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	adj.P.Val	Type
ENSG00000226555.1	ENSG00000226555.1	AGKP1	ENSG00000226555	7.270752	1.837492e-256	gene
ENSG00000229236.1	ENSG00000229236.1	TTTY10	ENSG00000229236	7.417472	3.642704e-249	gene
ENSG00000176728.7	ENSG00000176728.7	TTTY14	ENSG00000176728	8.813730	1.140965e-247	gene
ENSG00000260197.1	ENSG00000260197.1	NaN	ENSG00000260197	7.018888	5.725810e-244	gene
ENSG00000241859.6	ENSG00000241859.6	ANOS2P	ENSG00000241859	7.637736	5.448917e-236	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	-7.899335	1.688898e-213	transcript
ENST00000440408.5	ENST00000440408.5	TTTY15	ENSG00000233864	6.571364	2.058171e-212	transcript
ENST00000429829.5	ENST00000429829.5	XIST	ENSG00000229807	-10.273385	2.101470e-212	transcript
ENST00000382872.5	ENST00000382872.5	NLGN4Y	ENSG00000165246	6.520316	1.372850e-206	transcript
ENST00000416330.1	ENST00000416330.1	XIST	ENSG00000229807	-7.752337	3.398175e-202	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
e1160404	e1160404	XIST	ENSG00000229807	-7.911385	4.092584e-247	exon
e1160425	e1160425	XIST	ENSG00000229807	-7.004742	8.243948e-241	exon
e1180839	e1180839	KDM5D	ENSG00000012817	8.771863	8.980414e-241	exon
e1160439	e1160439	XIST	ENSG00000229807	-8.113401	9.234178e-241	exon
e1180866	e1180866	KDM5D	ENSG00000012817	8.449606	9.234178e-241	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:73831275-73833237(-)  chrX:73831275-73833237(-)   XIST  ENSG00000229807
chrX:73827985-73829067(-)  chrX:73827985-73829067(-)   XIST  ENSG00000229807
chrX:73822217-73826114(-)  chrX:73822217-73826114(-)   XIST  ENSG00000229807
chrX:73829232-73831065(-)  chrX:73829232-73831065(-)   XIST  ENSG00000229807

                                 logFC      adj.P.Val      Type
chrX:73833375-73837439(-) -8.206159  1.174942e-221  junction
chrX:73831275-73833237(-) -8.941121  1.930993e-215  junction
chrX:73827985-73829067(-) -8.126867  2.239959e-208  junction
chrX:73822217-73826114(-) -6.282980  4.678196e-206  junction
chrX:73829232-73831065(-) -8.995840  1.985000e-205  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:           573
Transcript:     422
Exon:           3566
Junction:       1809
```

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:           573
Transcript:     288
Exon:           901
Junction:       700
```

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:           510
Transcript:     288
Exon:           840
Junction:       700
```

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 67 unique Genes with abs(log2FC) >= 0.5
There are 31 unique Genes with abs(log2FC) >= 1

There are 223 unique Transcript with abs(log2FC) >= 0.5
There are 142 unique Transcript with abs(log2FC) >= 1

There are 767 unique Exons with abs(log2FC) >= 0.5
There are 394 unique Exons with abs(log2FC) >= 1

There are 479 unique Junctions with abs(log2FC) >= 0.5
There are 220 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 67 unique Genes with abs(log2FC) >= 0.5
There are 31 unique Genes with abs(log2FC) >= 1

There are 120 unique Transcripts with abs(log2FC) >= 0.5
There are 72 unique Transcripts with abs(log2FC) >= 1

There are 76 unique Exons with abs(log2FC) >= 0.5
There are 32 unique Exons with abs(log2FC) >= 1

There are 59 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	adj.P.Val	Type
37	ENSG00000205611.4	chr20	LINC01597	ENSG00000205611	1.219798	8.648719e-19	gene
40	ENSG00000149531.15	chr20	FRG1BP	ENSG00000149531	0.683514	9.850828e-17	gene
41	ENSG00000115297.10	chr2	TLX2	ENSG00000115297	-0.952295	1.621359e-15	gene
42	ENSG00000283443.1	chr20	NaN	ENSG00000283443	1.311288	6.869947e-15	gene
44	ENSG00000255346.9	chr15	NOX5	ENSG00000255346	0.915333	2.464554e-14	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]:
```
(481, 7)
```

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

[18]:
```
(481, 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 = trans[['Feature', 'chr', 'Symbol', 'ensemblID', 'logFC', 'adj.P.Val']]#.
↪rename(columns={'gene_name': 'Symbol'})
trans['Type'] = 'transcript'
trans.head()
```

[19]:

	Feature	chr	Symbol	ensemblID	\
ENST00000602495.1	ENST00000602495.1	chrX	XIST	ENSG00000229807	
ENST00000440408.5	ENST00000440408.5	chrY	TTTY15	ENSG00000233864	
ENST00000429829.5	ENST00000429829.5	chrX	XIST	ENSG00000229807	
ENST00000382872.5	ENST00000382872.5	chrY	NLGN4Y	ENSG00000165246	
ENST00000416330.1	ENST00000416330.1	chrX	XIST	ENSG00000229807	

	logFC	adj.P.Val	Type
ENST00000602495.1	-7.899335	1.688898e-213	transcript
ENST00000440408.5	6.571364	2.058171e-212	transcript
ENST00000429829.5	-10.273385	2.101470e-212	transcript
ENST00000382872.5	6.520316	1.372850e-206	transcript
ENST00000416330.1	-7.752337	3.398175e-202	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]:
```
(422, 7)
```

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

[22]:
```
(288, 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	e1002925	chr18	PHLPP1	ENSG00000081913	-0.178539	0.047767	exon
1	e1003849	chr18	RTTN	ENSG00000176225	-0.206977	0.039195	exon
2	e1003851	chr18	RTTN	ENSG00000176225	-0.259873	0.014349	exon
3	e1004000	chr18	RTTN	ENSG00000176225	-0.280113	0.016140	exon
4	e1004005	chr18	RTTN	ENSG00000176225	-0.283366	0.046143	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]:
```
(2551, 7)
```

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

[26]:
```
(801, 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	chr10:101002947-101003506(+)	chr10	LZTS2	ENSG00000107816	-0.211514	0.021862	junction
1	chr10:102594066-102597139(+)	chr10	SUFU	ENSG00000107882	-0.286279	0.019524	junction
2	chr10:103396033-103396408(-)	chr10	USMG5	ENSG00000173915	0.217285	0.011480	junction
3	chr10:124682380-124696157(-)	chr10	RP11-12J10.3	ENSG00000258539	-0.151267	0.047859	junction
4	chr10:125807488-125812213(-)	chr10	UROS	ENSG00000188690	0.165628	0.004315	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]:
```
(1245, 7)
```

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

[30]:
```
(626, 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:           481
Transcript:     422
Exon:           2551
Junction:       1245
```

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:           481
Transcript:     288
Exon:           801
Junction:       626
```

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:           423
Transcript:     288
Exon:           750
Junction:       626
```

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

There are 223 unique Transcript with abs(log2FC) >= 0.5
There are 142 unique Transcript with abs(log2FC) >= 1

There are 209 unique Exons with abs(log2FC) >= 0.5
There are 4 unique Exons with abs(log2FC) >= 1

There are 184 unique Junctions with abs(log2FC) >= 0.5
There are 34 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 25 unique Genes with abs(log2FC) >= 0.5
There are 2 unique Genes with abs(log2FC) >= 1

There are 120 unique Transcripts with abs(log2FC) >= 0.5
There are 72 unique Transcripts with abs(log2FC) >= 1

There are 29 unique Exons with abs(log2FC) >= 0.5
There are 3 unique Exons with abs(log2FC) >= 1

There are 26 unique Junctions with abs(log2FC) >= 0.5
```

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