main

August 2, 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	adj.P.Val	Type
ENSG00000229236.1	ENSG00000229236.1	TTTY10	ENSG00000229236	6.919904	5.186692e-243	gene
ENSG00000154620.5	ENSG00000154620.5	TMSB4Y	ENSG00000154620	7.017845	4.942051e-238	gene
ENSG00000226555.1	ENSG00000226555.1	AGKP1	ENSG00000226555	7.083112	9.807573e-236	gene
ENSG00000176728.7	ENSG00000176728.7	TTTY14	ENSG00000176728	8.090491	4.895668e-231	gene
ENSG00000260197.1	ENSG00000260197.1	NaN	ENSG00000260197	6.302909	8.726814e-229	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', 'Symbol', 'ensemblID', 'logFC', 'adj.P.Val']]#.
 ↪rename(columns={'gene_name': 'Symbol'})
trans['Type'] = 'transcript'
trans.head()
```

[3]:

	Feature	Symbol	ensemblID	logFC	adj.P.Val	Type
ENST00000602495.1	ENST00000602495.1	XIST	ENSG00000229807	-7.591232	1.889639e-242	transcript
ENST00000416330.1	ENST00000416330.1	XIST	ENSG00000229807	-7.791829	7.035924e-237	transcript
ENST00000429829.5	ENST00000429829.5	XIST	ENSG00000229807	-9.537955	6.043774e-225	transcript
ENST00000440408.5	ENST00000440408.5	TTTY15	ENSG00000233864	6.006990	3.153685e-208	transcript
ENST00000469599.6	ENST00000469599.6	KDM5D	ENSG00000012817	7.765403	5.908909e-179	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
e1160419	e1160419	XIST	ENSG00000229807	-8.089614	6.305380e-266	exon
e1160425	e1160425	XIST	ENSG00000229807	-6.939845	2.715346e-265	exon
e1160404	e1160404	XIST	ENSG00000229807	-7.779011	4.404419e-264	exon
e1160437	e1160437	XIST	ENSG00000229807	-8.654052	5.915596e-260	exon
e1160443	e1160443	XIST	ENSG00000229807	-8.621807	8.389574e-260	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:73829232-73831065(-)  chrX:73829232-73831065(-)   XIST  ENSG00000229807
chrX:73833375-73837439(-)  chrX:73833375-73837439(-)   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:73829232-73831065(-) -8.690958  6.863963e-239  junction
chrX:73833375-73837439(-) -8.135806  6.070401e-237  junction
chrX:73837504-73841381(-) -8.008542  9.772347e-231  junction
chrX:73831275-73833237(-) -8.753608  8.771081e-228  junction
chrX:73822217-73826114(-) -6.300175  9.334965e-218  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:           105
Transcript:     252
Exon:           952
Junction:       687
```

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:           105
Transcript:     137
Exon:           114
Junction:       159
```

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:           88
Transcript:     137
Exon:           96
Junction:       159
```

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 56 unique Genes with abs(log2FC) >= 0.5
There are 35 unique Genes with abs(log2FC) >= 1

There are 183 unique Transcript with abs(log2FC) >= 0.5
There are 130 unique Transcript with abs(log2FC) >= 1

There are 529 unique Exons with abs(log2FC) >= 0.5
There are 393 unique Exons with abs(log2FC) >= 1

There are 313 unique Junctions with abs(log2FC) >= 0.5
There are 206 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 56 unique Genes with abs(log2FC) >= 0.5
There are 35 unique Genes with abs(log2FC) >= 1

There are 89 unique Transcripts with abs(log2FC) >= 0.5
There are 60 unique Transcripts with abs(log2FC) >= 1

There are 58 unique Exons with abs(log2FC) >= 0.5
There are 37 unique Exons with abs(log2FC) >= 1

There are 46 unique Junctions with abs(log2FC) >= 0.5
There are 21 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	adj.P.Val	Type
41	ENSG00000205611.4	chr20	LINC01597	ENSG00000205611	1.177854	1.668358e-14	gene
43	ENSG00000283443.1	chr20	NaN	ENSG00000283443	1.230724	4.320235e-14	gene
45	ENSG00000282826.1	chr20	FRG1CP	ENSG00000282826	0.555011	8.233830e-14	gene
46	ENSG00000149531.15	chr20	FRG1BP	ENSG00000149531	0.649257	2.029046e-13	gene
48	ENSG00000258484.3	chr15	SPESP1	ENSG00000258484	0.759901	2.382040e-10	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]:
```
(42, 7)
```

