main

July 13, 2021

1 Tissue comparison for differential expression analysis

[1]:
```
import functools
import numpy as np
import pandas as pd
from plotnine import *
from scipy.stats import binom_test, fisher_exact, linregress

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': '../../../caudate/_m/genes/diffExpr_maleVfemale_full.txt',
    'dlpfc': '../../../dlpfc/_m/genes/diffExpr_maleVfemale_full.txt',
    'hippo': '../../../hippocampus/_m/genes/diffExpr_maleVfemale_full.txt',
    'cmc_dlpfc': '../../../cmc_dlpfc/_m/mssm_penn_pitt_maleVfemale.tsv',
    'cmc_hbcc': "../../../cmc_dlpfc/_m/nimh_hbcc_maleVfemale.tsv",
}
```

[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'])
    dft['ensemblID'] = dft.Feature.str.replace('\\..*', '', regex=True)
    return dft[['Feature', 'ensemblID', 'adj.P.Val', 'logFC', 't', 'Dir']]


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

```
@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):
    fdr = 0.05
    return get_deg_sig(config[tissue1], fdr).merge(get_deg_sig(config[tissue2],␣
 ↪fdr),
                                                   on='Feature',
                                                   suffixes=['_%s' % tissue1,␣
 ↪'_%s' % tissue2])
```

[4]:
```
def enrichment_binom(tissue1, tissue2, merge_fnc):
    df = merge_fnc(tissue1, tissue2)
    df['agree'] = df['Dir_%s' % tissue1] * df['Dir_%s' % tissue2]
    dft = df.groupby('agree').size().reset_index()
    print(dft)
    return binom_test(dft[0].iloc[1], dft[0].sum()) if dft.shape[0] != 1 else␣
 ↪print("All directions agree!")


def cal_fishers(tissue1, tissue2):
    df = merge_dataframes(tissue1, tissue2)
    fdr = 0.05
    table = [[np.sum((df['adj.P.Val_%s' % tissue1]<fdr) &
                     ((df['adj.P.Val_%s' % tissue2]<fdr))),
              np.sum((df['adj.P.Val_%s' % tissue1]<fdr) &
                     ((df['adj.P.Val_%s' % tissue2]>=fdr)))],
             [np.sum((df['adj.P.Val_%s' % tissue1]>=fdr) &
                     ((df['adj.P.Val_%s' % tissue2]<fdr))),
              np.sum((df['adj.P.Val_%s' % tissue1]>=fdr) &
                     ((df['adj.P.Val_%s' % tissue2]>=fdr)))]]
    print(table)
    return fisher_exact(table)


def calculate_corr(xx, yy):
    '''This calculates R^2 correlation via linear regression:
         - used to calculate relationship between 2 arrays
         - the arrays are principal components 1 or 2 (PC1, PC2) AND gender
         - calculated on a scale of 0 to 1 (with 0 being no correlation)
        Inputs:
```

```
            x: array of Gender (converted to binary output)
            y: array of PC
        Outputs:
            1. r2
            2. p-value, two-sided test
              - whose null hypothesis is that two sets of data are uncorrelated
            3. slope (beta): directory of correlations
    '''
    slope, intercept, r_value, p_value, std_err = linregress(xx, yy)
    return r_value, p_value


def corr_annotation(tissue1, tissue2, merge_fnc):
    dft = merge_fnc(tissue1, tissue2)
    xx = dft['t_%s' % tissue1]
    yy = dft['t_%s' % tissue2]
    r_value1, p_value1 = calculate_corr(xx, yy)
    return 'R2: %.2f\nP-value: %.2e' % (r_value1**2, p_value1)


def tissue_annotation(tissue):
    return {'dlpfc': 'DLPFC', 'hippo': 'Hippocampus',
            'caudate': 'Caudate', 'cmc_dlpfc': 'CMC DLPFC',
            "cmc_hbcc": "CMC DLPFC: HBCC"}[tissue]
```

[5]:
```
def plot_corr_impl(tissue1, tissue2, merge_fnc):
    dft = merge_fnc(tissue1, tissue2)
    title = '\n'.join([corr_annotation(tissue1, tissue2, merge_fnc)])
    xlab = 'T-statistic (%s)' % tissue_annotation(tissue1)
    ylab = 'T-statistic (%s)' % tissue_annotation(tissue2)
    pp = ggplot(dft, aes(x='t_%s'%tissue1, y='t_%s' % tissue2))\
    + geom_point(alpha=0.75, size=3)\
    + theme_matplotlib()\
    + theme(axis_text=element_text(size=18),
            axis_title=element_text(size=20, face='bold'),
            plot_title=element_text(size=22))
    pp += labs(x=xlab, y=ylab, title=title)
    return pp


def plot_corr(tissue1, tissue2, merge_fnc):
    return plot_corr_impl(tissue1, tissue2, merge_fnc)


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 Tissue Comparison

[6]:
```
caudate = get_deg(config['caudate'])
caudate.groupby('Dir').size()
```

[6]:
```
Dir
-1.0    11133
 1.0    12355
dtype: int64
```

[7]:
```
caudate[(caudate['adj.P.Val'] < 0.05)].shape
```

[7]:
```
(380, 6)
```

[8]:
```
dlpfc = get_deg(config['dlpfc'])
dlpfc.groupby('Dir').size()
```

[8]:
```
Dir
-1.0    11240
 1.0    11799
dtype: int64
```

[9]:
```
dlpfc[(dlpfc['adj.P.Val'] < 0.05)].shape
```

[9]:
```
(573, 6)
```

[10]:
```
hippo = get_deg(config['hippo'])
hippo.groupby('Dir').size()
```

[10]:
```
Dir
-1.0    11840
 1.0    11150
dtype: int64
```

[11]:
```
hippo[(hippo['adj.P.Val'] < 0.05)].shape
```

[11]:
```
(105, 6)
```

1.1.1 Enrichment of DEG

[12]:
```
cal_fishers('caudate', 'dlpfc')
```

```
[[117, 236], [428, 21171]]
```

[12]:
```
(24.52287937589102, 3.2462962516016504e-100)
```

[13]:
```
cal_fishers('caudate', 'hippo')
```

```
[[85, 270], [16, 21688]]
```

[13]: (426.73148148148147, 9.812269867491168e-140)

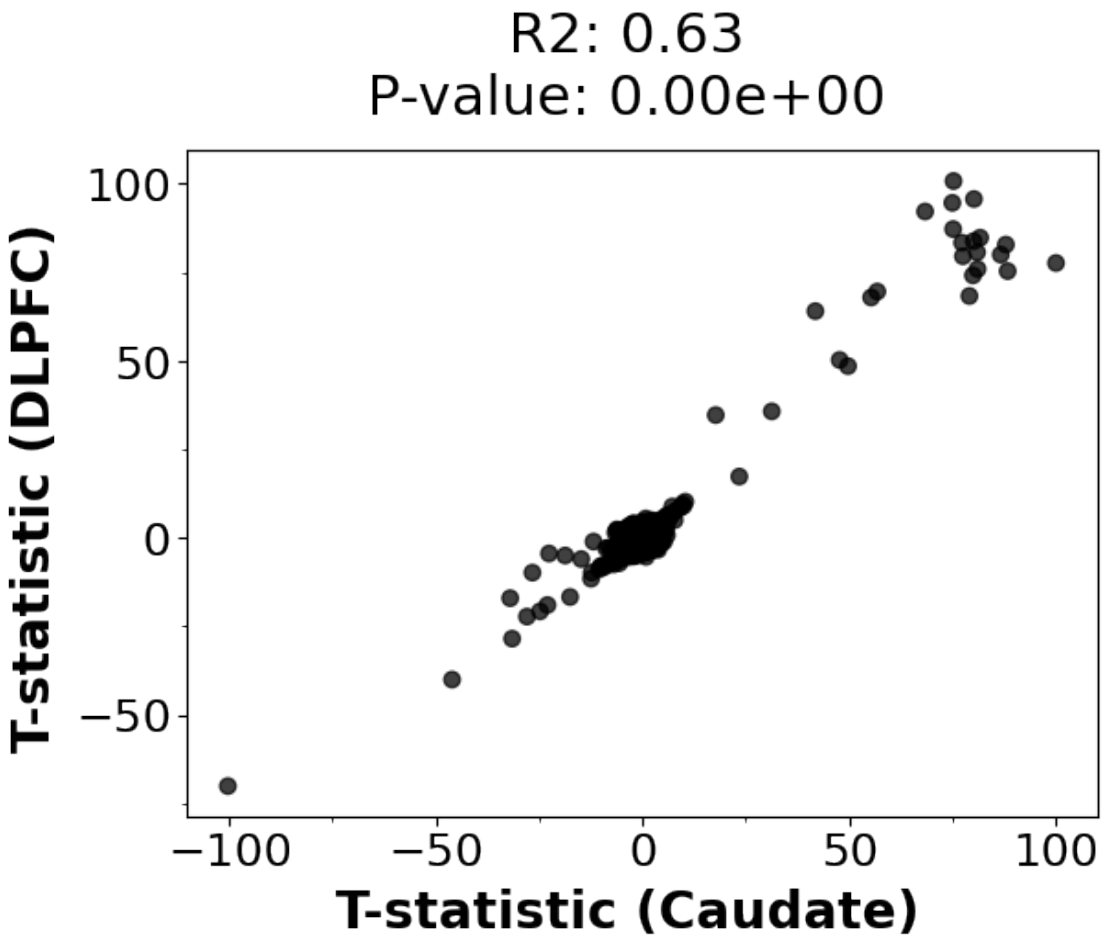
[14]:
```
cal_fishers('dlpfc', 'hippo')
```

