main

February 20, 2023

1 Feature summary of differential expression analysis

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

[2]:
```
def annotate_DE(feature):
    # Annotate DE results
    df = pd.read_csv(f'../../_m/{feature.lower()}s/diffExpr_maleVfemale_full.
 ↪txt',
                     sep='\t', index_col=0)\
           .rename(columns={"gene_id": "gencodeID", "gencodeGeneID":␣
 ↪"gencodeID",
                            "gene_name": "Symbol"})
    df = df[(df['adj.P.Val'] < 0.05)].sort_values('adj.P.Val')
    df['Feature'] = df.index
    df['ensemblID'] = df.gencodeID.str.replace("\\..*", "", regex=True)
    df['Type'] = feature; df["Region"] = "Caudate"
    return df[['Feature', 'Symbol', 'ensemblID',
               'logFC', 'SE', 'adj.P.Val', "Type"]]
```

1.1 Summary plots

1.1.1 Genes

[3]:
```
genes = annotate_DE("Gene")
genes.head(2)
```

[3]:

	Feature	Symbol	ensemblID	logFC	SE	adj.P.Val	Type
USP9Y|ENSG00000114374.13	USP9Y|ENSG00000114374.13	USP9Y	ENSG00000114374	13.682636	0.034572	0.000000e+00	Gene
TXLNGY|ENSG00000131002.14	TXLNGY|ENSG00000131002.14	TXLNGY	ENSG00000131002	11.725115	0.020161	7.767527e-304	Gene

1.1.2 Transcripts

[4]:
```
trans = annotate_DE("Transcript")
trans.head(2)
```

[4]:
```
                                                Feature Symbol  \
XIST-204|ENST00000429829.6    XIST-204|ENST00000429829.6   XIST
USP9Y-204|ENST00000440408.5  USP9Y-204|ENST00000440408.5  USP9Y

                                    ensemblID     logFC        SE  \
XIST-204|ENST00000429829.6   ENSG00000229807 -9.473817  0.057125
USP9Y-204|ENST00000440408.5  ENSG00000114374  4.701063  0.143236

                                   adj.P.Val        Type
XIST-204|ENST00000429829.6   5.256559e-260  Transcript
USP9Y-204|ENST00000440408.5  3.694803e-242  Transcript
```

1.1.3 Exons

[5]:
```
exons = annotate_DE("Exon")
exons.head(2)
```

[5]:
```
                                        Feature  Symbol        ensemblID  \
chrY:19588370-19590419+  chrY:19588370-19590419+  TXLNGY  ENSG00000131002
chrY:12912640-12913062+  chrY:12912640-12913062+   DDX3Y  ENSG00000067048

                              logFC        SE      adj.P.Val  Type
chrY:19588370-19590419+  10.531845  0.096627  1.016898e-289  Exon
chrY:12912640-12913062+  10.274268  0.122691  8.044671e-289  Exon
```

1.1.4 Junctions

[6]:
```
juncs = annotate_DE("Junction")
juncs.head(2)
```

[6]:
```
                                          Feature   Symbol        ensemblID  \
chrY:12912883-12912962:+  chrY:12912883-12912962:+  DDX11L1  ENSG00000223972
chrY:2854772-2865087:+      chrY:2854772-2865087:+  DDX11L1  ENSG00000223972

                              logFC        SE      adj.P.Val      Type
chrY:12912883-12912962:+  8.505714  0.119848  5.331736e-201  Junction
chrY:2854772-2865087:+    9.702493  0.095575  5.643537e-194  Junction
```

1.2 DE summary

1.2.1 DE (feature)

[7]:
```
gg = len(set(genes['Feature']))
tt = len(set(trans['Feature']))
ee = len(set(exons['Feature']))
jj = len(set(juncs['Feature']))

print(f"\nGene:\t\t{gg}\nTranscript:\t{tt}\nExon:\t\t{ee}\nJunction:\t{jj}")
```

```
Gene:           689
Transcript:     587
Exon:           2814
Junction:       1024
```

DE (EnsemblID)

[8]:
```
gg = len(set(genes['ensemblID']))
tt = len(set(trans['ensemblID']))
ee = len(set(exons['ensemblID']))
jj = len(set(juncs['ensemblID']))

print(f"\nGene:\t\t{gg}\nTranscript:\t{tt}\nExon:\t\t{ee}\nJunction:\t{jj}")
```

```
Gene:           689
Transcript:     326
Exon:           526
Junction:       12
```

DE (Gene Symbol)

[9]:
```
gg = len(set(genes['Symbol']))
tt = len(set(trans['Symbol']))
ee = len(set(exons['Symbol']))
jj = len(set(juncs['Symbol']))

print(f"\nGene:\t\t{gg}\nTranscript:\t{tt}\nExon:\t\t{ee}\nJunction:\t{jj}")
```

```
Gene:           689
Transcript:     326
Exon:           530
Junction:       14
```

1.2.2 Feature effect size summary

[10]:
```
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(f"\nThere are {half} unique {feature_list[ii]} with abs(log2FC) >= 0.
 ↪5")
    print(f"There are {one} unique {feature_list[ii]} with abs(log2FC) >= 1")
```

```
There are 127 unique Genes with abs(log2FC) >= 0.5
There are 60 unique Genes with abs(log2FC) >= 1

There are 335 unique Transcript with abs(log2FC) >= 0.5
There are 214 unique Transcript with abs(log2FC) >= 1

There are 1169 unique Exons with abs(log2FC) >= 0.5
There are 673 unique Exons with abs(log2FC) >= 1

There are 454 unique Junctions with abs(log2FC) >= 0.5
There are 288 unique Junctions with abs(log2FC) >= 1
```

[11]:
```
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(f"\nThere are {half} unique {feature_list[ii]} with abs(log2FC) >= 0.
 ↪5")
    print(f"There are {one} unique {feature_list[ii]} with abs(log2FC) >= 1")
```

```
There are 127 unique Genes with abs(log2FC) >= 0.5
There are 60 unique Genes with abs(log2FC) >= 1

There are 140 unique Transcripts with abs(log2FC) >= 0.5
There are 81 unique Transcripts with abs(log2FC) >= 1

There are 132 unique Exons with abs(log2FC) >= 0.5
There are 63 unique Exons with abs(log2FC) >= 1

There are 7 unique Junctions with abs(log2FC) >= 0.5
There are 7 unique Junctions with abs(log2FC) >= 1
```

1.3 Autosomal only

[12]:
```
from pyhere import here
from functools import lru_cache
```

[13]:
```
@lru_cache()
def get_annotation(feature):
    feat_lt = {"gene": "gene", "transcript": "tx",
               "exon": "exon", "junction": "jxn"}
    new_feature = feat_lt[feature]
    fn = here(f"input/counts/text_files_counts/_m/caudate/
 ↪{new_feature}_annotation.txt")
    return pd.read_csv(fn, sep='\t')
```

[14]:
```
def annotate_autosomes(feature):
    # Get annotation
    annot = get_annotation(feature.lower())
    # Annotate DE results
    df = pd.read_csv(f'../../_m/{feature.lower()}s/diffExpr_maleVfemale_full.
 ↪txt',
                     sep='\t', index_col=0)\
           .rename(columns={"gene_id": "gencodeID", "gencodeGeneID":␣
 ↪"gencodeID",
                            "gene_name": "Symbol"})
    df = df[(df['adj.P.Val'] < 0.05)].sort_values('adj.P.Val')
    df['name'] = df.index
    df['ensemblID'] = df.gencodeID.str.replace("\\..*", "", regex=True)
    df = annot.merge(df, on='name').rename(columns={"name": "Feature"})
    df = df[["Feature", "seqnames", "start", "end", "width", "gencodeID",␣
 ↪"ensemblID",
             "Symbol", "logFC", "AveExpr", "t", "P.Value", "adj.P.Val", "B",␣
 ↪"SE"]]
    df['Type'] = feature; df["Region"] = "Caudate"
    # Save annotated file
    df.sort_values('adj.P.Val').to_csv(f'chrom_annotation_{feature.lower()}.
 ↪txt',
                                       sep='\t', index=False)
    df = df[(df.seqnames.str.contains('chr\d+'))].copy()
    # Save autosomal DE features
    df.to_csv(f'{feature.lower()}_autosomal_DE.csv', index=False, header=True)
    return df[['Feature', 'seqnames', 'Symbol', 'ensemblID', 'logFC', 'SE',␣
 ↪'adj.P.Val', "Type"]]
```

1.3.1 Genes

[15]:
```
feature = "Gene"
genes = annotate_autosomes(feature)
genes.head(2)
```

[15]:
```
                     Feature seqnames   Symbol        ensemblID     logFC  \
1  CYP26B1|ENSG00000003137.9     chr2  CYP26B1  ENSG00000003137 -0.255704
2   SKAP2|ENSG00000005020.13     chr7    SKAP2  ENSG00000005020  0.155882

         SE  adj.P.Val  Type
1  0.017475    0.00004  Gene
2  0.054142    0.00343  Gene
```

[16]:
```
genes.shape
```

[16]:
```
(576, 8)
```

[17]:
```
genes.groupby('ensemblID').first().reset_index().shape
```

[17]:
```
(576, 8)
```

1.3.2 Transcripts

[18]:
```
trans = annotate_autosomes("Transcript")
trans.head(2)
trans.shape
```

[18]:
```
(252, 8)
```

[19]:
```
trans.groupby('ensemblID').first().reset_index().shape
```

[19]:
```
(232, 8)
```

1.3.3 Exons

[20]:
```
exons = annotate_autosomes("Exon")
exons.head(2)
exons.shape
```

[20]:
```
(1291, 8)
```

[21]:
```
exons.groupby('ensemblID').first().reset_index().shape
```

[21]:
```
(417, 8)
```

1.3.4 Junctions

[22]:
```
juncs = annotate_autosomes("Junction")
juncs.head(2)
juncs.shape
```

[22]:
```
(399, 8)
```

[23]:
```
juncs.groupby('ensemblID').first().reset_index().shape
```

[23]:
```
(4, 8)
```

1.4 DE summary

1.4.1 DE (feature)

[24]:
```
gg = len(set(genes['Feature']))
tt = len(set(trans['Feature']))
ee = len(set(exons['Feature']))
jj = len(set(juncs['Feature']))

print(f"\nGene:\t\t{gg}\nTranscript:\t{tt}\nExon:\t\t{ee}\nJunction:\t{jj}")
```

```
Gene:           576
Transcript:     252
Exon:           1291
Junction:       399
```

DE (EnsemblID)

[25]:
```
gg = len(set(genes.groupby('ensemblID').first().reset_index()['ensemblID']))
tt = len(set(trans.groupby('ensemblID').first().reset_index()['ensemblID']))
ee = len(set(exons.groupby('ensemblID').first().reset_index()['ensemblID']))
jj = len(set(juncs.groupby('ensemblID').first().reset_index()['ensemblID']))

print(f"\nGene:\t\t{gg}\nTranscript:\t{tt}\nExon:\t\t{ee}\nJunction:\t{jj}")
```

```
Gene:           576
Transcript:     232
Exon:           417
Junction:       4
```

DE (Gene Symbol)

[26]:
```
gg = len(set(genes.groupby('Symbol').first().reset_index()['Symbol']))
tt = len(set(trans.groupby('Symbol').first().reset_index()['Symbol']))
ee = len(set(exons.groupby('Symbol').first().reset_index()['Symbol']))
```

```
jj = len(set(juncs.groupby('Symbol').first().reset_index()['Symbol']))

print(f"\nGene:\t\t{gg}\nTranscript:\t{tt}\nExon:\t\t{ee}\nJunction:\t{jj}")
```

```
Gene:           576
Transcript:     232
Exon:           419
Junction:       4
```

1.4.2 Feature effect size summary

[27]:
```
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(f"\nThere are {half} unique {feature_list[ii]} with abs(log2FC) >= 0.
 ↪5")
    print(f"There are {one} unique {feature_list[ii]} with abs(log2FC) >= 1")
```

```
There are 70 unique Genes with abs(log2FC) >= 0.5
There are 16 unique Genes with abs(log2FC) >= 1

There are 91 unique Transcript with abs(log2FC) >= 0.5
There are 36 unique Transcript with abs(log2FC) >= 1

There are 209 unique Exons with abs(log2FC) >= 0.5
There are 34 unique Exons with abs(log2FC) >= 1

There are 104 unique Junctions with abs(log2FC) >= 0.5
There are 43 unique Junctions with abs(log2FC) >= 1
```

[28]:
```
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(f"\nThere are {half} unique {feature_list[ii]} with abs(log2FC) >= 0.
 ↪5")
    print(f"There are {one} unique {feature_list[ii]} with abs(log2FC) >= 1")
```

```
There are 70 unique Genes with abs(log2FC) >= 0.5
There are 16 unique Genes with abs(log2FC) >= 1
```

```
There are 82 unique Transcripts with abs(log2FC) >= 0.5
There are 34 unique Transcripts with abs(log2FC) >= 1

There are 61 unique Exons with abs(log2FC) >= 0.5
There are 14 unique Exons with abs(log2FC) >= 1

There are 3 unique Junctions with abs(log2FC) >= 0.5
There are 3 unique Junctions with abs(log2FC) >= 1
```

1.5 Session information

[29]:
```
import session_info
session_info.show()
```

[29]:
```
<IPython.core.display.HTML object>
```