main

August 6, 2021

1 Comparison with other datasets

[1]:
```
import functools
import numpy as np
import pandas as pd
```

1.1 BrainSeq functions

[2]:
```
config = {
    'caudate': '../../../caudate/male_analysis/metrics_summary/_m/
 ↪male_specific_DE_4features.txt',
    'dlpfc': '../../../dlpfc/male_analysis/metrics_summary/_m/
 ↪male_specific_DE_4features.txt',
    'hippo': '../../../hippocampus/male_analysis/metrics_summary/_m/
 ↪male_specific_DE_4features.txt',
}

config2 = {
    'caudate': '../../../caudate/male_analysis/_m/genes/diffExpr_szVctl_full.
 ↪txt',
    'dlpfc': '../../../dlpfc/male_analysis/_m/genes/diffExpr_szVctl_full.txt',
    'hippo': '../../../hippocampus/male_analysis/_m/genes/diffExpr_szVctl_full.
 ↪txt',
}
```

[3]:
```
@functools.lru_cache()
def get_deg(filename):
    dft = pd.read_csv(filename, sep='\t', index_col=0)
    if 'Type' in dft.columns:
        dft = dft[(dft['Type'] == 'gene')].copy()
    dft['Feature'] = dft.index
    dft['Dir'] = np.sign(dft['t'])
    if 'gene_id' in dft.columns:
        dft['ensemblID'] = dft.gene_id.str.replace('\\..*', '', regex=True)
    return dft[['Feature', 'ensemblID', 'Symbol', 'adj.P.Val', 'logFC', 't',␣
 ↪'Dir']]
```

```
@functools.lru_cache()
def get_deg_sig(filename, fdr):
    dft = get_deg(filename)
    return dft[(dft['adj.P.Val'] < fdr)]
```

[4]:
```
def tissue_annotation(tissue):
    return {'dlpfc': 'DLPFC', 'hippo': 'Hippocampus',
            'caudate': 'Caudate', 'cmc_dlpfc': 'CMC DLPFC'}[tissue]
```

1.2 Qin comparison

[5]:
```
qin_file = '/ceph/users/jbenja13/projects/sex_sz_ria/input/public_results/_m/
 ↪qin/qin_results_probesets.csv'
qin = pd.read_csv(qin_file)
tissue = 'caudate'
qin.head(2)
```

[5]:
```
     Probe set  Gene symbol     Locus   \
0    209735_at        ABCG2      4q22
1  208868_s_at    GABARAPL1   12p13.2

                                        Description  Fold difference   \
0  ATP-binding cassette, sub-family G (WHITE), me…           -1.29
1        GABA(A) receptor-associated protein like 1             -1.17

   q-value
0    0.031
1    0.043
```

[6]:
```
for tissue in ['caudate', 'dlpfc', 'hippo']:
    fdr = 0.05 if tissue != 'dlpfc' else 0.01
    tot = len(set(qin.loc[:, 'Gene symbol ']))
    overlap = len(set(get_deg_sig(config[tissue], fdr).Symbol) &
                  set(qin.loc[:, 'Gene symbol '].str.replace(' ','')))
    xx = overlap / tot
    print("There is %d (%.1f%%) overlap between %s and PFC!" %
          (overlap, xx* 100, tissue_annotation(tissue)))
```

```
There is 2 (4.3%) overlap between Caudate and PFC!
There is 0 (0.0%) overlap between DLPFC and PFC!
There is 1 (2.2%) overlap between Hippocampus and PFC!
```

[7]:
```
shared = set(get_deg_sig(config['caudate'], 0.05).Symbol) & set(qin.loc[:,␣
 ↪'Gene symbol '].str.replace(' ',''))
shared
```

[7]:
```
{'BBX', 'USE1'}
```

[8]:
```
shared = set(get_deg_sig(config['hippo'], 0.05).Symbol) & set(qin.loc[:, 'Gene␣
↪symbol '].str.replace(' ',''))
shared
```

[8]:
```
{'USE1'}
```

[9]:
```
qin[qin['Gene symbol '].isin(['USE1 ', 'BBX '])]
```

[9]:
```
      Probe set  Gene symbol      Locus   \
17  221706_s_at         USE1   19p13.11
31    213016_at          BBX     3q13.1

                                        Description  Fold difference   \
17  unconventional SNARE in the ER 1 homolog (S. c…           -1.07
31                    bobby sox homolog (Drosophila)            1.21

    q-value
17    0.042
31    0.031
```

[10]:
```
get_deg_sig(config['caudate'], 0.05)[get_deg_sig(config['caudate'], 0.05).
↪Symbol.isin(["USE1", "BBX"])]
```

[10]:
```
                              Feature        ensemblID Symbol  adj.P.Val  \
Feature
ENSG00000053501.12  ENSG00000053501.12  ENSG00000053501   USE1   0.012722
ENSG00000114439.18  ENSG00000114439.18  ENSG00000114439    BBX   0.016255

                        logFC         t  Dir
Feature
ENSG00000053501.12 -0.070904 -3.234511 -1.0
ENSG00000114439.18  0.052563  3.134511  1.0
```

[11]:
```
get_deg_sig(config['hippo'], 0.05)[get_deg_sig(config['hippo'], 0.05).Symbol ==␣
↪'USE1']
```

[11]:
```
                              Feature        ensemblID Symbol  adj.P.Val  \
Feature
ENSG00000053501.12  ENSG00000053501.12  ENSG00000053501   USE1   0.031459

                        logFC         t  Dir
Feature
ENSG00000053501.12 -0.120799 -3.953364 -1.0
```

[12]:
```
for tissue in ['caudate', 'dlpfc', 'hippo']:
    fdr = 0.05 if tissue != 'dlpfc' else 0.05
    tot = len(set(qin.loc[:, 'Gene symbol ']))
    overlap = len(set(get_deg_sig(config2[tissue], fdr).Symbol) &
```

```
                    set(qin.loc[:, 'Gene symbol '].str.replace(' ','')))
    xx = overlap / tot
    print("There is %d (%.1f%%) overlap between %s and PFC!" %
          (overlap, xx* 100, tissue_annotation(tissue)))
```

```
There is 5 (10.9%) overlap between Caudate and PFC!
There is 1 (2.2%) overlap between DLPFC and PFC!
There is 2 (4.3%) overlap between Hippocampus and PFC!
```

[13]:
```
shared = set(get_deg_sig(config2['caudate'], 0.05).Symbol) & set(qin.loc[:,␣
↪'Gene symbol '].str.replace(' ',''))
shared
```

[13]:
```
{'ABCG2', 'BBX', 'GABARAPL1', 'PARD3', 'USE1'}
```

[14]:
```
qin[qin['Gene symbol '].isin(['ABCG2 ', 'GABARAPL1 ', 'PARD3 ', 'USE1 ', 'BBX␣
↪'])]
```

[14]:

	Probe set	Gene symbol	Locus	\
0	209735_at	ABCG2	4q22	
1	208868_s_at	GABARAPL1	12p13.2	
17	221706_s_at	USE1	19p13.11	
31	213016_at	BBX	3q13.1	
49	210094_s_at	PARD3	10p11.21	

	Description	Fold difference	\
0	ATP-binding cassette, sub-family G (WHITE), me…	-1.29	
1	GABA(A) receptor-associated protein like 1	-1.17	
17	unconventional SNARE in the ER 1 homolog (S. c…	-1.07	
31	bobby sox homolog (Drosophila)	1.21	
49	par-3 partitioning defective 3 homolog (C. ele…	1.08	

	q-value
0	0.031
1	0.043
17	0.042
31	0.031
49	0.041

[15]:
```
get_deg_sig(config2['caudate'], 0.05)[get_deg_sig(config2['caudate'], 0.05)\
    .Symbol.isin(['ABCG2', 'GABARAPL1', 'PARD3', 'USE1', "BBX"])]
```

[15]:

	Feature	ensemblID	Symbol	adj.P.Val	\
ENSG00000139112.10	ENSG00000139112.10	ENSG00000139112	GABARAPL1	0.000093	
ENSG00000118777.10	ENSG00000118777.10	ENSG00000118777	ABCG2	0.009206	
ENSG00000053501.12	ENSG00000053501.12	ENSG00000053501	USE1	0.012722	
ENSG00000114439.18	ENSG00000114439.18	ENSG00000114439	BBX	0.016255	
ENSG00000148498.15	ENSG00000148498.15	ENSG00000148498	PARD3	0.036528	

```
                          logFC         t  Dir
ENSG00000139112.10  0.142220  4.872570  1.0
ENSG00000118777.10 -0.255668 -3.364630 -1.0
ENSG00000053501.12 -0.070904 -3.234511 -1.0
ENSG00000114439.18  0.052563  3.134511  1.0
ENSG00000148498.15  0.059527  2.776951  1.0
```

[16]:
```
set(get_deg_sig(config2['dlpfc'], 0.05).Symbol) & set(qin.loc[:, 'Gene symbol␣
↪'].str.replace(' ',''))
```

[16]:
```
{'ABCG2'}
```

[17]:
```
qin[qin['Gene symbol '].isin(['ABCG2 '])]
```

[17]:
```
   Probe set  Gene symbol  Locus   \
0  209735_at        ABCG2   4q22

                                        Description  Fold difference   \
0  ATP-binding cassette, sub-family G (WHITE), me…           -1.29

   q-value
0    0.031
```

[18]:
```
get_deg_sig(config2['dlpfc'], 0.05)[get_deg_sig(config2['dlpfc'], 0.05)\
    .Symbol.isin(['ABCG2'])]
```

[18]:
```
                              Feature        ensemblID Symbol  adj.P.Val  \
ENSG00000118777.10  ENSG00000118777.10  ENSG00000118777  ABCG2   0.002825

                          logFC        t  Dir
ENSG00000118777.10 -0.391751 -4.78399 -1.0
```

[19]:
```
set(get_deg_sig(config2['hippo'], 0.05).Symbol) & set(qin.loc[:, 'Gene symbol␣
↪'].str.replace(' ',''))
```

[19]:
```
{'ABCG2', 'USE1'}
```

[20]:
```
qin[qin['Gene symbol '].isin(['ABCG2 ', 'USE1 '])]
```

[20]:
```
      Probe set  Gene symbol      Locus   \
0     209735_at        ABCG2       4q22
17  221706_s_at         USE1   19p13.11

                                        Description  Fold difference   \
0   ATP-binding cassette, sub-family G (WHITE), me…           -1.29
17  unconventional SNARE in the ER 1 homolog (S. c…           -1.07
```

```
    q-value
0       0.031
17      0.042
```

[21]:
```
get_deg_sig(config2['hippo'], 0.05)[get_deg_sig(config2['hippo'], 0.05)\
    .Symbol.isin(['ABCG2', 'USE1'])]
```

[21]:
```
                              Feature        ensemblID Symbol  adj.P.Val  \
ENSG00000118777.10  ENSG00000118777.10  ENSG00000118777  ABCG2   0.017282
ENSG00000053501.12  ENSG00000053501.12  ENSG00000053501   USE1   0.031459

                         logFC         t  Dir
ENSG00000118777.10 -0.383008 -4.289096 -1.0
ENSG00000053501.12 -0.120799 -3.953364 -1.0
```

GABARAPL1 direction does not agree

[ ]:
```

```