main

July 12, 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_EAvsAA_full.txt',
    'dlpfc': '../../../dlpfc/_m/genes/diffExpr_EAvsAA_full.txt',
    'hippo': '../../../hippocampus/_m/genes/diffExpr_EAvsAA_full.txt',
    'gyrus': '../../../dentateGyrus/_m/genes/diffExpr_EAvsAA_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, 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):
    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='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)
    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):
    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', 'gyrus': 'Dentate Gyrus'}[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 Sample summary

[6]:
```
pheno_file = '../../../../input/phenotypes/merged/_m/merged_phenotypes.csv'
pheno = pd.read_csv(pheno_file, index_col=0)
pheno = pheno[(pheno['Age'] > 17) &
              (pheno['Dx'].isin(['Schizo', 'Control'])) &
              (pheno['Race'].isin(['AA', "CAUC"]))].copy()
pheno.head(2)
```

[6]:
```
          BrNum    RNum   Region  RIN    Age Sex Race      Dx  mitoRate  \
R12864  Br1303  R12864  Caudate  9.6  42.98   F   AA  Schizo  0.032654
R12865  Br1320  R12865  Caudate  9.5  53.12   M   AA  Schizo  0.019787

        rRNA_rate  overallMapRate
R12864   0.000087        0.909350
R12865   0.000070        0.873484
```

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

[7]:
```
Region
Caudate          394
DLPFC            360
DentateGyrus     161
HIPPO            376
dtype: int64
```

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

[8]:
```
Region        Race
Caudate       AA      205
              CAUC    189
DLPFC         AA      200
              CAUC    160
DentateGyrus  AA       78
              CAUC     83
HIPPO         AA      207
              CAUC    169
dtype: int64
```

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

[9]:
```
Region        Race  Sex
Caudate       AA    F       78
```

```
                      M      127
               CAUC   F       43
                      M      146
DLPFC          AA     F       75
                      M      125
               CAUC   F       39
                      M      121
DentateGyrus   AA     F       27
                      M       51
               CAUC   F       21
                      M       62
HIPPO          AA     F       81
                      M      126
               CAUC   F       40
                      M      129
dtype: int64
```

1.2 BrainSeq Tissue Comparison

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

[10]:
```
Dir
-1.0    10767
 1.0    11607
dtype: int64
```

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

[11]:
```
(2970, 6)
```

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

[12]:
```
Dir
-1.0    11691
 1.0    10707
dtype: int64
```

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

[13]:
```
(2760, 6)
```

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

[14]:
```
Dir
-1.0    11213
 1.0    11056
dtype: int64
```

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

[15]:
```
(2956, 6)
```

[16]:
```
gyrus = get_deg(config['gyrus'])
gyrus.groupby('Dir').size()
```

[16]:
```
Dir
-1.0    10855
 1.0    10285
dtype: int64
```

[17]:
```
gyrus[(gyrus['adj.P.Val'] < 0.05)].shape
```

[17]:
```
(786, 6)
```

1.2.1 Enrichment of DEG

[18]:
```
cal_fishers('caudate', 'dlpfc')
```

```
[[1115, 1692], [1507, 16814]]
```

[18]:
```
(7.352453718737303, 0.0)
```

[19]:
```
cal_fishers('caudate', 'hippo')
```

```
[[1142, 1681], [1726, 16648]]
```

[19]:
```
(6.552690661010558, 0.0)
```

[20]:
```
cal_fishers('dlpfc', 'hippo')
```

```
[[1251, 1437], [1610, 17300]]
```

[20]:
```
(9.354504078113045, 0.0)
```

[21]:
```
cal_fishers('caudate', 'gyrus')
```

```
[[311, 2231], [415, 16472]]
```

[21]:
```
(5.532979430046497, 1.069007184730363e-91)
```

[22]:
```
cal_fishers('dlpfc', 'gyrus')
```

```
[[342, 2117], [386, 16989]]
```

[22]: (7.110264549746562, 1.0684240001957168e-122)

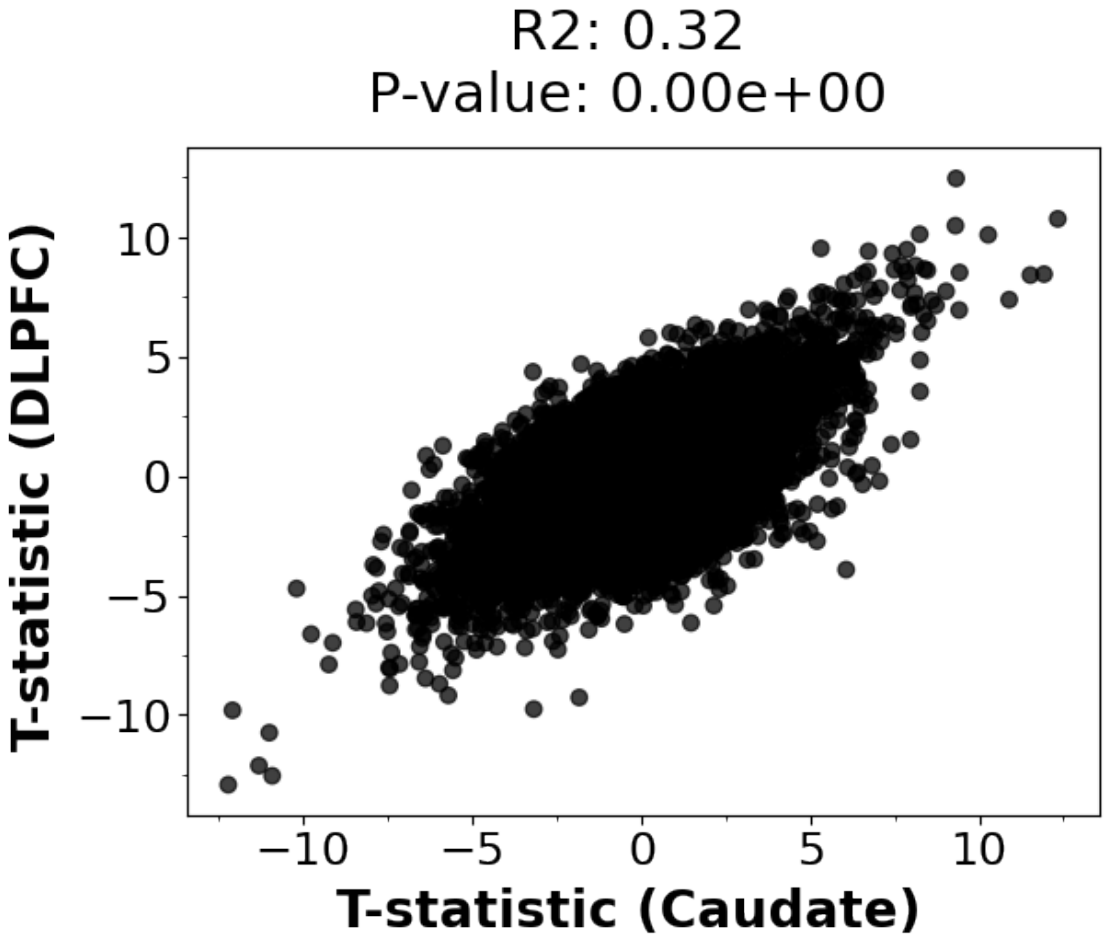
[23]:
```
cal_fishers('hippo', 'gyrus')
```

```
[[361, 2267], [382, 16834]]
```

[23]: (7.0174550862939, 3.432494766723409e-126)

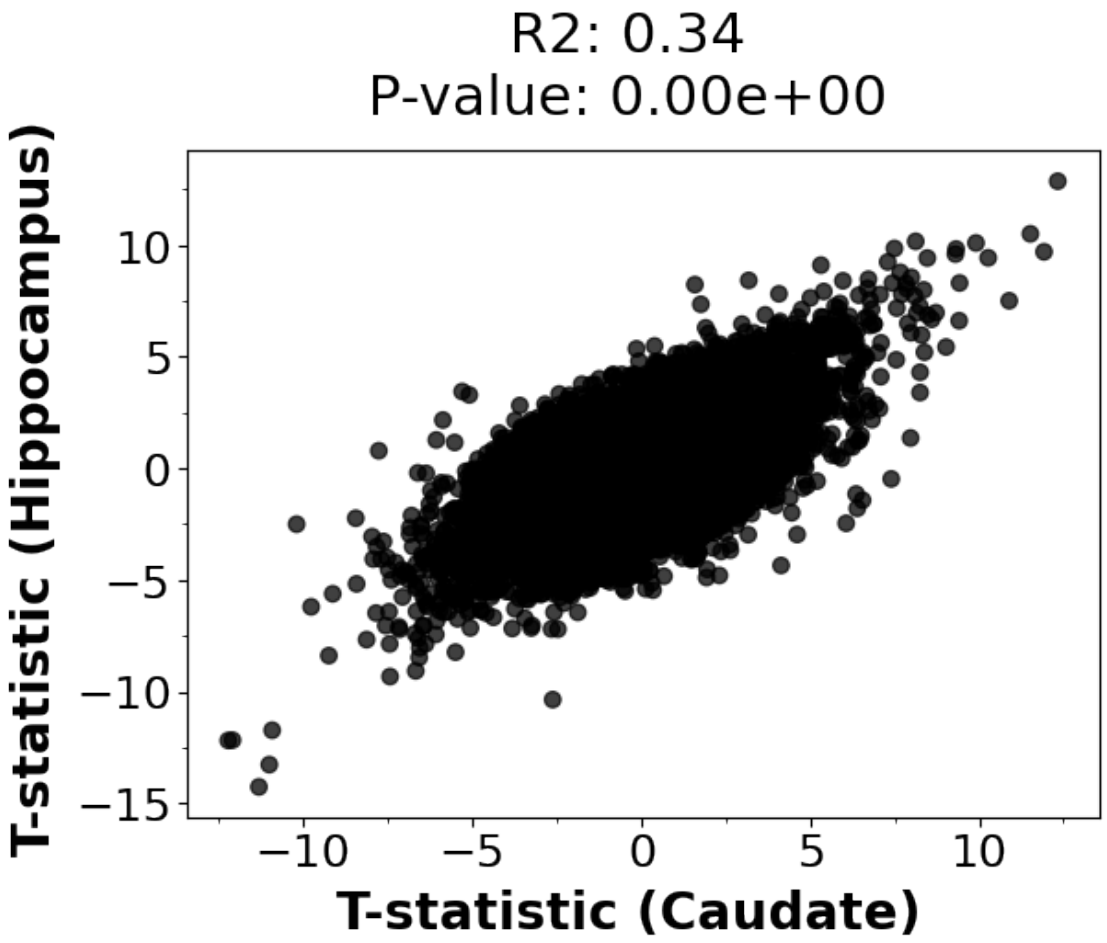
1.2.2 Correlation

[24]:
```
pp = plot_corr('caudate', 'dlpfc', merge_dataframes)
pp
```



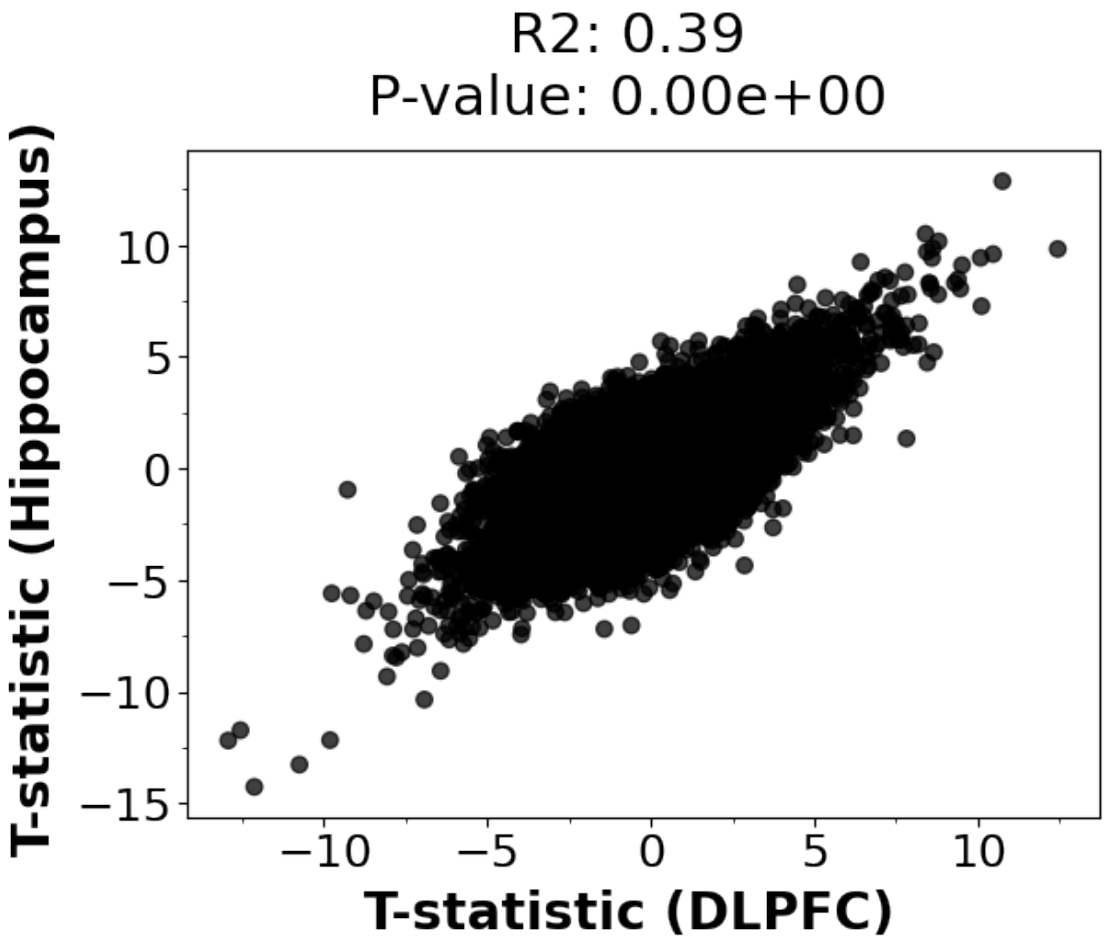
[24]: <ggplot: (8747080484307)>

[25]:
```
qq = plot_corr('caudate', 'hippo', merge_dataframes)
qq
```



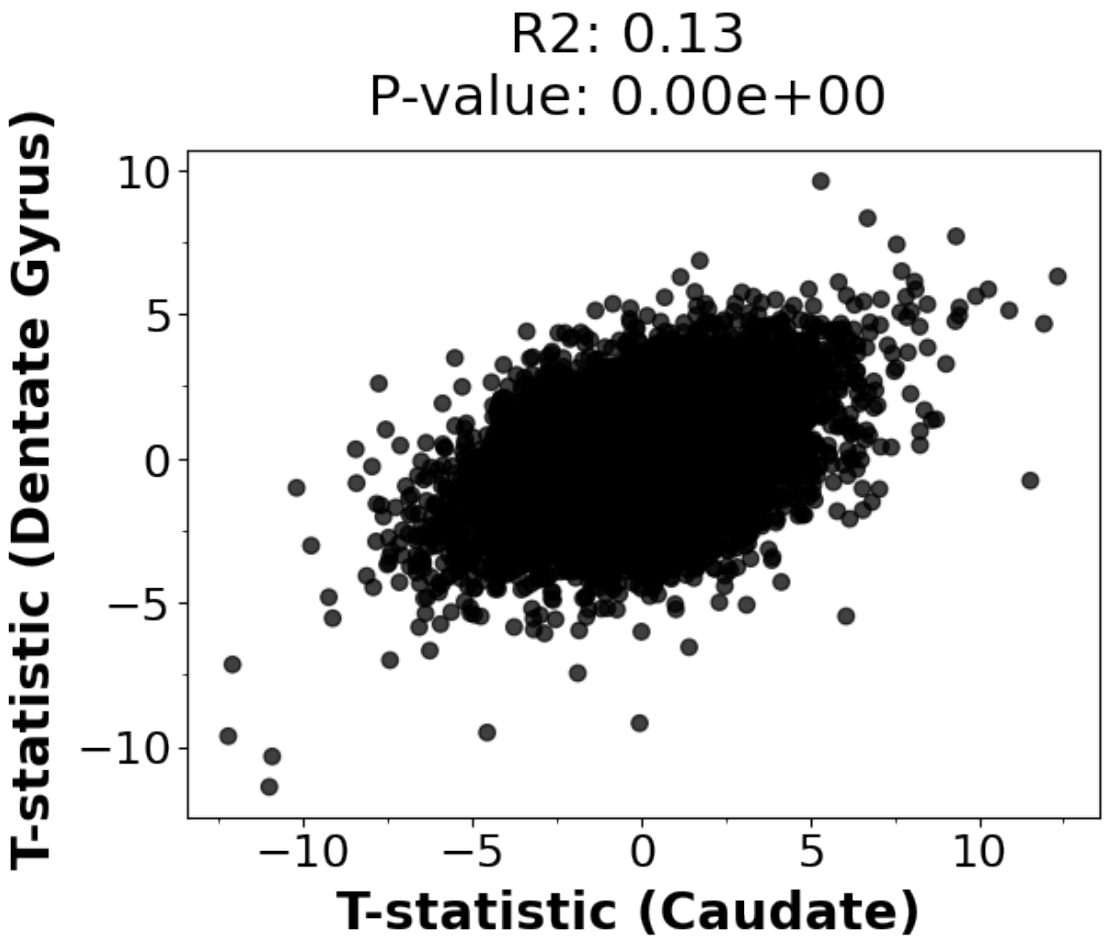
[25]:
```
<ggplot: (8747080016375)>
```

[26]:
```
ww = plot_corr('dlpfc', 'hippo', merge_dataframes)
ww
```



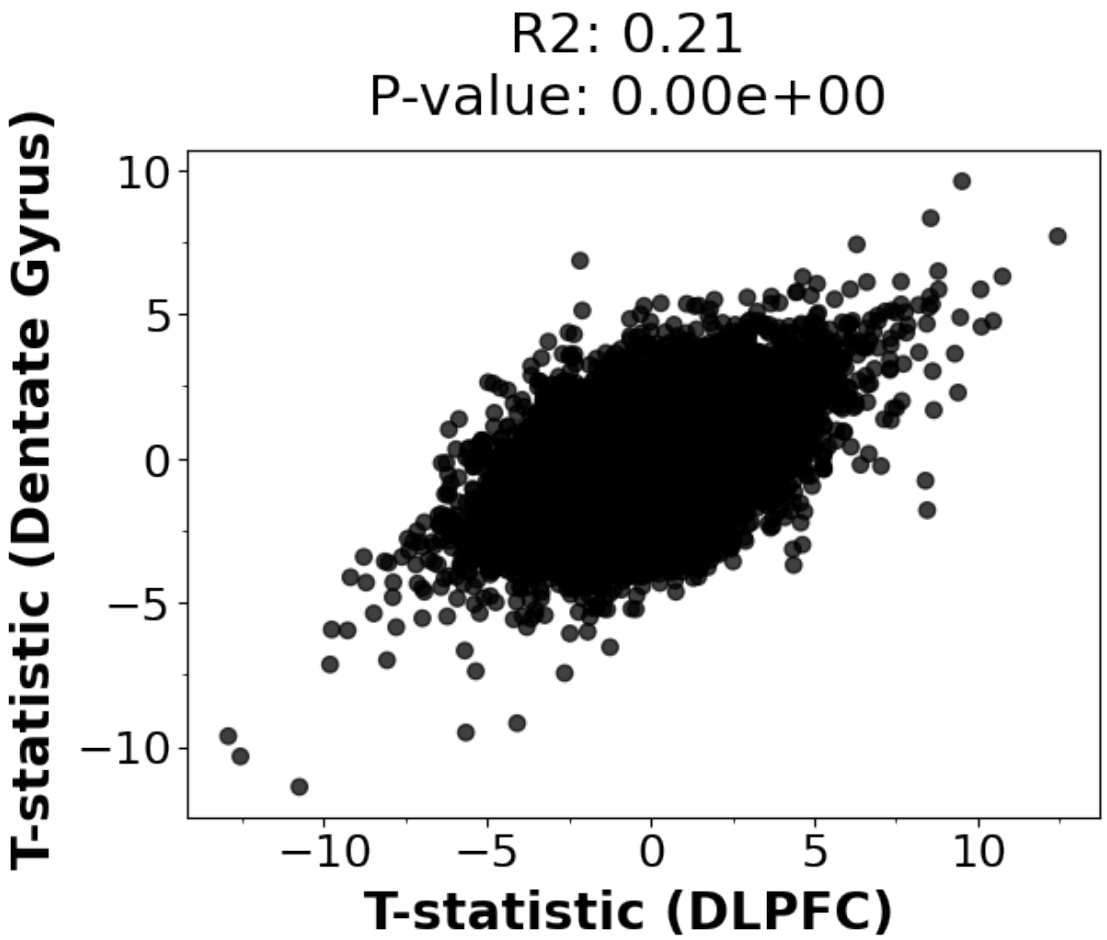
[26]: <ggplot: (8747082315048)>

[27]:
```
rr = plot_corr('caudate', 'gyrus', merge_dataframes)
rr
```



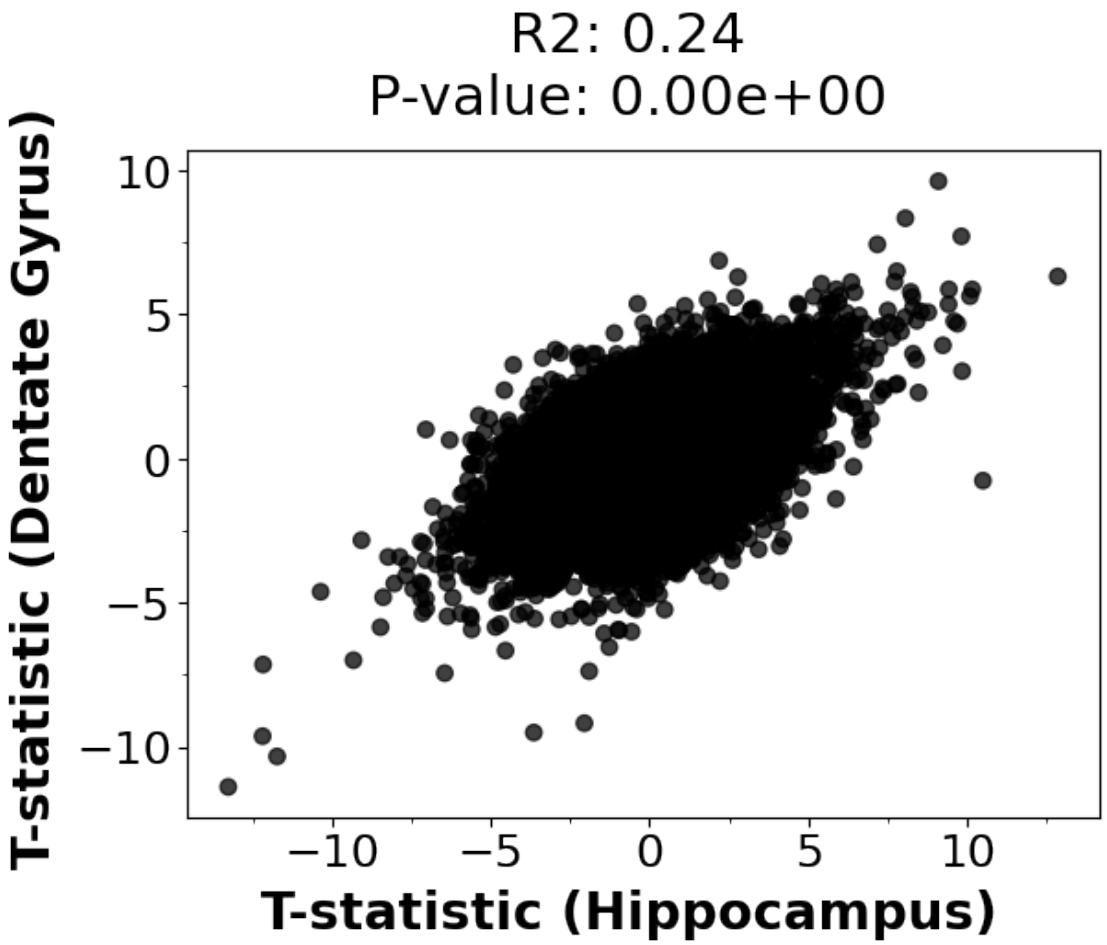
[27]:
```
<ggplot: (8747082316358)>
```

[28]:
```
ss = plot_corr('dlpfc', 'gyrus', merge_dataframes)
ss
```



[28]:
```
<ggplot: (8747082083853)>
```

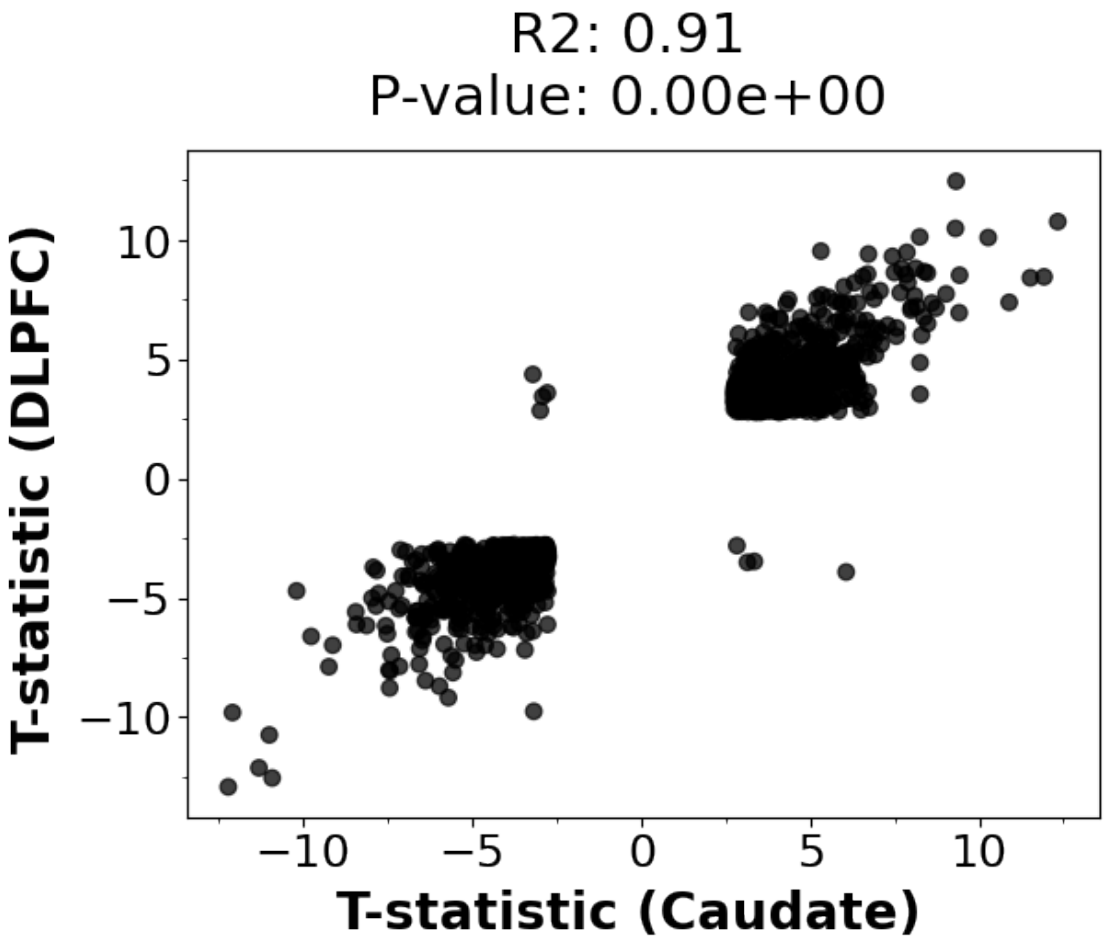
[29]:
```
tt = plot_corr('hippo', 'gyrus', merge_dataframes)
tt
```



[29]: <ggplot: (8747080212577)>

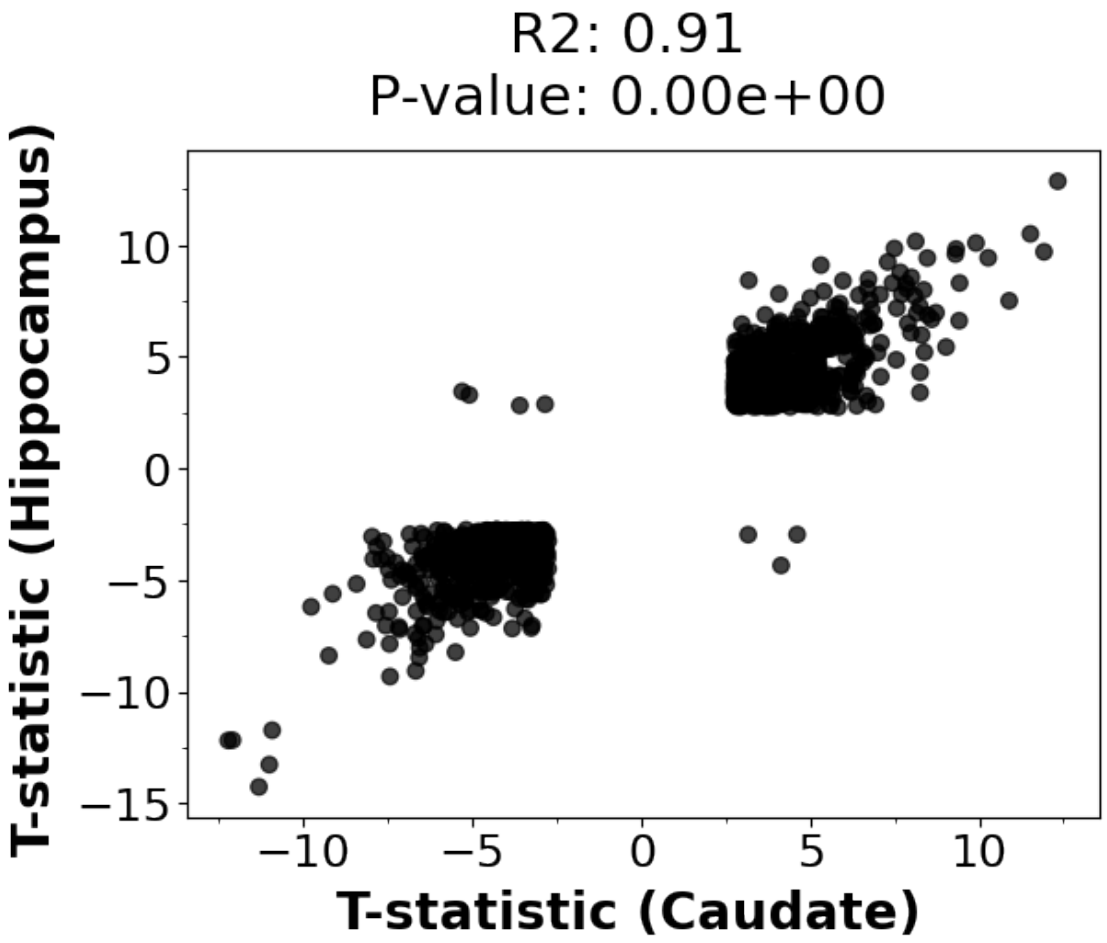
1.2.3 Significant correlation, FDR < 0.05

[30]:
```
pp = plot_corr('caudate', 'dlpfc', merge_dataframes_sig)
pp
```



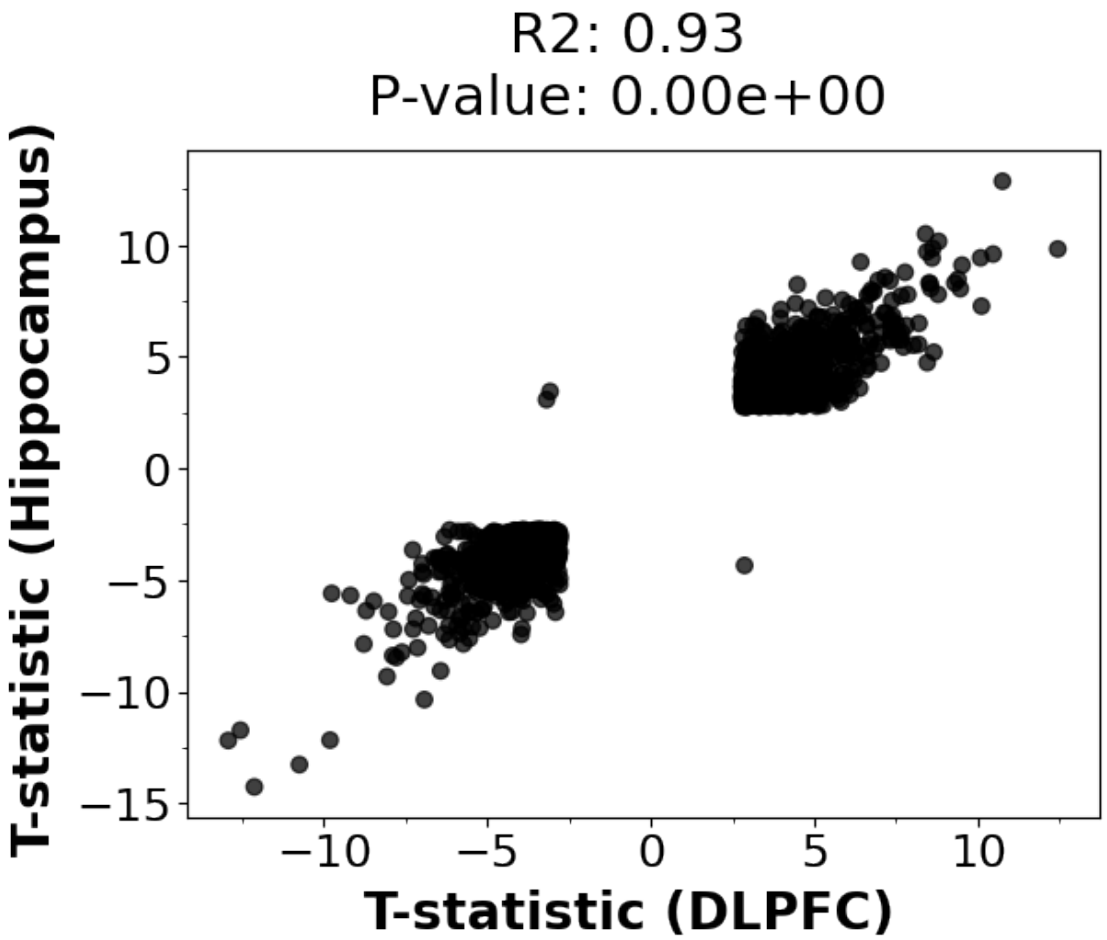
[30]: <ggplot: (8747080841280)>

[31]:
```
qq = plot_corr('caudate', 'hippo', merge_dataframes_sig)
qq
```



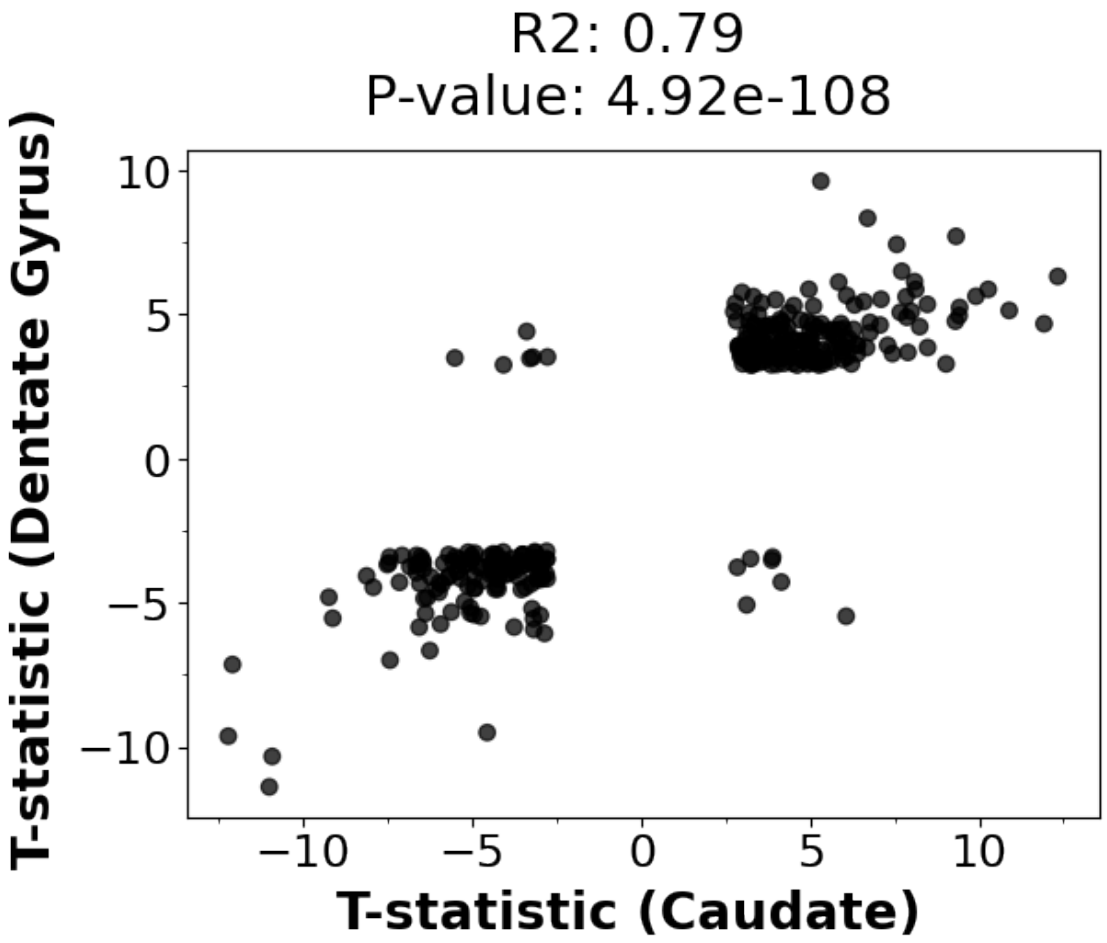
[31]: <ggplot: (8747080843632)>

[32]:
```
ww = plot_corr('dlpfc', 'hippo', merge_dataframes_sig)
ww
```



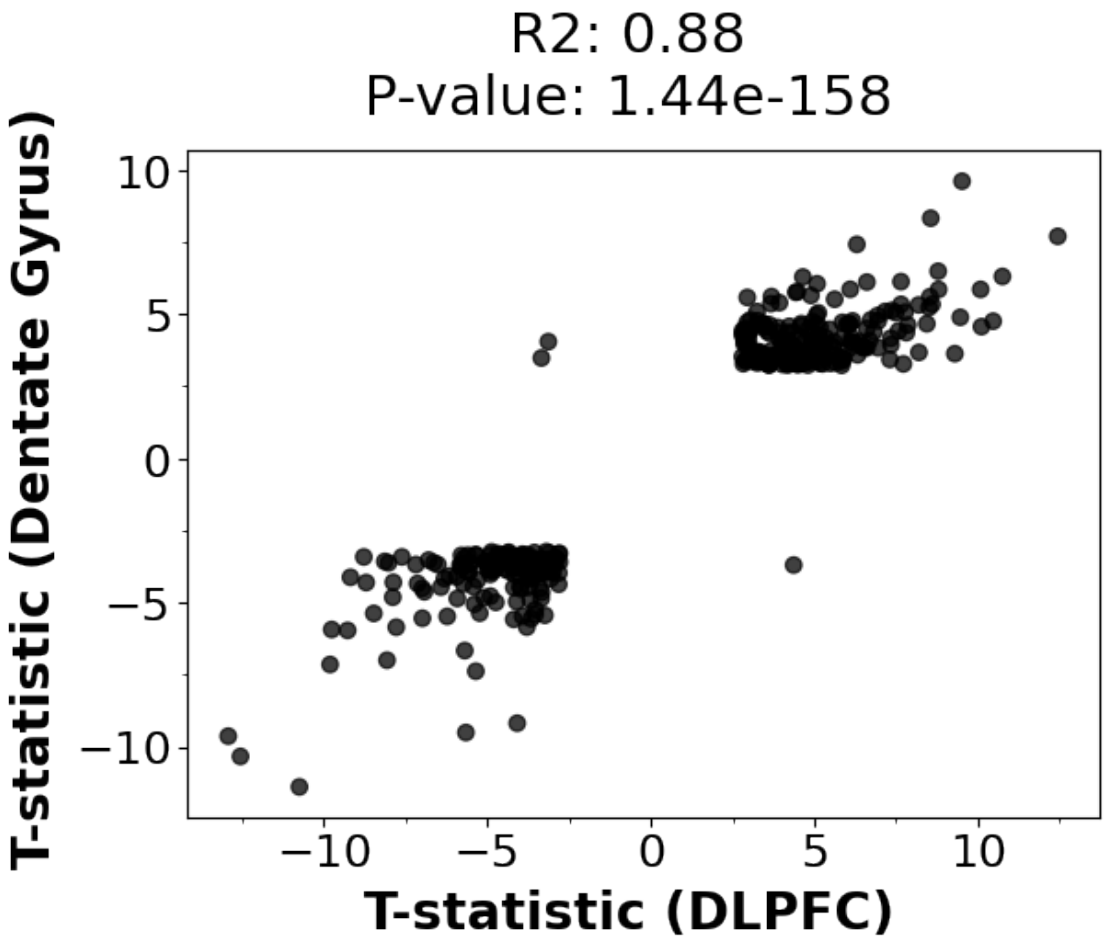
[32]:
```
<ggplot: (8747076498493)>
```

[33]:
```
rr = plot_corr('caudate', 'gyrus', merge_dataframes_sig)
rr
```



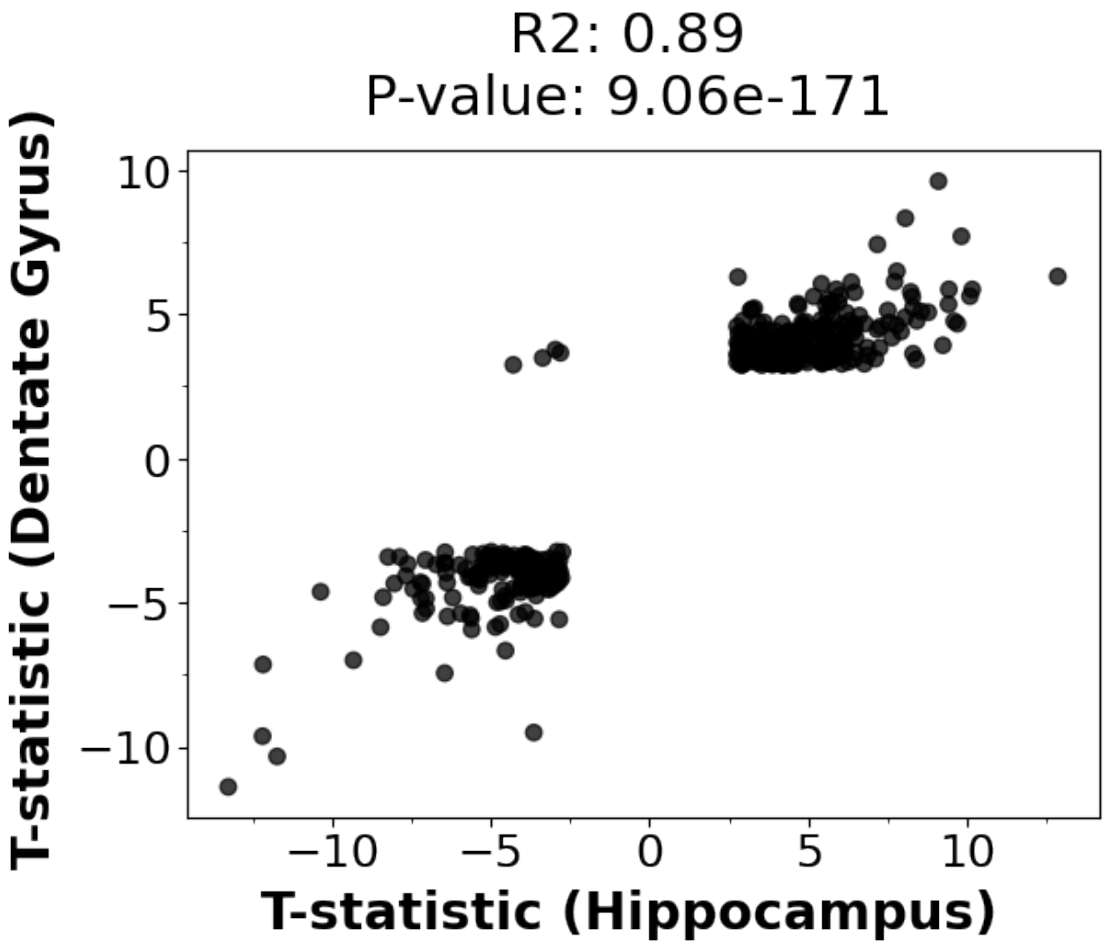
[33]: <ggplot: (8747079853162)>

[34]:
```
ss = plot_corr('dlpfc', 'gyrus', merge_dataframes_sig)
ss
```



[34]:
```
<ggplot: (8747082341392)>
```

[35]:
```
tt = plot_corr('hippo', 'gyrus', merge_dataframes_sig)
tt
```



[35]: <ggplot: (8747082294938)>

[36]:
```
#save_plot(pp, 'dlpfc_caudate_tstatistic_corr_sig')
#save_plot(qq, 'hippo_caudate_tstatistic_corr_sig')
#save_plot(ww, 'hippo_dlpfc_tstatistic_corr_sig')
```

1.2.4 Directionality test

All genes

[37]:
```
enrichment_binom('caudate', 'dlpfc', merge_dataframes)
```

```
   agree      0
0   -1.0   7508
1    1.0  13620
```

[37]: 5e-324

[38]:
```
enrichment_binom('caudate', 'hippo', merge_dataframes)
```

```
   agree      0
0   -1.0   7219
1    1.0  13978
```

[38]: 5e-324

[39]:
```
enrichment_binom('dlpfc', 'hippo', merge_dataframes)
```

```
   agree      0
0   -1.0   7251
1    1.0  14347
```

[39]: 5e-324

[40]:
```
enrichment_binom('caudate', 'gyrus', merge_dataframes)
```

```
   agree      0
0   -1.0   8063
1    1.0  11366
```

[40]: 1.1857793882825218e-124

[41]:
```
enrichment_binom('dlpfc', 'gyrus', merge_dataframes)
```

```
   agree      0
0   -1.0   7509
1    1.0  12325
```

[41]: 9.716255782985859e-259

[42]:
```
enrichment_binom('hippo', 'gyrus', merge_dataframes)
```

```
   agree      0
0   -1.0   7108
1    1.0  12736
```

[42]: 5e-324

Significant DEG (FDR < 0.05)

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

```
   agree     0
0   -1.0     8
1    1.0  1107
```

[43]: 2.61503106e-316

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

```
   agree     0
0   -1.0     7
1    1.0  1135
```

[44]: 0.0

[45]:
```
enrichment_binom('dlpfc', 'hippo', merge_dataframes_sig)
```

```
   agree     0
0   -1.0     3
1    1.0  1248
```

[45]: 0.0

[46]:
```
enrichment_binom('caudate', 'gyrus', merge_dataframes_sig)
```

```
   agree    0
0   -1.0   13
1    1.0  298
```

[46]: 1.589014927874492e-71

[47]:
```
enrichment_binom('dlpfc', 'gyrus', merge_dataframes_sig)
```

```
   agree    0
0   -1.0    3
1    1.0  339
```

[47]: 1.488391483735955e-96

[48]:
```
enrichment_binom('hippo', 'gyrus', merge_dataframes_sig)
```

```
   agree    0
0   -1.0    4
1    1.0  357
```

[48]: 2.996700665181341e-100

[ ]: