main

August 21, 2021

1 Generate supplementary data for DEGs

[1]:
```
import functools
import numpy as np
import pandas as pd
from scipy.stats import fisher_exact
from statsmodels.stats.multitest import multipletests
```

1.1 Function

1.1.1 Cached functions

[2]:
```
@functools.lru_cache()
def get_de():
    return pd.read_csv("../../deg_summary/_m/diffExpr_ancestry_full_4regions.
 ↪tsv", sep='\t')


@functools.lru_cache()
def feature_map(feature):
    return {"genes": "Gene", "transcripts": "Transcript",
            "exons": "Exon", "junctions": "Junction"}[feature]
```

1.1.2 Simple functinos

[3]:
```
def get_tissues_DS(tissue):
    cols = ["Feature", "gencodeID", "ensemblID", "Symbol", "logFC",
            "AveExpr", "t", "P.Value", "adj.P.Val", "Type"]
    tissue_map = {"Caudate": "caudate", "Dentate Gyrus": "dentateGyrus",
                  "DLPFC": "dlpfc", "Hippocampus": "hippocampus"}
    ## Load local splicing results
    ds = pd.read_csv("../../../%s/localsplicing/visualization/_m/
 ↪cluster_ds_results_annotated.txt" %
                     tissue_map[tissue], sep='\t')\
           .rename(columns={"gene_name": "Symbol"})
    ds['chr'] = ds.coord.str.split(":", expand=True)[0]
    ds["Type"] = "DTU"
    ds["Tissue"] = tissue
```

```
    return ds


def print_overlap(feature, tissue, ds):
    dx = get_de()[(get_de()["adj.P.Val"] < 0.05) &
                  (get_de()["Type"] == feature_map(feature)) &
                  (get_de()["Tissue"] == tissue)].copy()
    dft = ds[(ds["Tissue"] == tissue)].copy()
    overlap = len(set(dx.Symbol) & set(dft.gene))
    print("There are {} ({:.1%}) DS overlapping DE {}."\
          .format(overlap, overlap/len(dft.gene.unique()), feature))


def subset_de(feature, tissue):
    dx = get_de()[(get_de()["Type"] == feature_map(feature)) &
                  (get_de()["Tissue"] == tissue)].copy()
    return dx.loc[:, ["Symbol", "adj.P.Val"]].dropna()


def merge_data(feature, tissue, ds):
    df = ds[(ds["Tissue"] == tissue)].copy()
    return df.loc[:, ["gene", "FDR", "annotation"]].
 ↪drop_duplicates(subset="gene")\
             .merge(subset_de(feature, tissue), left_on="gene",␣
 ↪right_on="Symbol", how="outer")\
             .fillna(1)


def cal_fishers(feature, tissue, ds):
    dt = merge_data(feature, tissue, ds)
    table = [[np.sum((dt['adj.P.Val']<0.05) & ((dt['FDR']<0.05))),
              np.sum((dt['adj.P.Val']<0.05) & ((dt['FDR']>=0.05)))],
             [np.sum((dt['adj.P.Val']>=0.05) & ((dt['FDR']<0.05))),
              np.sum((dt['adj.P.Val']>=0.05) & ((dt['FDR']>=0.05)))]]
    #print(table)
    return fisher_exact(table)
```

1.2 Summary

[4]:
```
caudate = get_tissues_DS("Caudate")
dlpfc = get_tissues_DS("DLPFC")
gyrus = get_tissues_DS("Dentate Gyrus")
hippo = get_tissues_DS("Hippocampus")
ds = pd.concat([caudate, dlpfc, gyrus, hippo], axis=0)
print(ds.shape)
ds.head(2)
```

```
(5233, 9)
```

[4]:
```
      clusterID   N                       coord gene annotation            FDR  \
0  clu_128031_-  14  chr12:124911899-124913724  UBC    cryptic   0.000000e+00
1  clu_105375_?  13  chr12:124911899-124913724  UBC    cryptic  2.510000e-170

     chr Type   Tissue
0  chr12  DTU  Caudate
1  chr12  DTU  Caudate
```

[5]:
```
ds.groupby(["Tissue"]).size()
```

[5]:
```
Tissue
Caudate          1901
DLPFC            1345
Dentate Gyrus     655
Hippocampus      1332
dtype: int64
```

[6]:
```
ds.groupby(["Tissue", "gene"]).first().reset_index().groupby(["Tissue"]).size()
```

[6]:
```
Tissue
Caudate          1541
DLPFC            1105
Dentate Gyrus     532
Hippocampus      1124
dtype: int64
```

1.2.1 Save files

[7]:
```
ds.to_csv("diffSplicing_ancestry_FDR05_4regions.tsv", sep='\t', index=False)
```

1.3 Overlap with DE

[8]:
```
for tissue in ["Caudate", "Dentate Gyrus", "DLPFC", "Hippocampus"]:
    print(tissue)
    for feature in ["genes", "transcripts", "exons", "junctions"]:
        print_overlap(feature, tissue, ds)
    print("")
```

```
Caudate
There are 357 (23.2%) DS overlapping DE genes.
There are 594 (38.5%) DS overlapping DE transcripts.
There are 684 (44.4%) DS overlapping DE exons.
There are 654 (42.4%) DS overlapping DE junctions.

Dentate Gyrus
There are 48 (9.0%) DS overlapping DE genes.
```

```
There are 62 (11.7%) DS overlapping DE transcripts.
There are 89 (16.7%) DS overlapping DE exons.
There are 92 (17.3%) DS overlapping DE junctions.

DLPFC
There are 239 (21.6%) DS overlapping DE genes.
There are 388 (35.1%) DS overlapping DE transcripts.
There are 479 (43.3%) DS overlapping DE exons.
There are 411 (37.2%) DS overlapping DE junctions.

Hippocampus
There are 216 (19.2%) DS overlapping DE genes.
There are 399 (35.5%) DS overlapping DE transcripts.
There are 513 (45.6%) DS overlapping DE exons.
There are 410 (36.5%) DS overlapping DE junctions.
```

1.4 Enrichment analysis

[9]:
```
feature_lt = []; pval_lt = []; oddratio_lt = []; tissue_lt = []
for tissue in ["Caudate", "Dentate Gyrus", "DLPFC", "Hippocampus"]:
    print(tissue)
    for feature in ["genes", "transcripts", "exons", "junctions"]:
        oddratio, pval = cal_fishers(feature, tissue, ds)
        feature_lt.append(feature_map(feature))
        pval_lt.append(pval)
        oddratio_lt.append(oddratio)
        tissue_lt.append(tissue)
        print("Enrichment of DS within DE for {}:\nOdd Ratio: {:.2f}; P-value:␣
 ↪{:.1e}\n"\
              .format(feature_map(feature), oddratio, pval))
    print("")
pval_lt = [1e-323 if x == 0 else x for x in pval_lt]
_, fdr, _, _ = multipletests(pval_lt, method='fdr_bh')
df = pd.DataFrame({"Tissue": tissue_lt, "Feature": feature_lt,
                   "OR": oddratio_lt, "PValue": pval_lt, "FDR": fdr})
df.to_csv('diffSplice_enrichment_analysis.txt', sep='\t', index=False)
```

```
Caudate
Enrichment of DS within DE for Gene:
Odd Ratio: 2.11; P-value: 1.1e-27

Enrichment of DS within DE for Transcript:
Odd Ratio: 2.84; P-value: 1.5e-139

Enrichment of DS within DE for Exon:
Odd Ratio: 2.77; P-value: 0.0e+00
```

Enrichment of DS within DE for Junction:
Odd Ratio: 3.38; P-value: 0.0e+00

Dentate Gyrus
Enrichment of DS within DE for Gene:
Odd Ratio: 2.66; P-value: 2.1e-08

Enrichment of DS within DE for Transcript:
Odd Ratio: 4.17; P-value: 1.0e-23

Enrichment of DS within DE for Exon:
Odd Ratio: 3.90; P-value: 7.2e-106

Enrichment of DS within DE for Junction:
Odd Ratio: 5.08; P-value: 2.7e-66

DLPFC
Enrichment of DS within DE for Gene:
Odd Ratio: 1.98; P-value: 2.9e-17

Enrichment of DS within DE for Transcript:
Odd Ratio: 3.32; P-value: 4.8e-113

Enrichment of DS within DE for Exon:
Odd Ratio: 3.11; P-value: 0.0e+00

Enrichment of DS within DE for Junction:
Odd Ratio: 4.06; P-value: 1.2e-252

Hippocampus
Enrichment of DS within DE for Gene:
Odd Ratio: 1.57; P-value: 3.4e-08

Enrichment of DS within DE for Transcript:
Odd Ratio: 2.65; P-value: 1.6e-83

Enrichment of DS within DE for Exon:
Odd Ratio: 2.52; P-value: 0.0e+00

Enrichment of DS within DE for Junction:
Odd Ratio: 3.50; P-value: 8.8e-221

[ ]: