni.FILE) & set(gandal_twas.GeneID)))
      #list(set(caudate_bonferroni.ID) & set(gandal_twas.gene_name))

      print("There are %d caudate only genes present in the Gandal." %
            len(set(caudate.FILE) & set(gandal_twas.GeneID)))
```

```
np.array(list(set(caudate.FILE) & set(gandal_twas.GeneID)))
```

```
There are 21 caudate only genes present in the Gandal at Bonferroni < 0.05.
There are 29 caudate only genes present in the Gandal.
```

[17]:

```
array(['ENSG00000175662', 'ENSG00000219891', 'ENSG00000186522',
       'ENSG00000197279', 'ENSG00000205702', 'ENSG00000100162',
       'ENSG00000088808', 'ENSG00000006744', 'ENSG00000186470',
       'ENSG00000189298', 'ENSG00000204264', 'ENSG00000161896',
       'ENSG00000163938', 'ENSG00000231389', 'ENSG00000187987',
       'ENSG00000259404', 'ENSG00000158406', 'ENSG00000226314',
       'ENSG00000124613', 'ENSG00000219392', 'ENSG00000262074',
       'ENSG00000111237', 'ENSG00000272009', 'ENSG00000110492',
       'ENSG00000198315', 'ENSG00000231925', 'ENSG00000204963',
       'ENSG00000216901', 'ENSG00000168405'], dtype='<U15')
```

[18]:

```
print("There are %d caudate only genes (no MHC) present in the Gandal at␣
 ↪Bonferroni < 0.05." %
      len(set(caudate_noMHC_bonferroni.FILE) & set(gandal_twas.GeneID)))
#list(set(caudate_bonferroni.ID) & set(gandal_twas.gene_name))

print("There are %d caudate only genes (no MHC) present in the Gandal." %
      len(set(caudate_noMHC.FILE) & set(gandal_twas.GeneID)))
np.array(list(set(caudate_noMHC.FILE) & set(gandal_twas.GeneID)))
```

```
There are 9 caudate only genes (no MHC) present in the Gandal at Bonferroni <
0.05.
There are 13 caudate only genes (no MHC) present in the Gandal.
```

[18]:

```
array(['ENSG00000175662', 'ENSG00000262074', 'ENSG00000006744',
       'ENSG00000111237', 'ENSG00000163938', 'ENSG00000186522',
       'ENSG00000110492', 'ENSG00000259404', 'ENSG00000205702',
       'ENSG00000100162', 'ENSG00000088808', 'ENSG00000204963',
       'ENSG00000168405'], dtype='<U15')
```

MHC is present within Gandal analysis

[19]:

```
gandal[(gandal["TWAS.Bonferroni"] < 0.05)].shape
```

[19]:

```
(193, 31)
```

[20]:

```
## Not MHC
gandal[~(gandal['gene_name'].isin(list(set(mhc_genes)))) &
       (gandal["TWAS.Bonferroni"] < 0.05)].shape
```

[20]:

```
(177, 31)
```

[21]:

```
print("There are {} MHC genes within Gandal significant TWAS (Bonferroni < 0.
 ↪05)!"\
      .format(len(gandal[(gandal['gene_name'].isin(list(set(mhc_genes)))) &
       (gandal["TWAS.Bonferroni"] < 0.05)].gene_name)))
gandal[(gandal['gene_name'].isin(list(set(mhc_genes)))) &
       (gandal["TWAS.Bonferroni"] < 0.05)].gene_name
```

```
There are 16 MHC genes within Gandal significant TWAS (Bonferroni < 0.05)!
```

[21]:

```
2       HIST1H4J
3       HIST1H3C
5      ZSCAN12P1
10       ZSCAN23
15        BTN3A2
22       ZKSCAN3
29       ZKSCAN8
38        ZNF165
39      HIST1H4H
59      ZNF192P1
67         PSMB8
70         IP6K3
71        BTN2A2
124       ZNF391
146     HLA-DPA1
166        TAPBP
Name: gene_name, dtype: object
```

1.4.1 Calculated enrichment with Gandal

[22]:

```
caudate.shape
```

[22]:

```
(309, 47)
```

[23]:

```
dft = caudate.loc[:, ['FILE', 'ID', 'Bonferroni']]\
             .merge(gandal.loc[:, ['GeneID', 'gene_name', 'TWAS.Bonferroni']],
                    left_on='FILE', right_on='GeneID',
                    suffixes=["_Benjamin", "_Gandal"])
dft.shape
```

[23]:

```
(226, 6)
```

[24]:

```
table =  [[np.sum((dft['Bonferroni']<0.05) & ((dft['TWAS.Bonferroni']<.05))),
           np.sum((dft['Bonferroni']<0.05) & ((dft['TWAS.Bonferroni']>=.05)))],
          [np.sum((dft['Bonferroni']>=0.05) & ((dft['TWAS.Bonferroni']<.05))),
           np.sum((dft['Bonferroni']>=0.05) & ((dft['TWAS.Bonferroni']>=.05)))]]
print(table)
fisher_exact(table)
```

```
[[21, 30], [8, 167]]
```

[24]:

```
(14.6125, 7.051905747948124e-10)
```

1.4.2 Extract and save overlapping genes

Bonferroni < 0.05

[25]:
```
overlapping_twas = np.append(np.
 ↪array(caudate_bonferroni[(caudate_bonferroni['FILE'].
 ↪isin(list(set(caudate_bonferroni.FILE) &
                                                                                           ␣
 ↪set(gandal_twas.GeneID))))].ID),
                             np.array(list(set(caudate_bonferroni.ID) &␣
 ↪set(twas_hub))))
len(overlapping_twas)
```

[25]:
```
23
```

[26]:
```
caudate_bonferroni[~(caudate_bonferroni['ID'].isin(overlapping_twas))].
 ↪to_csv("caudate_only_twasList_genes_bonferroni.txt",
                                                        sep='\t', index=False)
caudate_bonferroni[~(caudate_bonferroni['ID'].isin(overlapping_twas))].shape
```

[26]:
```
(64, 47)
```

[27]:
```
drop_caudate = caudate_bonferroni[~(caudate_bonferroni['ID'].
 ↪isin(overlapping_twas))].copy()
drop_caudate[(drop_caudate['P'] > 5e-8)].shape
```

[27]:
```
(5, 47)
```

[28]:
```
drop_caudate[(drop_caudate['P'] > 5e-8)].sort_values('FDR')\
                                        .loc[:, ['ID', 'our_snp_id',␣
 ↪'CHR_TWAS', 'FDR', 'P', 'FILE']]
```

[28]:
```
          ID           our_snp_id  CHR_TWAS       FDR             P  \
270    CROCC     chr1:16497972:T:C         1  0.000042  1.200000e-04
85   ANKRD45    chr1:173743105:T:C         1  0.000054  5.050000e-07
0       STRC    chr15:43782086:A:G        15  0.000183  2.750000e-06
70    ADRA2A  chr10:111513602:C:T        10  0.000244  1.650000e-05
56    CCDC92  chr12:123431550:A:G        12  0.000334  7.100000e-07

                FILE
270  ENSG00000058453
85   ENSG00000183831
0    ENSG00000242866
70   ENSG00000150594
56   ENSG00000119242
```

FDR < 0.05

[29]:
```
overlapping_twas = np.append(np.array(caudate[(caudate['FILE'].
 ↪isin(list(set(caudate.FILE) &
                                                                                         ␣
 ↪set(gandal_twas.GeneID))))].ID),
                             np.array(list(set(caudate.ID) & set(twas_hub))))
len(overlapping_twas)
```

```
[29]: 35
```

```
[30]: caudate[~(caudate['ID'].isin(overlapping_twas))].
       ↪to_csv("caudate_only_twasList_genes.txt",
                                                        sep='\t', index=False)
      caudate[~(caudate['ID'].isin(overlapping_twas))].shape
```

```
[30]: (277, 47)
```

```
[31]: drop_caudate = caudate[~(caudate['ID'].isin(overlapping_twas))].copy()
      drop_caudate[(drop_caudate['P'] > 5e-8)].shape
```

```
[31]: (173, 47)
```

```
[32]: drop_caudate[(drop_caudate['P'] > 5e-8)].sort_values('FDR')\
                                                .loc[:, ['ID', 'our_snp_id',␣
       ↪'CHR_TWAS', 'FDR', 'P', 'FILE']].head(25)
```

```
[32]:                  ID            our_snp_id  CHR_TWAS       FDR             P  \
      270           CROCC     chr1:16497972:T:C         1  0.000042  1.200000e-04
      85          ANKRD45    chr1:173743105:T:C         1  0.000054  5.050000e-07
      0              STRC    chr15:43782086:A:G        15  0.000183  2.750000e-06
      70           ADRA2A   chr10:111513602:C:T        10  0.000244  1.650000e-05
      56           CCDC92   chr12:123431550:A:G        12  0.000334  7.100000e-07
      183          TBC1D5     chr3:16834975:C:T         3  0.000449  5.770000e-08
      71            STIM2     chr4:26781831:A:G         4  0.000558  4.210000e-07
      205            NTN5    chr19:48746940:T:C        19  0.000645  1.100000e-06
      182        SLC39A10    chr2:195559315:A:C         2  0.000664  3.000000e-05
      21    ENSG00000263715    chr17:46055092:G:A        17  0.000707  1.140000e-05
      212           ASIC1    chr12:50258497:G:A        12  0.000725  2.390000e-05
      215          UBQLN4    chr1:155906822:G:A         1  0.000816  9.600000e-06
      232           PPM1E    chr17:58681018:A:G        17  0.001272  2.390000e-06
      222            SNF8    chr17:48963251:G:A        17  0.001370  5.910000e-06
      40           PCDHB7    chr5:140841554:G:A         5  0.001405  8.660000e-07
      18            TRPS1     chr8:115455975:A:G         8  0.002413  4.280000e-07
      92             SSR2    chr1:155906822:G:A         1  0.002756  9.600000e-06
      25          SNORD13     chr8:33920637:G:A         8  0.002756  1.330000e-05
      68           ANAPC1    chr2:111884332:C:T         2  0.002925  1.200000e-03
      134       LRRC37A17P    chr17:46784796:G:A        17  0.003155  3.050000e-05
      86              CPD    chr17:30556723:C:T        17  0.003230  4.400000e-06
      12             EGFR     chr7:54797031:G:A         7  0.003444  1.740000e-04
      193         TUBGCP4    chr15:43782086:A:G        15  0.004538  2.750000e-06
```

```
147          PABPC1L   chr20:45053910:T:G        20  0.004817  3.320000e-06
63           RPS26P8   chr17:46055092:G:A        17  0.004940  1.140000e-05

                FILE
270  ENSG00000058453
85   ENSG00000183831
0    ENSG00000242866
70   ENSG00000150594
56   ENSG00000119242
183  ENSG00000131374
71   ENSG00000109689
205  ENSG00000142233
182  ENSG00000196950
21   ENSG00000263715
212  ENSG00000110881
215  ENSG00000160803
232  ENSG00000175175
222  ENSG00000159210
40   ENSG00000113212
18   ENSG00000104447
92   ENSG00000163479
25   ENSG00000239039
68   ENSG00000153107
134  ENSG00000263142
86   ENSG00000108582
12   ENSG00000146648
193  ENSG00000137822
147  ENSG00000101104
63   ENSG00000204652
```

[33]:
```
drop_caudate[(drop_caudate["ID"] == "MIAT")].loc[:, ['ID', 'our_snp_id',␣
 ↪'CHR_TWAS', 'FDR', 'P', 'FILE']]
```

[33]:
```
Empty DataFrame
Columns: [ID, our_snp_id, CHR_TWAS, FDR, P, FILE]
Index: []
```

1.5 TWAS tissue summary

[34]:
```
brainseq = pd.read_csv("../../libd_twas_comparison/_m/TWAS_gene_tissue_summary.
 ↪csv")
brainseq.shape
```

[34]:
```
(8348, 11)
```

[35]:
```
brainseq.head(2)
```

[35]:

```
            Geneid     Symbol  Caudate_TWAS.Z   Caudate_FDR Caudate_GWAS.SNP  \
0  ENSG00000219891  ZSCAN12P1       12.375993  1.871099e-31         Risk SNP
1  ENSG00000204338   CYP21A1P       12.163700  1.287468e-30         Risk SNP

   DLPFC_TWAS.Z  DLPFC_FDR DLPFC_GWAS.SNP  HIPPO_TWAS.Z  HIPPO_FDR  \
0           NaN        NaN            NaN           NaN        NaN
1           NaN        NaN            NaN           NaN        NaN

  HIPPO_GWAS.SNP
0            NaN
1            NaN
```

[36]:

```
bb = brainseq.merge(pd.DataFrame({'Symbol': twas_hub, 'inTWAS_HUB': 1}),␣
 ↪on='Symbol', how='left')\
             .merge(pd.DataFrame({'Geneid': gandal_twas.GeneID, 'inGandal': 1}),
                    on="Geneid", how='left')

bb.to_csv('TWAS_gene_tissue_summary.csv', index=False)
```