main

July 13, 2021

1 Annotated predictive feature with start and end information

[1]:

```
import functools
import numpy as np
import pandas as pd
from gtfparse import read_gtf
```

1.1 Functions

[2]:

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


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


def get_tissue_map(tissue):
    return {"Caudate": 'caudate', 'DLPFC': 'dlpfc',
            'Dentate Gyrus': 'dentateGyrus', 'Hippocampus':␣
 ↪'hippocampus'}[tissue]
```

1.2 Generate gene annotation

[3]:

```
gtf_file = "/ceph/genome/human/gencode25/gtf.CHR/_m/gencode.v25.annotation.gtf"
annot = gene_annotation(gtf_file).rename(columns={'seqname': 'chr'})
annot['ensemblID'] = annot.gene_id.str.replace("\\..*", "", regex=True)
annot["length"] = np.abs(annot.start - annot.end)
annot.head(2)
```

```
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']
```

```
[3]:               gene_id gene_name                           gene_type   chr  \
     0   ENSG00000223972.5   DDX11L1  transcribed_unprocessed_pseudogene  chr1
     12  ENSG00000227232.5    WASH7P              unprocessed_pseudogene  chr1

         start    end strand        ensemblID  length
     0   11869  14409      +  ENSG00000223972    2540
     12  14404  29570      -  ENSG00000227232   15166
```

```
[4]: annot = annot[["gene_id", "ensemblID", "gene_name", "chr",
                    "start", "end", "length", "strand"]]\
         .set_index("gene_id")
     annot.head(2)
```

```
[4]:                          ensemblID gene_name   chr  start    end  length  \
     gene_id
     ENSG00000223972.5  ENSG00000223972   DDX11L1  chr1  11869  14409    2540
     ENSG00000227232.5  ENSG00000227232    WASH7P  chr1  14404  29570   15166

                       strand
     gene_id
     ENSG00000223972.5      +
     ENSG00000227232.5      -
```

1.3 Load DEG summary

```
[5]: deg_file = "../../differential_analysis/tissue_comparison/deg_summary/"+\
               "_m/diffExpr_ancestry_full_4regions.tsv"
     df = pd.read_csv(deg_file, sep='\t', index_col=0)\
            .loc[:, ["logFC", "AveExpr", "t", "adj.P.Val", "Type", "Tissue"]]
     df = df[(df["Type"] == "Gene")].copy()
     print(df.groupby(["Tissue"]).size())
     df.tail(2)
```

```
Tissue
Caudate          22374
DLPFC            22398
Dentate Gyrus    21140
Hippocampus      22269
dtype: int64
```

```
[5]:                        logFC   AveExpr         t  adj.P.Val  Type       Tissue
     Feature
     ENSG00000147118.11 -0.000004  2.869577 -0.000160   0.999915  Gene  Hippocampus
     ENSG00000077942.18 -0.000012  4.470327 -0.000107   0.999915  Gene  Hippocampus
```

1.4 Merge files and clean data

[6]:

```
dft = annot.merge(df, left_index=True, right_index=True)\
            .sort_values(["Tissue", "adj.P.Val"])
dft["New_Tissue"] = [get_tissue_map(x) for x in dft.Tissue]
print(dft.shape)
dft.head(10)
```

```
(88181, 14)
```

[6]:

```
                           ensemblID      gene_name    chr      start  \
ENSG00000272977.1    ENSG00000272977  CTA-390C10.10  chr22   25476218
ENSG00000233913.7    ENSG00000233913   CTC-575D19.1   chr5  168616352
ENSG00000259479.6    ENSG00000259479         SORD2P  chr15   44825747
ENSG00000068654.15   ENSG00000068654         POLR1A   chr2   86020216
ENSG00000084628.9    ENSG00000084628         NKAIN1   chr1   31179745
ENSG00000204894.4    ENSG00000204894  RP11-208G20.2   chr7  152367171
ENSG00000226278.1    ENSG00000226278         PSPHP1   chr7   55764797
ENSG00000271361.1    ENSG00000271361       HTATSF1P2   chr6    3023142
ENSG00000230076.1    ENSG00000230076     AC016708.2   chr2  214847128
ENSG00000140263.13   ENSG00000140263           SORD  chr15   45023104

                           end  length strand     logFC    AveExpr          t  \
ENSG00000272977.1    25479971    3753      +  2.197155  1.176962  12.328222
ENSG00000233913.7   168616996     644      + -2.941671  3.106682 -12.213021
ENSG00000259479.6    44884694   58947      - -2.338783 -0.546410 -12.087500
ENSG00000068654.15   86106155   85939      -  0.292087  5.940820  11.922914
ENSG00000084628.9    31239554   59809      -  1.891807  1.657673  11.518655
ENSG00000204894.4   152367260      89      + -4.696103 -1.835114 -11.306024
ENSG00000226278.1    55773288    8491      + -5.659256 -0.002860 -10.998267
ENSG00000271361.1     3023772     630      -  3.418221 -2.870529  10.931060
ENSG00000230076.1   214847445     317      + -4.536309  0.001767 -10.914092
ENSG00000140263.13   45077185   54081      +  0.626047  3.449221  10.892127

                        adj.P.Val  Type   Tissue New_Tissue
ENSG00000272977.1    1.293546e-22  Gene  Caudate    caudate
ENSG00000233913.7    1.511451e-22  Gene  Caudate    caudate
ENSG00000259479.6    2.536508e-22  Gene  Caudate    caudate
ENSG00000068654.15   6.364724e-22  Gene  Caudate    caudate
ENSG00000084628.9    9.739085e-21  Gene  Caudate    caudate
ENSG00000204894.4    3.795494e-20  Gene  Caudate    caudate
ENSG00000226278.1    2.993636e-19  Gene  Caudate    caudate
ENSG00000271361.1    4.243489e-19  Gene  Caudate    caudate
ENSG00000230076.1    4.260014e-19  Gene  Caudate    caudate
ENSG00000140263.13   4.487638e-19  Gene  Caudate    caudate
```

[7]:

```
dft[(dft["adj.P.Val"] < 0.05)].to_csv("degs_annotation.txt",
                                      sep='\t', index=True, header=True)
```

1.5 Get random genes based on adjusted P-value

1.5.1 Select genes

[8]:

```
dft.loc[(dft["adj.P.Val"] < 0.05), ["gene_name", "Tissue"]]\
    .groupby("Tissue").count()
```

[8]:

```
               gene_name
Tissue
Caudate             2970
DLPFC               2760
Dentate Gyrus        786
Hippocampus         2956
```

[9]:

```
caudate = dft[(dft["adj.P.Val"] > 0.05) & (dft["Tissue"] == "Caudate")]\
        .sort_values(["adj.P.Val"], ascending=False).head(2970)
print(caudate.shape)
dlpfc = dft[(dft["adj.P.Val"] > 0.05) & (dft["Tissue"] == "DLPFC")]\
        .sort_values(["adj.P.Val"], ascending=False).head(2760)
print(dlpfc.shape)
hippo = dft[(dft["adj.P.Val"] > 0.05) & (dft["Tissue"] == "Hippocampus")]\
        .sort_values(["adj.P.Val"], ascending=False).head(2956)
print(hippo.shape)
gyrus = dft[(dft["adj.P.Val"] > 0.05) & (dft["Tissue"] == "Dentate Gyrus")]\
        .sort_values(["adj.P.Val"], ascending=False).head(786)
print(gyrus.shape)
```

```
(2970, 14)
(2760, 14)
(2956, 14)
(786, 14)
```

1.5.2 Merge data and save

[10]:

```
ran_df = pd.concat([caudate, gyrus, dlpfc, hippo], axis=0)
print(ran_df.shape)
ran_df.head()
```

```
(9472, 14)
```

[10]:

```
                          ensemblID      gene_name    chr     start       end  \
ENSG00000145734.18  ENSG00000145734           BDP1   chr5  71455615  71567820
ENSG00000179262.9   ENSG00000179262         RAD23A  chr19  12945855  12953642
ENSG00000177076.5   ENSG00000177076          ACER2   chr9  19409059  19452020
ENSG00000277954.1   ENSG00000277954  RP11-679B19.1  chr16  79202624  79206739
ENSG00000104228.12  ENSG00000104228         TRIM35   chr8  27284887  27311319

                    length strand         logFC   AveExpr         t  \
ENSG00000145734.18  112205      + -7.434530e-07  7.263799 -0.000035
```

```
ENSG00000179262.9     7787      + -2.245137e-06  5.726987 -0.000098
ENSG00000177076.5    42961      + -9.871997e-06  2.633614 -0.000197
ENSG00000277954.1     4115      -  3.533458e-05  1.977845  0.000542
ENSG00000104228.12   26432      -  1.857056e-05  4.853768  0.000635

                    adj.P.Val  Type   Tissue New_Tissue
ENSG00000145734.18   0.999972  Gene  Caudate    caudate
ENSG00000179262.9    0.999967  Gene  Caudate    caudate
ENSG00000177076.5    0.999932  Gene  Caudate    caudate
ENSG00000277954.1    0.999702  Gene  Caudate    caudate
ENSG00000104228.12   0.999673  Gene  Caudate    caudate
```

[11]:

```
ran_df.to_csv("randomGenes_annotation.txt", sep='\t', index=True, header=True)
```

[ ]: