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"] = "DLPFC"
         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  \
     BCORP1|ENSG00000215580.12                      BCORP1|ENSG00000215580.12
     ENSG00000260197|ENSG00000260197.1   ENSG00000260197|ENSG00000260197.1


                                           Symbol          ensemblID      logFC  \
     BCORP1|ENSG00000215580.12              BCORP1   ENSG00000215580   8.446420
     ENSG00000260197|ENSG00000260197.1   ENSG00000260197   ENSG00000260197   8.269605


                                            SE       adj.P.Val  Type
     BCORP1|ENSG00000215580.12        0.057509   1.716612e-207  Gene
```

```
ENSG00000260197|ENSG00000260197.1   0.025685   5.824188e-203   Gene
```

1.1.2 Transcripts

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

```
[4]:                                                    Feature  Symbol  \
     USP9Y-204|ENST00000440408.5   USP9Y-204|ENST00000440408.5   USP9Y
     XIST-208|ENST00000602495.1    XIST-208|ENST00000602495.1    XIST

                                            ensemblID     logFC        SE  \
     USP9Y-204|ENST00000440408.5   ENSG00000114374   5.106249   0.068411
     XIST-208|ENST00000602495.1    ENSG00000229807  -7.901838   0.059550

                                         adj.P.Val        Type
     USP9Y-204|ENST00000440408.5   4.806531e-272   Transcript
     XIST-208|ENST00000602495.1    1.829393e-211   Transcript
```

1.1.3 Exons

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

```
[5]:                                    Feature    Symbol       ensemblID  \
     chr9:88016300-88016793+   chr9:88016300-88016793+   RPS10P3   ENSG00000217716
     chr5:56504635-56505072-   chr5:56504635-56505072-   RPL26P19  ENSG00000226221

                                   logFC        SE     adj.P.Val  Type
     chr9:88016300-88016793+  -4.657763   0.108053   1.743948e-148  Exon
     chr5:56504635-56505072-  -3.413456   0.114270   7.084867e-130  Exon
```

1.1.4 Junctions

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

```
[6]:                                    Feature    Symbol       ensemblID  \
     chrY:2845744-2854599:+   chrY:2845744-2854599:+   RPS4Y1,XGY2  ENSG00000129824
     chrY:2854772-2865087:+   chrY:2854772-2865087:+   DDX11L1  ENSG00000223972

                                   logFC        SE     adj.P.Val        Type
     chrY:2845744-2854599:+   9.207380   0.113165   7.371018e-178  Junction
     chrY:2854772-2865087:+   9.240353   0.113205   7.371018e-178  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:           256
Transcript:     605
Exon:           1628
Junction:       2030
```

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:           256
Transcript:     377
Exon:           375
Junction:       19
```

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:           256
Transcript:     377
Exon:           379
Junction:       22
```

1.2.2 Feature effect size summary

```python
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")
```

[10]:

```
There are 107 unique Genes with abs(log2FC) >= 0.5
There are 51 unique Genes with abs(log2FC) >= 1

There are 340 unique Transcript with abs(log2FC) >= 0.5
There are 216 unique Transcript with abs(log2FC) >= 1

There are 992 unique Exons with abs(log2FC) >= 0.5
There are 628 unique Exons with abs(log2FC) >= 1

There are 625 unique Junctions with abs(log2FC) >= 0.5
There are 347 unique Junctions with abs(log2FC) >= 1
```

[11]:
```python
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 107 unique Genes with abs(log2FC) >= 0.5
There are 51 unique Genes with abs(log2FC) >= 1

There are 166 unique Transcripts with abs(log2FC) >= 0.5
There are 88 unique Transcripts with abs(log2FC) >= 1

There are 130 unique Exons with abs(log2FC) >= 0.5
There are 64 unique Exons with abs(log2FC) >= 1

There are 8 unique Junctions with abs(log2FC) >= 0.5
There are 8 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/dlpfc/
       ↪{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"] = "DLPFC"
          # 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  \
      4   HEBP1|ENSG00000013583.10    chr12  HEBP1  ENSG00000013583 -0.144560
      5  NLRP2|ENSG00000022556.16    chr19  NLRP2  ENSG00000022556 -0.695142

                SE  adj.P.Val  Type
      4   0.021626   0.002068  Gene
      5   0.048143   0.002549  Gene
```

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

```
[16]: (186, 8)
```

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

```
[17]: (186, 8)
```

1.3.2 Transcripts

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

```
[18]: (307, 8)
```

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

```
[19]: (291, 8)
```

1.3.3 Exons

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

```
[20]: (903, 8)
```

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

```
[21]: (309, 8)
```

6

1.3.4 Junctions

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

```
[22]: (1421, 8)
```

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

```
[23]: (12, 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:           186
Transcript:     307
Exon:           903
Junction:       1421
```

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:           186
Transcript:     291
Exon:           309
Junction:       12
```

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']))
```

```python
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:           186
Transcript:     291
Exon:           310
Junction:       13
```

1.4.2 Feature effect size summary

```python
[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 58 unique Genes with abs(log2FC) >= 0.5
There are 14 unique Genes with abs(log2FC) >= 1

There are 114 unique Transcript with abs(log2FC) >= 0.5
There are 48 unique Transcript with abs(log2FC) >= 1

There are 279 unique Exons with abs(log2FC) >= 0.5
There are 31 unique Exons with abs(log2FC) >= 1

There are 302 unique Junctions with abs(log2FC) >= 0.5
There are 125 unique Junctions with abs(log2FC) >= 1
```

```python
[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 58 unique Genes with abs(log2FC) >= 0.5
There are 14 unique Genes with abs(log2FC) >= 1
```

```
There are 110 unique Transcripts with abs(log2FC) >= 0.5
There are 48 unique Transcripts with abs(log2FC) >= 1

There are 65 unique Exons with abs(log2FC) >= 0.5
There are 17 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

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

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