main

August 11, 2021

1 Generate supplemental data for TWAS, caudate, across all features

```
[1]: import pandas as pd
```

1.1 With MHC

```
[2]: genes = pd.read_csv('/ceph/projects/v4_phase3_paper/analysis/twas/'+\
                         'gene_weights/fusion_pgc2/summary_stats/_m/
      ↪fusion_associations.txt', sep='\t')
     annot = pd.read_csv('../../../../differential_expression/_m/genes/
      ↪diffExpr_szVctl_full.txt', sep='\t')
     genes = annot[['ensemblID']].merge(genes, left_on='ensemblID', right_on='FILE')
     genes = genes[['FILE', 'ensemblID', 'ID', 'HSQ', 'BEST.GWAS.ID', 'EQTL.ID',
                    'TWAS.Z', 'TWAS.P', 'FDR', 'Bonferroni']]
     genes['Type'] = 'Gene'
     genes.rename(columns={'FILE': 'Feature'}, inplace=True)
     genes.sort_values('TWAS.P').head(2)
```

```
[2]:               Feature        ensemblID        ID       HSQ  \
     6805  ENSG00000158691  ENSG00000158691   ZSCAN12  0.070262
     7214  ENSG00000219891  ENSG00000219891  ZSCAN12P1  0.266109

              BEST.GWAS.ID            EQTL.ID     TWAS.Z        TWAS.P  \
     6805  chr6:28744470:A:G  chr6:28744886:A:G -12.627320  1.492752e-36
     7214  chr6:28426903:C:T  chr6:27883095:G:A  12.353178  4.682431e-35

                    FDR    Bonferroni  Type
     6805  1.225699e-32  1.225699e-32  Gene
     7214  1.922372e-31  3.844744e-31  Gene
```

```
[3]: trans = pd.read_csv('/ceph/projects/v4_phase3_paper/analysis/twas/'+\
                         'transcript_weights/fusion_pgc2/summary_stats/_m/
      ↪fusion_associations.txt', sep='\t')
     annot = pd.read_csv('../../../../differential_expression/_m/transcripts/
      ↪diffExpr_szVctl_full.txt', sep='\t')
     annot['ensemblID'] = annot.gene_id.str.replace('\\..*', '', regex=True)
```

```
annot['FILE'] = annot.transcript_id.str.replace('\\..*', '', regex=True)
trans = annot[['ensemblID', 'FILE']].merge(trans, on='FILE')
trans = trans[['FILE', 'ensemblID', 'ID', 'HSQ', 'BEST.GWAS.ID', 'EQTL.ID',
               'TWAS.Z', 'TWAS.P', 'FDR', 'Bonferroni']]
trans['Type'] = 'Transcript'
trans.rename(columns={'FILE': 'Feature'}, inplace=True)
trans.sort_values('TWAS.P').head(2)
```

```
[3]:               Feature          ensemblID      ID       HSQ       BEST.GWAS.ID  \
     12743  ENST00000421553  ENSG00000197062  ZSCAN26  0.040779  chr6:28744470:A:G
     14531  ENST00000508906  ENSG00000186470   BTN3A2  0.187261  chr6:26463346:G:T

                      EQTL.ID     TWAS.Z        TWAS.P           FDR    Bonferroni  \
     12743  chr6:28650974:A:G  12.745212  3.314893e-37  4.880849e-33  4.880849e-33
     14531  chr6:26354866:G:A  11.909938  1.050557e-32  7.734201e-29  1.546840e-28

                  Type
     12743  Transcript
     14531  Transcript
```

```
[4]: exons = pd.read_csv('/ceph/projects/v4_phase3_paper/analysis/twas/'+\
                         'exon_weights/fusion_pgc2/summary_stats/_m/
      ↪fusion_associations.txt', sep='\t')
     annot = pd.read_csv('../../../../differential_expression/_m/exons/
      ↪diffExpr_szVctl_full.txt',
                         sep='\t', index_col=0)
     exons = annot[['ensemblID']].merge(exons, left_index=True, right_on='FILE')
     exons = exons[['FILE', 'ensemblID', 'ID', 'HSQ', 'BEST.GWAS.ID', 'EQTL.ID',
                    'TWAS.Z', 'TWAS.P', 'FDR', 'Bonferroni']]
     exons['Type'] = 'Exon'
     exons.rename(columns={'FILE': 'Feature'}, inplace=True)
     exons.sort_values('TWAS.P').head(2)
```

```
[4]:       Feature        ensemblID    ID       HSQ       BEST.GWAS.ID  \
     62254  e385121  ENSG00000168477  TNXB  0.043518  chr6:31793436:G:A
     62253  e385001  ENSG00000168477  TNXB  0.044636  chr6:31793436:G:A

                      EQTL.ID     TWAS.Z        TWAS.P           FDR    Bonferroni  \
     62254  chr6:32253775:G:A  12.941234  2.633644e-38  1.783056e-33  1.783056e-33
     62253  chr6:32253775:G:A  12.728702  4.095902e-37  1.386524e-32  2.773049e-32

            Type
     62254  Exon
     62253  Exon
```

```
[5]: dj_file = '../../../../differential_expression/_m/junctions/
      ↪diffExpr_szVctl_full.txt'
```

```
dj = pd.read_csv(dj_file, sep='\t', index_col=0)
dj = dj[['Symbol', 'ensemblID']]

jannot_file = '/ceph/projects/v4_phase3_paper/analysis/twas/_m/junctions/
 ↪jxn_annotation.tsv'
jannot = pd.read_csv(jannot_file, sep='\t', index_col=1)

jannot = jannot[['JxnID']]
annot = pd.merge(jannot, dj, left_index=True, right_index=True)

juncs = pd.read_csv('/ceph/projects/v4_phase3_paper/analysis/twas/'+\
                    'junction_weights/fusion_pgc2/summary_stats/_m/
 ↪fusion_associations.txt', sep='\t')
juncs = pd.merge(annot, juncs, left_on='JxnID', right_on='FILE')
juncs = juncs[['FILE', 'ensemblID', 'Symbol', 'HSQ', 'BEST.GWAS.ID', 'EQTL.ID',
               'TWAS.Z', 'TWAS.P', 'FDR', 'Bonferroni']]
juncs['Type'] = 'Junction'
juncs.rename(columns={'Symbol': 'ID', 'FILE': 'Feature'}, inplace=True)
juncs.sort_values('TWAS.P').head(2)
```

```
/usr/lib/python3.9/site-packages/IPython/core/interactiveshell.py:3146:
DtypeWarning: Columns (2) have mixed types.Specify dtype option on import or set
low_memory=False.
  has_raised = await self.run_ast_nodes(code_ast.body, cell_name,
```

```
[5]:        Feature        ensemblID     ID       HSQ       BEST.GWAS.ID  \
     19664  j125659              NaN    NaN  0.148096  chr6:31204374:T:C
     18979  j122115  ENSG00000137411  VARS2  0.118039  chr6:31348749:T:C

                      EQTL.ID     TWAS.Z        TWAS.P           FDR    Bonferroni  \
     19664  chr6:31229085:G:A -12.920964  3.428198e-38  8.003127e-34  8.003127e-34
     18979  chr6:30951614:G:A  12.375775  3.534662e-35  3.201745e-31  8.251668e-31

                Type
     19664  Junction
     18979  Junction
```

```
[6]: df = pd.concat([genes, trans, exons, juncs], axis=0)
     print(df.shape)
     df.head(2)
```

```
(113983, 11)
```

```
[6]:            Feature        ensemblID        ID       HSQ        BEST.GWAS.ID  \
     0  ENSG00000138944  ENSG00000138944  KIAA1644  0.185313  chr22:43809985:A:G
     1  ENSG00000185052  ENSG00000185052   SLC24A3  0.178962  chr20:18949619:C:T

                  EQTL.ID     TWAS.Z     TWAS.P       FDR  Bonferroni  Type
```

```
0  chr22:44052458:G:A -0.025951  0.979296  0.992839         1.0  Gene
1  chr20:19234998:A:G  0.210426  0.833335  0.940940         1.0  Gene
```

[7]:

```
df.to_csv('BrainSeq_Phase3_Caudate_TWAS_associations_allFeatures.txt.gz',␣
 ↪index=False, header=True, sep='\t')
```

1.2 Without MHC

[8]:

```
genes = pd.read_csv('/ceph/projects/v4_phase3_paper/analysis/twas/'+\
                    'gene_weights/fusion_pgc2/summary_stats/_m/
 ↪fusion_associations_noMHC.txt', sep='\t')
annot = pd.read_csv('../../../../differential_expression/_m/genes/
 ↪diffExpr_szVctl_full.txt', sep='\t')
genes = annot[['ensemblID']].merge(genes, left_on='ensemblID', right_on='FILE')
genes = genes[['FILE', 'ensemblID', 'ID', 'HSQ', 'BEST.GWAS.ID', 'EQTL.ID',
               'TWAS.Z', 'TWAS.P', 'FDR', 'Bonferroni']]
genes['Type'] = 'Gene'
genes.rename(columns={'FILE': 'Feature'}, inplace=True)
genes.sort_values('TWAS.P').head(2)
```

[8]:

```
              Feature        ensemblID        ID       HSQ  \
3154  ENSG00000100138  ENSG00000100138     SNU13  0.071722
4190  ENSG00000088808  ENSG00000088808  PPP1R13B  0.269173

            BEST.GWAS.ID             EQTL.ID    TWAS.Z        TWAS.P  \
3154   chr22:41944840:T:C   chr22:42069256:T:C -8.100041  5.494072e-16
4190  chr14:103847845:G:A  chr14:103756555:C:T  7.012638  2.338656e-12

               FDR    Bonferroni  Type
3154  4.437562e-12  4.437562e-12  Gene
4190  8.058985e-09  1.888933e-08  Gene
```

[9]:

```
trans = pd.read_csv('/ceph/projects/v4_phase3_paper/analysis/twas/'+\
                    'transcript_weights/fusion_pgc2/summary_stats/_m/
 ↪fusion_associations_noMHC.txt', sep='\t')
annot = pd.read_csv('../../../../differential_expression/_m/transcripts/
 ↪diffExpr_szVctl_full.txt', sep='\t')
annot['ensemblID'] = annot.gene_id.str.replace('\\..*', '', regex=True)
annot['FILE'] = annot.transcript_id.str.replace('\\..*', '', regex=True)
trans = annot[['ensemblID', 'FILE']].merge(trans, on='FILE')
trans = trans[['FILE', 'ensemblID', 'ID', 'HSQ', 'BEST.GWAS.ID', 'EQTL.ID',
               'TWAS.Z', 'TWAS.P', 'FDR', 'Bonferroni']]
trans['Type'] = 'Transcript'
trans.rename(columns={'FILE': 'Feature'}, inplace=True)
trans.sort_values('TWAS.P').head(2)
```

[9]:
```
                Feature         ensemblID     ID       HSQ          BEST.GWAS.ID  \
2276   ENST00000433628  ENSG00000148842  CNNM2  0.077605  chr10:103092132:T:C
13474  ENST00000553286  ENSG00000126214   KLC1  0.430065  chr14:103847845:G:A

                   EQTL.ID    TWAS.Z        TWAS.P           FDR  \
2276   chr10:103085115:T:C  7.652389  1.972789e-14  2.180951e-10
13474  chr14:103673689:C:T -7.597796  3.012155e-14  2.180951e-10

         Bonferroni        Type
2276   2.856796e-10  Transcript
13474  4.361902e-10  Transcript
```

[10]:
```python
exons = pd.read_csv('/ceph/projects/v4_phase3_paper/analysis/twas/'+\
                    'exon_weights/fusion_pgc2/summary_stats/_m/
 ↪fusion_associations.txt', sep='\t')
annot = pd.read_csv('../../../../differential_expression/_m/exons/
 ↪diffExpr_szVctl_full.txt',
                    sep='\t', index_col=0)
exons = annot[['ensemblID']].merge(exons, left_index=True, right_on='FILE')
exons = exons[['FILE', 'ensemblID', 'ID', 'HSQ', 'BEST.GWAS.ID', 'EQTL.ID',
               'TWAS.Z', 'TWAS.P', 'FDR', 'Bonferroni']]
exons['Type'] = 'Exon'
exons.rename(columns={'FILE': 'Feature'}, inplace=True)
exons.sort_values('TWAS.P').head(2)
```

[10]:
```
       Feature        ensemblID    ID       HSQ       BEST.GWAS.ID  \
62254  e385121  ENSG00000168477  TNXB  0.043518  chr6:31793436:G:A
62253  e385001  ENSG00000168477  TNXB  0.044636  chr6:31793436:G:A

                 EQTL.ID     TWAS.Z        TWAS.P           FDR    Bonferroni  \
62254  chr6:32253775:G:A  12.941234  2.633644e-38  1.783056e-33  1.783056e-33
62253  chr6:32253775:G:A  12.728702  4.095902e-37  1.386524e-32  2.773049e-32

       Type
62254  Exon
62253  Exon
```

[11]:
```python
dj_file = '../../../../differential_expression/_m/junctions/
 ↪diffExpr_szVctl_full.txt'
dj = pd.read_csv(dj_file, sep='\t', index_col=0)
dj = dj[['Symbol', 'ensemblID']]

jannot_file = '/ceph/projects/v4_phase3_paper/analysis/twas/_m/junctions/
 ↪jxn_annotation.tsv'
jannot = pd.read_csv(jannot_file, sep='\t', index_col=1)

jannot = jannot[['JxnID']]
```

```
annot = pd.merge(jannot, dj, left_index=True, right_index=True)

juncs = pd.read_csv('/ceph/projects/v4_phase3_paper/analysis/twas/'+\
                    'junction_weights/fusion_pgc2/summary_stats/_m/
 ↪fusion_associations_noMHC.txt', sep='\t')
juncs = pd.merge(annot, juncs, left_on='JxnID', right_on='FILE')
juncs = juncs[['FILE', 'ensemblID', 'Symbol', 'HSQ', 'BEST.GWAS.ID', 'EQTL.ID',
               'TWAS.Z', 'TWAS.P', 'FDR', 'Bonferroni']]
juncs['Type'] = 'Junction'
juncs.rename(columns={'Symbol': 'ID', 'FILE': 'Feature'}, inplace=True)
juncs.sort_values('TWAS.P').head(2)
```

```
/usr/lib/python3.9/site-packages/IPython/core/interactiveshell.py:3146:
DtypeWarning: Columns (2) have mixed types.Specify dtype option on import or set
low_memory=False.
  has_raised = await self.run_ast_nodes(code_ast.body, cell_name,
```

```
[11]:      Feature        ensemblID           ID       HSQ          BEST.GWAS.ID  \
      2595  j17393  ENSG00000270316  BORCS7-ASMT  0.150648  chr10:103092132:T:C
      2593  j17391              NaN          NaN  0.226750  chr10:103092132:T:C

                       EQTL.ID    TWAS.Z        TWAS.P           FDR    Bonferroni  \
      2595  chr10:102911075:C:A -9.579719  9.730918e-22  2.237430e-17  2.237430e-17
      2593  chr10:102911075:C:A -8.094280  5.760397e-16  5.242528e-12  1.324488e-11

                Type
      2595  Junction
      2593  Junction
```

```
[12]: df = pd.concat([genes, trans, exons, juncs], axis=0)
      print(df.shape)
      df.head(2)
```

```
      (113254, 11)
```

```
[12]:            Feature        ensemblID        ID       HSQ         BEST.GWAS.ID  \
      0  ENSG00000138944  ENSG00000138944  KIAA1644  0.185313  chr22:43809985:A:G
      1  ENSG00000185052  ENSG00000185052   SLC24A3  0.178962  chr20:18949619:C:T

                    EQTL.ID    TWAS.Z    TWAS.P       FDR  Bonferroni  Type
      0  chr22:44052458:G:A -0.025951  0.979296  0.993066         1.0  Gene
      1  chr20:19234998:A:G  0.210426  0.833335  0.942432         1.0  Gene
```

```
[13]: df.to_csv('BrainSeq_Phase3_Caudate_TWAS_associations_allFeatures_noMHC.txt.gz',
                index=False, header=True, sep='\t')
```