main

July 9, 2021

1 Feature summary of differential expression analysis

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

1.1 Summary plots

1.1.1 Genes

[2]:
```
genes = pd.read_csv('../../_m/genes/diffExpr_maleVfemale_full.txt', sep='\t',␣
 ↪index_col=0)
genes = genes[(genes['adj.P.Val'] < 0.05)].sort_values('adj.P.Val')
genes['Feature'] = genes.index
genes = genes[['Feature', 'Symbol', 'ensemblID', 'logFC', 'adj.P.Val']]
genes['Type'] = 'gene'
genes.head()
```

[2]:
```
                               Feature Symbol        ensemblID     logFC  \
ENSG00000229807.10  ENSG00000229807.10   XIST  ENSG00000229807 -9.296137
ENSG00000114374.12  ENSG00000114374.12  USP9Y  ENSG00000114374  8.683679
ENSG00000183878.15  ENSG00000183878.15    UTY  ENSG00000183878  8.597152
ENSG00000012817.15  ENSG00000012817.15  KDM5D  ENSG00000012817  8.693010
ENSG00000067048.16  ENSG00000067048.16  DDX3Y  ENSG00000067048  8.587803

                        adj.P.Val  Type
ENSG00000229807.10  1.953623e-272  gene
ENSG00000114374.12  1.953623e-272  gene
ENSG00000183878.15  8.133127e-253  gene
ENSG00000012817.15  3.593495e-252  gene
ENSG00000067048.16  5.035188e-250  gene
```

1.1.2 Transcripts

[3]:
```
trans = pd.read_csv('../../_m/transcripts/diffExpr_maleVfemale_full.txt',␣
 ↪sep='\t', index_col=0)
trans = trans[(trans['adj.P.Val'] < 0.05)].sort_values('adj.P.Val')
trans['Feature'] = trans.index
trans['ensemblID'] = trans.gene_id.str.replace('\\.\d+', '', regex=True)
```

```
trans = trans[['Feature', 'gene_name', 'ensemblID', 'logFC', 'adj.P.Val']].
 ↪rename(columns={'gene_name': 'Symbol'})
trans['Type'] = 'transcript'
trans.head()
```

```
[3]:                                 Feature  Symbol        ensemblID     logFC  \
     ENST00000429829.5  ENST00000429829.5    XIST  ENSG00000229807 -9.421571
     ENST00000336079.7  ENST00000336079.7   DDX3Y  ENSG00000067048  5.981154
     ENST00000440408.5  ENST00000440408.5  TTTY15  ENSG00000233864  3.429488
     ENST00000253320.8  ENST00000253320.8  TXLNGY  ENSG00000131002  5.239253
     ENST00000382872.5  ENST00000382872.5  NLGN4Y  ENSG00000165246  5.419085

                            adj.P.Val        Type
     ENST00000429829.5  1.116837e-282  transcript
     ENST00000336079.7  2.976789e-259  transcript
     ENST00000440408.5  3.212763e-244  transcript
     ENST00000253320.8  8.738864e-237  transcript
     ENST00000382872.5  5.938183e-234  transcript
```

1.1.3 Exons

```
[4]: exons = pd.read_csv('../../_m/exons/diffExpr_maleVfemale_full.txt', sep='\t',␣
      ↪index_col=0)
     exons = exons[(exons['adj.P.Val'] < 0.05)].sort_values('adj.P.Val')
     exons['Feature'] = exons.index
     exons = exons[['Feature', 'Symbol', 'ensemblID', 'logFC', 'adj.P.Val']]
     exons['Type'] = 'exon'
     exons.head()
```

```
[4]:              Feature Symbol        ensemblID     logFC      adj.P.Val  Type
     e1160408  e1160408   XIST  ENSG00000229807 -9.187694  4.748966e-265  exon
     e1160419  e1160419   XIST  ENSG00000229807 -8.410733  6.215720e-265  exon
     e1160425  e1160425   XIST  ENSG00000229807 -7.139194  2.819076e-258  exon
     e1160412  e1160412   XIST  ENSG00000229807 -8.550789  1.340577e-257  exon
     e1160415  e1160415   XIST  ENSG00000229807 -8.666911  1.340577e-257  exon
```

1.1.4 Junctions

```
[5]: juncs = pd.read_csv('../../_m/junctions/diffExpr_maleVfemale_full.txt',␣
      ↪sep='\t', index_col=0)
     juncs = juncs[(juncs['adj.P.Val'] < 0.05)].sort_values('adj.P.Val')
     juncs['Feature'] = juncs.index
     juncs = juncs[['Feature', 'Symbol', 'ensemblID', 'logFC', 'adj.P.Val']]
     juncs['Type'] = 'junction'
     juncs.head()
```

[5]:

```
                                              Feature Symbol        ensemblID  \
chrX:73833375-73837439(-)  chrX:73833375-73837439(-)   XIST  ENSG00000229807
chrX:73829232-73831065(-)  chrX:73829232-73831065(-)   XIST  ENSG00000229807
chrX:73837504-73841381(-)  chrX:73837504-73841381(-)   XIST  ENSG00000229807
chrX:73831275-73833237(-)  chrX:73831275-73833237(-)   XIST  ENSG00000229807
chrX:73822217-73826114(-)  chrX:73822217-73826114(-)   XIST  ENSG00000229807

                              logFC      adj.P.Val      Type
chrX:73833375-73837439(-) -8.479058  9.243134e-237  junction
chrX:73829232-73831065(-) -8.745313  1.084182e-230  junction
chrX:73837504-73841381(-) -8.204010  1.187605e-229  junction
chrX:73831275-73833237(-) -8.938933  3.055212e-224  junction
chrX:73822217-73826114(-) -6.485295  5.171709e-210  junction
```

1.2 DE summary

1.2.1 DE (feature)

[6]:

```python
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:           380
Transcript:     462
Exon:           1479
Junction:       772
```

DE (EnsemblID)

[7]:

```python
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:           380
Transcript:     286
Exon:           267
Junction:       138
```

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:           333
Transcript:     277
Exon:           231
Junction:       138
```

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 77 unique Genes with abs(log2FC) >= 0.5
There are 41 unique Genes with abs(log2FC) >= 1

There are 239 unique Transcript with abs(log2FC) >= 0.5
There are 151 unique Transcript with abs(log2FC) >= 1

There are 639 unique Exons with abs(log2FC) >= 0.5
There are 411 unique Exons with abs(log2FC) >= 1

There are 372 unique Junctions with abs(log2FC) >= 0.5
There are 226 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 77 unique Genes with abs(log2FC) >= 0.5
There are 41 unique Genes with abs(log2FC) >= 1

There are 120 unique Transcripts with abs(log2FC) >= 0.5
There are 70 unique Transcripts with abs(log2FC) >= 1

There are 78 unique Exons with abs(log2FC) >= 0.5
There are 41 unique Exons with abs(log2FC) >= 1

There are 43 unique Junctions with abs(log2FC) >= 0.5
There are 22 unique Junctions with abs(log2FC) >= 1
```

1.3 Autosomal only

[11]:

```
import functools
from gtfparse import read_gtf
```

[12]:

```
@functools.lru_cache()
def get_gtf(gtf_file):
    return read_gtf(gtf_file)
```

[13]:

```
def gene_annotation(gtf_file, feature):
    gtf0 = get_gtf(gtf_file)
    gtf = gtf0[gtf0["feature"] == feature]
    return gtf[["gene_id", "gene_name", "transcript_id", "exon_id",␣
↪"gene_type", "seqname", "start", "end", "strand"]]
```

[14]:

```
gtf_file = '/ceph/genome/human/gencode25/gtf.CHR/_m/gencode.v25.annotation.gtf'
```

1.3.1 Genes

[15]:

```
gtf_annot = gene_annotation(gtf_file, 'gene')

genes = pd.read_csv('../../_m/genes/diffExpr_maleVfemale_full.txt', sep='\t',␣
↪index_col=0)
genes = genes[(genes['adj.P.Val'] < 0.05)].sort_values('adj.P.Val')
genes['Feature'] = genes.index
genes = pd.merge(gtf_annot[['gene_id', 'seqname']], genes, left_on='gene_id',␣
↪right_on='Feature', how='right')
genes.loc[:, 'seqname'] = genes.seqname.fillna('chr?')
genes.sort_values('adj.P.Val').to_csv('chrom_annotation_genes.txt', sep='\t',␣
↪index=False)
```

```
genes = genes[(genes.seqname.str.contains('chr\d+')) | (genes['seqname'] ==␣
↪'chr?')].copy().rename(columns={'seqname': 'chr'})
genes = genes[['Feature', 'chr', 'Symbol', 'ensemblID', 'logFC', 'adj.P.Val']]
genes['Type'] = 'gene'
genes.head()
```

```
INFO:root:Extracted GTF attributes: ['gene_id', 'gene_type', 'gene_status',
'gene_name', 'level', 'havana_gene', 'transcript_id', 'transcript_type',
'transcript_status', 'transcript_name', 'transcript_support_level', 'tag',
'havana_transcript', 'exon_number', 'exon_id', 'ont', 'protein_id', 'ccdsid']
```

[15]:

```
                Feature    chr     Symbol        ensemblID     logFC  \
50   ENSG00000205611.4  chr20  LINC01597  ENSG00000205611  1.307990
52   ENSG00000283443.1  chr20        NaN  ENSG00000283443  1.416253
54  ENSG00000149531.15  chr20     FRG1BP  ENSG00000149531  0.671576
56   ENSG00000095932.6  chr19     SMIM24  ENSG00000095932 -0.901329
57   ENSG00000282826.1  chr20     FRG1CP  ENSG00000282826  0.555138

       adj.P.Val  Type
50  6.877867e-20  gene
52  2.838705e-18  gene
54  1.190836e-17  gene
56  1.022672e-15  gene
57  1.474014e-15  gene
```

[16]:

```
genes[(genes.chr == 'chr?')]
```

[16]:

```
Empty DataFrame
Columns: [Feature, chr, Symbol, ensemblID, logFC, adj.P.Val, Type]
Index: []
```

1.3.2 Annotate unknown by hand

There are none.

[17]:

```
#genes = genes[~(genes['Symbol'].isin(['NLGN4Y', 'JPX', 'PCDH11X', 'GABRE']))]
genes.to_csv('autosomal_DEG.csv', index=False, header=True)
genes.shape
```

[17]:

```
(300, 7)
```

[18]:

```
genes.groupby('ensemblID').first().reset_index().shape
```

[18]:

```
(300, 7)
```

1.3.3 Transcripts

[19]:
```
gtf_annot = gene_annotation(gtf_file, 'transcript')

trans = pd.read_csv('../../_m/transcripts/diffExpr_maleVfemale_full.txt',␣
 ↪sep='\t', index_col=0)
trans = trans[(trans['adj.P.Val'] < 0.05)].sort_values('adj.P.Val')
trans.loc[:, 'Feature'] = trans.index
trans.loc[:, 'ensemblID'] = trans.gene_id.str.replace('\\.\d+', '', regex=True)
trans = pd.merge(gtf_annot[['transcript_id', 'seqname']], trans,␣
 ↪left_on='transcript_id', right_on='Feature', how='right')
trans.loc[:, 'seqname'] = trans.seqname.fillna('chr?')
trans = trans[(trans.seqname.str.contains('chr\d+')) | (trans['seqname'] ==␣
 ↪'chr?')].copy().rename(columns={'seqname': 'chr'})
trans = trans[['Feature', 'chr', 'gene_name', 'ensemblID', 'logFC', 'adj.P.
 ↪Val']].rename(columns={'gene_name': 'Symbol'})
trans['Type'] = 'transcript'
trans.head()
```

[19]:
```
               Feature    chr       Symbol        ensemblID     logFC  \
41   ENST00000550058.1  chr12      METTL25  ENSG00000127720  3.901876
69   ENST00000609745.1  chr20  SDCBP2-AS1  ENSG00000234684 -0.974643
115  ENST00000474345.5   chr1         FDPS  ENSG00000160752  2.344502
130  ENST00000551722.1  chr12      METTL25  ENSG00000127720  0.763093
132  ENST00000414784.1   chr2   AC012442.5  ENSG00000243389  0.697684

         adj.P.Val        Type
41   3.911560e-124  transcript
69    1.869780e-69  transcript
115   2.024836e-32  transcript
130   4.297611e-22  transcript
132   6.993611e-20  transcript
```

[20]:
```
trans[(trans.chr == 'chr?')]
```

[20]:
```
Empty DataFrame
Columns: [Feature, chr, Symbol, ensemblID, logFC, adj.P.Val, Type]
Index: []
```

1.3.4 Annotate unknown by hand

There are none.

[21]:
```
#trans = trans[~(trans['Symbol'].isin(['NLGN4Y']))]
trans.to_csv('transcripts_autosomal_DE.csv', index=False, header=True)
trans.shape
```

[21]:
```
(184, 7)
```

[22]:
```
trans.groupby('ensemblID').first().reset_index().shape
```

[22]:
```
(176, 7)
```

1.3.5 Exons

[23]:
```
gtf_annot = gene_annotation(gtf_file, 'exon')
gtf_annot['ensemblID'] = gtf_annot.gene_id.str.replace('\\.\d+', '', regex=True)

exons = pd.read_csv('../../_m/exons/diffExpr_maleVfemale_full.txt', sep='\t',␣
 ↪index_col=0)
exons = exons[(exons['adj.P.Val'] < 0.05)].sort_values('adj.P.Val')
exons['Feature'] = exons.index
exons = pd.merge(gtf_annot[['ensemblID', 'seqname']], exons, on='ensemblID',␣
 ↪how='right')
exons.loc[:, 'seqname'] = exons.seqname.fillna('chr?')
exons = exons[(exons.seqname.str.contains('chr\d+')) | (exons['seqname'] ==␣
 ↪'chr?')].copy().rename(columns={'seqname': 'chr'})
exons = exons[['Feature', 'chr', 'Symbol', 'ensemblID', 'logFC', 'adj.P.Val']].
 ↪groupby('Feature').first().reset_index()
exons['Type'] = 'exon'
exons.head()
```

[23]:
```
    Feature    chr  Symbol        ensemblID     logFC     adj.P.Val  Type
0  e1011449  chr19  SMIM24  ENSG00000095932 -0.908524  1.769384e-14  exon
1  e1011451  chr19  SMIM24  ENSG00000095932 -0.903646  3.351434e-14  exon
2  e1011454  chr19  SMIM24  ENSG00000095932 -0.894382  5.286518e-14  exon
3  e1013243  chr19   PLIN5  ENSG00000214456 -0.461966  9.464117e-03  exon
4  e1013248  chr19   PLIN5  ENSG00000214456 -0.293909  3.391309e-04  exon
```

[24]:
```
exons[(exons['chr'] == 'chr?')].groupby('ensemblID').first().reset_index()
```

[24]:
```
Empty DataFrame
Columns: [ensemblID, Feature, chr, Symbol, logFC, adj.P.Val, Type]
Index: []
```

1.3.6 Annotate unknown by hand

There are none.

[25]:
```
#exons = exons[~(exons['ensemblID'].isin(['ENSG00000269941']))]
exons.to_csv('exons_autosomal_DE.csv', index=False, header=True)
exons.shape
```

[25]:
```
(492, 7)
```

[26]:
```
exons.groupby('ensemblID').first().reset_index().shape
```

```
[26]: (189, 7)
```

1.3.7 Junctions

```
[27]: juncs = pd.read_csv('../../_m/junctions/diffExpr_maleVfemale_full.txt',␣
       ↪sep='\t', index_col=0)
      juncs = juncs[(juncs['adj.P.Val'] < 0.05)].sort_values('adj.P.Val')
      juncs['Feature'] = juncs.index
      juncs = pd.merge(gtf_annot[['ensemblID', 'seqname']], juncs, on='ensemblID',␣
       ↪how='right')
      juncs.loc[:, 'seqname'] = juncs.seqname.fillna('chr?')
      juncs = juncs[(juncs.seqname.str.contains('chr\d+')) | (juncs['seqname'] ==␣
       ↪'chr?')].copy().rename(columns={'seqname': 'chr'})
      juncs = juncs[['Feature', 'chr', 'Symbol', 'ensemblID', 'logFC', 'adj.P.Val']].
       ↪groupby('Feature').first().reset_index()
      juncs['Type'] = 'junction'
      juncs.head()
```

```
[27]:                      Feature    chr  Symbol        ensemblID     logFC  \
      0  chr10:11314271-11320856(+)  chr10   CELF2  ENSG00000048740  0.547582
      1  chr10:46911502-46943917(+)  chr10  PTPN20  ENSG00000204179  0.613532
      2  chr10:46946676-46999911(+)  chr10  PTPN20  ENSG00000204179  0.648941
      3  chr10:60106060-60108829(-)  chr10    ANK3  ENSG00000151150  0.131277
      4  chr10:60264021-60270130(-)  chr10    ANK3  ENSG00000151150  0.112018

         adj.P.Val      Type
      0   0.017608  junction
      1   0.005269  junction
      2   0.000035  junction
      3   0.027027  junction
      4   0.020419  junction
```

```
[28]: juncs[(juncs['chr'] == 'chr?')].groupby('ensemblID').first()
```

```
[28]: Empty DataFrame
      Columns: [Feature, chr, Symbol, logFC, adj.P.Val, Type]
      Index: []
```

1.3.8 Annotate unknown by hand

None unknown

```
[29]: juncs.to_csv('junctions_autosomal_DE.csv', index=False, header=True)
      juncs.shape
```

```
[29]: (236, 7)
```

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

```
[30]: (85, 7)
```

1.4 DE summary

1.4.1 DE (feature)

```
[31]: 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:           300
Transcript:     184
Exon:           492
Junction:       236
```

DE (EnsemblID)

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

```
Gene:           300
Transcript:     176
Exon:           189
Junction:       85
```

DE (Gene Symbol)

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

```
Gene:           259
Transcript:     176
Exon:           165
Junction:       85
```

1.4.2 Feature effect size summary

[34]:

```
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 32 unique Genes with abs(log2FC) >= 0.5
There are 5 unique Genes with abs(log2FC) >= 1

There are 56 unique Transcript with abs(log2FC) >= 0.5
There are 17 unique Transcript with abs(log2FC) >= 1

There are 81 unique Exons with abs(log2FC) >= 0.5
There are 6 unique Exons with abs(log2FC) >= 1

There are 92 unique Junctions with abs(log2FC) >= 0.5
There are 40 unique Junctions with abs(log2FC) >= 1
```

[35]:

```
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 32 unique Genes with abs(log2FC) >= 0.5
There are 5 unique Genes with abs(log2FC) >= 1

There are 53 unique Transcripts with abs(log2FC) >= 0.5
There are 17 unique Transcripts with abs(log2FC) >= 1

There are 27 unique Exons with abs(log2FC) >= 0.5
There are 4 unique Exons with abs(log2FC) >= 1

There are 13 unique Junctions with abs(log2FC) >= 0.5
There are 1 unique Junctions with abs(log2FC) >= 1
```