main

August 12, 2021

1 Feature summary analysis of schizophrenia differential expression for the caudate nucleus

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

1.1 Summary plots

1.1.1 Genes

[2]:
```
genes0 = pd.read_csv('../../_m/genes/diffExpr_szVctl_full.txt',
                     sep='\t', index_col=0)
genes0['Feature'] = genes0.index
genes0 = genes0[['Feature', 'gencodeID', 'ensemblID', 'Symbol', 'logFC',
                 'AveExpr', 't', 'P.Value', 'adj.P.Val', 'B']]
genes0['Type'] = 'Gene'
genes = genes0[(genes0['adj.P.Val'] < 0.05)].sort_values('adj.P.Val')
genes.head(2)
```

[2]:

	Feature	gencodeID	ensemblID	Symbol	logFC	AveExpr	t	P.Value	adj.P.Val	B	Type
ENSG00000248587.7	ENSG00000248587.7	ENSG00000248587.7	ENSG00000248587	GDNF-AS1	0.801502	1.657783	12.696887	6.044699e-31	1.387742e-26	58.250922	Gene
ENSG00000138944.7	ENSG00000138944.7	ENSG00000138944.7	ENSG00000138944	KIAA1644	0.563733	4.807890	12.073351	1.487513e-28	1.707516e-24	54.072890	Gene

1.1.2 Transcripts

[3]:
```
trans0 = pd.read_csv('../../_m/transcripts/diffExpr_szVctl_full.txt',
                     sep='\t', index_col=0)
trans0['Feature'] = trans0.index
trans0['ensemblID'] = trans0.gene_id.str.replace('\\.\d+', '', regex=True)
trans0 = trans0[['Feature', 'gene_id', 'ensemblID', 'gene_name', 'logFC',
```

```
                'AveExpr', 't', 'P.Value', 'adj.P.Val', 'B']]
trans0['Type'] = 'Transcript'
trans0.rename(columns={'gene_id': 'gencodeID', 'gene_name': 'Symbol'},␣
 ↪inplace=True)
trans = trans0[(trans0['adj.P.Val'] < 0.05)].sort_values('adj.P.Val')
trans.head(2)
```

[3]:

	Feature	gencodeID	ensemblID	Symbol	logFC	AveExpr	t	P.Value	adj.P.Val	B	Type
ENST00000381176.4	ENST00000381176.4	ENSG00000138944.7	ENSG00000138944	KIAA1644	0.540563	4.639461	12.232444	3.872832e-29	3.957686e-24	55.307961	Transcript
ENST00000637926.1	ENST00000637926.1	ENSG00000248587.7	ENSG00000248587	GDNF-AS1	0.936477	-0.650623	11.723729	3.229728e-27	1.650246e-22	47.309512	Transcript

1.1.3 Exons

[4]:
```
exons0 = pd.read_csv('../../_m/exons/diffExpr_szVctl_full.txt',
                     sep='\t', index_col=0)
exons0['Feature'] = exons0.index
exons0 = exons0[['Feature', 'gencodeID', 'ensemblID', 'Symbol', 'logFC',
                 'AveExpr', 't', 'P.Value', 'adj.P.Val', 'B']]
exons0['Type'] = 'Exon'
exons = exons0[(exons0['adj.P.Val'] < 0.05)].sort_values('adj.P.Val')
exons.head(2)
```

[4]:

	Feature	gencodeID	ensemblID	Symbol	logFC	AveExpr	t	P.Value	adj.P.Val	B	Type
e1138732	e1138732	ENSG00000138944.7	ENSG00000138944	KIAA1644	0.572350	3.242573	12.041351	1.715280e-28	6.079621e-23	53.731706	Exon
e326758	e326758	ENSG00000248587.7	ENSG00000248587	GDNF-AS1	1.172798	-2.699521	11.775570	1.742174e-27	3.087472e-22	44.496329	Exon

1.1.4 Junctions

[5]:
```
juncs0 = pd.read_csv('../../_m/junctions/diffExpr_szVctl_full.txt',
                     sep='\t', index_col=0)
juncs0['Feature'] = juncs0.index
juncs0 = juncs0[['Feature', 'newGeneID', 'ensemblID', 'newGeneSymbol', 'logFC',
                 'AveExpr', 't', 'P.Value', 'adj.P.Val', 'B']]
juncs0['Type'] = 'Junction'
```

```
juncs0.rename(columns={'newGeneID': 'gencodeID', 'newGeneSymbol': 'Symbol'},␣
 ↪inplace=True)
juncs = juncs0[(juncs0['adj.P.Val'] < 0.05)].sort_values('adj.P.Val')
juncs.head()
```

[5]:

	Feature	gencodeID	ensemblID	Symbol	logFC	AveExpr	t	P.Value	adj.P.Val	B	Type
chr6:144797991-144803035(+)	chr6:144797991-144803035(+)	ENSG00000152818.18	ENSG00000152818	UTRN	0.325750	3.108503	9.445835	3.741731e-19	5.753623e-14	32.626131	Junction
chr20:19696912-19698567(+)	chr20:19696912-19698567(+)	ENSG00000185052.11	ENSG00000185052	SLC24A3	0.381360	1.176929	8.875803	2.820088e-17	2.168210e-12	27.638405	Junction
chr20:19684337-19685099(+)	chr20:19684337-19685099(+)	ENSG00000185052.11	ENSG00000185052	SLC24A3	0.364192	1.525751	8.796880	5.063055e-17	2.595136e-12	27.389888	Junction
chr6:144583601-144678405(+)	chr6:144583601-144678405(+)	ENSG00000152818.18	ENSG00000152818	UTRN	0.452747	2.458287	8.518893	3.872394e-16	1.035700e-11	25.850103	Junction
chr6:144821019-144827347(+)	chr6:144821019-144827347(+)	ENSG00000152818.18	ENSG00000152818	UTRN	0.306557	3.345261	8.502598	4.357161e-16	1.035700e-11	25.860288	Junction

1.2 DE summary

1.2.1 DE (feature)

[6]:
```
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:           2701
Transcript:     1920
```

```
Exon:           22445
Junction:       7525
```

DE (EnsemblID)

[7]:
```
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:           2701
Transcript:     1609
Exon:           3783
Junction:       2126
```

DE (Gene Symbol)

[8]:
```
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:           2495
Transcript:     1608
Exon:           3527
Junction:       2169
```

1.2.2 Feature effect size summary

[9]:
```
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 33 unique Genes with abs(log2FC) >= 0.5
```

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

There are 167 unique Transcript with abs(log2FC) >= 0.5
There are 54 unique Transcript with abs(log2FC) >= 1

There are 100 unique Exons with abs(log2FC) >= 0.5
There are 5 unique Exons with abs(log2FC) >= 1

There are 79 unique Junctions with abs(log2FC) >= 0.5
There are 11 unique Junctions with abs(log2FC) >= 1
```

[10]:
```
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 33 unique Genes with abs(log2FC) >= 0.5
There are 0 unique Genes with abs(log2FC) >= 1

There are 158 unique Transcripts with abs(log2FC) >= 0.5
There are 53 unique Transcripts with abs(log2FC) >= 1

There are 46 unique Exons with abs(log2FC) >= 0.5
There are 2 unique Exons with abs(log2FC) >= 1

There are 34 unique Junctions with abs(log2FC) >= 0.5
There are 4 unique Junctions with abs(log2FC) >= 1
```

1.3 Save results

[11]:
```
df = pd.concat([genes0, trans0, exons0, juncs0], axis=0)
print(df.shape)
df.to_csv('BrainSeq_Phase3_Caudate_DifferentialExpression_DxSZ_all.txt.gz',
          sep='\t', index=False, header=True)
```

```
(633357, 11)
```

[ ]: