main

July 9, 2021

1 Tissue comparison for differential expression analysis

[1]:
```
import functools
import numpy as np
import pandas as pd
from plotnine import *

from warnings import filterwarnings
from matplotlib.cbook import mplDeprecation
filterwarnings('ignore', category=mplDeprecation)
filterwarnings('ignore', category=UserWarning, module='plotnine.*')
filterwarnings('ignore', category=DeprecationWarning, module='plotnine.*')
```

[2]:
```
config = {
    'caudate': '../../../_m/genes/diffExpr_szVctl_full.txt',
    'dlpfc': '/ceph/projects/v4_phase3_paper/inputs/public_data/_m/phase2/
 ↪dlpfc_diffExpr_szVctl_full.txt',
    'hippo': '/ceph/projects/v4_phase3_paper/inputs/public_data/_m/phase2/
 ↪hippo_diffExpr_szVctl_full.txt',
}
```

[3]:
```
@functools.lru_cache()
def get_deg(filename):
    dft = pd.read_csv(filename, sep='\t', index_col=0)
    dft['Feature'] = dft.index
    dft['Dir'] = np.sign(dft['t'])
    if 'gene_id' in dft.columns:
        dft['ensemblID'] = dft.gene_id.str.replace('\\..*', '', regex=True)
    elif 'ensembl_gene_id' in dft.columns:
        dft.rename(columns={'ensembl_gene_id': 'ensemblID'}, inplace=True)
    return dft[['Feature', 'ensemblID', 'adj.P.Val', 'logFC', 't', 'Dir']]

@functools.lru_cache()
def get_deg_sig(filename):
    dft = get_deg(filename)
    return dft[(dft['adj.P.Val'] < 0.05)]
```

```
@functools.lru_cache()
def merge_dataframes(tissue1, tissue2):
    return get_deg(config[tissue1]).merge(get_deg(config[tissue2]),
                                          on='Feature',
                                          suffixes=['_%s' % tissue1, '_%s' %␣
 ↪tissue2])


@functools.lru_cache()
def merge_dataframes_sig(tissue1, tissue2):
    return get_deg_sig(config[tissue1]).merge(get_deg_sig(config[tissue2]),
                                              on='Feature',
                                              suffixes=['_%s' % tissue1, '_%s'␣
 ↪% tissue2])
```

[4]:
```
def tissue_annotation(tissue):
    return {'dlpfc': 'DLPFC', 'hippo': 'Hippocampus',
            'caudate': 'Caudate'}[tissue]


def save_plot(p, fn, width=7, height=7):
    '''Save plot as svg, png, and pdf with specific label and dimension.'''
    for ext in ['.svg', '.png', '.pdf']:
        p.save(fn+ext, width=width, height=height)
```

1.1 BrainSeq Comparison

[5]:
```
caudate = get_deg(config['caudate'])
caudate.groupby('Dir').size()
```

[5]:
```
Dir
-1.0    12061
 1.0    10897
dtype: int64
```

[6]:
```
caudate[(caudate['adj.P.Val'] < 0.05)].shape
```

[6]:
```
(2701, 6)
```

[7]:
```
dlpfc = get_deg(config['dlpfc'])
dlpfc.groupby('Dir').size()
```

[7]:
```
Dir
-1.0    13207
 1.0    11445
dtype: int64
```

[8]:
```
dlpfc[(dlpfc['adj.P.Val'] < 0.05)].shape
```

[8]:
```
(245, 6)
```

[9]:
```
hippo = get_deg(config['hippo'])
hippo.groupby('Dir').size()
```

[9]:
```
Dir
-1.0    12852
 1.0    11800
dtype: int64
```

[10]:
```
hippo[(hippo['adj.P.Val'] < 0.05)].shape
```

[10]:
```
(48, 6)
```

1.1.1 Upset Plot

[11]:
```
phase2_dlpfc = dlpfc[(dlpfc['adj.P.Val'] < 0.05)].copy()
phase2_dlpfc['DLPFC'] = 1
phase2_dlpfc = phase2_dlpfc[['ensemblID', 'DLPFC']]

phase2_hippo = hippo[(hippo['adj.P.Val'] < 0.05)].copy()
phase2_hippo['Hippocampus'] = 1
phase2_hippo = phase2_hippo[['ensemblID', 'Hippocampus']]

phase3_caudate = caudate[(caudate['adj.P.Val'] < 0.05)].copy()
phase3_caudate['Caudate'] = 1
phase3_caudate = phase3_caudate[['ensemblID', 'Caudate']]
```

[12]:
```
geneList = pd.merge(phase3_caudate[['ensemblID']],
                    phase2_dlpfc[['ensemblID']],
                    on=['ensemblID'], how='outer')\
             .merge(phase2_hippo[['ensemblID']],
                    on=['ensemblID'], how='outer')\
             .groupby(['ensemblID']).first().reset_index()
## Caudate
newC = pd.merge(geneList, phase3_caudate, on=['ensemblID'],
                how='outer').fillna(0)
newC['Caudate'] = newC['Caudate'].astype('int')
## DLPFC
newD1 = pd.merge(geneList, phase2_dlpfc, on=['ensemblID'],
                 how='outer').fillna(0)
newD1['DLPFC'] = newD1['DLPFC'].astype('int')
## Hippocampus
newH = pd.merge(geneList, phase2_hippo, on=['ensemblID'],
                how='outer').fillna(0)
```

```
newH['Hippocampus'] = newH['Hippocampus'].astype('int')

print(newC.shape, newH.shape, newD1.shape)
```

```
(2929, 2) (2929, 2) (2929, 2)
```

[13]:
```
df = pd.concat([newC.set_index(['ensemblID']),
                newD1.set_index(['ensemblID']),
                newH.set_index(['ensemblID'])],
               axis=1, join='outer')
df.head(2)
```

[13]:

ensemblID	Caudate	DLPFC	Hippocampus
ENSG00000001084	1	0	0
ENSG00000001497	1	0	0

[14]:
```
%load_ext rpy2.ipython
```

[15]:
```
%%R
#library(UpSetR)
#upset(df, order.by="freq", text.scale=c(3, 2.5, 2.4, 2.25, 2.6, 2.6), point.
 ↪size=3.6, line.size=1.4)
library(ComplexHeatmap)
subset_pvalue <- function(filename, fdr_cutoff){
    df <- subset(read.delim(filename, row.names=1, stringsAsFactors = F),
                 adj.P.Val < fdr_cutoff)
    if('gene_id' %in% colnames(df)){
        df$ensemblID <- gsub('\\..*', '', df$gene_id)
    } else if('ensembl_gene_id' %in% colnames(df)){
        df <- dplyr::rename(df, ensemblID=ensembl_gene_id)
    }
    return(df$ensemblID)
}

caudate = subset_pvalue('../../../_m/genes/diffExpr_szVctl_full.txt', 0.05)
dlpfc = subset_pvalue('/ceph/projects/v4_phase3_paper/inputs/public_data/_m/
 ↪phase2/dlpfc_diffExpr_szVctl_full.txt', 0.05)
hippo = subset_pvalue('/ceph/projects/v4_phase3_paper/inputs/public_data/_m/
 ↪phase2/hippo_diffExpr_szVctl_full.txt', 0.05)

lt = list(Caudate = caudate,
          DLPFC = dlpfc,
          Hippocampus = hippo)

m = make_comb_mat(lt)
cbb_palette <- c("#000000", "#E69F00", "#56B4E9", "#009E73",
```

```
                "#F0E442", "#0072B2", "#D55E00", "#CC79A7")
```

```
R[write to console]: Loading required package: grid

R[write to console]: ========================================
ComplexHeatmap version 2.6.2
Bioconductor page: http://bioconductor.org/packages/ComplexHeatmap/
Github page: https://github.com/jokergoo/ComplexHeatmap
Documentation: http://jokergoo.github.io/ComplexHeatmap-reference

If you use it in published research, please cite:
Gu, Z. Complex heatmaps reveal patterns and correlations in multidimensional
  genomic data. Bioinformatics 2016.

This message can be suppressed by:
  suppressPackageStartupMessages(library(ComplexHeatmap))
========================================
```

[16]:
```
%%R
right_annot = upset_right_annotation(
    m, ylim = c(0, 3000),
    gp = gpar(fill = "black"),
    annotation_name_side = "top",
    axis_param = list(side = "top"))

top_annot = upset_top_annotation(
    m, height=unit(7, "cm"),
    ylim = c(0, 3000),
    gp=gpar(fill=cbb_palette[comb_degree(m)]),
    annotation_name_rot = 90)

pdf('BrainSeq_dx_tissue_upsetR_DEgenes.pdf', width=8, height=4)
ht = draw(UpSet(m, pt_size=unit(4, "mm"), lwd=3,
                comb_col=cbb_palette[comb_degree(m)],
                set_order = c("Caudate", "DLPFC", "Hippocampus"),
                comb_order = order(-comb_size(m)),
                row_names_gp = gpar(fontsize = 14, fontface='bold'),
                right_annotation = right_annot,
                top_annotation = top_annot))
od = column_order(ht)
cs = comb_size(m)
decorate_annotation("intersection_size", {
    grid.text(cs[od], x = seq_along(cs), y = unit(cs[od], "native") +
              unit(6, "pt"),
        default.units = "native", just = "bottom", gp = gpar(fontsize = 11))
```

```
})
dev.off()

svg('BrainSeq_dx_tissue_upsetR_DEgenes.svg', width=8, height=4)
ht = draw(UpSet(m, pt_size=unit(4, "mm"), lwd=3,
                comb_col=cbb_palette[comb_degree(m)],
                set_order = c("Caudate", "DLPFC", "Hippocampus"),
                comb_order = order(-comb_size(m)),
                row_names_gp = gpar(fontsize = 14, fontface='bold'),
                right_annotation = right_annot,
                top_annotation = top_annot))
od = column_order(ht)
cs = comb_size(m)
decorate_annotation("intersection_size", {
    grid.text(cs[od], x = seq_along(cs), y = unit(cs[od], "native") +
              unit(6, "pt"),
        default.units = "native", just = "bottom", gp = gpar(fontsize = 11))
})
dev.off()
```

```
png
  2
```

[17]:
```
%%R
right_ha = rowAnnotation(
    "Intersection\nsize" = anno_barplot(comb_size(m), border=F,
                                        ylim = c(0, 3000),
                                        ␣
 ↪gp=gpar(fill=cbb_palette[comb_degree(m)]),
                                        width = unit(7, "cm")))
top_ha = HeatmapAnnotation(
    "Set size" = anno_barplot(set_size(m), border=F,
                              ylim = c(0, 3000),
                              gp = gpar(fill = "black"),
                              height = unit(2, "cm")),
    gap = unit(2, "mm"), annotation_name_side = "left",
    annotation_name_rot = 90)


pdf("BrainSeq_dx_tissue_upsetR_DEgenes_transpose.pdf", width=5, height=10)
ht = draw(UpSet(t(m), pt_size=unit(5, "mm"), lwd=3,
                comb_order = order(-comb_size(m)),
                comb_col=cbb_palette[comb_degree(m)],
                set_order = c("Caudate", "DLPFC", "Hippocampus"),
                column_names_gp = gpar(fontsize = 16, fontface='bold'),
                right_annotation = right_ha, top_annotation=top_ha))
```

```
od = rev(row_order(ht))
cs = comb_size(m)
decorate_annotation("Intersection\nsize", {
    grid.text(cs[od], y = seq_along(cs), x = unit(cs[od], "native") +
              unit(6, "pt"),
        default.units = "native", just = "left", gp = gpar(fontsize = 11))
})
dev.off()

svg("BrainSeq_dx_tissue_upsetR_DEgenes_transpose.svg", width=5, height=10)
ht = draw(UpSet(t(m), pt_size=unit(5, "mm"), lwd=3,
                comb_order = order(-comb_size(m)),
                comb_col=cbb_palette[comb_degree(m)],
                set_order = c("Caudate", "DLPFC", "Hippocampus"),
                column_names_gp = gpar(fontsize = 16, fontface='bold'),
                right_annotation = right_ha, top_annotation=top_ha))

od = rev(row_order(ht))
cs = comb_size(m)
decorate_annotation("Intersection\nsize", {
    grid.text(cs[od], y = seq_along(cs), x = unit(cs[od], "native") +
              unit(6, "pt"),
        default.units = "native", just = "left", gp = gpar(fontsize = 11))
})
dev.off()
```

```
png
  2
```

[18]:
```
import functools
from gtfparse import read_gtf
```

[19]:
```
@functools.lru_cache()
def get_gtf(gtf_file):
    return read_gtf(gtf_file)
```

[20]:
```
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"]]

gtf_file = '/ceph/genome/human/gencode25/gtf.CHR/_m/gencode.v25.annotation.gtf'
gtf_annot = gene_annotation(gtf_file, 'gene')
gtf_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']
```

[20]:
```
              gene_id gene_name transcript_id exon_id  \
0   ENSG00000223972.5   DDX11L1
12  ENSG00000227232.5    WASH7P

                            gene_type seqname  start    end strand
0   transcribed_unprocessed_pseudogene    chr1  11869  14409      +
12              unprocessed_pseudogene    chr1  14404  29570      -
```

[21]:
```
dft = caudate.merge(gtf_annot[['gene_id', 'gene_name', 'seqname']],
                    left_index=True, right_on='gene_id')
dft.head()
```

[21]:
```
                    Feature        ensemblID     adj.P.Val     logFC  \
699321    ENSG00000248587.7  ENSG00000248587  1.387742e-26  0.801502
2481489   ENSG00000138944.7  ENSG00000138944  1.707516e-24  0.563733
2362233  ENSG00000185052.11  ENSG00000185052  3.972599e-21  0.291763
1705114  ENSG00000140015.19  ENSG00000140015  6.716497e-18  0.515655
2551916  ENSG00000171004.17  ENSG00000171004  3.196203e-16  0.302105

                 t  Dir             gene_id gene_name seqname
699321   12.696887  1.0   ENSG00000248587.7  GDNF-AS1    chr5
2481489  12.073351  1.0   ENSG00000138944.7  KIAA1644   chr22
2362233  11.122852  1.0  ENSG00000185052.11   SLC24A3   chr20
1705114  10.185331  1.0  ENSG00000140015.19     KCNH5   chr14
2551916   9.670025  1.0  ENSG00000171004.17    HS6ST2    chrX
```

[22]:
```
shared_df = dft.loc[:, ['gene_id', 'ensemblID', 'seqname', 'gene_name', 'Dir']]\
        .merge(pd.DataFrame({'ensemblID': list(set(phase2_dlpfc['ensemblID']) &
                                               set(phase2_hippo['ensemblID']) &
                                               ␣
 ↪set(phase3_caudate['ensemblID']))}),
               on='ensemblID')
#shared_df.to_csv('BrainSeq_shared_degs_annotation.txt', sep='\t', index=False,␣
 ↪header=True)
shared_df.head()
```

[22]:
```
Empty DataFrame
Columns: [gene_id, ensemblID, seqname, gene_name, Dir]
Index: []
```

[23]:
```
gtf_annot['ensemblID'] = gtf_annot.gene_id.str.replace("\\..*", "", regex=True)
gtf_annot[['gene_id', 'ensemblID', 'gene_name', 'seqname', 'gene_type']]\
    .merge(df, left_on='ensemblID', right_index=True)\
    .to_csv('brainseq_deg_across_tissues_comparison.csv', index=False)
```

[ ]: