main

July 13, 2021

1 Feature summary analysis

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

1.1 Summary plots

1.1.1 MSSM Penn Pitt

[2]:
```
mpp = pd.read_csv('../../_m/mssm_penn_pitt_maleVfemale.tsv', sep='\t')
mpp = mpp[(mpp['adj.P.Val'] < 0.05)].sort_values('adj.P.Val')
mpp.head()
```

[2]:

	Geneid	logFC	AveExpr	t	P.Value	adj.P.Val	\
0	ENSG00000241859.7	7.426322	0.211930	106.905258	0.0	0.0	
1	ENSG00000206159.11	6.829826	-0.454650	101.687435	0.0	0.0	
2	ENSG00000183878.15	8.566599	2.821199	96.624132	0.0	0.0	
3	ENSG00000099725.14	7.234079	0.094070	96.796064	0.0	0.0	
4	ENSG00000215580.11	6.602361	-0.667862	95.449115	0.0	0.0	

	B	Coef	Symbol	Entrez	Chrom
0	794.024282	Reported_GenderMale	ANOS2P	NaN	Y
1	769.944857	Reported_GenderMale	GYG2P1	NaN	Y
2	747.659203	Reported_GenderMale	UTY	7404.0	Y
3	747.089270	Reported_GenderMale	PRKY	NaN	Y
4	739.995713	Reported_GenderMale	BCORP1	NaN	Y

1.1.2 NIMH HBCC

[3]:
```
hbcc = pd.read_csv('../../_m/nimh_hbcc_maleVfemale.tsv', sep='\t')
hbcc = hbcc[(hbcc['adj.P.Val'] < 0.05)].sort_values('adj.P.Val')
hbcc.head()
```

[3]:

	Geneid	logFC	AveExpr	t	P.Value	\
0	ENSG00000229807.11	-11.504527	1.451875	-158.907479	9.677013e-244	
1	ENSG00000241859.7	8.165803	0.195418	98.530405	4.690014e-195	
2	ENSG00000206159.11	7.594584	-0.185954	92.012611	3.727481e-188	
4	ENSG00000215580.11	6.980341	-0.607569	89.120159	6.068671e-185	

```
3  ENSG00000067646.12   9.872471  2.014148   88.829183  1.293023e-184

        adj.P.Val           B                 Coef  Symbol  Entrez Chrom
0  1.849374e-239  541.117577  Reported_GenderMale    XIST     NaN     X
1  4.481543e-191  425.604156  Reported_GenderMale  ANOS2P     NaN     Y
2  2.374530e-184  411.071479  Reported_GenderMale  GYG2P1     NaN     Y
4  2.899459e-181  404.120608  Reported_GenderMale  BCORP1     NaN     Y
3  4.942193e-181  404.158301  Reported_GenderMale     ZFY  7544.0     Y
```

1.2 DE summary

1.2.1 DE (feature)

[4]:
```
gg1 = len(set(mpp['Geneid']))
gg2 = len(set(hbcc['Geneid']))

print("Gene MPP:\t%d\nGene HBCC:\t%d" % (gg1, gg2))
```

```
Gene MPP:       482
Gene HBCC:      148
```

1.2.2 Feature effect size summary

[5]:
```
feature_list = ['Genes: MPP', 'Genes: HBCC', 'Exons', 'Junctions']
feature_df = [mpp, hbcc]
ii = 0

for ii in [0,1]:
    ff = feature_df[ii]
    half = len(set(ff[(np.abs(ff['logFC']) >= 0.5)].Geneid))
    one = len(set(ff[(np.abs(ff['logFC']) >= 1)].Geneid))
    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 36 unique Genes: MPP with abs(log2FC) >= 0.5
There are 25 unique Genes: MPP with abs(log2FC) >= 1

There are 40 unique Genes: HBCC with abs(log2FC) >= 0.5
There are 27 unique Genes: HBCC with abs(log2FC) >= 1
```

1.3 Autosomal only

[6]:
```
mpp.Chrom.fillna('?', inplace=True)
auto1 = mpp[(mpp.Chrom.str.contains('\d+'))].copy()\
        .rename(columns={'Geneid': 'gene_id'})
auto1 = auto1[['gene_id', 'Chrom', 'Symbol', 'logFC', 'adj.P.Val']]
print(auto1.shape)
auto1.head()
```

```
(418, 5)
```

[6]:

	gene_id	Chrom	Symbol	logFC	adj.P.Val
33	ENSG00000205611.4	20	LINC01597	0.977532	3.625963e-25
35	ENSG00000255346.10	15	NOX5	0.964489	8.867594e-22
38	ENSG00000283443.1	20	AC018688.1	0.879612	8.234590e-18
39	ENSG00000258484.4	15	SPESP1	0.701390	3.003091e-17
41	ENSG00000149531.15	20	FRG1BP	0.528819	6.561622e-13

[7]:
```
auto1.sort_values('adj.P.Val').to_csv('autosomal_DEG_mpp.csv', index=False,␣
↪header=True)
```

[8]:
```
hbcc.Chrom.fillna('?', inplace=True)
auto2 = hbcc[(hbcc.Chrom.str.contains('\d+'))].copy()\
        .rename(columns={'Geneid': 'gene_id'})
auto2 = auto2[['gene_id', 'Chrom', 'Symbol', 'logFC', 'adj.P.Val']]
print(auto2.shape)
auto2.head()
```

```
(98, 5)
```

[8]:

	gene_id	Chrom	Symbol	logFC	adj.P.Val
35	ENSG00000095932.6	19	SMIM24	-0.813149	9.014157e-15
37	ENSG00000149531.15	20	FRG1BP	0.739522	6.859297e-12
41	ENSG00000283443.1	20	AC018688.1	1.149979	2.357916e-09
42	ENSG00000205611.4	20	LINC01597	1.015293	8.239587e-09
43	ENSG00000258484.4	15	SPESP1	0.776058	1.221874e-08

[9]:
```
auto2.sort_values('adj.P.Val').to_csv('autosomal_DEG_hbcc.csv', index=False,␣
↪header=True)
```

1.4 DE summary

1.4.1 DE (feature)

[10]:
```
gg1 = len(set(auto1['gene_id']))
gg2 = len(set(auto2['gene_id']))

print("Gene MPP:\t%d\nGene HBCC:\t%d" % (gg1, gg2))
```

```
Gene MPP:       418
Gene HBCC:      98
```

1.4.2 Feature effect size summary

[11]:
```
feature_list = ['Genes: MPP', 'Genes: HBCC', 'Exons', 'Junctions']
feature_df = [auto1, auto2]
ii = 0

for ii in [0,1]:
    ff = feature_df[ii]
    half = len(set(ff[(np.abs(ff['logFC']) >= 0.5)].gene_id))
    one = len(set(ff[(np.abs(ff['logFC']) >= 1)].gene_id))
    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 8 unique Genes: MPP with abs(log2FC) >= 0.5
There are 0 unique Genes: MPP with abs(log2FC) >= 1

There are 12 unique Genes: HBCC with abs(log2FC) >= 0.5
There are 2 unique Genes: HBCC with abs(log2FC) >= 1
```

[ ]: