main

September 20, 2021

1 Examine enrichment in psychiatric disorder TWAS and DEGs

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

1.1 Configuration

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

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",
    'cmc_file': '/ceph/projects/v3_phase3_paper/inputs/cmc/_m/
 ↪CMC_MSSM-Penn-Pitt_DLPFC_mRNA_IlluminaHiSeq2500'+\
    '_gene-adjustedSVA-differentialExpression-includeAncestry-DxSCZ-DE.tsv',
    'gandal_de_file': "/ceph/users/jbenja13/psychENCODE/expression_results/_m/
 ↪gandal2018_psychENCODE_DE_results.xlsx",
    'twas_asd_file': '/ceph/users/jbenja13/psychENCODE/_m/psychENCODE_twas_asd.
 ↪csv',
    'twas_sz_file': '/ceph/users/jbenja13/psychENCODE/_m/psychENCODE_twas_sz.
 ↪csv',
    'twas_bd_file': '/ceph/users/jbenja13/psychENCODE/_m/psychENCODE_twas_bd.
 ↪csv',
    'twas_bs_file': '/ceph/projects/v4_phase3_paper/analysis/twas/
 ↪public_twas_comp/_m/TWAS_gene_tissue_summary.csv'
}
```

1.2 Functions

1.2.1 Cached

[3]:
```
@functools.lru_cache()
def get_database():
    dataset = Dataset(name="hsapiens_gene_ensembl",
                      host="http://www.ensembl.org",
                      use_cache=True)
    db = dataset.query(attributes=["ensembl_gene_id",
                                   "external_gene_name",
                                   "entrezgene_id"],
                       use_attr_names=True).dropna(subset=['entrezgene_id'])
    return db


@functools.lru_cache()
def get_deg(fn):
    return pd.read_csv(fn, sep='\t')


@functools.lru_cache()
def get_sig_deg(fn):
    return get_deg(fn)[(get_deg(fn)["adj.P.Val"] < 0.05)]


@functools.lru_cache()
def get_gandal_deg():
    return pd.read_excel(config["gandal_de_file"], sheet_name="DGE")


@functools.lru_cache()
def get_eGenes(tissue):
    df = pd.read_csv("../../summary_table/_m/
 ↪BrainSeq_sexGenotypes_4features_3regions.txt.gz", sep='\t')
    df = df[(df["Type"] == "Gene") & (df["Tissue"] == tissue)]\
        .drop(["variant_id", "lfsr"], axis=1).drop_duplicates()
    df["ensemblID"] = df.gencodeID.str.replace("\\..*", "", regex=True)
    return df


@functools.lru_cache()
def get_bs_twas():
    return pd.read_csv(config["twas_bs_file"])


@functools.lru_cache()
def get_gandal_twas(fn):
```

```
    df = pd.read_csv(fn)
    return df[(df["TWAS.Bonferroni"] < 0.05)]
```

1.2.2 Simple

[4]:
```
def fet(a, b):
    """
    Calculates Fisher's Exact test (fet) with sets a and b in universe u.
    Inputs are sets.
    """
    u = set(get_database().ensembl_gene_id)
    a = set(a); b = set(b)
    yes_a = u.intersection(a)
    yes_b = u.intersection(b)
    no_a = u - a
    no_b = u - b
    m = [[len(yes_a.intersection(yes_b)), len(no_a.intersection(yes_b))],
         [len(yes_a.intersection(no_b)), len(no_a.intersection(no_b))]]
    return fisher_exact(m)
```

1.3 Load eGenes for sex-interacting analysis

[5]:
```
caudate = set(get_eGenes("Caudate").ensemblID)
dlpfc = set(get_eGenes("DLPFC").ensemblID)
hippocampus = set(get_eGenes("Hippocampus").ensemblID)
```

1.4 Differential expression

1.4.1 BrainSeq SZ case vs control

[6]:
```
bs_caudate_degs = set(get_sig_deg(config["caud8_file"]).ensemblID)
bs_dlpfc_degs = set(get_sig_deg(config["dlpfc_file"]).ensemblID)
bs_hippo_degs = set(get_sig_deg(config["hippo_file"]).ensemblID)
```

1.4.2 CommonMind SZ, DLPFC

[7]:
```
cmc_dlpfc_degs = set(get_sig_deg(config["cmc_file"]).genes)
```

1.4.3 PsychENCODE (Gandal)

[8]:
```
psy_sz = set(get_gandal_deg()[(get_gandal_deg()["SCZ.fdr"] < 0.05)].
 ↪ensembl_gene_id)
psy_asd = set(get_gandal_deg()[(get_gandal_deg()["ASD.fdr"] < 0.05)].
 ↪ensembl_gene_id)
psy_bd = set(get_gandal_deg()[(get_gandal_deg()["BD.fdr"] < 0.05)].
 ↪ensembl_gene_id)
```

1.5 TWAS

1.5.1 BrainSeq

[9]:
```
bs_caudate_twas = set(get_bs_twas()[(get_bs_twas()["Caudate_FDR"] < 0.05)].
↪Geneid)
bs_dlpfc_twas = set(get_bs_twas()[(get_bs_twas()["DLPFC_FDR"] < 0.05)].Geneid)
bs_hippo_twas = set(get_bs_twas()[(get_bs_twas()["HIPPO_FDR"] < 0.05)].Geneid)
```

1.5.2 PsychENCODE

[10]:
```
psy_asd_twas = set(get_gandal_twas(config["twas_asd_file"]).GeneID)
psy_sz_twas = set(get_gandal_twas(config["twas_sz_file"]).GeneID)
psy_bd_twas = set(get_gandal_twas(config["twas_bd_file"]).ID)
```

1.6 Enrichment analysis

[11]:
```
egenes_dict = {"Caudate": caudate, "DLPFC": dlpfc, "Hippocampus": hippocampus}
comp_dict = {"BS_Caudate_DEG": bs_caudate_degs, "BS_DLPFC_DEG": bs_dlpfc_degs,
             "BS_Hippocampus_DEG": bs_hippo_degs, "CMC_DLPFC_DEG":␣
↪cmc_dlpfc_degs,
             "PSY_SZ_DEG": psy_sz, "PSY_ASD_DEG": psy_asd, "PSY_BD_DEG": psy_bd,
             "BS_Caudate_TWAS": bs_caudate_twas, "BS_DLPFC_TWAS": bs_dlpfc_twas,
             "BS_Hippocampus_TWAS": bs_hippo_twas, "PSY_SZ_TWAS": psy_sz_twas,
             "PSY_ASD_TWAS": psy_asd_twas, "PSY_BD_TWAS": psy_bd_twas}
comp_list = ["BS_Caudate_DEG", "BS_DLPFC_DEG", "BS_Hippocampus_DEG",
             "CMC_DLPFC_DEG", "PSY_SZ_DEG", "PSY_ASD_DEG", "PSY_BD_DEG",
             "BS_Caudate_TWAS", "BS_DLPFC_TWAS", "BS_Hippocampus_TWAS",
             "PSY_SZ_TWAS", "PSY_ASD_TWAS", "PSY_BD_TWAS"]

or_lt = []; pval_lt = []; tissue_lt = []; comparison_lt = [];
for tissue in ["Caudate", "DLPFC", "Hippocampus"]:
    for comp in comp_list:
        oddratio, pvals = fet(egenes_dict[tissue], comp_dict[comp])
        or_lt.append(oddratio); pval_lt.append(pvals);
        tissue_lt.append(tissue); comparison_lt.append(comp)
fdr = multipletests(pval_lt, method='fdr_bh')[1]
dt = pd.DataFrame({"Tissue": tissue_lt, "Comparison": comparison_lt, "OR":␣
↪or_lt,
                   "P-value": pval_lt, "FDR": fdr})
```

[12]:
```
dt[(dt["FDR"] < 0.05)]
```

[12]:

	Tissue	Comparison	OR	P-value	FDR
0	Caudate	BS_Caudate_DEG	2.756170	1.392999e-25	5.432696e-24
1	Caudate	BS_DLPFC_DEG	2.349303	4.310847e-03	8.848580e-03
3	Caudate	CMC_DLPFC_DEG	1.812393	1.662972e-02	2.687730e-02
4	Caudate	PSY_SZ_DEG	2.063669	7.475636e-18	5.830996e-17

5	Caudate	PSY_ASD_DEG	2.491460	1.093601e-13	6.092922e-13
6	Caudate	PSY_BD_DEG	1.532679	1.004704e-02	1.781066e-02
7	Caudate	BS_Caudate_TWAS	2.500256	1.529816e-06	5.966284e-06
8	Caudate	BS_DLPFC_TWAS	2.251763	3.043822e-03	6.982886e-03
9	Caudate	BS_Hippocampus_TWAS	2.775863	2.302466e-03	5.986412e-03
11	Caudate	PSY_ASD_TWAS	10.722030	2.261958e-02	3.392937e-02
13	DLPFC	BS_Caudate_DEG	2.637220	2.073170e-21	4.042682e-20
14	DLPFC	BS_DLPFC_DEG	2.405059	5.530647e-03	1.078476e-02
17	DLPFC	PSY_SZ_DEG	2.131577	4.809498e-18	4.689261e-17
18	DLPFC	PSY_ASD_DEG	2.508374	6.512848e-13	3.175013e-12
19	DLPFC	PSY_BD_DEG	1.528323	1.292183e-02	2.191093e-02
20	DLPFC	BS_Caudate_TWAS	2.442884	1.185970e-05	4.204804e-05
21	DLPFC	BS_DLPFC_TWAS	2.326861	3.259251e-03	7.061710e-03
22	DLPFC	BS_Hippocampus_TWAS	3.062388	1.049313e-03	2.923086e-03
26	Hippocampus	BS_Caudate_DEG	2.566870	3.014768e-19	3.919198e-18
27	Hippocampus	BS_DLPFC_DEG	2.143821	1.722904e-02	2.687730e-02
30	Hippocampus	PSY_SZ_DEG	2.069228	7.035580e-16	4.573127e-15
31	Hippocampus	PSY_ASD_DEG	2.307929	3.482678e-10	1.509160e-09
32	Hippocampus	PSY_BD_DEG	1.612040	5.833417e-03	1.083349e-02
33	Hippocampus	BS_Caudate_TWAS	2.407614	2.612245e-05	8.489798e-05
34	Hippocampus	BS_DLPFC_TWAS	2.450188	2.526461e-03	6.158249e-03
35	Hippocampus	BS_Hippocampus_TWAS	3.223829	6.865169e-04	2.059551e-03

[13]:
```
dt.to_csv("clincial_phenotypes_enrichment_analysis_3brainRegions.tsv",␣
↪sep='\t', index=False)
```