main

July 14, 2021

1 Extract unique female specific SZ-associated genes

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

[2]:
```
environ['NUMEXPR_MAX_THREADS'] = '16'
```

[3]:
```
@functools.lru_cache()
def get_res_df(feature):
    return pd.read_csv('../../../../../interaction_model/hippocampus/_m/%s/
 ↪residualized_expression.tsv' %
                       feature, sep='\t').T


@functools.lru_cache()
def get_pheno_df():
    return pd.read_csv('/ceph/projects/v3_phase3_paper/inputs/phenotypes/_m/
 ↪hippo_phenotypes.csv',
                       index_col=0)


@functools.lru_cache()
def get_res_pheno_df(feature):
    return pd.merge(get_pheno_df(), get_res_df(feature), left_index=True,␣
 ↪right_index=True)


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

[4]:
```
def map_features(feature):
    return {"genes": "gene", 'transcripts': 'tx',
            'exons': 'exon', 'junctions': 'jxn'}[feature]


def gene_annot(feature):
    gtf_file = '/ceph/genome/human/gencode25/gtf.CHR/_m/gencode.v25.annotation.
 ↪gtf'
    gtf0 = get_gtf(gtf_file)
    gtf = gtf0[(gtf0["feature"] == feature)]
    return gtf[["gene_id", "gene_name", "transcript_id", "exon_id",
                "gene_type", "seqname", "start", "end", "strand"]]


def get_de(feature):
    f = pd.read_csv('../../../female_analysis/_m/%s/diffExpr_szVctl_full.txt' %␣
 ↪feature,
                    sep='\t', index_col=0)\
          .rename(columns={'gencodeGeneID': 'gencodeID'})
    m = pd.read_csv('../../../male_analysis/_m/%s/diffExpr_szVctl_full.txt' %␣
 ↪feature,
                    sep='\t', index_col=0)\
          .rename(columns={'gencodeGeneID': 'gencodeID'})
    return f, m


def get_unique(x, y, thres=0.05):
    return x.merge(pd.DataFrame(index = list(set(x[(x['adj.P.Val'] <= thres)].
 ↪index) -
                                             set(y[(y['adj.P.Val'] <= thres)].
 ↪index))),
                   left_index=True, right_index=True)


def subset_sz_male(feature):
    df = get_res_pheno_df(feature)
    ctl = df[(df['Dx'] == 'Control') & (df['Sex'] == 'M')].copy()
    sz = df[(df['Dx'] == 'Schizo') & (df['Sex'] == 'M')].copy()
    return ctl, sz


def add_pvals_adjustPval(feature, df):
    ctl, sz = subset_sz_male(feature)
    pval_df = []
    for gene_id in df.Feature:
        stat, pval = mannwhitneyu(ctl[gene_id], sz[gene_id])
        pval_df.append(pval)
```

```
    fdr_df = fdrcorrection(pval_df)
    return pd.concat([df.set_index('Feature'),
                      pd.DataFrame({'Male_Pval': pval_df, 'Male_FDR':␣
↪fdr_df[1]},
                                    index=df.Feature)], axis=1)
```

1.1 Genes

[5]:
```
gtf_annot = gene_annot('gene')
f, m = get_de('genes')
```

```
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']
```

[6]:
```
f['Feature'] = f.index
#genes = get_unique(get_unique(f, m), a)
genes = get_unique(f, m)
genes = pd.merge(gtf_annot[['gene_id', 'seqname']], genes, left_on='gene_id',
                 right_on='Feature', how='right').rename(columns={'seqname':␣
 ↪'Chrom'})
genes = genes[['Feature', 'gencodeID', 'Symbol', 'ensemblID',
               'Chrom', 'logFC', 't', 'adj.P.Val']].sort_values('adj.P.Val')
genes = add_pvals_adjustPval('genes', genes)
genes = genes[~(genes['Male_Pval'] <= 0.05)].sort_values('adj.P.Val').
 ↪reset_index() ## Stringents
genes['Type'] = 'gene'
genes.head(2)
```

[6]:
```
Empty DataFrame
Columns: [Feature, gencodeID, Symbol, ensemblID, Chrom, logFC, t, adj.P.Val,
Male_Pval, Male_FDR, Type]
Index: []
```

1.2 Transcripts

[7]:
```
gtf_annot = gene_annot('transcript')
```

[8]:
```
f, m = get_de('transcripts')
f['Feature'] = f.index
f['ensemblID'] = f.gene_id.str.replace('\\.\d+', '', regex=True)
#trans = get_unique(get_unique(f, m), a)
trans = get_unique(f, m)
trans = pd.merge(gtf_annot[['transcript_id', 'seqname']], trans,
                 left_on='transcript_id', right_on='Feature',
```

```
                how='right').rename(columns={'seqname': 'Chrom'}).
↪sort_values('adj.P.Val')
trans = trans[['Feature', 'gene_id', 'Symbol', 'ensemblID', 'Chrom',
               'logFC', 't', 'adj.P.Val']].rename(columns={'gene_id':␣
↪'gencodeID'})
trans = add_pvals_adjustPval('transcripts', trans)
trans = trans[~(trans['Male_Pval'] <= 0.05)].sort_values('adj.P.Val').
↪reset_index() ## Stringents
trans['Type'] = 'transcript'
trans.head(2)
```

[8]:
```
Empty DataFrame
Columns: [Feature, gencodeID, Symbol, ensemblID, Chrom, logFC, t, adj.P.Val,
Male_Pval, Male_FDR, Type]
Index: []
```

1.2.1 Exons

[9]:
```
gtf_annot = gene_annot('exon')
gtf_annot['ensemblID'] = gtf_annot.gene_id.str.replace('\\.\d+', '', regex=True)
```

[10]:
```
f, m = get_de('exons')
f['Feature'] = f.index
#exons = get_unique(get_unique(f, m), a)
exons = get_unique(f, m)
exons = pd.merge(gtf_annot[['ensemblID', 'seqname']], exons,
                 on='ensemblID', how='right').rename(columns={'seqname':␣
↪'Chrom'})
exons = exons[['Feature', 'gencodeID', 'Symbol', 'ensemblID',
               'Chrom', 'logFC', 't', 'adj.P.Val']].groupby('Feature')\
            .first().reset_index().sort_values('adj.P.Val')
exons = add_pvals_adjustPval('exons', exons)
exons = exons[~(exons['Male_Pval'] <= 0.05)].sort_values('adj.P.Val').
↪reset_index()
exons['Type'] = 'exon'
exons.head(2)
```

[10]:
```
Empty DataFrame
Columns: [Feature, gencodeID, Symbol, ensemblID, Chrom, logFC, t, adj.P.Val,
Male_Pval, Male_FDR, Type]
Index: []
```

1.2.2 Junctions

[11]:
```
f, m = get_de('junctions')
f['Feature'] = f.index
#juncs = get_unique(get_unique(f, m), a)
juncs = get_unique(f, m)
juncs = pd.merge(gtf_annot[['ensemblID', 'seqname']], juncs,
                 on='ensemblID', how='right').rename(columns={'seqname':␣
 ↪'Chrom'})
juncs = juncs[['Feature', 'gencodeID', 'Symbol', 'ensemblID', 'Chrom',
               'logFC', 't', 'adj.P.Val']].groupby('Feature')\
        .first().reset_index().sort_values('adj.P.Val')
juncs = add_pvals_adjustPval('junctions', juncs)
juncs = juncs[~(juncs['Male_Pval'] <= 0.05)].sort_values('adj.P.Val').
 ↪reset_index() ## Stringents
juncs['Type'] = 'junction'
juncs.head(2)
```

[11]:

	Feature	gencodeID	Symbol	ensemblID	Chrom	logFC	t	adj.P.Val	Male_Pval	Male_FDR	Type
0	chr1:26558938-26560725(+)	ENSG00000117676.13	RPS6KA1	ENSG00000117676	chr1	-0.692728	-5.571689	0.023666	0.372987	0.402706	junction
1	chr1:40303677-40303806(-)	ENSG00000049089.13	COL9A2	ENSG00000049089	chr1	-0.774649	-5.330986	0.034602	0.127582	0.318956	junction

1.3 DE summary

1.3.1 DE (feature)

[12]:
```
gg = len(set(genes['Feature']))
tt = len(set(trans['Feature']))
ee = len(set(exons['Feature']))
jj = len(set(juncs['Feature']))

print("\nGene:\t\t%d\nTranscript:\t%d\nExon:\t\t%d\nJunction:\t%d" % (gg, tt,␣
 ↪ee, jj))
```

```
Gene:           0
Transcript:     0
Exon:           0
Junction:       5
```

DE (EnsemblID)

[13]:
```
gg = len(set(genes['ensemblID']))
tt = len(set(trans['ensemblID']))
ee = len(set(exons['ensemblID']))
```

```
jj = len(set(juncs['ensemblID']))

print("\nGene:\t\t%d\nTranscript:\t%d\nExon:\t\t%d\nJunction:\t%d" % (gg, tt,␣
 ↪ee, jj))
```

```
Gene:           0
Transcript:     0
Exon:           0
Junction:       5
```

DE (Gene Symbol)

[14]:
```
gg = len(set(genes['Symbol']))
tt = len(set(trans['Symbol']))
ee = len(set(exons['Symbol']))
jj = len(set(juncs['Symbol']))

print("\nGene:\t\t%d\nTranscript:\t%d\nExon:\t\t%d\nJunction:\t%d" % (gg, tt,␣
 ↪ee, jj))
```

```
Gene:           0
Transcript:     0
Exon:           0
Junction:       5
```

1.3.2 Feature effect size summary

[15]:
```
feature_list = ['Genes', 'Transcript', 'Exons', 'Junctions']
feature_df = [genes, trans, exons, juncs]
for ii in range(4):
    ff = feature_df[ii]
    half = len(set(ff[(np.abs(ff['logFC']) >= 0.5)].Feature))
    one = len(set(ff[(np.abs(ff['logFC']) >= 1)].Feature))
    print("\nThere are %d unique %s with abs(log2FC) >= 0.5" % (half,␣
 ↪feature_list[ii]))
    print("There are %d unique %s with abs(log2FC) >= 1" % (one,␣
 ↪feature_list[ii]))
```

```
There are 0 unique Genes with abs(log2FC) >= 0.5
There are 0 unique Genes with abs(log2FC) >= 1

There are 0 unique Transcript with abs(log2FC) >= 0.5
There are 0 unique Transcript with abs(log2FC) >= 1

There are 0 unique Exons with abs(log2FC) >= 0.5
There are 0 unique Exons with abs(log2FC) >= 1
```

```
There are 4 unique Junctions with abs(log2FC) >= 0.5
There are 1 unique Junctions with abs(log2FC) >= 1
```

[16]:
```
feature_list = ['Genes', 'Transcripts', 'Exons', 'Junctions']
feature_df = [genes, trans, exons, juncs]
for ii in range(4):
    ff = feature_df[ii]
    half = len(set(ff[(np.abs(ff['logFC']) >= 0.5)].ensemblID))
    one = len(set(ff[(np.abs(ff['logFC']) >= 1)].ensemblID))
    print("\nThere are %d unique %s with abs(log2FC) >= 0.5" % (half,␣
 ↪feature_list[ii]))
    print("There are %d unique %s with abs(log2FC) >= 1" % (one,␣
 ↪feature_list[ii]))
```

```
There are 0 unique Genes with abs(log2FC) >= 0.5
There are 0 unique Genes with abs(log2FC) >= 1

There are 0 unique Transcripts with abs(log2FC) >= 0.5
There are 0 unique Transcripts with abs(log2FC) >= 1

There are 0 unique Exons with abs(log2FC) >= 0.5
There are 0 unique Exons with abs(log2FC) >= 1

There are 4 unique Junctions with abs(log2FC) >= 0.5
There are 1 unique Junctions with abs(log2FC) >= 1
```

[17]:
```
df = pd.concat([genes, trans, exons, juncs], axis=0)
df.to_csv('female_specific_DE_4features.txt', sep='\t', index=False,␣
 ↪header=True)
```

1.4 Number of DEGs on allosomes

[18]:
```
df[(df['Chrom'].isin(['chrX', 'chrY']))].groupby(['Type', 'Chrom']).size()
```

[18]:
```
Series([], dtype: int64)
```

[19]:
```
df
```

[19]:

	Feature	gencodeID	Symbol	ensemblID	\
0	chr1:26558938-26560725(+)	ENSG00000117676.13	RPS6KA1	ENSG00000117676	
1	chr1:40303677-40303806(-)	ENSG00000049089.13	COL9A2	ENSG00000049089	
2	chr1:236045684-236048689(-)	ENSG00000116962.14	NID1	ENSG00000116962	
3	chr5:132210524-132213915(-)	ENSG00000072682.18	P4HA2	ENSG00000072682	
4	chr1:245059459-245082082(+)	ENSG00000203666.12	EFCAB2	ENSG00000203666	

Chrom	logFC	t	adj.P.Val	Male_Pval	Male_FDR	Type

```
0  chr1 -0.692728 -5.571689   0.023666   0.372987  0.402706  junction
1  chr1 -0.774649 -5.330986   0.034602   0.127582  0.318956  junction
2  chr1  1.115432  5.134431   0.040765   0.241606  0.402677  junction
3  chr5  0.517820  5.163802   0.040765   0.066026  0.318956  junction
4  chr1 -0.420878 -5.047418   0.047387   0.402706  0.402706  junction
```

[ ]: