main

September 4, 2021

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/'+\
                        "gene_weights/fusion_pgc2/_m/PGC2.SCZ.Caudate.6.dat.
 ↪MHC", sep='\t').ID
len(set(mhc_genes))
```

[3]:
```
136
```

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]:
```
(445, 47)
```

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

[5]:
```
356
```

[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
2	ENSG00000100138	SNU13	22	41673930	41690504.0
3	ENSG00000204963	PCDHA7	5	140834248	141012344.0

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 8 caudate only genes present in the TWAS hub
```

[12]:
```
{'AKT3', 'C2orf47', 'ELAC2', 'ERCC8', 'MGAT3', 'PPP1R13B', 'SNAP91', 'VPS29'}
```

[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 8 caudate only genes (no MHC) present in the TWAS hub
```

[13]:
```
{'AKT3', 'C2orf47', 'ELAC2', 'ERCC8', 'MGAT3', 'PPP1R13B', 'SNAP91', 'VPS29'}
```

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]:
```
108
```

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

[16]:
```
45
```

[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 33 caudate only genes present in the Gandal at Bonferroni < 0.05.
There are 42 caudate only genes present in the Gandal.
```

[17]:
```
array(['ENSG00000185829', 'ENSG00000088808', 'ENSG00000205702',
       'ENSG00000175662', 'ENSG00000110492', 'ENSG00000177096',
       'ENSG00000124610', 'ENSG00000226314', 'ENSG00000197238',
       'ENSG00000030110', 'ENSG00000272009', 'ENSG00000085788',
       'ENSG00000228223', 'ENSG00000186470', 'ENSG00000162972',
       'ENSG00000219392', 'ENSG00000249839', 'ENSG00000006744',
       'ENSG00000168237', 'ENSG00000198315', 'ENSG00000231389',
       'ENSG00000065609', 'ENSG00000111237', 'ENSG00000261353',
       'ENSG00000259404', 'ENSG00000204963', 'ENSG00000262074',
       'ENSG00000197279', 'ENSG00000187987', 'ENSG00000100162',
       'ENSG00000204252', 'ENSG00000158406', 'ENSG00000117020',
       'ENSG00000168405', 'ENSG00000161896', 'ENSG00000219891',
       'ENSG00000186522', 'ENSG00000216901', 'ENSG00000163938',
       'ENSG00000174939', 'ENSG00000189298', 'ENSG00000124613'],
      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 16 caudate only genes (no MHC) present in the Gandal at Bonferroni <
0.05.
There are 22 caudate only genes (no MHC) present in the Gandal.
```

[18]:
```
array(['ENSG00000185829', 'ENSG00000088808', 'ENSG00000205702',
       'ENSG00000175662', 'ENSG00000110492', 'ENSG00000177096',
       'ENSG00000085788', 'ENSG00000162972', 'ENSG00000006744',
       'ENSG00000249839', 'ENSG00000168237', 'ENSG00000065609',
       'ENSG00000111237', 'ENSG00000259404', 'ENSG00000262074',
       'ENSG00000204963', 'ENSG00000100162', 'ENSG00000117020',
       'ENSG00000168405', 'ENSG00000186522', 'ENSG00000163938',
       'ENSG00000174939'], 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]:
```
(174, 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 19 MHC genes within Gandal significant TWAS (Bonferroni < 0.05)!
```

[21]:
```
2       HIST1H4J
3       HIST1H3C
5      ZSCAN12P1
6       HIST1H1A
10       ZSCAN23
15        BTN3A2
22       ZKSCAN3
29       ZKSCAN8
34         HCG11
38        ZNF165
39      HIST1H4H
59      ZNF192P1
70         IP6K3
71        BTN2A2
74       HLA-DOA
94      HIST1H3A
124       ZNF391
146     HLA-DPA1
174         BAK1
Name: gene_name, dtype: object
```

1.4.1 Calculated enrichment with Gandal

[22]:
```
caudate.shape
```

[22]:
```
(445, 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]: (312, 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)
```

```
[[33, 31], [9, 239]]
```

[24]: (28.268817204301076, 6.91682701358258e-19)

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]: 37

[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]: (75, 47)

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

[27]: (6, 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	\
192	STRC	chr15:43782086:A:G	15	0.000062	2.750000e-06	
132	ANKRD45	chr1:173743105:T:C	1	0.000129	5.050000e-07	
358	ENSG00000269938	chr12:123996254:A:G	12	0.000156	7.380000e-07	

```
220           ZNF852    chr3:44034110:G:A         3  0.000220  1.870000e-06
259  ENSG00000283361  chr13:114134675:T:C        13  0.000241  2.780000e-07
71            SPHKAP   chr2:228452776:C:T         2  0.000298  7.070000e-08

                FILE
192  ENSG00000242866
132  ENSG00000183831
358  ENSG00000269938
220  ENSG00000178917
259  ENSG00000283361
71   ENSG00000153820
```

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]:
```
50
```

[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]:
```
(401, 47)
```

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

[31]:
```
(250, 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  \
192             STRC   chr15:43782086:A:G        15  0.000062  2.750000e-06
132          ANKRD45   chr1:173743105:T:C         1  0.000129  5.050000e-07
358  ENSG00000269938  chr12:123996254:A:G        12  0.000156  7.380000e-07
220           ZNF852    chr3:44034110:G:A         3  0.000220  1.870000e-06
259  ENSG00000283361  chr13:114134675:T:C        13  0.000241  2.780000e-07
71            SPHKAP   chr2:228452776:C:T         2  0.000298  7.070000e-08
333  ENSG00000279726   chr5:140841554:G:A         5  0.000384  8.660000e-07
52          SYNDIG1L   chr14:74416653:A:C        14  0.000454  3.090000e-05
```

301	NTN5	chr19:48746940:T:C	19	0.000477	1.100000e-06
106	TBC1D5	chr3:16834975:C:T	3	0.000628	5.770000e-08
213	PCDHAC1	chr5:140841554:G:A	5	0.000643	8.660000e-07
425	CCDC92	chr12:123431550:A:G	12	0.000703	7.100000e-07
200	GSTO1	chr10:104732662:T:C	10	0.000712	2.010000e-05
380	NLRP1	chr17:5267575:T:C	17	0.000735	1.300000e-05
138	CEBPZOS	chr2:37017150:C:T	2	0.000815	3.140000e-06
91	IFT57	chr3:108058204:C:A	3	0.000961	2.010000e-05
153	TRPS1	chr8:115455975:A:G	8	0.000972	4.280000e-07
289	PDIA3	chr15:43782086:A:G	15	0.000974	2.750000e-06
250	RPF2	chr6:111221185:A:G	6	0.001017	8.300000e-07
90	FAM134A	chr2:219196879:T:C	2	0.001115	6.300000e-06
198	DND1	chr5:140841554:G:A	5	0.001170	8.660000e-07
409	RN7SKP101	chr15:47038124:C:T	15	0.001186	2.670000e-06
241	ZBTB37	chr1:173743105:T:C	1	0.001201	5.050000e-07
48	NPIPB3	chr16:21680256:A:G	16	0.001547	1.790000e-07
12	ADAM10	chr15:58749813:T:C	15	0.001766	1.610000e-06

	FILE
192	ENSG00000242866
132	ENSG00000183831
358	ENSG00000269938
220	ENSG00000178917
259	ENSG00000283361
71	ENSG00000153820
333	ENSG00000279726
52	ENSG00000183379
301	ENSG00000142233
106	ENSG00000131374
213	ENSG00000248383
425	ENSG00000119242
200	ENSG00000148834
380	ENSG00000091592
138	ENSG00000218739
91	ENSG00000114446
153	ENSG00000104447
289	ENSG00000167004
250	ENSG00000197498
90	ENSG00000144567
198	ENSG00000256453
409	ENSG00000223308
241	ENSG00000185278
48	ENSG00000169246
12	ENSG00000137845

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

[33]:

	ID	our_snp_id	CHR_TWAS	FDR	P	FILE
300	MIAT	chr22:26649934:G:T	22	0.014559	0.000111	ENSG00000225783

1.5 TWAS tissue summary

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

[34]:
```
(10387, 11)
```

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

[35]:

	Geneid	Symbol	Caudate_TWAS.Z	Caudate_FDR	Caudate_GWAS.SNP	DLPFC_TWAS.Z	DLPFC_FDR	DLPFC_GWAS.SNP	HIPPO_TWAS.Z	HIPPO_FDR	HIPPO_GWAS.SNP
0	ENSG00000000457	SCYL3	1.090068	0.597981	Other	NaN	NaN	NaN	NaN	NaN	NaN
1	ENSG00000000460	C1orf112	-0.372763	0.892471	Other	NaN	NaN	NaN	NaN	NaN	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)
```