[18]:
```
genes.groupby('ensemblID').first().reset_index().shape
```

[18]:
```
(42, 7)
```

1.3.3 Transcripts

[19]:
```
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 = trans[(trans.chr.str.contains('chr\d+'))].copy()
trans = trans[['Feature', 'chr', 'Symbol', 'ensemblID', 'logFC', 'adj.P.Val']]#.
↪rename(columns={'gene_name': 'Symbol'})
trans['Type'] = 'transcript'
trans.head()
```

[19]:

	Feature	chr	Symbol	ensemblID	\
ENST00000421320.1	ENST00000421320.1	chr10	RP11-124O11.1	ENSG00000234944	
ENST00000358464.9	ENST00000358464.9	chr20	FRG1CP	ENSG00000282826	
ENST00000380888.4	ENST00000380888.4	chr20	LINC01597	ENSG00000205611	
ENST00000636528.1	ENST00000636528.1	chr20	RP11-462H3.2	ENSG00000283443	
ENST00000612308.1	ENST00000612308.1	chr9	FAM27D1	ENSG00000275493	

	logFC	adj.P.Val	Type
ENST00000421320.1	-1.114327	6.986165e-16	transcript
ENST00000358464.9	0.543532	3.805030e-13	transcript
ENST00000380888.4	0.744357	3.576296e-12	transcript
ENST00000636528.1	1.286835	7.740744e-12	transcript
ENST00000612308.1	1.045770	8.059660e-11	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]:
```
(71, 7)
```

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

[22]:
```
(67, 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	e1028339	chr19	DCAF15	ENSG00000132017	-0.238367	0.007060	exon
1	e1038350	chr19	ZNF208	ENSG00000160321	-0.436529	0.002094	exon
2	e1038351	chr19	ZNF208	ENSG00000160321	-0.445934	0.003064	exon
3	e1038352	chr19	ZNF208	ENSG00000160321	-0.434854	0.011539	exon
4	e1038361	chr19	ZNF208	ENSG00000160321	-0.343576	0.040657	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]:
```
(133, 7)
```

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

[26]:
```
(46, 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	adj.P.Val	Type
0	chr10:112426360-112426787(+)	chr10	ACSL5	ENSG00000197142	-0.386362	0.032155	junction
1	chr10:20889950-20896957(-)	chr10	NEBL	ENSG00000078114	-0.188722	0.002322	junction
2	chr10:46946676-46999911(+)	chr10	PTPN20	ENSG00000204179	0.548855	0.047015	junction
3	chr10:95238686-95247214(-)	chr10	PDLIM1	ENSG00000107438	-0.556318	0.022140	junction
4	chr11:12473984-12477846(+)	chr11	PARVA	ENSG00000197702	-0.152227	0.023287	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]:
```
(218, 7)
```

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

[30]:
```
(109, 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:           42
Transcript:     71
Exon:           133
Junction:       218
```

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:           42
Transcript:     67
Exon:           46
Junction:       109
```

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:           31
Transcript:     67
Exon:           40
Junction:       109
```

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 15 unique Genes with abs(log2FC) >= 0.5
There are 2 unique Genes with abs(log2FC) >= 1

There are 44 unique Transcript with abs(log2FC) >= 0.5
There are 26 unique Transcript with abs(log2FC) >= 1

There are 55 unique Exons with abs(log2FC) >= 0.5
There are 6 unique Exons with abs(log2FC) >= 1

There are 80 unique Junctions with abs(log2FC) >= 0.5
There are 27 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 15 unique Genes with abs(log2FC) >= 0.5
There are 2 unique Genes with abs(log2FC) >= 1

There are 42 unique Transcripts with abs(log2FC) >= 0.5
There are 26 unique Transcripts with abs(log2FC) >= 1

There are 10 unique Exons with abs(log2FC) >= 0.5
There are 3 unique Exons with abs(log2FC) >= 1

There are 16 unique Junctions with abs(log2FC) >= 0.5
```

```
There are 2 unique Junctions with abs(log2FC) >= 1
```