```
[[81, 474], [18, 21662]]
```

[14]: (205.65189873417722, 6.409439839482686e-114)

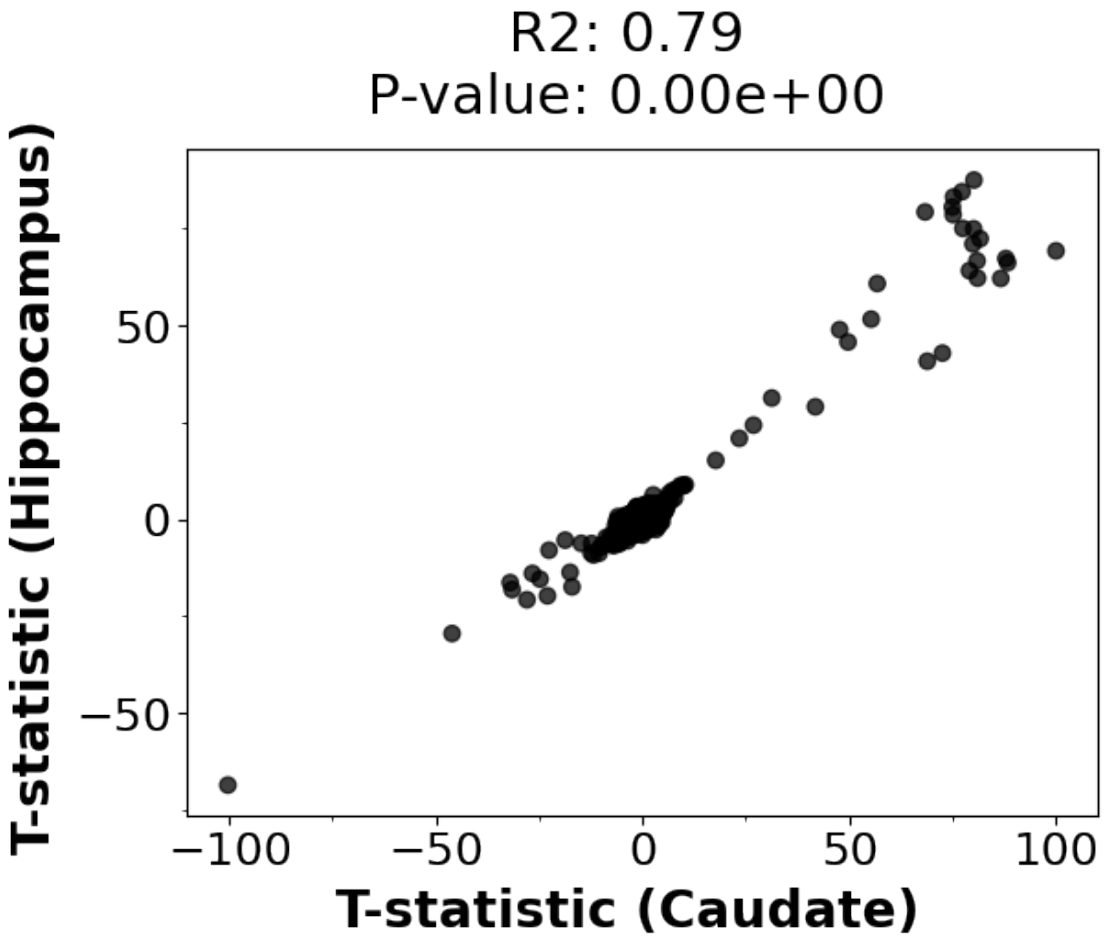
1.1.2 Correlation

[15]:
```
pp = plot_corr('caudate', 'dlpfc', merge_dataframes)
pp
```



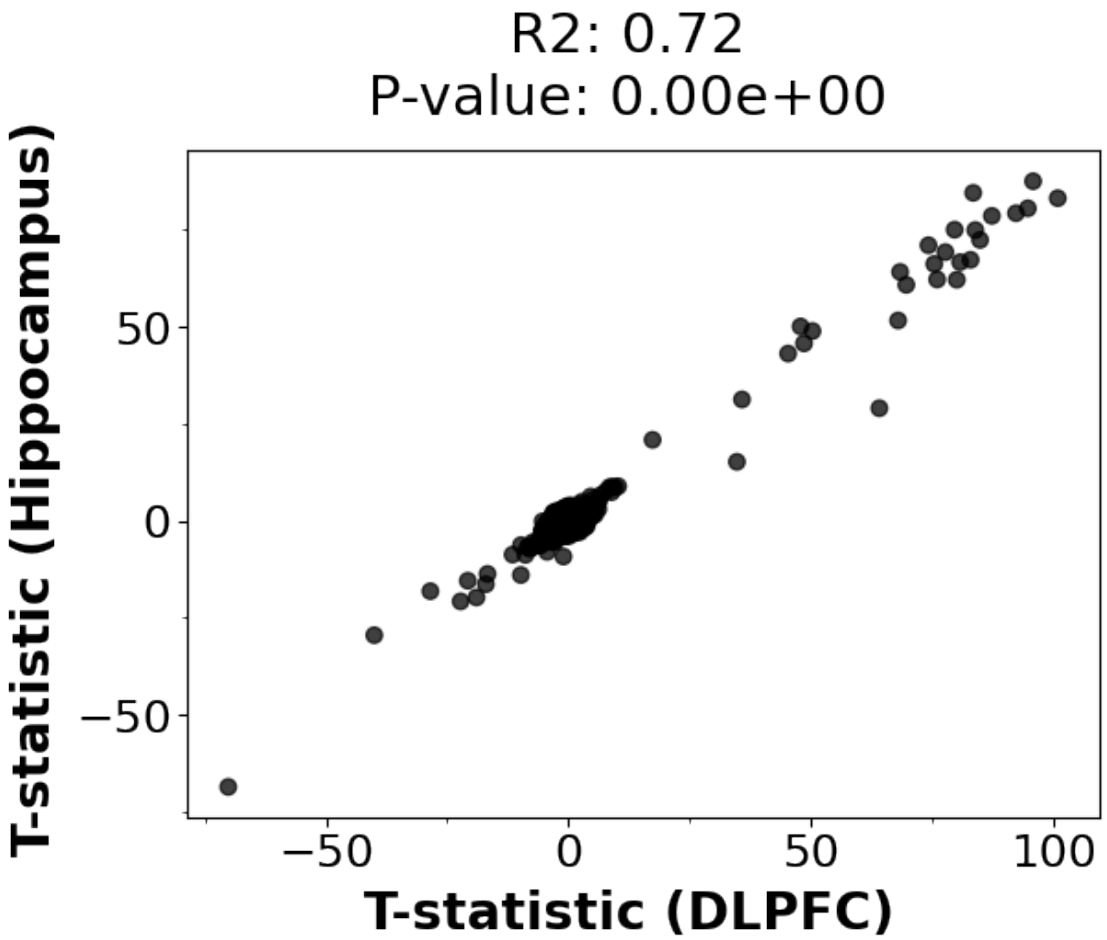
[15]: <ggplot: (8751092462159)>

[16]:
```
qq = plot_corr('caudate', 'hippo', merge_dataframes)
qq
```



[16]:
```
<ggplot: (8750340322838)>
```

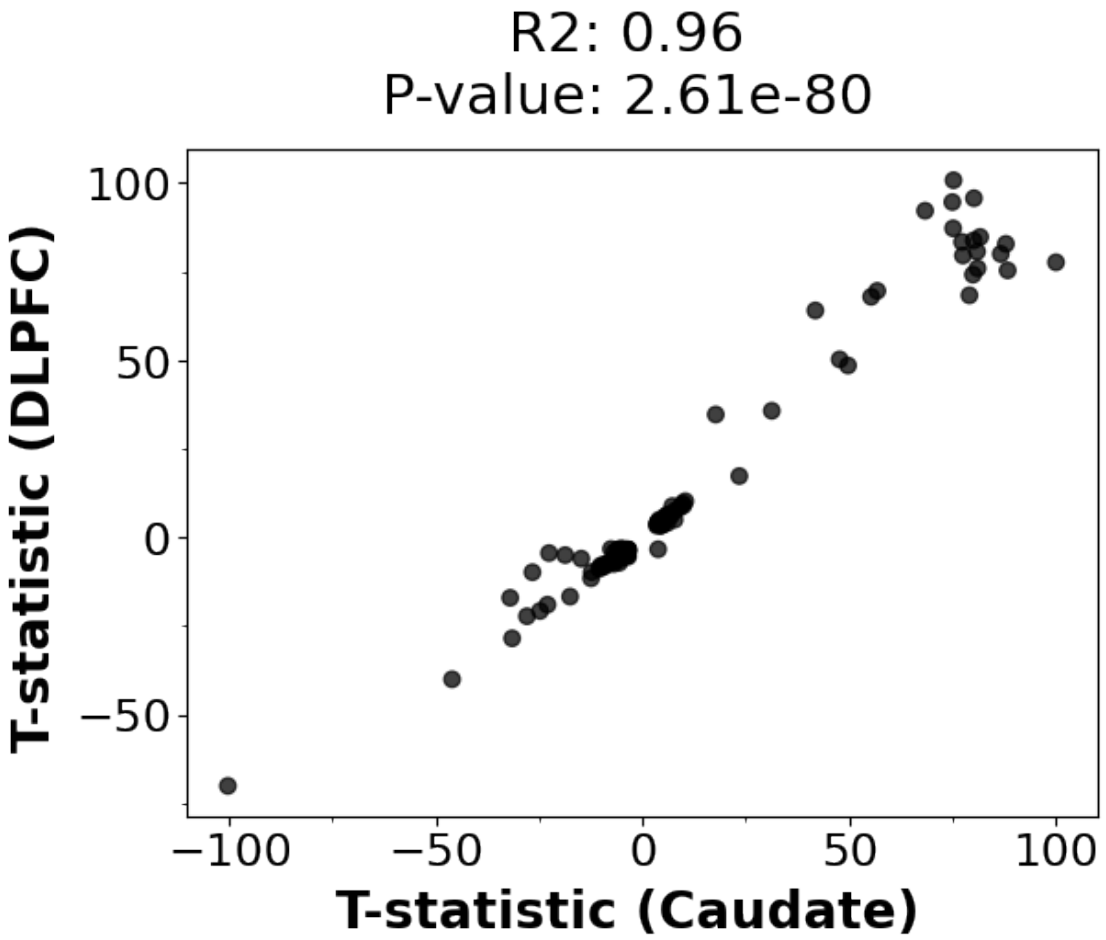
[17]:
```
ww = plot_corr('dlpfc', 'hippo', merge_dataframes)
ww
```



[17]: <ggplot: (8750341430233)>

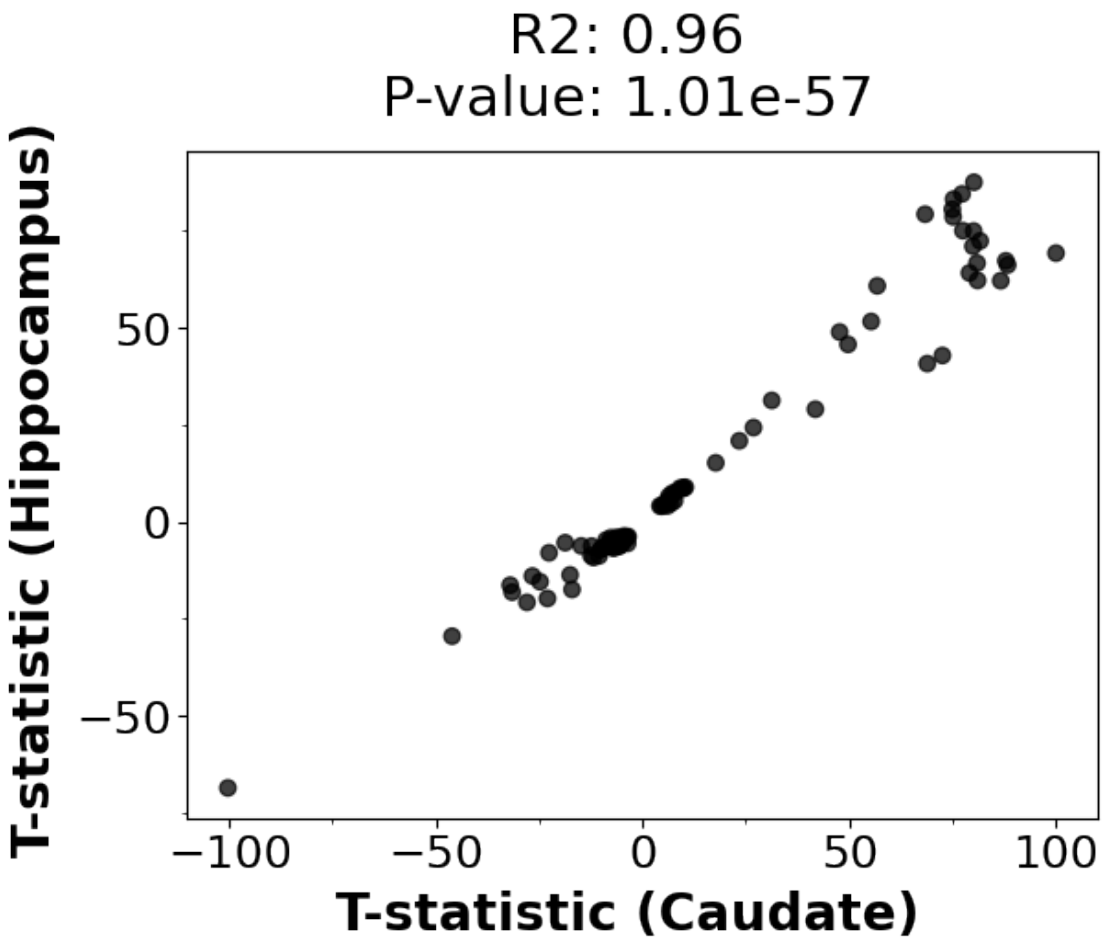
1.1.3 Significant correlation, FDR < 0.05

[18]:
```
pp = plot_corr('caudate', 'dlpfc', merge_dataframes_sig)
pp
```



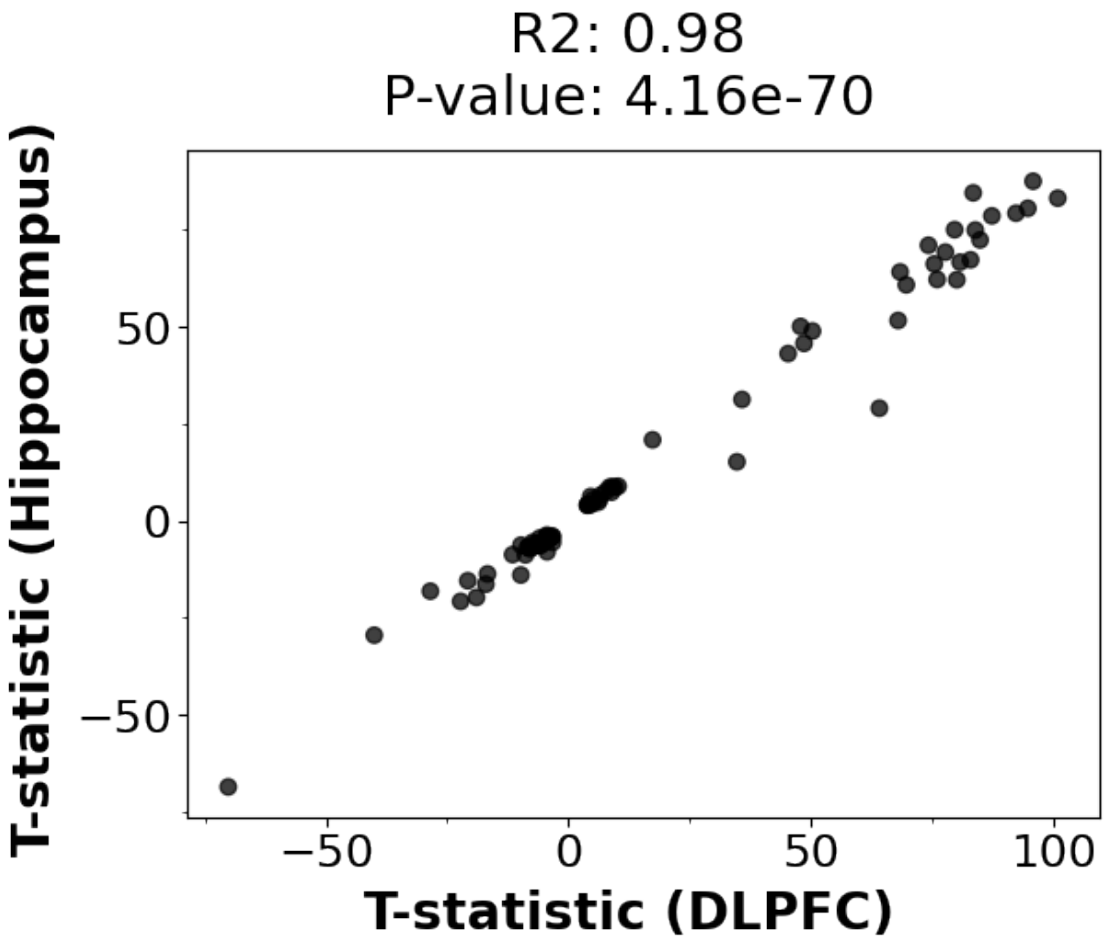
[18]: <ggplot: (8750340906617)>

[19]:
```
qq = plot_corr('caudate', 'hippo', merge_dataframes_sig)
qq
```



[19]: <ggplot: (8750341669503)>

[20]:
```
ww = plot_corr('dlpfc', 'hippo', merge_dataframes_sig)
ww
```



[20]: <ggplot: (8750340907463)>

[21]:
```
#save_plot(pp, 'dlpfc_caudate_tstatistic_corr_sig')
#save_plot(qq, 'hippo_caudate_tstatistic_corr_sig')
#save_plot(ww, 'hippo_dlpfc_tstatistic_corr_sig')
```

1.1.4 Directionality test

All genes

[22]:
```
enrichment_binom('caudate', 'dlpfc', merge_dataframes)
```

```
   agree      0
0   -1.0   9764
1    1.0  12188
```

[22]: 3.115367597709529e-60

[23]:
```
enrichment_binom('caudate', 'hippo', merge_dataframes)
```

```
   agree      0
0   -1.0   7835
1    1.0  14224
```

[23]:
```
5e-324
```

[24]:
```
enrichment_binom('dlpfc', 'hippo', merge_dataframes)
```

```
   agree      0
0   -1.0   8879
1    1.0  13356
```

[24]:
```
2.6476758684712667e-199
```

Significant DEG (FDR < 0.05)

[25]:
```
enrichment_binom('caudate', 'dlpfc', merge_dataframes_sig)
```

```
   agree    0
0   -1.0    1
1    1.0  116
```

[25]:
```
1.420373333985586e-33
```

[26]:
```
df = merge_dataframes_sig("caudate", "dlpfc")
df[(df['agree']<0)]
```

[26]:
```
                Feature ensemblID_caudate  adj.P.Val_caudate  logFC_caudate  \
101  ENSG00000066629.16   ENSG00000066629           0.014861       0.089089

     t_caudate  Dir_caudate  ensemblID_dlpfc  adj.P.Val_dlpfc  logFC_dlpfc  \
101   3.830688          1.0  ENSG00000066629         0.043284    -0.080965

      t_dlpfc  Dir_dlpfc  agree
101 -3.361968       -1.0   -1.0
```

[27]:
```
enrichment_binom('caudate', 'hippo', merge_dataframes_sig)
```

```
   agree   0
0    1.0  85
All directions agree!
```

[28]:
```
enrichment_binom('dlpfc', 'hippo', merge_dataframes_sig)
```

```
   agree   0
0    1.0  81
All directions agree!
```

1.2 Common Mind Comparison: MSSM Penn Pitt

[29]:
```
cmc_dlpfc = get_deg(config['cmc_dlpfc'])
cmc_dlpfc.groupby('Dir').size()
```

[29]:
```
Dir
-1.0     8613
 1.0    10498
dtype: int64
```

[30]:
```
cmc_dlpfc[(cmc_dlpfc['adj.P.Val'] < 0.05)].shape
```

[30]:
```
(482, 6)
```

1.2.1 Enrichment of DEG

[31]:
```
cal_fishers('dlpfc', 'cmc_dlpfc')
```

```
[[53, 162], [137, 8011]]
```

[31]:
```
(19.130530774083084, 1.8185445727989242e-41)
```

[32]:
```
cal_fishers('hippo', 'cmc_dlpfc')
```

```
[[30, 6], [158, 8060]]
```

[32]:
```
(255.0632911392405, 8.440029094936809e-45)
```

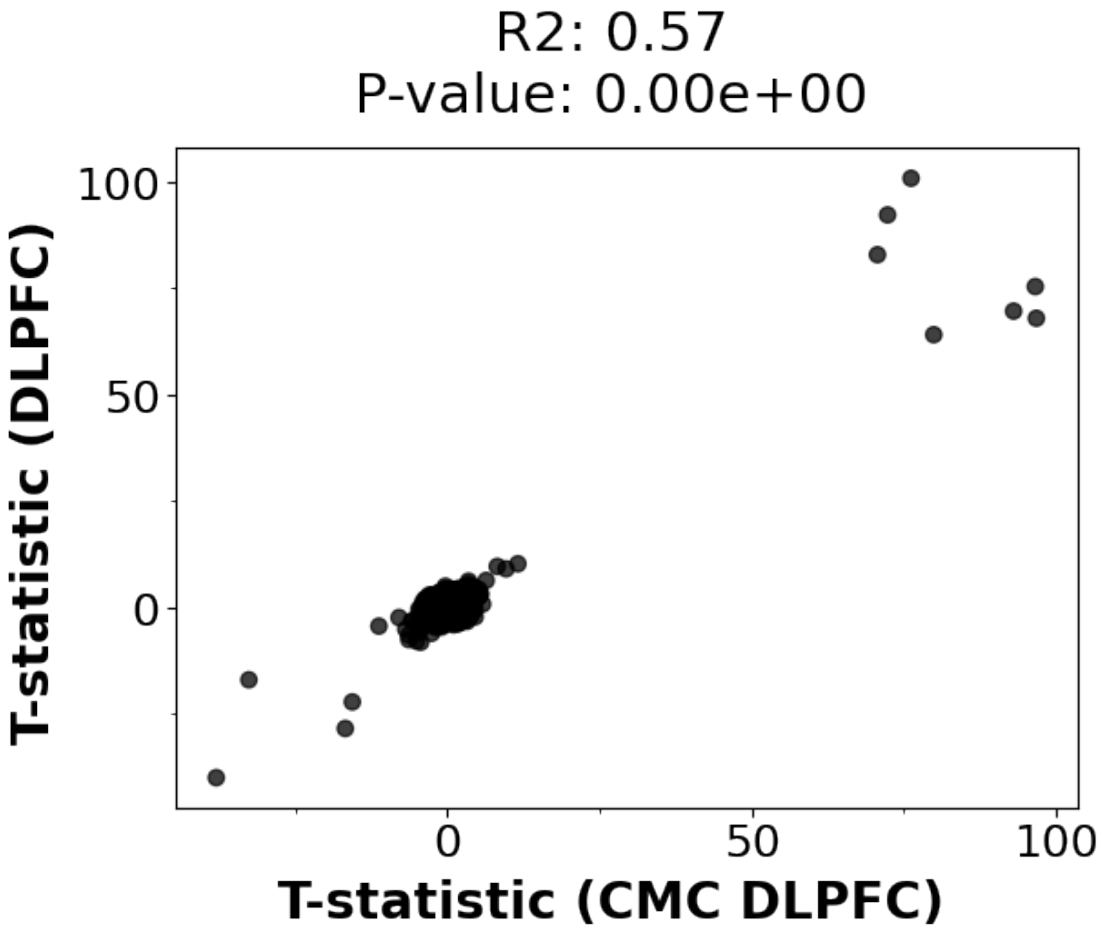
[33]:
```
cal_fishers('caudate', 'cmc_dlpfc')
```

```
[[41, 79], [144, 7856]]
```

[33]:
```
(28.313642756680732, 2.330504365005766e-38)
```

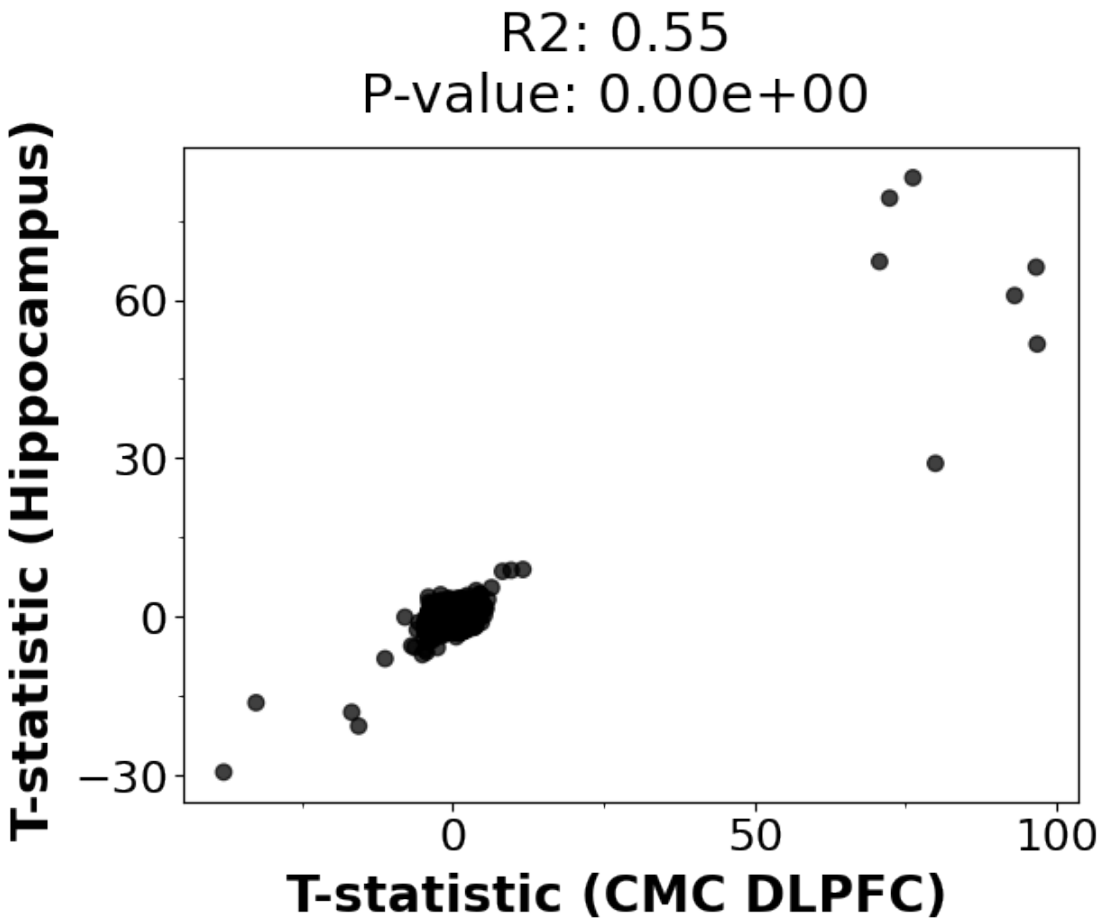
1.2.2 Correlation

[34]:
```
pp = plot_corr('cmc_dlpfc', 'dlpfc', merge_dataframes)
pp
```



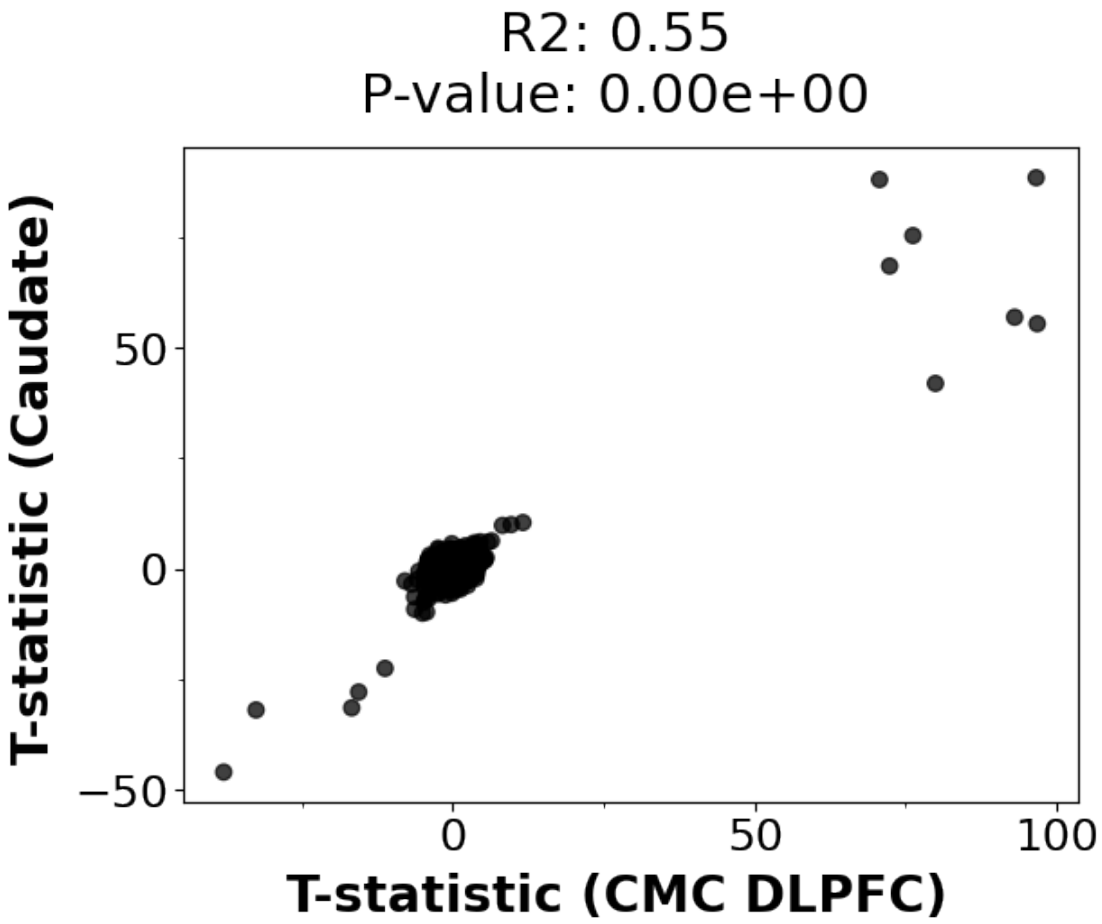
[34]:
```
<ggplot: (8750340941812)>
```

[35]:
```
qq = plot_corr('cmc_dlpfc', 'hippo', merge_dataframes)
qq
```



[35]: <ggplot: (8750340138102)>

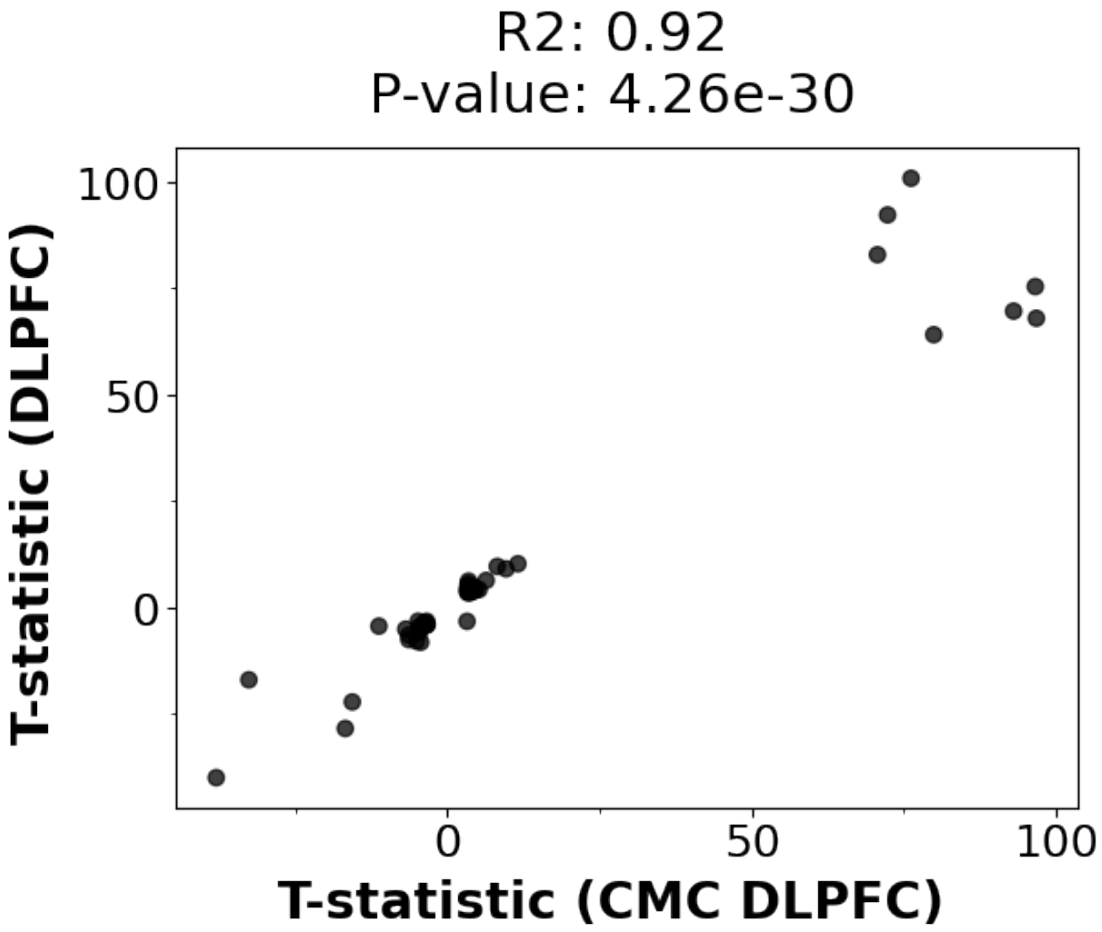
[36]:
```
ww = plot_corr('cmc_dlpfc', 'caudate', merge_dataframes)
ww
```



[36]: <ggplot: (8750340905054)>

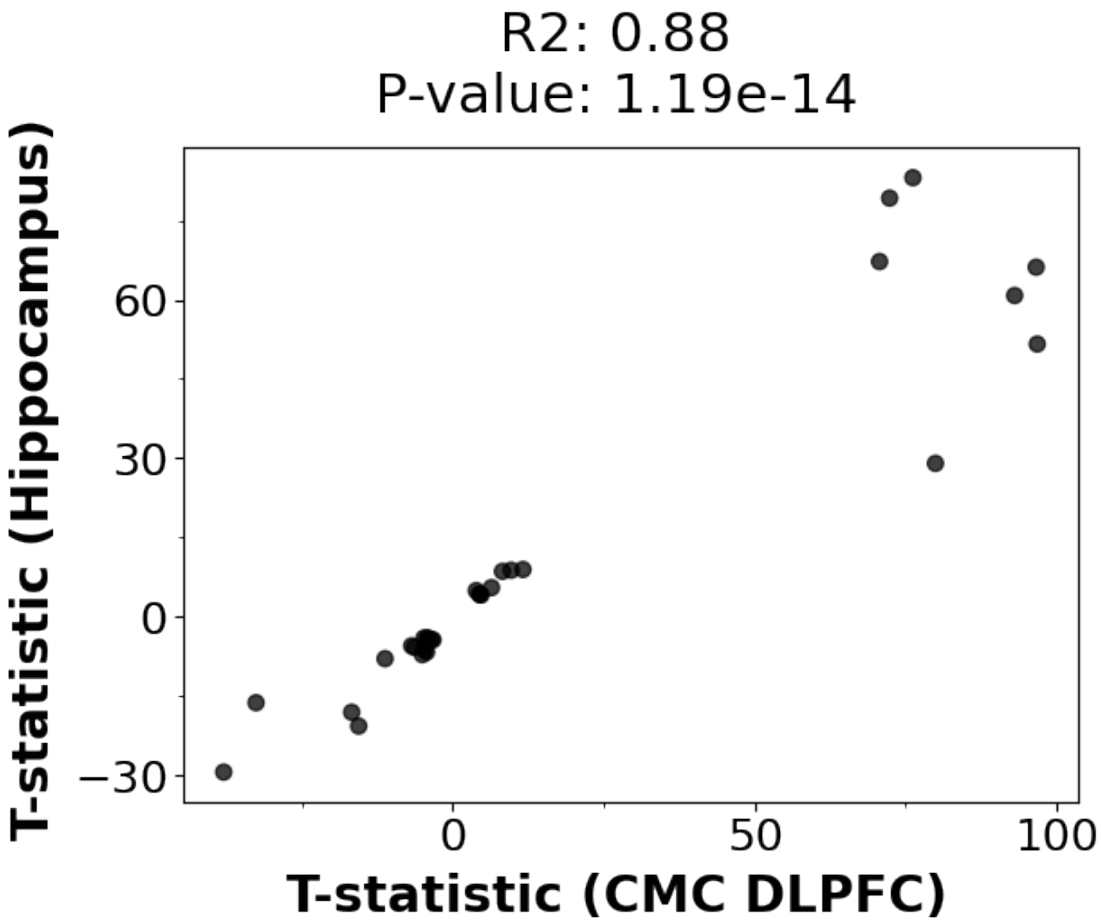
1.2.3 Significant correlation, FDR < 0.05

[37]:
```
pp = plot_corr('cmc_dlpfc', 'dlpfc', merge_dataframes_sig)
pp
```



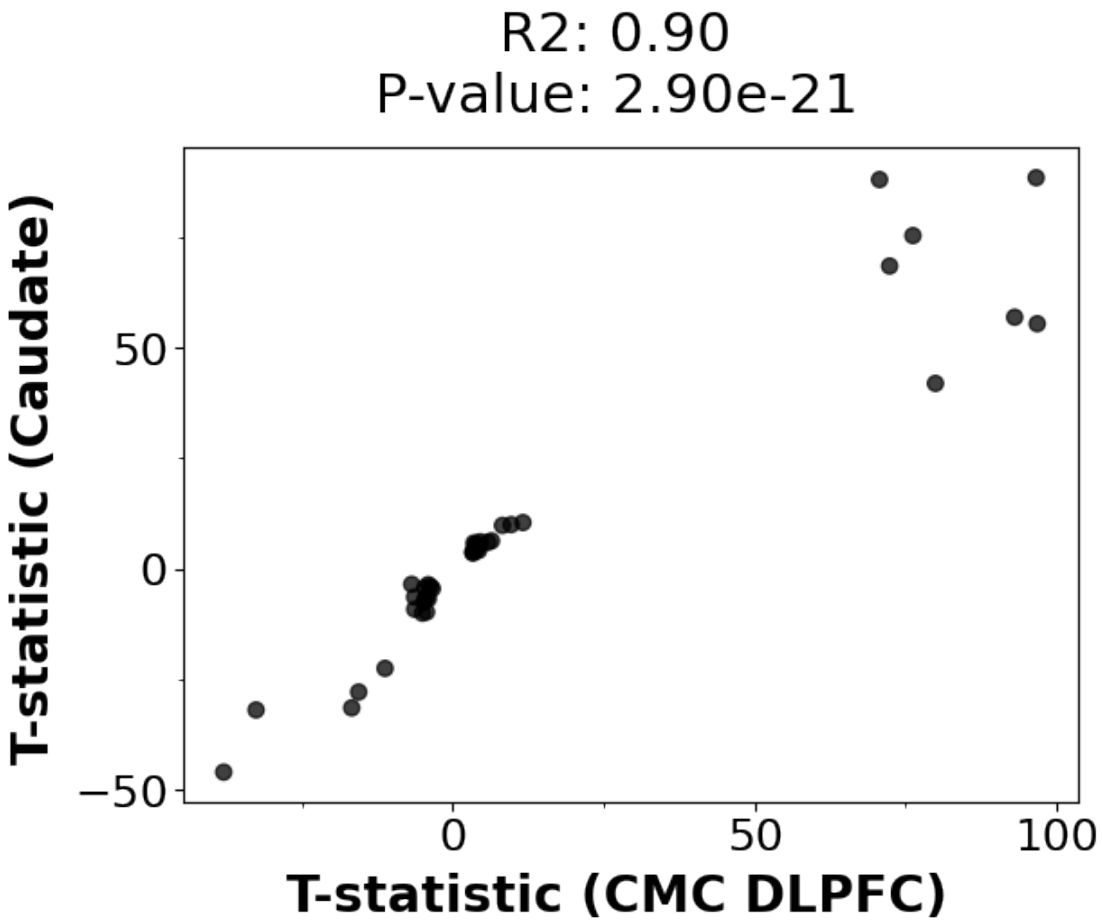
[37]: <ggplot: (8750340939234)>

[38]:
```
qq = plot_corr('cmc_dlpfc', 'hippo', merge_dataframes_sig)
qq
```



[38]: <ggplot: (8750340138018)>

[39]:
```
ww = plot_corr('cmc_dlpfc', 'caudate', merge_dataframes_sig)
ww
```



[39]: <ggplot: (8750340916824)>

1.2.4 Directionality

All genes

[40]:
```
enrichment_binom('cmc_dlpfc', 'dlpfc', merge_dataframes)
```

```
   agree     0
0   -1.0  3739
1    1.0  4624
```

[40]: 3.8327419817236607e-22

[41]:
```
enrichment_binom('cmc_dlpfc', 'hippo', merge_dataframes)
```

```
   agree     0
0   -1.0  4241
1    1.0  4013
```

[41]: 0.012464472090755149

[42]:
```
enrichment_binom('cmc_dlpfc', 'caudate', merge_dataframes)
```

```
   agree     0
0   -1.0  3864
1    1.0  4256
```

[42]: 1.4255704462859754e-05

Significant DEG (FDR < 0.05)

[43]:
```
enrichment_binom('cmc_dlpfc', 'dlpfc', merge_dataframes_sig)
```

```
   agree   0
0   -1.0   1
1    1.0  52
```

[43]: 1.199040866595169e-14

[44]:
```
enrichment_binom('cmc_dlpfc', 'hippo', merge_dataframes_sig)
```

```
   agree   0
0    1.0  30
All directions agree!
```

[45]:
```
enrichment_binom('cmc_dlpfc', 'caudate', merge_dataframes_sig)
```

```
   agree   0
0    1.0  41
All directions agree!
```

1.3 Common Mind Comparison: NIMH HBCC

[46]:
```
cmc_dlpfc = get_deg(config['cmc_hbcc'])
cmc_dlpfc.groupby('Dir').size()
```

```
Dir
-1.0    10712
 1.0     8399
dtype: int64
```

[47]:
```
cmc_dlpfc[(cmc_dlpfc['adj.P.Val'] < 0.05)].shape
```

[47]: (148, 6)

1.3.1 Enrichment of DEG

[48]:
```
cal_fishers('dlpfc', 'cmc_hbcc')
```

```
[[33, 182], [17, 8131]]
```

[48]: (86.72365869424694, 1.862789275842866e-41)

[49]:
```
cal_fishers('hippo', 'cmc_hbcc')
```

```
[[25, 11], [23, 8195]]
```

[49]: (809.7826086956521, 3.5194406812336975e-51)

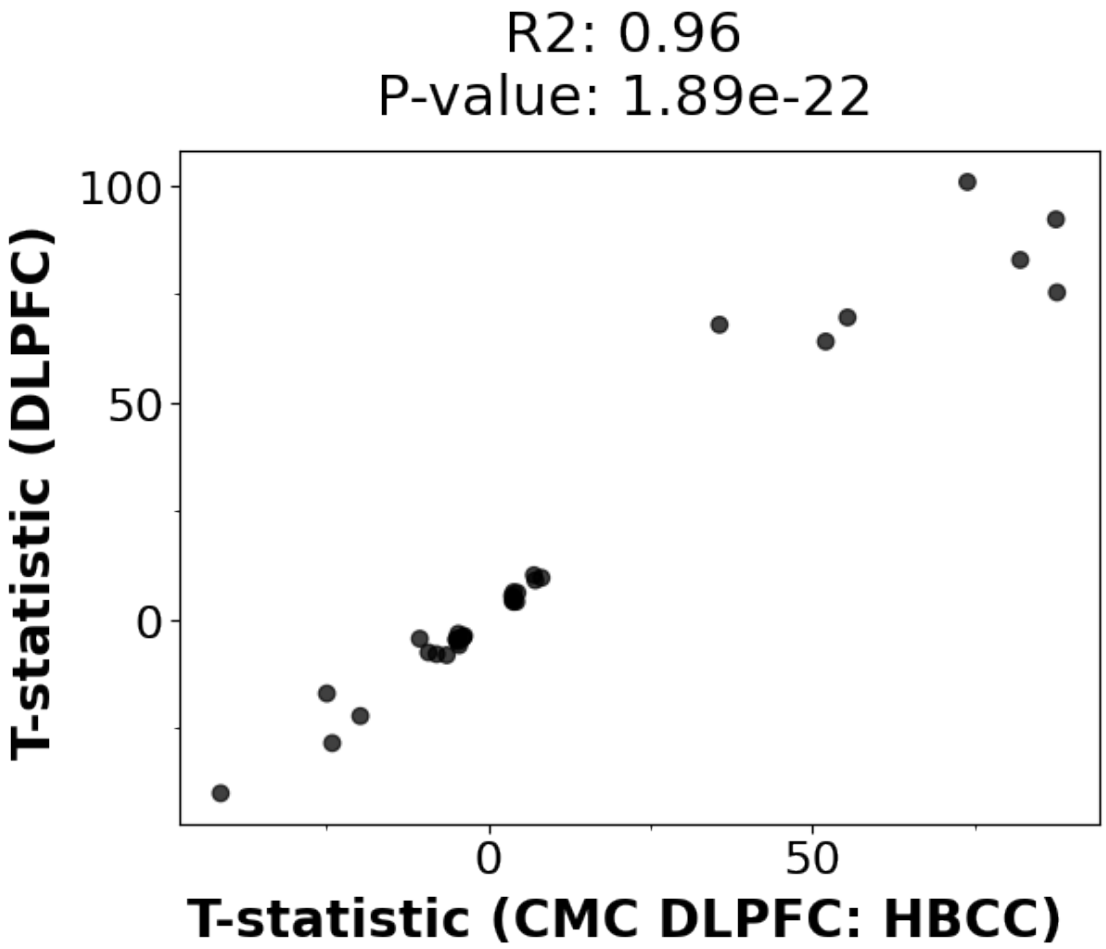
[50]:
```
cal_fishers('caudate', 'cmc_hbcc')
```

```
[[31, 89], [18, 7982]]
```

[50]: (154.458177278402, 2.6017102121050127e-46)

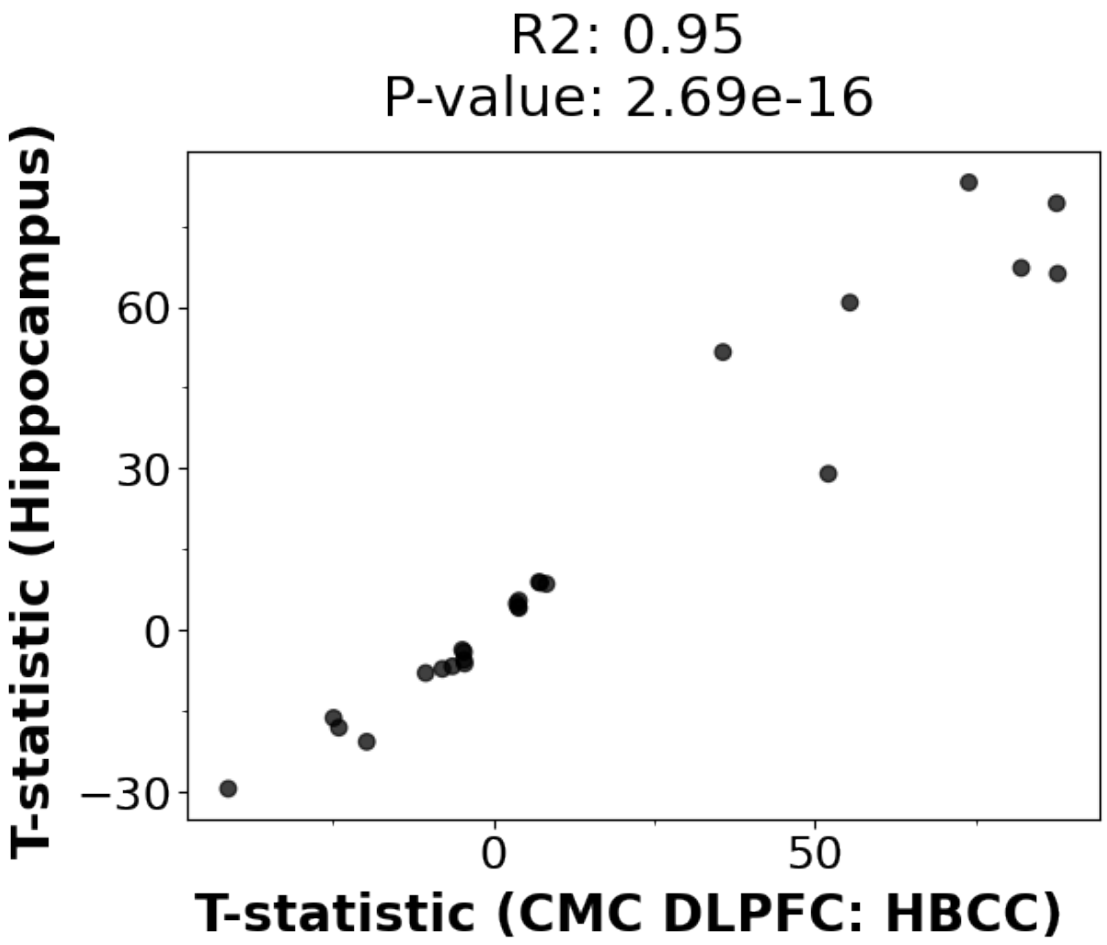
1.3.2 Significant correlation, FDR < 0.05

[51]:
```
pp = plot_corr('cmc_hbcc', 'dlpfc', merge_dataframes_sig)
pp
```



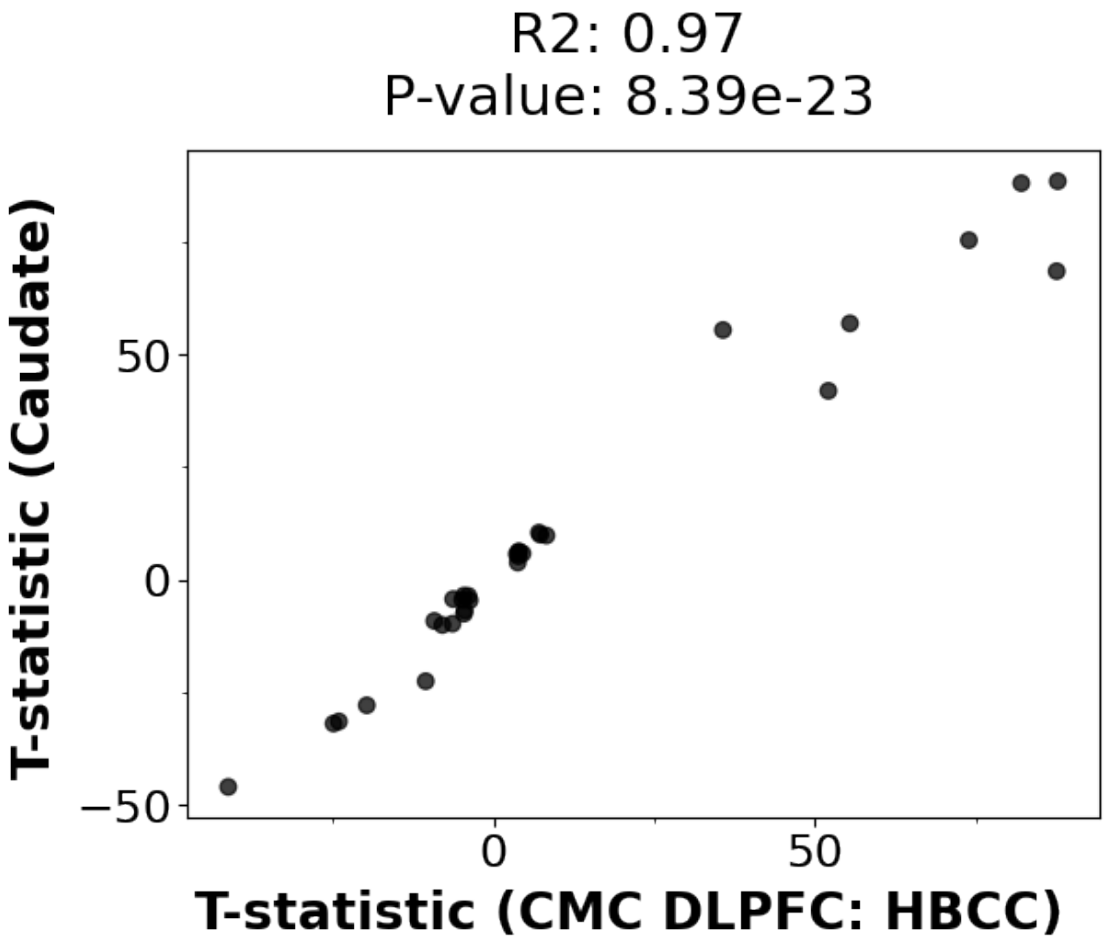
[51]: <ggplot: (8751092464065)>

[52]:
```
qq = plot_corr('cmc_hbcc', 'hippo', merge_dataframes_sig)
qq
```



[52]: <ggplot: (8750341673569)>

[53]:
```
ww = plot_corr('cmc_hbcc', 'caudate', merge_dataframes_sig)
ww
```



[53]: <ggplot: (8750339790147)>

1.3.3 Directionality

All genes

[54]:
```
enrichment_binom('cmc_hbcc', 'dlpfc', merge_dataframes)
```

```
   agree     0
0   -1.0  3606
1    1.0  4757
```

[54]: 2.2473183550989796e-36

[55]:
```
enrichment_binom('cmc_hbcc', 'hippo', merge_dataframes)
```

```
   agree     0
0   -1.0  3464
1    1.0  4790
```

[55]: 2.2305500097037318e-48

[56]:
```
enrichment_binom('cmc_hbcc', 'caudate', merge_dataframes)
```

```
   agree     0
0   -1.0  3271
1    1.0  4849
```

[56]: 5.331426043076026e-69

Significant DEG (FDR < 0.05)

[57]:
```
enrichment_binom('cmc_hbcc', 'dlpfc', merge_dataframes_sig)
```

```
   agree   0
0    1.0  33
All directions agree!
```

[58]:
```
enrichment_binom('cmc_hbcc', 'hippo', merge_dataframes_sig)
```

```
   agree   0
0    1.0  25
All directions agree!
```

[59]:
```
enrichment_binom('cmc_hbcc', 'caudate', merge_dataframes_sig)
```

```
   agree   0
0    1.0  31
All directions agree!
```

[ ]: