main

August 4, 2021

1 Feature summary analysis of gender only differential expression

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

1.1 Summary plots

```python
[2]: def subset_deg(fn, feature):
         feature_dict = {
             'genes': 'Gene',
             'transcripts': 'Transcript',
             'exons': 'Exon',
             'junctions': 'Junction'}
         df = pd.read_csv(fn, sep='\t', index_col=0)
         if feature == 'transcripts':
             df['Feature'] = df.index
             df['ensemblID'] = df.gene_id.str.replace('\\.\d+', '', regex=True)
             df.rename(columns={'gene_name': 'Symbol'}, inplace=True)
         elif feature == 'genes':
             df['Feature'] = df.gencodeID
         else:
             df['Feature'] = df.index
         df = df[['Feature', 'Symbol', 'ensemblID', 'logFC', 'adj.P.Val']].copy()
         df['Type'] = feature_dict[feature]
         return df


     def extract_deg(feature):
         config = {
             'CvD': '../../_m/%s/diffExpr_CvD_sex_FDR05.txt' % feature,
             'CvH': '../../_m/%s/diffExpr_CvH_sex_FDR05.txt' % feature,
             'DvH': '../../_m/%s/diffExpr_DvH_sex_FDR05.txt' % feature,
         }
         cvd = subset_deg(config['CvD'], feature)
         cvd['Comparison'] = 'CvD'
         cvh = subset_deg(config['CvH'], feature)
         cvh['Comparison'] = 'CvH'
```

```
    dvh = subset_deg(config['DvH'], feature)
    dvh['Comparison'] = 'DvH'
    return cvd, cvh, dvh
```

1.1.1 Genes

```
[3]: feature = 'genes'
     genes_cvd, genes_cvh, genes_dvh = extract_deg(feature)
     genes = pd.concat([genes_cvd, genes_cvh, genes_dvh])
     genes.head(2)
```

```
[3]:                           Feature  Symbol        ensemblID      logFC  \
     ENSG00000184991.2    ENSG00000184991.2  TTTY13  ENSG00000184991 -4.360657
     ENSG00000092377.13  ENSG00000092377.13   TBL1Y  ENSG00000092377 -3.627633

                            adj.P.Val  Type Comparison
     ENSG00000184991.2    5.005190e-133  Gene        CvD
     ENSG00000092377.13   6.609422e-76  Gene        CvD
```

```
[4]: shared = set(genes_cvd.Feature) & set(genes_cvh.Feature)
     print("There are %d genes that interact for caudate between DLPFC and␣
      ↪Hippocampus!" % len(shared))
     genes_cvd.set_index('Feature').loc[np.array(list(shared)), :]
```

There are 17 genes that interact for caudate between DLPFC and Hippocampus!

```
[4]:                       Symbol        ensemblID      logFC       adj.P.Val  Type  \
     Feature
     ENSG00000121743.3       GJA3  ENSG00000121743 -0.518816   3.419185e-02  Gene
     ENSG00000099715.14    PCDH11Y  ENSG00000099715  2.519740   3.588357e-32  Gene
     ENSG00000252766.1   RNU6-255P  ENSG00000252766  1.353570   1.141351e-13  Gene
     ENSG00000197181.11     PIWIL2  ENSG00000197181 -0.393875   9.253554e-03  Gene
     ENSG00000277438.1         NaN  ENSG00000277438  0.919121   2.146899e-07  Gene
     ENSG00000273906.1         NaN  ENSG00000273906 -3.585686   9.649875e-67  Gene
     ENSG00000173809.16     TDRD12  ENSG00000173809 -0.462759   1.872604e-03  Gene
     ENSG00000228223.2       HCG11  ENSG00000228223  0.131063   4.230495e-02  Gene
     ENSG00000146674.14     IGFBP3  ENSG00000146674  0.949975   1.836787e-12  Gene
     ENSG00000185275.6       CD24P4  ENSG00000185275  1.269074   1.052913e-06  Gene
     ENSG00000238241.1       CCR12P  ENSG00000238241 -0.588489   7.882787e-03  Gene
     ENSG00000232348.2     LINC00279  ENSG00000232348 -1.334907   5.750720e-13  Gene
     ENSG00000170011.13      MYRIP  ENSG00000170011  0.238429   3.461297e-03  Gene
     ENSG00000231436.1      RBMY3AP  ENSG00000231436 -3.634175   1.955784e-50  Gene
     ENSG00000092377.13      TBL1Y  ENSG00000092377 -3.627633   6.609422e-76  Gene
     ENSG00000066629.16       EML1  ENSG00000066629 -0.222427   1.412821e-04  Gene
     ENSG00000184991.2       TTTY13  ENSG00000184991 -4.360657   5.005190e-133  Gene

                        Comparison
     Feature
```

```
ENSG00000121743.3        CvD
ENSG00000099715.14       CvD
ENSG00000252766.1        CvD
ENSG00000197181.11       CvD
ENSG00000277438.1        CvD
ENSG00000273906.1        CvD
ENSG00000173809.16       CvD
ENSG00000228223.2        CvD
ENSG00000146674.14       CvD
ENSG00000185275.6        CvD
ENSG00000238241.1        CvD
ENSG00000232348.2        CvD
ENSG00000170011.13       CvD
ENSG00000231436.1        CvD
ENSG00000092377.13       CvD
ENSG00000066629.16       CvD
ENSG00000184991.2        CvD
```

```python
[5]: shared = set(genes_cvd.Feature) & set(genes_dvh.Feature)
     print("There are %d genes that interact for DLPFC between caudate and␣
      ↪Hippocampus!" % len(shared))
     genes_cvd.set_index('Feature').loc[np.array(list(shared)), :]
```

There are 20 genes that interact for DLPFC between caudate and Hippocampus!

```
[5]:                      Symbol      ensemblID     logFC    adj.P.Val Type  \
     Feature
     ENSG00000239893.1   ZNF736P9Y  ENSG00000239893 -3.128749  3.952339e-33 Gene
     ENSG00000176728.7      TTTY14  ENSG00000176728  0.559170  3.879491e-02 Gene
     ENSG00000183878.15        UTY  ENSG00000183878  1.156821  1.252025e-06 Gene
     ENSG00000227494.2     USP9YP14  ENSG00000227494  1.398617  3.384241e-12 Gene
     ENSG00000185275.6       CD24P4  ENSG00000185275  1.269074  1.052913e-06 Gene
     ENSG00000205664.10        NaN  ENSG00000205664  0.366608  6.489451e-09 Gene
     ENSG00000067048.16       DDX3Y  ENSG00000067048  0.976405  3.043073e-05 Gene
     ENSG00000131002.11      TXLNGY  ENSG00000131002  0.843662  3.393715e-03 Gene
     ENSG00000092377.13       TBL1Y  ENSG00000092377 -3.627633  6.609422e-76 Gene
     ENSG00000273906.1         NaN  ENSG00000273906 -3.585686  9.649875e-67 Gene
     ENSG00000233864.7       TTTY15  ENSG00000233864  0.933324  4.252748e-05 Gene
     ENSG00000099725.14        PRKY  ENSG00000099725  0.912710  1.195206e-04 Gene
     ENSG00000012817.15       KDM5D  ENSG00000012817  1.018661  4.159236e-06 Gene
     ENSG00000114374.12       USP9Y  ENSG00000114374  1.253950  3.318947e-10 Gene
     ENSG00000129824.15      RPS4Y1  ENSG00000129824  0.809539  3.393715e-03 Gene
     ENSG00000006757.11      PNPLA4  ENSG00000006757  0.316647  2.421224e-14 Gene
     ENSG00000234449.2         NaN  ENSG00000234449  0.393613  1.380234e-03 Gene
     ENSG00000260197.1         NaN  ENSG00000260197  0.604253  6.546653e-03 Gene
     ENSG00000151923.17       TIAL1  ENSG00000151923 -0.090797  3.656674e-02 Gene
     ENSG00000229807.10        XIST  ENSG00000229807 -0.956805  1.145939e-04 Gene
```

```
                    Comparison
Feature
ENSG00000239893.1          CvD
ENSG00000176728.7          CvD
ENSG00000183878.15         CvD
ENSG00000227494.2          CvD
ENSG00000185275.6          CvD
ENSG00000205664.10         CvD
ENSG00000067048.16         CvD
ENSG00000131002.11         CvD
ENSG00000092377.13         CvD
ENSG00000273906.1          CvD
ENSG00000233864.7          CvD
ENSG00000099725.14         CvD
ENSG00000012817.15         CvD
ENSG00000114374.12         CvD
ENSG00000129824.15         CvD
ENSG00000006757.11         CvD
ENSG00000234449.2          CvD
ENSG00000260197.1          CvD
ENSG00000151923.17         CvD
ENSG00000229807.10         CvD
```

```python
[6]: shared = set(genes_cvh.Feature) & set(genes_dvh.Feature)
     print("There are %d genes that interact for Hippocampus between caudate and␣
      ↪DLPFC!" % len(shared))
     genes_cvh.set_index('Feature').loc[np.array(list(shared)), :]
```

There are 3 genes that interact for Hippocampus between caudate and DLPFC!

```
[6]:                  Symbol       ensemblID      logFC    adj.P.Val  Type  \
     Feature
     ENSG00000273906.1    NaN  ENSG00000273906 -2.789483  4.378328e-45  Gene
     ENSG00000092377.13  TBL1Y  ENSG00000092377 -2.825989  6.275529e-52  Gene
     ENSG00000185275.6  CD24P4  ENSG00000185275 -0.837700  6.893577e-04  Gene


                    Comparison
     Feature
     ENSG00000273906.1          CvH
     ENSG00000092377.13         CvH
     ENSG00000185275.6          CvH
```

1.1.2 Transcripts

```
[7]: feature = 'transcripts'
     trans_cvd, trans_cvh, trans_dvh = extract_deg(feature)
     trans = pd.concat([trans_cvd, trans_cvh, trans_dvh])
     trans.head(2)
```

```
[7]:                              Feature   Symbol       ensemblID    logFC  \
     ENST00000418290.1  ENST00000418290.1    CDY4P  ENSG00000228411 -5.77571
     ENST00000400275.2  ENST00000400275.2  PSMA6P1  ENSG00000215414 -5.64374

                            adj.P.Val        Type Comparison
     ENST00000418290.1  6.693728e-291  Transcript        CvD
     ENST00000400275.2  1.051031e-288  Transcript        CvD
```

1.1.3 Exons

```
[8]: feature = 'exons'
     exons_cvd, exons_cvh, exons_dvh = extract_deg(feature)
     exons = pd.concat([exons_cvd, exons_cvh, exons_dvh])
     exons.head(2)
```

```
[8]:            Feature  Symbol       ensemblID     logFC      adj.P.Val  Type  \
     e1181198  e1181198  TTTY13  ENSG00000184991 -4.158493  2.766886e-127  Exon
     e1179056  e1179056   TBL1Y  ENSG00000092377 -4.247636  1.583091e-117  Exon

              Comparison
     e1181198        CvD
     e1179056        CvD
```

1.1.4 Junctions

```
[9]: feature = 'junctions'
     juncs_cvd, juncs_cvh, juncs_dvh = extract_deg(feature)
     juncs = pd.concat([juncs_cvd, juncs_cvh, juncs_dvh])
     juncs.head(2)
```

```
[9]:                                         Feature   Symbol       ensemblID  \
     chrY:5002759-5031905(+)  chrY:5002759-5031905(+)  PCDH11Y  ENSG00000099715
     chrY:5100708-5104343(+)  chrY:5100708-5104343(+)  PCDH11Y  ENSG00000099715

                                 logFC      adj.P.Val      Type Comparison
     chrY:5002759-5031905(+)  2.978971  1.725570e-32  Junction        CvD
     chrY:5100708-5104343(+)  3.068750  2.870987e-30  Junction        CvD
```

1.1.5 Merge data frame

```
[10]: df = pd.concat([genes, trans, exons, juncs], axis=0)
      df.to_csv('differential_expression_region_interaction_sex_4features.txt',␣
        ↪sep='\t', index=False)
      df.shape
```

```
[10]: (4507, 7)
```

1.2 DE summary

1.2.1 Caudate and DLPFC

DE (feature)

```
[11]: gg = len(set(genes_cvd['Feature']))
      tt = len(set(trans_cvd['Feature']))
      ee = len(set(exons_cvd['Feature']))
      jj = len(set(juncs_cvd['Feature']))

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

```
Gene:           668
Transcript:     642
Exon:           2254
Junction:       324
```

DE (EnsemblID)

```
[12]: gg = len(set(genes_cvd['ensemblID']))
      tt = len(set(trans_cvd['ensemblID']))
      ee = len(set(exons_cvd['ensemblID']))
      jj = len(set(juncs_cvd['ensemblID']))

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

```
Gene:           668
Transcript:     569
Exon:           649
Junction:       147
```

DE (Gene Symbol)

```
[13]: gg = len(set(genes_cvd['Symbol']))
      tt = len(set(trans_cvd['Symbol']))
      ee = len(set(exons_cvd['Symbol']))
      jj = len(set(juncs_cvd['Symbol']))
```

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

```
Gene:          564
Transcript:    569
Exon:          597
Junction:      147
```

1.2.2 Caudate and Hippocampus

DE (feature)

```
[14]: gg = len(set(genes_cvh['Feature']))
      tt = len(set(trans_cvh['Feature']))
      ee = len(set(exons_cvh['Feature']))
      jj = len(set(juncs_cvh['Feature']))

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

```
Gene:          23
Transcript:    177
Exon:          129
Junction:      47
```

DE (EnsemblID)

```
[15]: gg = len(set(genes_cvh['ensemblID']))
      tt = len(set(trans_cvh['ensemblID']))
      ee = len(set(exons_cvh['ensemblID']))
      jj = len(set(juncs_cvh['ensemblID']))

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

```
Gene:          23
Transcript:    129
Exon:          38
Junction:      17
```

DE (Gene Symbol)

```
[16]: gg = len(set(genes_cvh['Symbol']))
      tt = len(set(trans_cvh['Symbol']))
      ee = len(set(exons_cvh['Symbol']))
      jj = len(set(juncs_cvh['Symbol']))
```

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

```
Gene:           19
Transcript:     129
Exon:           36
Junction:       17
```

1.2.3   DLPFC and Hippocampus

DE (feature)

```
[17]: gg = len(set(genes_dvh['Feature']))
      tt = len(set(trans_dvh['Feature']))
      ee = len(set(exons_dvh['Feature']))
      jj = len(set(juncs_dvh['Feature']))

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

```
Gene:           26
Transcript:     28
Exon:           131
Junction:       58
```

DE (EnsemblID)

```
[18]: gg = len(set(genes_dvh['ensemblID']))
      tt = len(set(trans_dvh['ensemblID']))
      ee = len(set(exons_dvh['ensemblID']))
      jj = len(set(juncs_dvh['ensemblID']))

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

```
Gene:           26
Transcript:     19
Exon:           24
Junction:       16
```

DE (Gene Symbol)

```
[19]: gg = len(set(genes_dvh['Symbol']))
      tt = len(set(trans_dvh['Symbol']))
      ee = len(set(exons_dvh['Symbol']))
      jj = len(set(juncs_dvh['Symbol']))
```

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

```
Gene:           23
Transcript:     19
Exon:           23
Junction:       16
```

[ ]: