main

August 6, 2021

1 Extract overlapping genes with CMC

[1]:
```
import functools
import numpy as np
import pandas as pd
from gtfparse import read_gtf
```

[2]:
```
@functools.lru_cache()
def get_gtf(gtf_file):
    return read_gtf(gtf_file)
```

[3]:
```
def gene_annotation(gtf_file, feature):
    gtf0 = get_gtf(gtf_file)
    gtf = gtf0[gtf0["feature"] == feature]
    return gtf[["gene_id", "gene_name", "gene_type",
                "seqname", "start", "end", "strand"]]
```

[4]:
```
gtf_file = '/ceph/genome/human/gencode25/gtf.CHR/_m/gencode.v25.annotation.gtf'
gtf_annot = gene_annotation(gtf_file, 'gene')
gtf_annot['ensemblID'] = gtf_annot.gene_id.str.replace("\\..*", "", regex=True)
```

```
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']
```

1.1 Male specific

[5]:
```
cmc_file = "../../_m/cmc_all_deg_across_tissues_maleSpecific.csv"
df = gtf_annot.merge(pd.read_csv(cmc_file), on='ensemblID')
df.head(2)
```

[5]:

	gene_id_x	gene_name_x	gene_type_x	seqname_x	start	end	strand	ensemblID	Unnamed: 0	gene_id_y
0	ENSG00000272512.1	RP11-540I7.17	lincRNA	chr1	995966	998051	-	ENSG00000272512	1386	ENSG00000272512.1
1	ENSG00000107404.18	DVL1	protein_coding	chr1	1335276					

```
1  1349350      -  ENSG00000107404        4087  ENSG00000107404.18

    gene_name_y seqname_y     gene_type_y  Caudate  DLPFC  Hippocampus  \
0  RP11-54O7.17      chr1         lincRNA        1      0            0
1          DVL1      chr1  protein_coding        1      0            0

   CMC DLPFC
0          0
1          0
```

1.1.1 CMC DLPFC overlapping Caudate

[6]:
```
df[(df['CMC DLPFC'] == 1) & (df['Caudate'] == 1)]
```

```
INFO:numexpr.utils:Note: NumExpr detected 64 cores but "NUMEXPR_MAX_THREADS" not
set, so enforcing safe limit of 8.
INFO:numexpr.utils:NumExpr defaulting to 8 threads.
```

[6]:
```
               gene_id_x   gene_name_x        gene_type_x seqname_x  \
29    ENSG00000189410.11         SH2D5     protein_coding      chr1
50    ENSG00000066185.12       ZMYND12     protein_coding      chr1
286   ENSG00000115170.13         ACVR1     protein_coding      chr2
1154   ENSG00000260400.1  RP11-119F7.5  sense_overlapping     chr10
1397  ENSG00000139372.14           TDG     protein_coding     chr12
1870   ENSG00000256463.8         SALL3     protein_coding     chr18
2116  ENSG00000100116.16          GCAT     protein_coding     chr22
2127  ENSG00000100266.18       PACSIN2     protein_coding     chr22
2193  ENSG00000134597.14         RBMX2     protein_coding      chrX

          start        end strand        ensemblID  Unnamed: 0  \
29     20719732   20732837      -  ENSG00000189410       30978
50     42430329   42456267      -  ENSG00000066185       64447
286   157736444  157875862      -  ENSG00000115170      341965
1154   68698500   68700794      +  ENSG00000260400     1260519
1397  103965804  103988874      +  ENSG00000139372     1584469
1870   78980275   79002677      +  ENSG00000256463     2184098
2116   37807905   37817176      +  ENSG00000100116     2469692
2127   42835412   43015145      -  ENSG00000100266     2479888
2193  130401969  130413343      +  ENSG00000134597     2550700

               gene_id_y   gene_name_y seqname_y        gene_type_y  Caudate  \
29    ENSG00000189410.11         SH2D5      chr1     protein_coding        1
50    ENSG00000066185.12       ZMYND12      chr1     protein_coding        1
286   ENSG00000115170.13         ACVR1      chr2     protein_coding        1
1154   ENSG00000260400.1  RP11-119F7.5     chr10  sense_overlapping        1
1397  ENSG00000139372.14           TDG     chr12     protein_coding        1
1870   ENSG00000256463.8         SALL3     chr18     protein_coding        1
```

```
2116  ENSG00000100116.16          GCAT     chr22     protein_coding        1
2127  ENSG00000100266.18       PACSIN2     chr22     protein_coding        1
2193  ENSG00000134597.14         RBMX2      chrX     protein_coding        1

      DLPFC  Hippocampus  CMC DLPFC
29        0            0          1
50        0            0          1
286       0            0          1
1154      0            0          1
1397      0            0          1
1870      0            0          1
2116      0            0          1
2127      0            0          1
2193      0            0          1
```

1.1.2 CMC DLPFC overlapping DLPFC

[7]:
```
df[(df['CMC DLPFC'] == 1) & (df['DLPFC'] == 1)]
```

[7]:
```
               gene_id_x gene_name_x                         gene_type_x  \
80    ENSG00000171488.14      LRRC8C                      protein_coding
104    ENSG00000231752.5       EMBP1  transcribed_unprocessed_pseudogene
1537  ENSG00000156414.18       TDRD9                      protein_coding

     seqname_x      start        end strand        ensemblID  Unnamed: 0  \
80        chr1   89633072   89769903      +  ENSG00000171488      104025
104       chr1  121519112  121571892      +  ENSG00000231752      128562
1537     chr14  103928462  104052667      +  ENSG00000156414     1748741

               gene_id_y gene_name_y seqname_y  \
80    ENSG00000171488.14      LRRC8C      chr1
104    ENSG00000231752.5       EMBP1      chr1
1537  ENSG00000156414.18       TDRD9     chr14

                           gene_type_y  Caudate  DLPFC  Hippocampus  \
80                      protein_coding        0      1            0
104   transcribed_unprocessed_pseudogene        0      1            0
1537                    protein_coding        0      1            0

      CMC DLPFC
80            1
104           1
1537          1
```

1.1.3 CMC DLPFC overlapping Hippocampus

[8]:
```
df[(df['CMC DLPFC'] == 1) & (df['Hippocampus'] == 1)]
```

[8]:
```
Empty DataFrame
Columns: [gene_id_x, gene_name_x, gene_type_x, seqname_x, start, end, strand,
ensemblID, Unnamed: 0, gene_id_y, gene_name_y, seqname_y, gene_type_y, Caudate,
DLPFC, Hippocampus, CMC DLPFC]
Index: []
```

1.1.4 CMC DLPFC overlapping Caudate & DLPFC

[9]:
```
df[(df['CMC DLPFC'] == 1) & (df['Caudate'] == 1) & (df['DLPFC'] == 1)]
```

[9]:
```
Empty DataFrame
Columns: [gene_id_x, gene_name_x, gene_type_x, seqname_x, start, end, strand,
ensemblID, Unnamed: 0, gene_id_y, gene_name_y, seqname_y, gene_type_y, Caudate,
DLPFC, Hippocampus, CMC DLPFC]
Index: []
```

1.2 Female specific

[10]:
```
cmc_file = "../../_m/cmc_all_deg_across_tissues_femaleSpecific.csv"
df = gtf_annot.merge(pd.read_csv(cmc_file), on='ensemblID')
df.head(2)
```

[10]:

	gene_id_x	gene_name_x	gene_type_x	seqname_x	start	end	strand	ensemblID	Unnamed: 0	gene_id_y	gene_name_y	seqname_y	gene_type_y	Caudate	DLPFC	Hippocampus	CMC DLPFC
0	ENSG00000272455.1	RP4-758J18.13	lincRNA	chr1	1409096	1410618	+	ENSG00000272455	4753	ENSG00000272455.1	RP4-758J18.13	chr1	lincRNA	0	0	0	1
1	ENSG00000142609.17	CFAP74	protein_coding	chr1	1921951	2003837	-	ENSG00000142609	7511	ENSG00000142609.17	CFAP74	chr1	protein_coding	0	0	0	1

1.2.1 CMC DLPFC overlapping Caudate

[11]:
```
df[(df['CMC DLPFC'] == 1) & (df['Caudate'] == 1)]
```

[11]:
```
              gene_id_x   gene_name_x                       gene_type_x  \
208   ENSG00000249669.8         CARMN                           lincRNA
474  ENSG00000100814.17      CCNB1IP1                    protein_coding
529   ENSG00000269937.1  RP11-20I23.8                         antisense
555  ENSG00000167703.14       SLC43A2                    protein_coding
588   ENSG00000263006.6       ROCK1P1  transcribed_unprocessed_pseudogene

    seqname_x      start        end strand        ensemblID  Unnamed: 0  \
208      chr5  149406689  149432835      +  ENSG00000249669      770108
474     chr14   20311368   20333312      -  ENSG00000100814     1661524
529     chr16    2561471    2565096      -  ENSG00000269937     1867195
555     chr17    1569267    1628886      -  ENSG00000167703     1979014
588     chr18     109065     122219      +  ENSG00000263006     2140998

              gene_id_y   gene_name_y seqname_y  \
208   ENSG00000249669.8         CARMN      chr5
474  ENSG00000100814.17      CCNB1IP1     chr14
529   ENSG00000269937.1  RP11-20I23.8     chr16
555  ENSG00000167703.14       SLC43A2     chr17
588   ENSG00000263006.6       ROCK1P1     chr18

                           gene_type_y  Caudate  DLPFC  Hippocampus  \
208                            lincRNA        1      0            0
474                     protein_coding        1      0            0
529                          antisense        1      0            0
555                     protein_coding        1      0            0
588  transcribed_unprocessed_pseudogene        1      0            0

     CMC DLPFC
208          1
474          1
529          1
555          1
588          1
```

1.2.2 CMC DLPFC overlapping DLPFC

[12]:
```
df[(df['CMC DLPFC'] == 1) & (df['DLPFC'] == 1)]
```

[12]:
```
Empty DataFrame
Columns: [gene_id_x, gene_name_x, gene_type_x, seqname_x, start, end, strand,
ensemblID, Unnamed: 0, gene_id_y, gene_name_y, seqname_y, gene_type_y, Caudate,
DLPFC, Hippocampus, CMC DLPFC]
Index: []
```

1.2.3 CMC DLPFC overlapping Hippocampus

[13]:
```
df[(df['CMC DLPFC'] == 1) & (df['Hippocampus'] == 1)]
```

[13]:
```
Empty DataFrame
Columns: [gene_id_x, gene_name_x, gene_type_x, seqname_x, start, end, strand,
ensemblID, Unnamed: 0, gene_id_y, gene_name_y, seqname_y, gene_type_y, Caudate,
DLPFC, Hippocampus, CMC DLPFC]
Index: []
```

1.2.4 CMC DLPFC overlapping Caudate & DLPFC

[14]:
```
df[(df['CMC DLPFC'] == 1) & (df['Caudate'] == 1) & (df['DLPFC'] == 1)]
```

[14]:
```
Empty DataFrame
Columns: [gene_id_x, gene_name_x, gene_type_x, seqname_x, start, end, strand,
ensemblID, Unnamed: 0, gene_id_y, gene_name_y, seqname_y, gene_type_y, Caudate,
DLPFC, Hippocampus, CMC DLPFC]
Index: []
```

[ ]:
```

```