main_plot

August 23, 2021

1 Visualize GO analysis

[1]:
```
import numpy as np
import pandas as pd
```

[2]:
```
def get_top_GO(fn, label):
    df = pd.read_excel(fn).sort_values('p_uncorrected').head(10)
    df['Log10'] = -np.log10(df['p_fdr_bh'])
    df['Tissue'] = label
    return df
```

[3]:
```
config = {
    'Caudate': 'GO_analysis_mashr_caudate.xlsx',
    'Dentate Gyrus': 'GO_analysis_mashr_dentate_gyrus.xlsx',
    'DLPFC': 'GO_analysis_mashr_dlpfc.xlsx',
    'Hippocampus': 'GO_analysis_mashr_hippocampus.xlsx',
}

df = pd.DataFrame()
for bias in ["Caudate", "Dentate Gyrus", "DLPFC", "Hippocampus"]:
    df = pd.concat([df, get_top_GO(config[bias], bias)], axis=0)

fac = []
for ii in range(df.shape[0]):
    xx, yy = df[['ratio_in_study']].iloc[ii, 0].split('/')
    fac.append((int(xx) / int(yy)) * 2)

df['geneRatio'] = fac
print(np.min(fac), np.max(fac))
```

```
0.020497803806734993 1.302325581395349
```

[4]:
```
df.to_csv("GO_analysis_top10_mashr.tsv", sep='\t', index=False)
```

1.1 Plot

[5]:
```
%load_ext rpy2.ipython
```

[6]:
```
%%R -i df
library(ggplot2)
library(tidyverse)

save_plot <- function(p, fn, w, h){
    for(ext in c('.svg', '.png', '.pdf')){
        ggsave(file=paste0(fn,ext), plot=p, width=w, height=h)
    }
}

plot_GO <- function(){
    cbPalette <- ggpubr::get_palette(palette = "npg", 4)
    gg1 = df %>%
        ggplot(aes(x=Log10, y=name, color=Tissue, size=geneRatio)) +
        geom_point(shape=18, alpha=0.8) + labs(y='', x='-Log10 (FDR)') +
        theme_bw(base_size=15) +
        scale_colour_manual(name="Ancestry", values=cbPalette,
                            labels=c("Caudate", "Dentate Gyrus",
                                     "DLPFC", "Hippocampus")) +
        geom_vline(xintercept = -log10(0.05), linetype = "dotted") +
        theme(axis.title=element_text(face='bold'),
              strip.text=element_text(face='bold'))
    return(gg1)
}
```

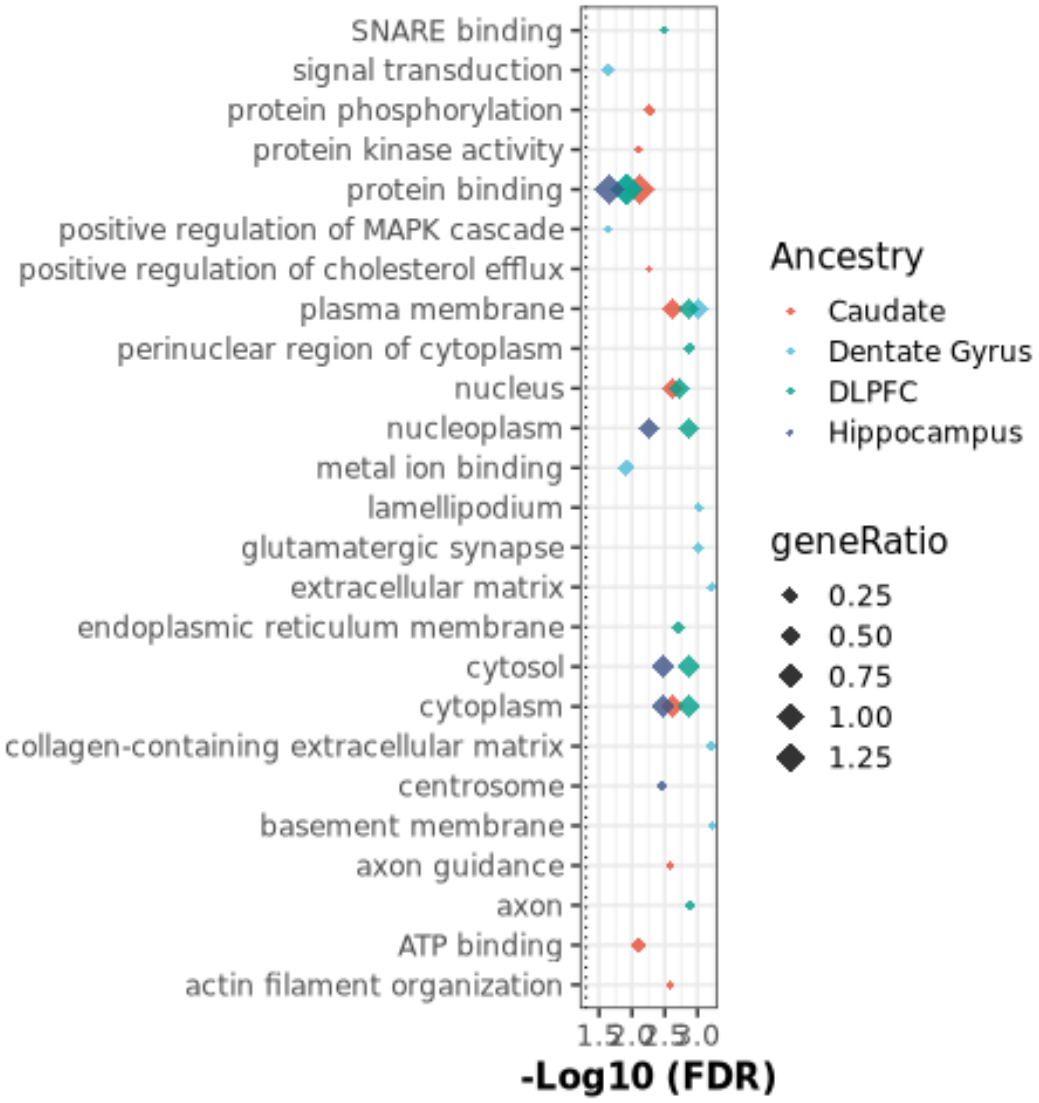
```
R[write to console]: Want to understand how all the pieces fit together? Read R
for Data
Science: https://r4ds.had.co.nz/

R[write to console]:   Attaching packages
                      tidyverse 1.3.1

R[write to console]:  tibble  3.1.2      dplyr   1.0.7
 tidyr   1.1.3      stringr 1.4.0
 readr   1.4.0      forcats 0.5.1
 purrr   0.3.4

R[write to console]:   Conflicts
tidyverse_conflicts()
 dplyr::filter() masks stats::filter()
 dplyr::lag()    masks stats::lag()
```

[7]:
```
%%R
gg1 = plot_GO()
print(gg1)
save_plot(gg1, "ancestry_mashr_GO_top10_stacked", 8, 6)
```


SNARE binding
signal transduction
protein phosphorylation
protein kinase activity
protein binding
positive regulation of MAPK cascade
positive regulation of cholesterol efflux
plasma membrane
perinuclear region of cytoplasm
nucleus
nucleoplasm
metal ion binding
lamellipodium
glutamatergic synapse
extracellular matrix
endoplasmic reticulum membrane
cytosol
cytoplasm
collagen-containing extracellular matrix
centrosome
basement membrane
axon guidance
axon
ATP binding
actin filament organization
1.5
2.0
2.5
3.0
-Log10 (FDR)
Ancestry
Caudate
Dentate Gyrus
DLPFC
Hippocampus
geneRatio
0.25
0.50
0.75
1.00
1.25

[ ]: