main

August 4, 2021

1 Examine if there is an enrichment of sex linked genes within BrainSeq Phase 2 and 3 (case-control; schizophrenia)

[1]:
```
import functools
import numpy as np
import pandas as pd
from os import environ
from gtfparse import read_gtf
from scipy.stats import fisher_exact
from statsmodels.stats.multitest import multipletests
```

[2]:
```
environ['NUMEXPR_MAX_THREADS'] = '32'
```

1.1 Load phase 2 results

[3]:
```
config = {
    "dlpfc_file": "/ceph/users/jbenja13/phase3_paper/phase2/extract_de/_m/
 ↪dlpfc_diffExpr_szVctl_full.txt",
    "caud8_file": "/ceph/projects/v4_phase3_paper/analysis/
 ↪differential_expression/_m/genes/diffExpr_szVctl_full.txt",
    "hippo_file": "/ceph/users/jbenja13/phase3_paper/phase2/extract_de/_m/
 ↪hippo_diffExpr_szVctl_full.txt",
    'xci_file': '../../../xci_enrichment/_h/xci_status_hg19.txt',
    "gtf_file": '/ceph/genome/human/gencode25/gtf.CHR/_m/gencode.v25.annotation.
 ↪gtf',
    'cmc_file': '/ceph/projects/v3_phase3_paper/inputs/cmc/_m/
 ↪CMC_MSSM-Penn-Pitt_DLPFC_mRNA_IlluminaHiSeq2500_gene-adjustedSVA-differentialExpression-inc
 ↪tsv'
}
```

[4]:
```
@functools.lru_cache()
def get_cmc():
    return pd.read_csv(config['cmc_file'], sep='\t')


@functools.lru_cache()
```

```
def get_xci():
    return pd.read_csv(config['xci_file'], sep='\t')


@functools.lru_cache()
def get_gtf(gtf_file):
    return read_gtf(gtf_file)


@functools.lru_cache()
def get_dlpfc():
    return pd.read_csv(config["dlpfc_file"], sep='\t')


@functools.lru_cache()
def get_hippo():
    return pd.read_csv(config["hippo_file"], sep='\t')


@functools.lru_cache()
def get_caudate():
    return pd.read_csv(config["caud8_file"], sep='\t')
```

[5]:
```
def gene_annotation():
    gtf0 = get_gtf(config["gtf_file"])
    gtf = gtf0[gtf0["feature"] == "gene"]
    return gtf[["gene_id", "gene_name", "seqname", "start", "end", "strand"]]
```

1.2 Hippocampus

[6]:
```
hippo = gene_annotation().merge(get_hippo()[(get_hippo()["type"] == 'gene')],
                                left_on="gene_id", right_on="gencodeID")
```

```
INFO:root:Extracted GTF attributes: ['gene_id', 'gene_type', 'gene_status',
'gene_name', 'level', 'havana_gene', 'transcript_id', 'transcript_type',
'transcript_status', 'transcript_name', 'transcript_support_level', 'tag',
'havana_transcript', 'exon_number', 'exon_id', 'ont', 'protein_id', 'ccdsid']
```

[7]:
```
table = [[np.sum((hippo['adj.P.Val']<0.20) & (hippo['seqname'] == 'chrX')),
          np.sum((hippo['adj.P.Val']<0.20) & (hippo['seqname'] != 'chrX'))],
         [np.sum((hippo['adj.P.Val']>=0.20) & (hippo['seqname'] == 'chrX')),
          np.sum((hippo['adj.P.Val']>=0.20) & (hippo['seqname'] != 'chrX'))]]
print(table)
fisher_exact(table)
```

```
[[7, 325], [829, 23491]]
```

[7]:
```
(0.6103256936067551, 0.22311752769934323)
```

[8]:
```
table = [[np.sum((hippo['adj.P.Val']<0.20) & (hippo['t']<0) & (hippo['seqname']␣
↪== 'chrX')),
          np.sum((hippo['adj.P.Val']<0.20) & (hippo['t']<0) & (hippo['seqname']␣
↪!= 'chrX'))],
         [np.sum((hippo['adj.P.Val']>=0.20) & (hippo['t']<0) &␣
↪(hippo['seqname'] == 'chrX')),
          np.sum((hippo['adj.P.Val']>=0.20) & (hippo['t']<0) &␣
↪(hippo['seqname'] != 'chrX'))]]
print(table)
fisher_exact(table)
```

```
[[4, 167], [461, 12220]]
```

[8]:
```
(0.6349123878057334, 0.5330876730005754)
```

[9]:
```
table = [[np.sum((hippo['adj.P.Val']<0.20) & (hippo['t']>0) & (hippo['seqname']␣
↪== 'chrX')),
          np.sum((hippo['adj.P.Val']<0.20) & (hippo['t']>0) & (hippo['seqname']␣
↪!= 'chrX'))],
         [np.sum((hippo['adj.P.Val']>=0.20) & (hippo['t']>0) &␣
↪(hippo['seqname'] == 'chrX')),
          np.sum((hippo['adj.P.Val']>=0.20) & (hippo['t']>0) &␣
↪(hippo['seqname'] != 'chrX'))]]
print(table)
fisher_exact(table)
```

```
[[3, 158], [368, 11271]]
```

[9]:
```
(0.5815389378095762, 0.493689880575509)
```

1.3 DLPFC

[10]:
```
dlpfc = gene_annotation().merge(get_dlpfc()[(get_dlpfc()["type"] == 'gene')],
                                left_on="gene_id", right_on="gencodeID")
```

[11]:
```
table = [[np.sum((dlpfc['adj.P.Val']<0.05) & (dlpfc['seqname'] == 'chrX')),
          np.sum((dlpfc['adj.P.Val']<0.05) & (dlpfc['seqname'] != 'chrX'))],
         [np.sum((dlpfc['adj.P.Val']>=0.05) & (dlpfc['seqname'] == 'chrX')),
          np.sum((dlpfc['adj.P.Val']>=0.05) & (dlpfc['seqname'] != 'chrX'))]]
print(table)
fisher_exact(table)
```

```
[[10, 235], [826, 23581]]
```

[11]:
```
(1.2148266446860028, 0.47891981920788795)
```

[12]:
```
table = [[np.sum((dlpfc['adj.P.Val']<0.05) & (dlpfc['t']<0) & (dlpfc['seqname']␣
↪== 'chrX')),
```

```
          np.sum((dlpfc['adj.P.Val']<0.05) & (dlpfc['t']<0) & (dlpfc['seqname']␣
 ↪!= 'chrX'))],
         [np.sum((dlpfc['adj.P.Val']>=0.05) & (dlpfc['t']<0) &␣
 ↪(dlpfc['seqname'] == 'chrX')),
          np.sum((dlpfc['adj.P.Val']>=0.05) & (dlpfc['t']<0) &␣
 ↪(dlpfc['seqname'] != 'chrX'))]]
print(table)
fisher_exact(table)
```

```
[[2, 140], [423, 12642]]
```

[12]: (0.4269503546099291, 0.3337250489445369)

[13]:
```
table = [[np.sum((dlpfc['adj.P.Val']<0.05) & (dlpfc['t']>0) & (dlpfc['seqname']␣
 ↪== 'chrX')),
          np.sum((dlpfc['adj.P.Val']<0.05) & (dlpfc['t']>0) & (dlpfc['seqname']␣
 ↪!= 'chrX'))],
         [np.sum((dlpfc['adj.P.Val']>=0.05) & (dlpfc['t']>0) &␣
 ↪(dlpfc['seqname'] == 'chrX')),
          np.sum((dlpfc['adj.P.Val']>=0.05) & (dlpfc['t']>0) &␣
 ↪(dlpfc['seqname'] != 'chrX'))]]
print(table)
fisher_exact(table)
```

```
[[8, 95], [403, 10939]]
```

[13]: (2.2858038396238736, 0.03174889585949236)

1.4 Caudate nucleus

[14]:
```
caudate = gene_annotation().merge(get_caudate(), left_on="gene_id",␣
 ↪right_on="gencodeID")
```

[15]:
```
table = [[np.sum((caudate['adj.P.Val']<0.05) & (caudate['seqname'] == 'chrX')),
          np.sum((caudate['adj.P.Val']<0.05) & (caudate['seqname'] != 'chrX'))],
         [np.sum((caudate['adj.P.Val']>=0.05) & (caudate['seqname'] == 'chrX')),
          np.sum((caudate['adj.P.Val']>=0.05) & (caudate['seqname'] !=␣
 ↪'chrX'))]]
print(table)
fisher_exact(table)
```

```
[[101, 2595], [653, 19609]]
```

[15]: (1.1687625218717819, 0.1506223372700362)

[16]:
```
table = [[np.sum((caudate['adj.P.Val']<0.05) & (caudate['t']<0) &␣
 ↪(caudate['seqname'] == 'chrX')),
```

```
          np.sum((caudate['adj.P.Val']<0.05) & (caudate['t']<0) &␣
 ↪(caudate['seqname'] != 'chrX'))],
         [np.sum((caudate['adj.P.Val']>=0.05) & (caudate['t']<0) &␣
 ↪(caudate['seqname'] == 'chrX')),
          np.sum((caudate['adj.P.Val']>=0.05) & (caudate['t']<0) &␣
 ↪(caudate['seqname'] != 'chrX'))]]
print(table)
fisher_exact(table)
```

```
[[49, 1348], [332, 10329]]
```

[16]:
```
(1.1309056701583784, 0.416285160776556)
```

[17]:
```
table = [[np.sum((caudate['adj.P.Val']<0.05) & (caudate['t']>0) &␣
 ↪(caudate['seqname'] == 'chrX')),
          np.sum((caudate['adj.P.Val']<0.05) & (caudate['t']>0) &␣
 ↪(caudate['seqname'] != 'chrX'))],
         [np.sum((caudate['adj.P.Val']>=0.05) & (caudate['t']>0) &␣
 ↪(caudate['seqname'] == 'chrX')),
          np.sum((caudate['adj.P.Val']>=0.05) & (caudate['t']>0) &␣
 ↪(caudate['seqname'] != 'chrX'))]]
print(table)
fisher_exact(table)
```

```
[[52, 1247], [321, 9280]]
```

[17]:
```
(1.2055350286169673, 0.2223566468113563)
```

1.5 CMC DLPFC (SVA)

[18]:
```
annot = gene_annotation()
annot["genes"] = annot.gene_id.str.replace("\\..*", "", regex=True)
cmc = annot.merge(get_cmc(), on="genes")
```

[19]:
```
table = [[np.sum((cmc['adj.P.Val']<0.05) & (cmc['seqname'] == 'chrX')),
          np.sum((cmc['adj.P.Val']<0.05) & (cmc['seqname'] != 'chrX'))],
         [np.sum((cmc['adj.P.Val']>=0.05) & (cmc['seqname'] == 'chrX')),
          np.sum((cmc['adj.P.Val']>=0.05) & (cmc['seqname'] != 'chrX'))]]
print(table)
fisher_exact(table)
```

```
[[13, 401], [546, 15026]]
```

[19]:
```
(0.8921743260895381, 0.7869905606431027)
```

[20]:
```
table = [[np.sum((cmc['adj.P.Val']<0.05) & (cmc['t']<0) & (cmc['seqname'] ==␣
 ↪'chrX')),
```

```
          np.sum((cmc['adj.P.Val']<0.05) & (cmc['t']<0) & (cmc['seqname'] !=␣
 ↪'chrX'))],
         [np.sum((cmc['adj.P.Val']>=0.05) & (cmc['t']<0) & (cmc['seqname'] ==␣
 ↪'chrX')),
          np.sum((cmc['adj.P.Val']>=0.05) & (cmc['t']<0) & (cmc['seqname'] !=␣
 ↪'chrX'))]]
print(table)
fisher_exact(table)
```

```
[[4, 225], [284, 8140]]
```

[20]:
```
(0.5095461658841941, 0.2578853679103255)
```

[21]:
```
table = [[np.sum((cmc['adj.P.Val']<0.05) & (cmc['t']>0) & (cmc['seqname'] ==␣
 ↪'chrX')),
          np.sum((cmc['adj.P.Val']<0.05) & (cmc['t']>0) & (cmc['seqname'] !=␣
 ↪'chrX'))],
         [np.sum((cmc['adj.P.Val']>=0.05) & (cmc['t']>0) & (cmc['seqname'] ==␣
 ↪'chrX')),
          np.sum((cmc['adj.P.Val']>=0.05) & (cmc['t']>0) & (cmc['seqname'] !=␣
 ↪'chrX'))]]
print(table)
fisher_exact(table)
```

```
[[9, 176], [262, 6886]]
```

[21]:
```
(1.3439885496183206, 0.4247922107896491)
```

[22]:
```
setA = set(dlpfc[(dlpfc['adj.P.Val']<0.05) & (dlpfc['t']>0) & (dlpfc['seqname']␣
 ↪== 'chrX')].gene_name)
setA
```

[22]:
```
{'CHRDL1',
 'CXorf40B',
 'EFHC2',
 'FHL1',
 'GABRQ',
 'MAMLD1',
 'PABPC5-AS1',
 'SLC16A2'}
```

[23]:
```
setB = set(cmc[(cmc['adj.P.Val']<0.05) & (cmc['t']>0) & (cmc['seqname'] ==␣
 ↪'chrX')].MAPPED_genes)
setB
```

[23]:
```
{'.',
 'ARHGAP6',
 'CHRDL1',
```

```
 'CXorf57',
 'IL13RA2',
 'SLC16A2',
 'SPRY3',
 'SYTL5',
 'TENM1'}
```

[24]:
```
setA & setB
```

[24]:
```
{'CHRDL1', 'SLC16A2'}
```

[ ]:
```

```