main

September 18, 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': '../../../_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',
    'cmc_sva': '/ceph/projects/v4_phase3_paper/inputs/public_data/_m/cmc/
 ↪CMC_MSSM-Penn-Pitt_DLPFC_mRNA_'+\
   ␣
 ↪'IlluminaHiSeq2500_gene-adjustedSVA-differentialExpression-includeAncestry-DxSCZ-DE.
 ↪tsv',
    'cmc': '/ceph/projects/v4_phase3_paper/inputs/public_data/_m/cmc/
 ↪CMC_MSSM-Penn-Pitt_DLPFC_mRNA_'+\
   ␣
 ↪'IlluminaHiSeq2500_gene-adjustedNoSVA-differentialExpression-includeAncestry-DxSCZ-DE.
 ↪tsv'
}
```

[3]:
```
@functools.lru_cache()
def get_cmc(SVA=True):
    if SVA:
        cmc_dlpfc = pd.read_csv(config["cmc_sva"], sep='\t')\
```

```
                        .rename(columns={'MAPPED_genes': 'Symbol', "genes":␣
↪"ensemblID"})
    else:
        cmc_dlpfc = pd.read_csv(config["cmc"], sep='\t')\
                        .rename(columns={'MAPPED_genes': "Symbol", "genes":␣
↪"ensemblID"})
    cmc_dlpfc['Dir'] = np.sign(cmc_dlpfc['t'])
    cmc_dlpfc["Feature"] = cmc_dlpfc.ensemblID
    return cmc_dlpfc[["Feature", "ensemblID", 'adj.P.Val', 't', 'Dir',␣
↪"Symbol"]]


@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, 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='ensemblID',
                                          suffixes=['_%s' % tissue1, '_%s' %␣
↪tissue2])


@functools.lru_cache()
def merge_dataframes_sig(tissue1, tissue2):
    fdr1 = 0.05 if tissue1 != 'dlpfc' else 0.05
    fdr2 = 0.05 if tissue2 != 'dlpfc' else 0.05
    return get_deg_sig(config[tissue1], fdr1).
↪merge(get_deg_sig(config[tissue2], fdr2),
                                                  on='ensemblID',
                                                  suffixes=['_%s' % tissue1,␣
↪'_%s' % tissue2])
```

```
@functools.lru_cache()
def merge_cmc(tissue1, sig=False, SVA=True):
    if sig:
        df1 = get_cmc(SVA)[(get_cmc(SVA)["adj.P.Val"] < 0.05)]
        df2 = get_deg_sig(config[tissue1], 0.05)
    else:
        df1 = get_cmc(SVA)
        df2 = get_deg(config[tissue1])
    return df2.merge(df1, on="ensemblID", suffixes=["_%s" % tissue1, '_cmc'])
```

[4]:
```
def enrichment_binom(tissue1, tissue2, merge_fnc, sig=False, sva=True):
    if tissue2 != "cmc":
        df = merge_fnc(tissue1, tissue2)
    else:
        df = merge_fnc(tissue1, sig, sva)
    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, fnc, sva=True):
    if tissue2 != 'cmc':
        df = fnc(tissue1, tissue2)
    else:
        df = fnc(tissue1, False, sva)
    fdr1 = 0.05 if tissue1 != 'dlpfc' else 0.05
    fdr2 = 0.05 if tissue2 != 'dlpfc' else 0.05
    table = [[np.sum((df['adj.P.Val_%s' % tissue1]<fdr1) &
                     ((df['adj.P.Val_%s' % tissue2]<fdr2))),
              np.sum((df['adj.P.Val_%s' % tissue1]<fdr1) &
                     ((df['adj.P.Val_%s' % tissue2]>=fdr2)))],
             [np.sum((df['adj.P.Val_%s' % tissue1]>=fdr1) &
                     ((df['adj.P.Val_%s' % tissue2]<fdr2))),
              np.sum((df['adj.P.Val_%s' % tissue1]>=fdr1) &
                     ((df['adj.P.Val_%s' % tissue2]>=fdr2)))]]
    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, sig=False, sva=True):
    if tissue2 != 'cmc':
        dft = merge_fnc(tissue1, tissue2)
    else:
        dft = merge_fnc(tissue1, sig, sva)
    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': "CMC DLPFC"}[tissue]
```

[5]:
```
def plot_corr_impl(tissue1, tissue2, merge_fnc, sig, sva):
    if tissue2 != "cmc":
        dft = merge_fnc(tissue1, tissue2)
        title = '\n'.join([corr_annotation(tissue1, tissue2, merge_fnc)])
    else:
        dft = merge_fnc(tissue1, sig, sva)
        title = '\n'.join([corr_annotation(tissue1, tissue2, merge_fnc, sig,␣
 ↪sva)])
    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, sig=False, sva=True):
    return plot_corr_impl(tissue1, tissue2, merge_fnc, sig, sva)


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 Sample summary

[6]:
```
pheno_file = '/ceph/projects/v4_phase3_paper/inputs/phenotypes/_m/
 ↪merged_phenotypes.csv'
pheno = pd.read_csv(pheno_file, index_col=0)
pheno = pheno[(pheno['Age'] > 17) &
              (pheno['Dx'].isin(['SZ', 'CTL'])) &
              (pheno['Race'].isin(['AA', "EA"]))].copy()
pheno.head(2)
```

[6]:

RNum	Sex	Race	Dx	Age	mitoRate	rRNA_rate	totalAssignedGene	RIN	\
R11135	Male	EA	CTL	18.77	0.257280	0.000169	0.523132	5.9	
R11137	Male	EA	CTL	41.44	0.384027	0.000088	0.593343	9.2	

RNum	ERCCsumLogErr	overallMapRate	snpPC1	snpPC2	snpPC3	snpPC4	\
R11135	-22.049787	0.8746	-0.036163	0.003232	0.000562	0.001725	
R11137	-29.498329	0.9149	-0.035985	0.003539	-0.000170	-0.001330	

RNum	snpPC5	Region	BrNum	antipsychotics	lifetime_antipsych	Protocol
R11135	-0.000807	HIPPO	Br2063	False	False	RiboZeroHMR
R11137	0.002003	HIPPO	Br2582	False	False	RiboZeroHMR

[7]:
```
pheno.groupby(['Region']).size()
```

[7]:
```
Region
Caudate    394
DLPFC      360
HIPPO      376
dtype: int64
```

[8]:
```
pheno.groupby(['Region', 'Race']).size()
```

[8]:
```
Region   Race
Caudate  AA      205
         EA      189
DLPFC    AA      200
         EA      160
HIPPO    AA      207
         EA      169
dtype: int64
```

[9]:
```
pheno.groupby(['Region', 'Race', 'Sex']).size()
```

[9]:
```
Region   Race  Sex
Caudate  AA    Female     78
               Male      127
         EA    Female     43
               Male      146
DLPFC    AA    Female     75
               Male      125
         EA    Female     39
               Male      121
HIPPO    AA    Female     81
               Male      126
         EA    Female     40
               Male      129
dtype: int64
```

1.2 BrainSeq Tissue Comparison

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

[10]:
```
Dir
-1.0    12061
 1.0    10897
dtype: int64
```

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

[11]:
```
(2701, 6)
```

[12]:
```
dlpfc = get_deg(config['dlpfc'])
dlpfc.groupby('Dir').size()
```

[12]:
```
Dir
-1.0    13207
 1.0    11445
dtype: int64
```

[13]:
```
dlpfc[(dlpfc['adj.P.Val'] < 0.05)].shape
```

[13]:
```
(245, 6)
```

[14]:
```
hippo = get_deg(config['hippo'])
hippo.groupby('Dir').size()
```

[14]:
```
Dir
-1.0    12852
 1.0    11800
dtype: int64
```

[15]:
```
hippo[(hippo['adj.P.Val'] < 0.05)].shape
```

[15]:
```
(48, 6)
```

1.2.1 Enrichment of DEG

[16]:
```
cal_fishers('caudate', 'dlpfc', merge_dataframes)
```

```
[[49, 2498], [180, 18132]]
```

[16]:
```
(1.975954096610622, 9.40458506586896e-05)
```

[17]:
```
cal_fishers('caudate', 'hippo', merge_dataframes)
```

```
[[10, 2537], [35, 18277]]
```

[17]:
```
(2.0583366180528184, 0.06245006401479434)
```

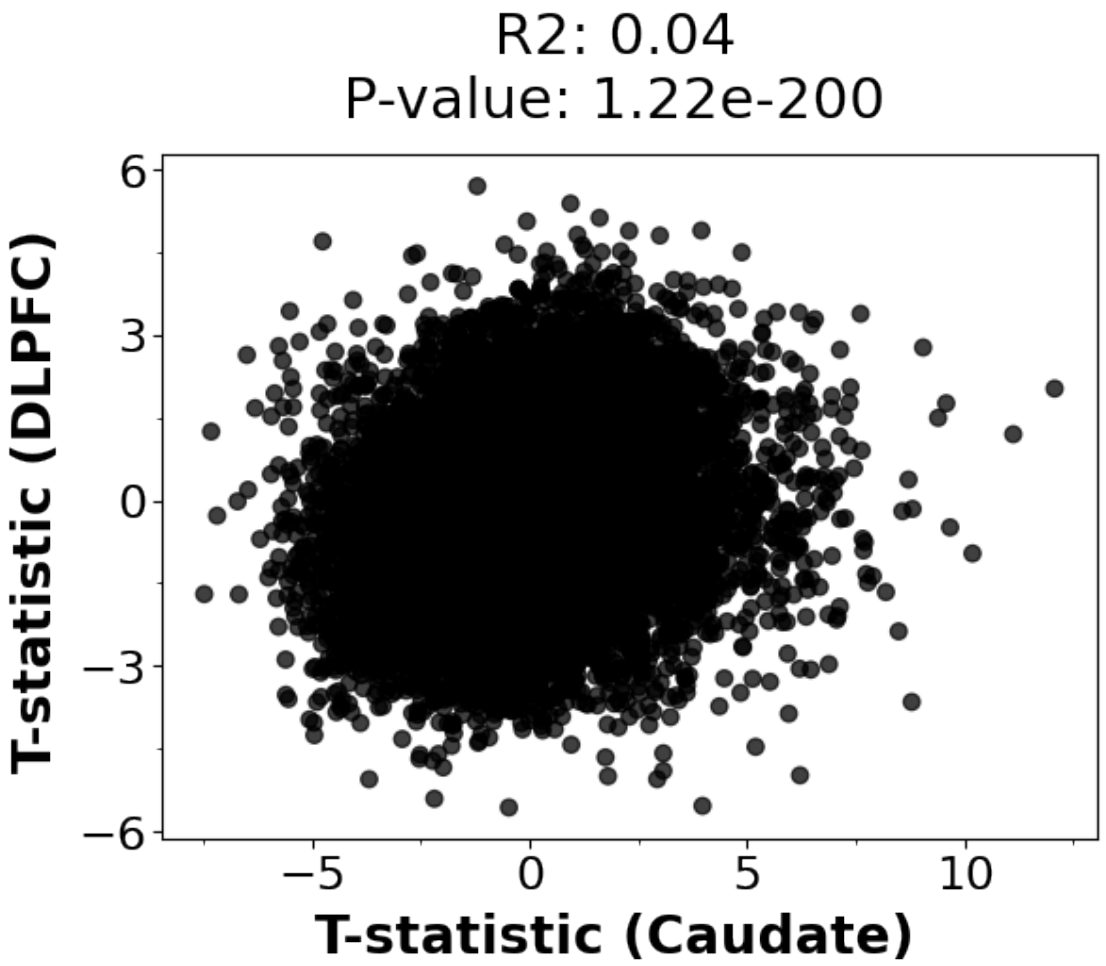
[18]:
```
cal_fishers('dlpfc', 'hippo', merge_dataframes)
```

```
[[6, 239], [42, 24365]]
```

[18]:
```
(14.563658099222954, 7.842543158014382e-06)
```

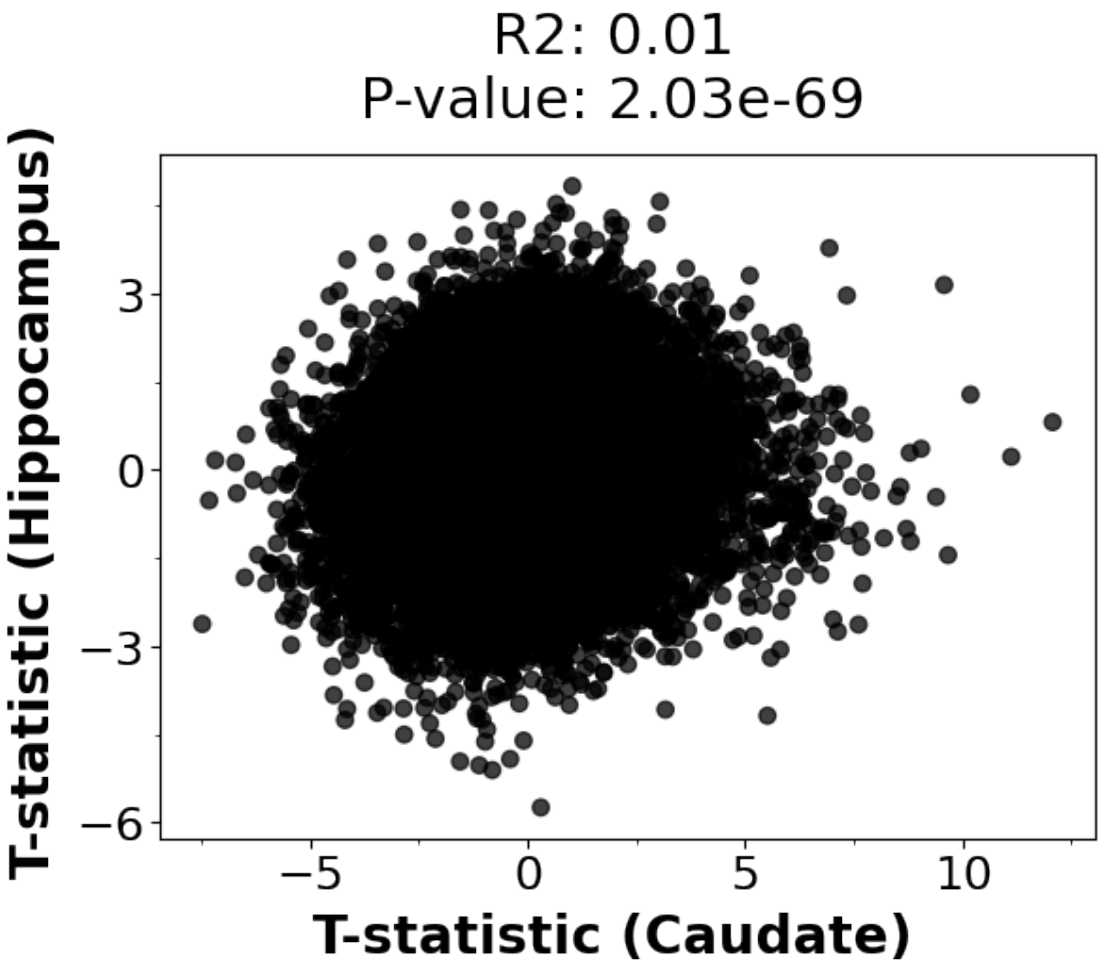
1.2.2 Correlation

[19]:
```
pp = plot_corr('caudate', 'dlpfc', merge_dataframes)
pp
```



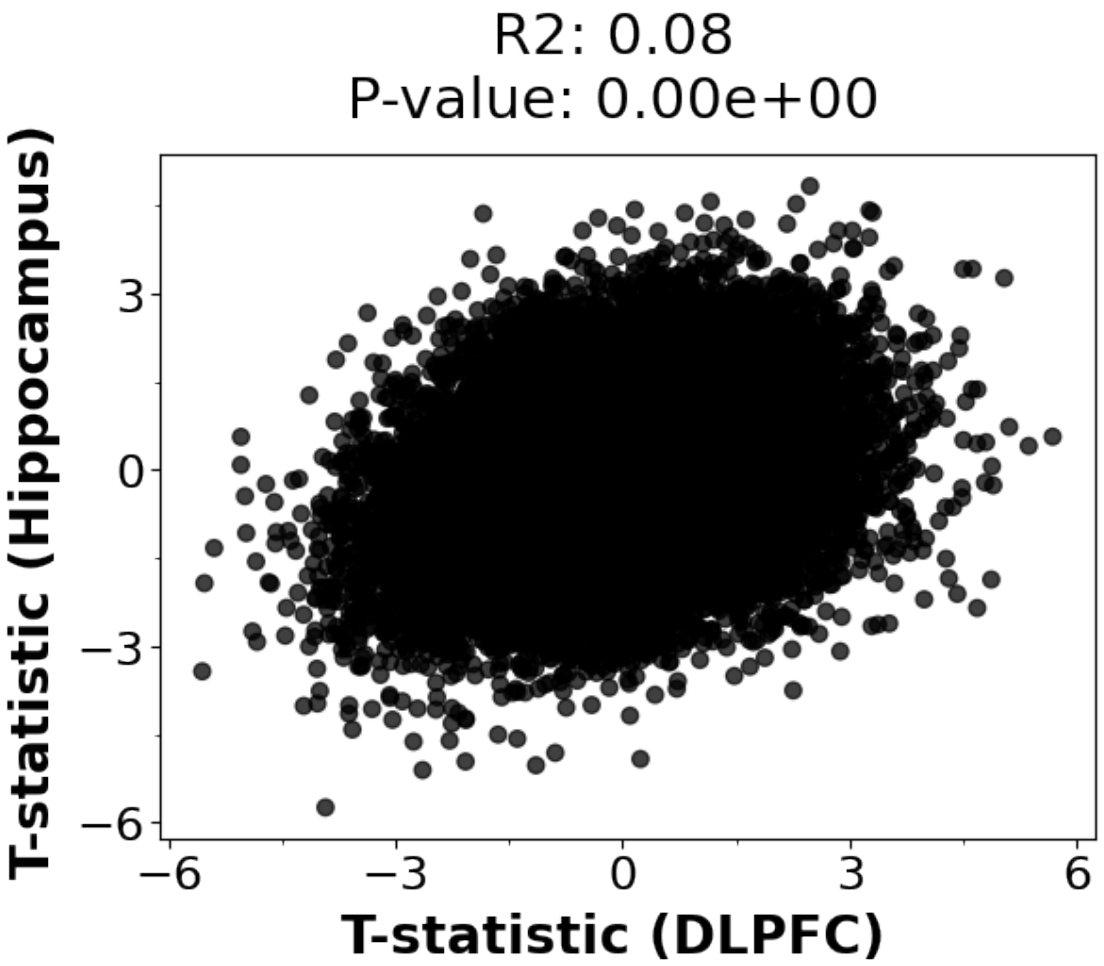
[19]: <ggplot: (8793606213373)>

[20]:
```
qq = plot_corr('caudate', 'hippo', merge_dataframes)
qq
```



[20]: <ggplot: (8793608008744)>

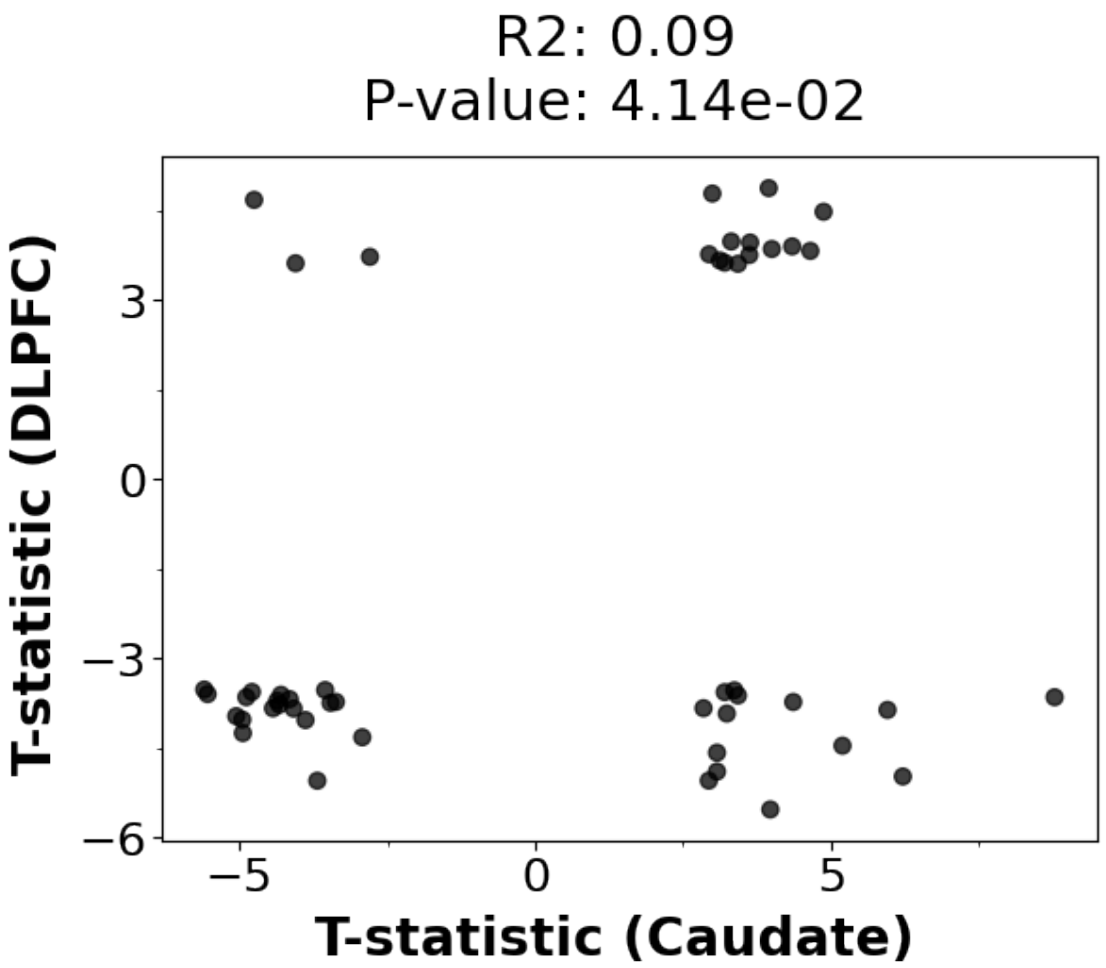
[21]:
```
ww = plot_corr('dlpfc', 'hippo', merge_dataframes)
ww
```



[21]: <ggplot: (8793607223796)>

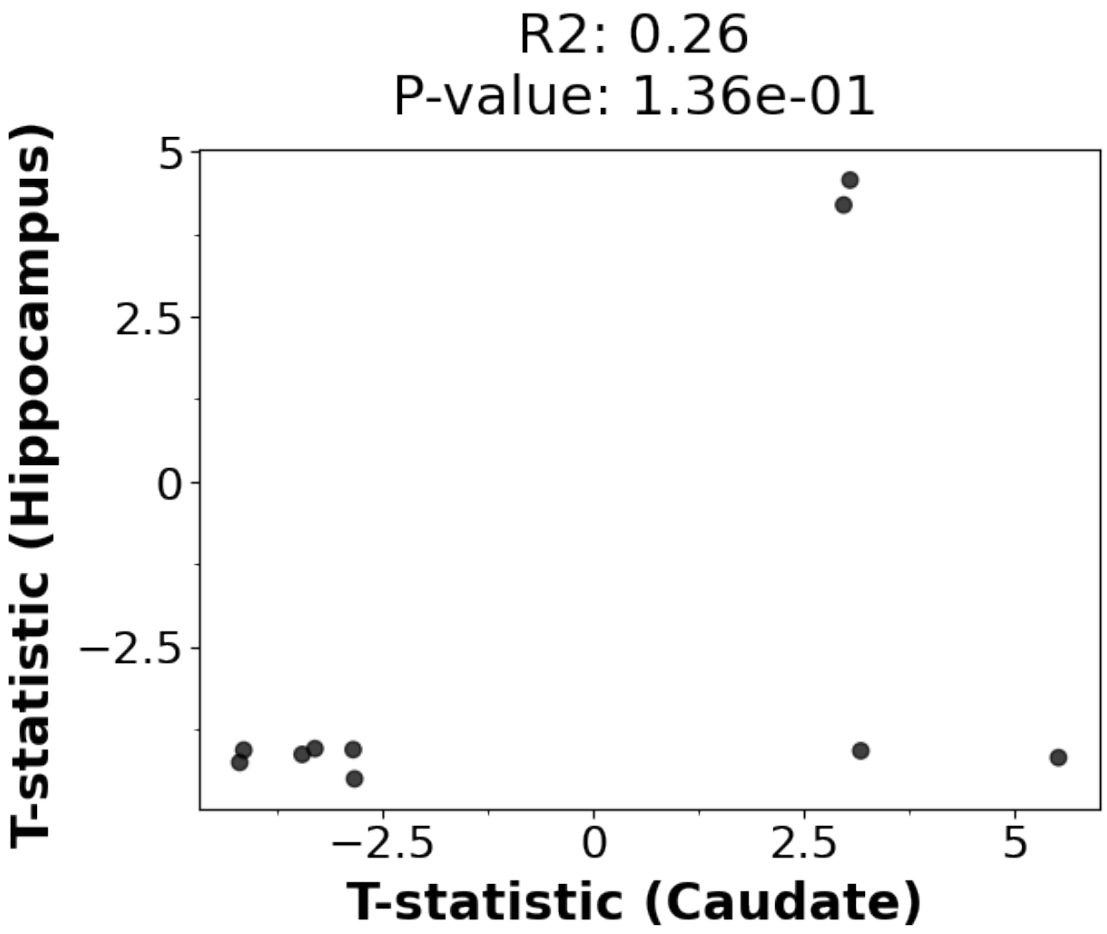
1.2.3 Significant correlation, FDR < 0.05

[22]:
```
pp = plot_corr('caudate', 'dlpfc', merge_dataframes_sig)
pp
```



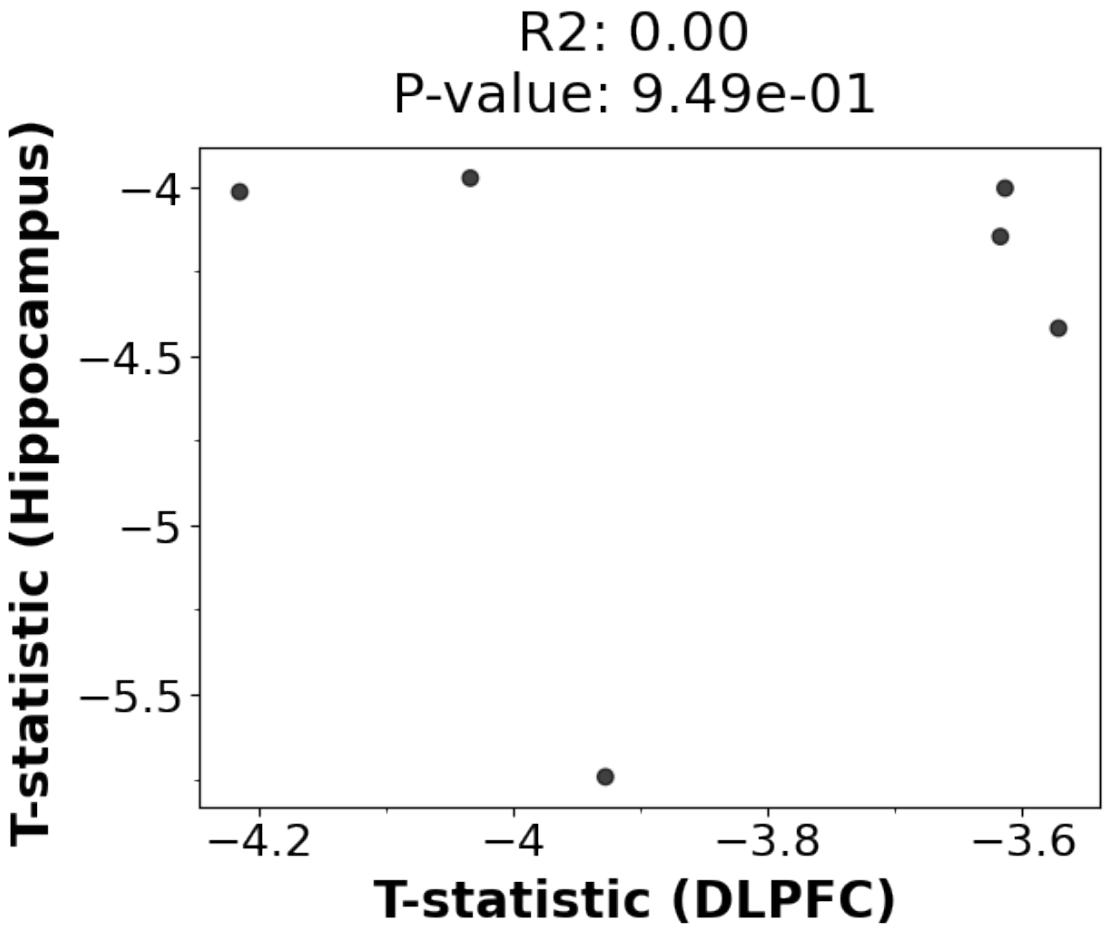
[22]: <ggplot: (8793608067991)>

[23]:
```
qq = plot_corr('caudate', 'hippo', merge_dataframes_sig)
qq
```



[23]: <ggplot: (8793608028916)>

[24]:
```
ww = plot_corr('dlpfc', 'hippo', merge_dataframes_sig)
ww
```



[24]: <ggplot: (8793608044566)>

1.2.4 Directionality test

All genes

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

	agree	0
0	-1.0	8821
1	1.0	12038

[25]: 2.8390706398892144e-110

[26]:
```
enrichment_binom('caudate', 'hippo', merge_dataframes)
```

	agree	0
0	-1.0	9545
1	1.0	11314

[26]: 1.704505022943847e-34

[27]:
```
enrichment_binom('dlpfc', 'hippo', merge_dataframes)
```

```
   agree      0
0   -1.0  10291
1    1.0  14361
```

[27]: 9.504008229391508e-149

Significant DEG (FDR < 0.05)

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

```
   agree   0
0   -1.0  17
1    1.0  32
```

[28]: 0.04438416098714981

[29]:
```
enrichment_binom('caudate', 'hippo', merge_dataframes_sig)
```

```
   agree  0
0   -1.0  2
1    1.0  8
```

[29]: 0.109375

[30]:
```
enrichment_binom('dlpfc', 'hippo', merge_dataframes_sig)
```

```
   agree  0
0    1.0  6
All directions agree!
```

1.3 CMC comparison

1.3.1 Adjusted SVA

[31]:
```
cmc = get_cmc(SVA=True)
cmc.groupby('Dir').size()
```

[31]:
```
Dir
-1.0    8898
 1.0    7525
dtype: int64
```

[32]:
```
cmc[(cmc['adj.P.Val'] < 0.05)].shape
```

[32]: (419, 6)

1.3.2 No adjusted SVA

[33]:
```
cmc_dlpfc2 = get_cmc(False)
cmc_dlpfc2.groupby('Dir').size()
```

[33]:
```
Dir
-1.0    8759
 1.0    7664
dtype: int64
```

[34]:
```
cmc_dlpfc2[(cmc_dlpfc2['adj.P.Val'] < 0.05)].shape
```

[34]:
```
(573, 6)
```

1.3.3 Enrichment of DEG

SVA corrected

[35]:
```
cal_fishers("caudate", "cmc", merge_cmc, True)
```

```
[[97, 2226], [303, 12761]]
```

[35]:
```
(1.8352222014654296, 1.2211322278814786e-06)
```

[36]:
```
cal_fishers("dlpfc", "cmc", merge_cmc, True)
```

```
[[30, 192], [374, 14598]]
```

[36]:
```
(6.098763368983957, 2.1306508938443574e-13)
```

[37]:
```
cal_fishers("hippo", "cmc", merge_cmc, True)
```

```
[[0, 42], [404, 14748]]
```

[37]:
```
(0.0, 0.6296359956197478)
```

No SVA correction

[38]:
```
cal_fishers("caudate", "cmc", merge_cmc, False)
```

```
[[98, 2225], [449, 12615]]
```

[38]:
```
(1.2374765396261356, 0.06783703737285668)
```

[39]:
```
cal_fishers("dlpfc", "cmc", merge_cmc, False)
```

```
[[16, 206], [533, 14439]]
```

[39]:
```
(2.1040820415672417, 0.00939503421659622)
```

[40]:
```
cal_fishers("hippo", "cmc", merge_cmc, False)
```

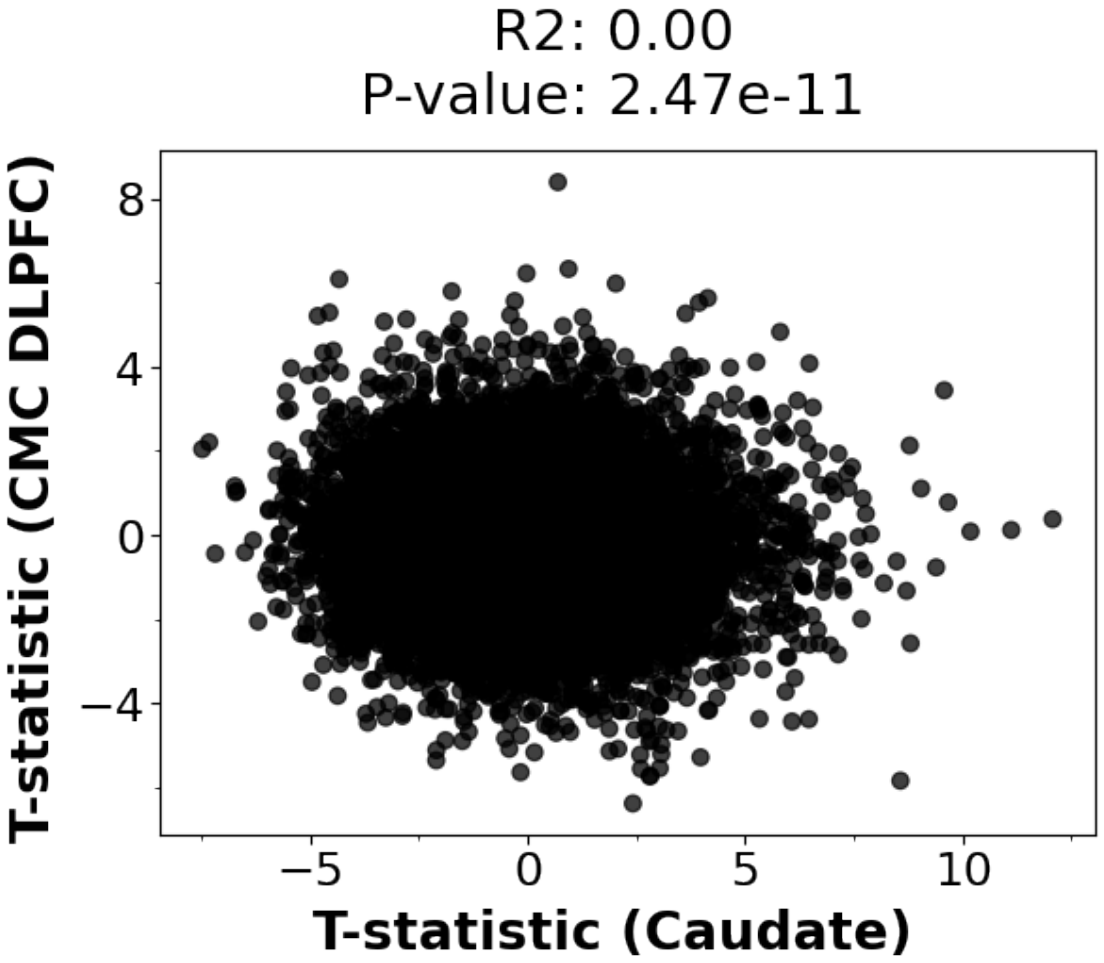
```
[[2, 40], [547, 14605]]
```

[40]: (1.3350091407678244, 0.6641462563057603)

1.3.4 Correlation

SVA correction

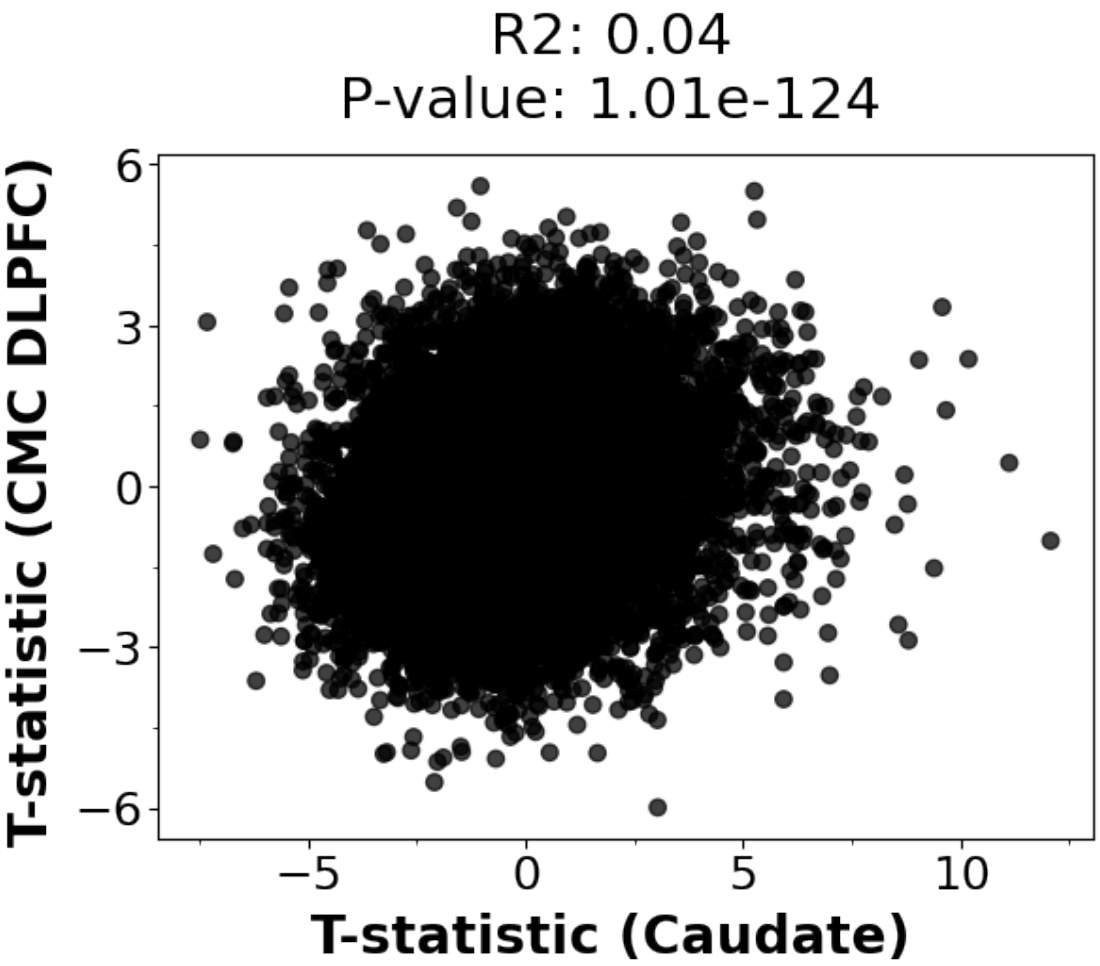
[41]:
```
pp = plot_corr('caudate', 'cmc', merge_cmc, False, True)
pp
```



[41]: <ggplot: (8793607265877)>

No SVA correction

[42]:
```
qq = plot_corr('caudate', 'cmc', merge_cmc, False, False)
qq
```

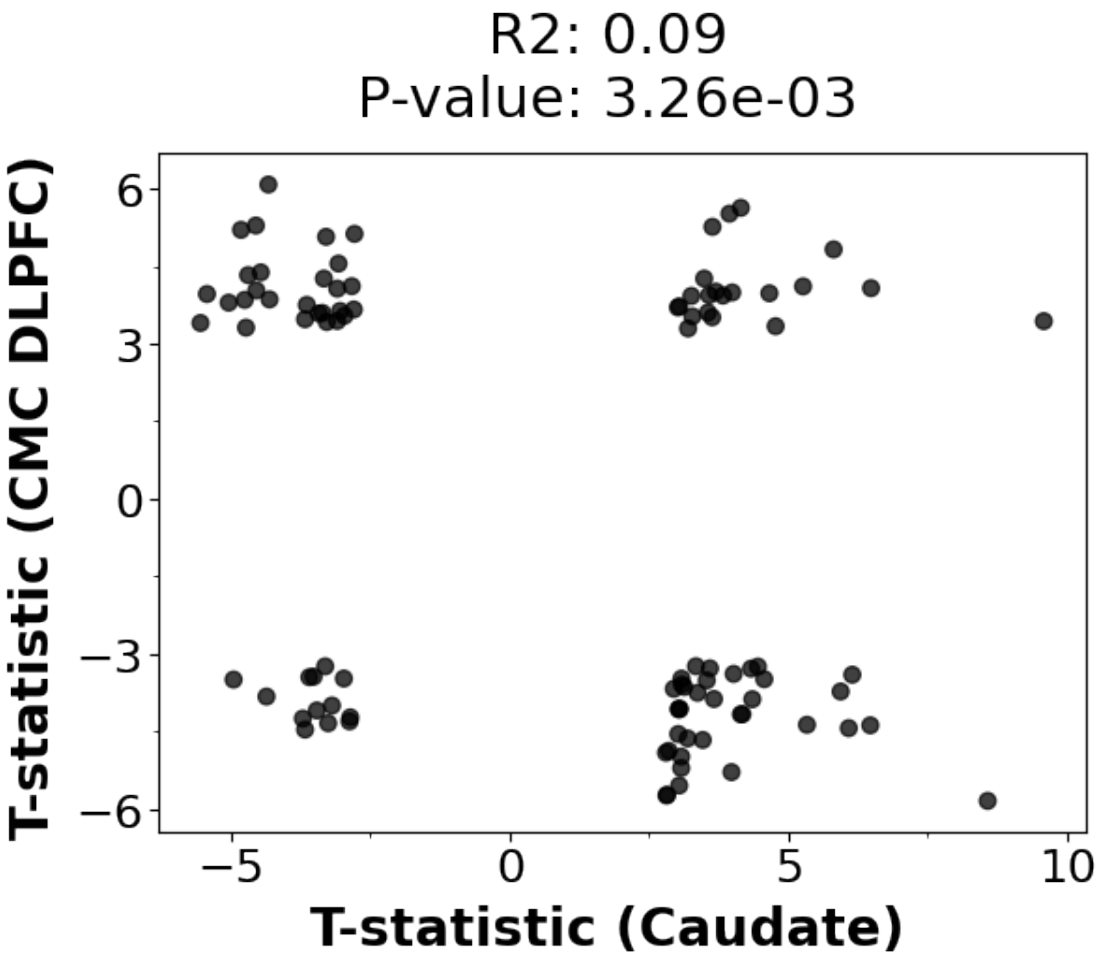


[42]: <ggplot: (8793605816715)>

1.3.5 Significant correlation, FDR < 0.05

SVA correction

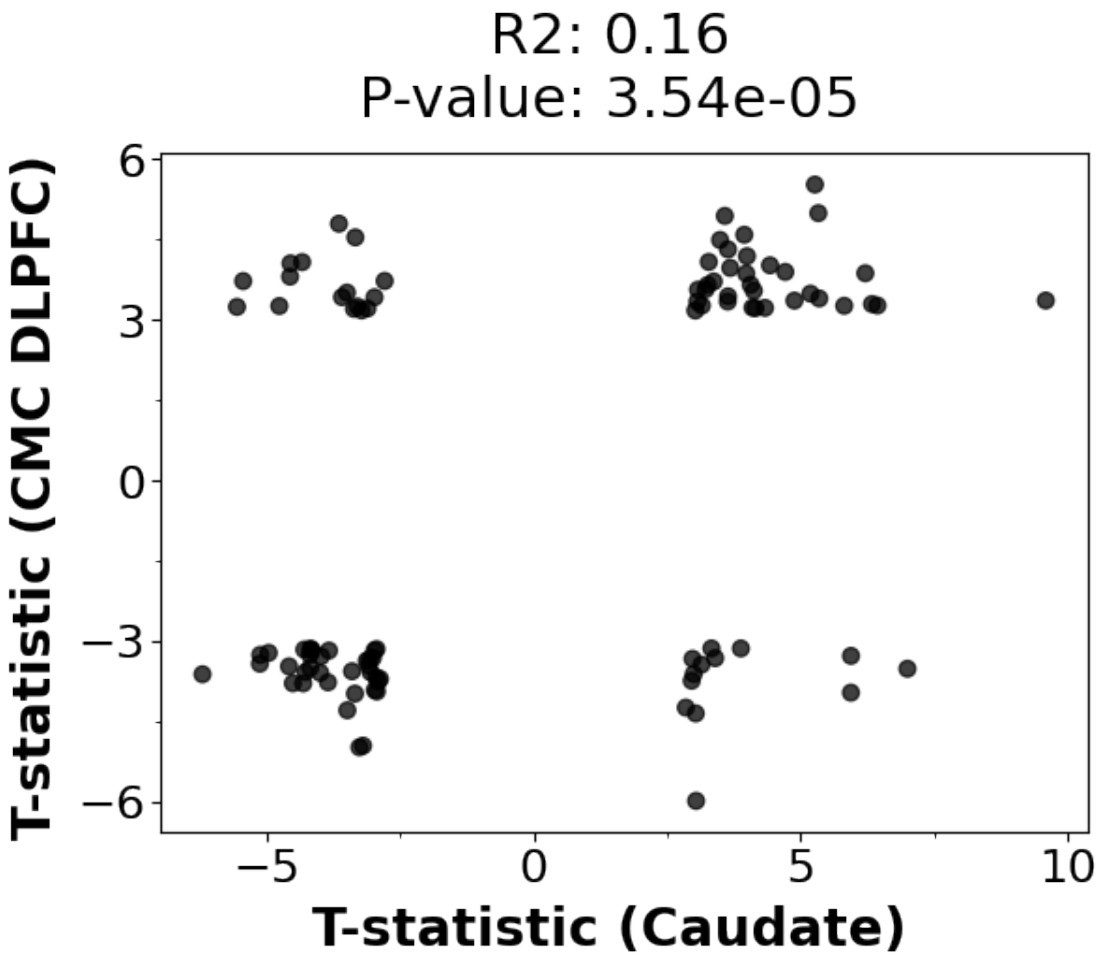
[43]:
```
pp = plot_corr('caudate', 'cmc', merge_cmc, True, True)
pp
```



[43]: <ggplot: (8793606824418)>

No SVA correction

[44]:
```
qq = plot_corr('caudate', 'cmc', merge_cmc, True, False)
qq
```



[44]: <ggplot: (8793607258185)>

1.3.6 Directionality test

All genes

SVA correction

[45]:
```
enrichment_binom('caudate', 'cmc', merge_cmc, False, True)
```

```
   agree     0
0   -1.0  7924
1    1.0  7463
```

[45]: 0.00020840229110907974

[46]:
```
enrichment_binom('dlpfc', 'cmc', merge_cmc, False, True)
```

```
   agree     0
0   -1.0  6455
1    1.0  8739
```

[46]:
```
7.258471671003183e-77
```

[47]:
```
enrichment_binom('hippo', 'cmc', merge_cmc, False, True)
```

```
   agree     0
0   -1.0  7433
1    1.0  7761
```

[47]:
```
0.007979600732433059
```

No SVA correction

[48]:
```
enrichment_binom('caudate', 'cmc', merge_cmc, False, False)
```

```
   agree     0
0   -1.0  6608
1    1.0  8779
```

[48]:
```
9.589616811740652e-69
```

[49]:
```
enrichment_binom('dlpfc', 'cmc', merge_cmc, False, False)
```

```
   agree     0
0   -1.0  6904
1    1.0  8290
```

[49]:
```
2.4847225859034744e-29
```

[50]:
```
enrichment_binom('hippo', 'cmc', merge_cmc, False, False)
```

```
   agree     0
0   -1.0  6856
1    1.0  8338
```

[50]:
```
2.6467222183158303e-33
```

Significant DEG (FDR $<$ 0.05)

SVA correction

[51]:
```
enrichment_binom('caudate', 'cmc', merge_cmc, True, True)
```

```
   agree   0
0   -1.0  63
1    1.0  34
```

[51]:
```
0.0042258039216827616
```

[52]:
```
enrichment_binom('dlpfc', 'cmc', merge_cmc, True, True)
```

```
   agree   0
0   -1.0   1
1    1.0  29
```

[52]:
```
5.774199962615967e-08
```

[53]:
```
#enrichment_binom('hippo', 'cmc', merge_cmc, True, True)
```

No SVA correction

[54]:
```
enrichment_binom('caudate', 'cmc', merge_cmc, True, False)
```

```
   agree   0
0   -1.0  29
1    1.0  69
```

[54]:
```
6.572240952992274e-05
```

[55]:
```
enrichment_binom('dlpfc', 'cmc', merge_cmc, True, False)
```

```
   agree   0
0    1.0  16
All directions agree!
```

[56]:
```
enrichment_binom('hippo', 'cmc', merge_cmc, True, False)
```

```
   agree  0
0    1.0  2
All directions agree!
```

[ ]:
```

```