main

June 8, 2023

1 Examine overlaps with published data

[1]:

```
import numpy as np
import pandas as pd
import session_info
from pyhere import here
```

1.1 Public si-eQTL analysis

[2]:

```
shen = ["GDAP2", "AIM2", "SLAMF6", "RLF", "ATG4C", "FUT7",
        "TMEM218", "C11orf74", "RAB35", "TMEM5", "HNRNPK",
        "CDCA3", "ERCC5", "GJB6", "SNTB2", "SPNS3",
        "XAF1", "RBBP8", "RUFY4", "CA2", "RAPGEF1"]
print("Shen et al.:")
print(len(shen))

kukurba = ["NOD2", "WDR36", "BSCL2", "MAP7D3", "RHOXF1", "DNAH1"]
print("Kukurba et al.:")
print(len(kukurba))

yao = ["NOD2", "HLA-DRB5", "HLA-DRB5", "KIAA0586", "PPP2R5A",
       "TSNAXIP1", "MUT", "GRIK2", "C15orf37", "LIMA1", "IL6ST",
       "HCG8", "BLOC1S3", "NKX3-1", "CXorf23"]
print("Yao et al.:")
print(len(np.unique(yao)))
print("Total of Yao + Kukurba:")
len(set(yao) | set(kukurba))
```

```
Shen et al.:
21
Kukurba et al.:
6
Yao et al.:
14
Total of Yao + Kukurba:
```

[2]:
```
19
```

1.2 Load BrainSeq si-eQTL results

1.2.1 Interacting variant-gene pairs

[3]:
```
bs0 = pd.read_csv("../../_m/BrainSeq_sexGenotypes_4features_3regions.txt.gz",␣
  ↪sep='\t')
bs0["ensembl_gene_id"] = bs0.gene_id.str.replace("\\..*", "", regex=True)
biomart = pd.read_csv("../_h/biomart.csv", index_col=0)
bs = bs0.merge(biomart, on="ensembl_gene_id").drop_duplicates(subset="gene_id")
print(bs.shape)
bs.tail(2)
```

```
(692, 15)
```

[3]:
```
       region             gene_id           variant_id          gencode_id  \
8285  Caudate   ENSG00000270605.1    chr1:28102893:G:C   ENSG00000270605.1
8286    DLPFC  ENSG00000187498.16  chr13:109650494:C:T  ENSG00000187498.16

            gene_name seqnames      start        end     lfsr  \
8285  ENSG00000270605     chr1   28239509   28241453  0.049873
8286           COL4A1    chr13  110148963  110307157  0.048863

      posterior_mean feature_type  ensembl_gene_id external_gene_name  \
8285       -0.261352         Gene  ENSG00000270605         AL353622.1
8286        0.193807         Gene  ENSG00000187498             COL4A1

      entrezgene                                         description
8285         NaN                                                 NaN
8286      1282.0  collagen type IV alpha 1 chain [Source:HGNC Sy…
```

[4]:
```
bs[(bs['external_gene_name'].isin(shen))].to_csv("siEQTL_Shen_comparison.csv",␣
  ↪index=False)
```

[5]:
```
bs[(bs['external_gene_name'].isin(kukurba))]
```

[5]:
```
Empty DataFrame
Columns: [region, gene_id, variant_id, gencode_id, gene_name, seqnames, start,
end, lfsr, posterior_mean, feature_type, ensembl_gene_id, external_gene_name,
entrezgene, description]
Index: []
```

[6]:
```
bs[(bs['external_gene_name'].isin(yao))]
```

```
[6]: Empty DataFrame
     Columns: [region, gene_id, variant_id, gencode_id, gene_name, seqnames, start,
     end, lfsr, posterior_mean, feature_type, ensembl_gene_id, external_gene_name,
     entrezgene, description]
     Index: []
```

```
[7]: bs[(bs['external_gene_name'].isin(shen+kukurba+yao))]
```

```
[7]:        region             gene_id         variant_id          gencode_id  \
     5787  Caudate  ENSG00000125703.15  chr1:63060301:G:A  ENSG00000125703.15
     8130  Caudate  ENSG00000104267.10  chr8:84966439:A:T  ENSG00000104267.10

          gene_name seqnames     start       end      lfsr  posterior_mean  \
     5787     ATG4C     chr1  62784132  62865516  0.019839        0.094136
     8130       CA2     chr8  85463968  85481493  0.045578        0.289557

          feature_type  ensembl_gene_id external_gene_name  entrezgene  \
     5787         Gene  ENSG00000125703              ATG4C     84938.0
     8130         Gene  ENSG00000104267                CA2       760.0

                                                 description
     5787  autophagy related 4C cysteine peptidase [Sourc…
     8130  carbonic anhydrase 2 [Source:HGNC Symbol;Acc:H…
```

1.3 GTEx comparison

```
[8]: gtex = pd.read_csv(here("input/public_results/gtex_results/_m",
                             "GTEx_Analysis_v8_sbeQTLs/GTEx_Analysis_v8_sbeQTLs.
      ↪txt"),
                        sep='\t')
     gtex.iloc[0:2, 0:10]
```

```
[8]:      ensembl_gene_id   hugo_gene_id               gene_type  \
     0  ENSG00000241860.6  RP11-34P13.13    processed_transcript
     1  ENSG00000227232.5         WASH7P  unprocessed_pseudogene

              variant_id        rs_id                 Tissue       maf  \
     0  chr1_14677_G_A_b38  rs201327123  Adipose_Subcutaneous  0.051635
     1  chr1_64764_C_T_b38  rs769952832  Adipose_Subcutaneous  0.061102

        pval_nominal_sb  slope_sb  slope_se_sb
     0         0.847114  0.055080     0.285537
     1         0.316881  0.222928     0.222511
```

```
[9]: gtex.iloc[0:2, 10:14]
```

```
[9]:    numtested  pvals.corrected      qval  pval_nominal_f
     0          1         0.847114  1.000000        0.022302
     1          1         0.316881  0.981254        0.003978
```

```
[10]: ## qval threshold equal to number of published sb-eQTL
      gtex[(gtex['qval'] < 0.25) & (gtex["Tissue"].str.contains("Brain"))]\
          .loc[:, ["ensembl_gene_id", "hugo_gene_id", "Tissue", "pvals.corrected",␣
       ↪'qval']].head(10)
```

```
[10]:           ensembl_gene_id  hugo_gene_id  \
      62155   ENSG00000026025.15           VIM
      116842  ENSG00000160818.16       GPATCH4
      121904  ENSG00000141562.17          NARF
      122123   ENSG00000267174.5  CTC-510F12.4

                                          Tissue  pvals.corrected      qval
      62155                       Brain_Amygdala         0.000004  0.012836
      116842  Brain_Nucleus_accumbens_basal_ganglia         0.000088  0.198445
      121904  Brain_Nucleus_accumbens_basal_ganglia         0.000056  0.198445
      122123  Brain_Nucleus_accumbens_basal_ganglia         0.000083  0.198445
```

```
[11]: ## qval threshold equal to number of published sb-eQTL
      gtex[(gtex['qval'] < 0.25) & (gtex["Tissue"].str.contains("Whole"))]\
          .loc[:, ["ensembl_gene_id", "hugo_gene_id", "Tissue", "pvals.corrected",␣
       ↪'qval']].head(10)
```

```
[11]:           ensembl_gene_id hugo_gene_id       Tissue  pvals.corrected  \
      362961   ENSG00000221571.3  RNU6ATAC35P  Whole_Blood         0.000039
      365043   ENSG00000196743.8         GM2A  Whole_Blood         0.000011
      367164  ENSG00000148459.15        PDSS1  Whole_Blood         0.000027

                  qval
      362961  0.139762
      365043  0.116825
      367164  0.139762
```

```
[12]: gtex_sig = gtex[(gtex['qval'] < 0.25)]
      gtex_sig.shape
```

```
[12]: (369, 22)
```

```
[13]: gtex_sig.head(10)
```

```
[13]:          ensembl_gene_id hugo_gene_id                           gene_type  \
      1096   ENSG00000076356.6       PLXNA2                      protein_coding
      5262  ENSG00000170632.13       ARMC10                      protein_coding
      5644  ENSG00000120907.17       ADRA1A                      protein_coding
```

```
6414   ENSG00000136830.11      FAM129B                      protein_coding
7220    ENSG00000166787.3        SAA3P  transcribed_unprocessed_pseudogene
8540    ENSG00000183463.5         URAD                      protein_coding
9191    ENSG00000282651.2   IGHV5-10-1                           IG_V_gene
14611  ENSG00000143933.16        CALM2                      protein_coding
15082   ENSG00000144410.4          CPO                      protein_coding
17452   ENSG00000211698.2        TRGV4                           TR_V_gene

                    variant_id        rs_id                    Tissue  \
1096    chr1_208030492_G_A_b38    rs3811383      Adipose_Subcutaneous   
5262    chr7_103076937_C_T_b38    rs6958836      Adipose_Subcutaneous   
5644     chr8_26839198_G_A_b38  rs117380715      Adipose_Subcutaneous   
6414    chr9_127584339_G_A_b38   rs10739693      Adipose_Subcutaneous   
7220    chr11_18269355_T_C_b38   rs34068567      Adipose_Subcutaneous   
8540    chr13_27990205_T_A_b38    rs7335293      Adipose_Subcutaneous   
9191   chr14_106114510_A_G_b38    rs4573838      Adipose_Subcutaneous   
14611    chr2_46225349_C_T_b38   rs12477148  Adipose_Visceral_Omentum   
15082   chr2_206822186_C_T_b38   rs12470278  Adipose_Visceral_Omentum   
17452    chr7_38361995_A_C_b38   rs10233345  Adipose_Visceral_Omentum   

            maf  pval_nominal_sb  slope_sb  slope_se_sb  …      qval  \
1096   0.123924     5.391600e-05  0.338278     0.083064  …  0.121068   
5262   0.169535     5.011130e-05  0.357403     0.087384  …  0.192900   
5644   0.216867     1.045890e-05 -0.323552     0.072676  …  0.084548   
6414   0.304647     7.387010e-07 -0.283660     0.056579  …  0.004976   
7220   0.278830     2.207290e-05  0.323030     0.075427  …  0.074347   
8540   0.500000     9.078700e-09 -0.444892     0.076123  …  0.000122   
9191   0.419105     2.025150e-05 -0.406760     0.094541  …  0.074347   
14611  0.072495     4.497930e-05 -0.480557     0.116471  …  0.161955   
15082  0.097015     3.204120e-05  0.682291     0.162191  …  0.115370   
17452  0.335821     6.438100e-05  0.427491     0.105837  …  0.139089   

       pval_nominal_f   slope_f  slope_se_f  pval_nominal_m   slope_m  \
1096     1.718880e-08  0.456729    0.075705    9.155700e-01  0.009739   
5262     4.933240e-01 -0.054539    0.079379    3.219220e-07 -0.429800   
5644     4.637410e-18 -0.779707    0.076596    3.976660e-10 -0.469672   
6414     1.978000e-06 -0.333315    0.066772    1.653380e-01 -0.082625   
7220     6.409400e-08  0.453034    0.078725    3.138000e-01  0.063002   
8540     9.982650e-21 -0.887723    0.078738    1.892290e-09 -0.457733   
9191     5.248710e-12 -0.682629    0.089412    2.805060e-03 -0.289091   
14611    4.747150e-04 -0.491287    0.134732    9.165740e-01  0.013043   
15082    1.165430e-01  0.280837    0.176978    4.431060e-06 -0.558002   
17452    7.011230e-06 -0.481758    0.100091    8.857670e-15 -1.068840   

       slope_se_m  pval_nominal     slope  slope_se  
1096     0.091682  2.747400e-05  0.171830  0.040604  
5262     0.079545  8.797530e-08 -0.216374  0.039857  
```

```
5644     0.069091  5.637370e-52 -0.568916  0.033334
6414     0.059205  1.393160e-08 -0.168762  0.029260
7220     0.062292  2.433600e-08  0.211910  0.037395
8540     0.070571  3.077310e-53 -0.640604  0.036976
9191     0.094806  3.458420e-21 -0.445408  0.045073
14611    0.124116  2.197750e-05 -0.246023  0.057281
15082    0.113158  7.896400e-06 -0.320288  0.070745
17452    0.112111  1.630390e-49 -0.838766  0.049090

[10 rows x 22 columns]
```

1.3.1 mashr

[14]:
```
gtex_overlap = bs[(bs['gene_id'].isin(gtex_sig.ensembl_gene_id))].
 ↪drop_duplicates()
print(gtex_overlap.shape)
gtex_overlap
```

```
(2, 15)
```

[14]:
```
       region            gene_id          variant_id         gencode_id  \
4943  Caudate  ENSG00000272977.1  chr22:25059120:A:C  ENSG00000272977.1
8285  Caudate  ENSG00000270605.1   chr1:28102893:G:C  ENSG00000270605.1

            gene_name seqnames     start       end      lfsr  posterior_mean  \
4943  ENSG00000272977    chr22  25476218  25479971  0.011928        0.323847
8285  ENSG00000270605     chr1  28239509  28241453  0.049873       -0.261352

     feature_type  ensembl_gene_id external_gene_name  entrezgene description
4943         Gene  ENSG00000272977         AL008721.2         NaN         NaN
8285         Gene  ENSG00000270605         AL353622.1         NaN         NaN
```

[15]:
```
gtex_overlap.shape[0]/bs.shape[0] * 100
```

[15]:
```
0.2890173410404624
```

[16]:
```
gtex_sig[(gtex_sig['ensembl_gene_id'].isin(bs.gene_id))]
```

[16]:
```
          ensembl_gene_id   hugo_gene_id       gene_type  \
297207  ENSG00000270605.1   RP5-1092A3.4       antisense
338770  ENSG00000272977.1  CTA-390C10.10  sense_intronic

                 variant_id      rs_id                         Tissue  \
297207   chr1_28223937_C_T_b38   rs481640  Skin_Not_Sun_Exposed_Suprapubic
338770  chr22_25459662_G_A_b38  rs6004655                          Spleen
```

```
              maf  pval_nominal_sb  slope_sb  slope_se_sb  …      qval  \
297207  0.323985         0.000237 -0.241524     0.065161  …  0.208195
338770  0.167401         0.000038  0.413128     0.097856  …  0.212883

        pval_nominal_f   slope_f  slope_se_f  pval_nominal_m   slope_m  \
297207    4.046590e-21 -0.760755    0.063142    6.395880e-13 -0.599143
338770    7.931210e-14 -1.129030    0.107327    5.155260e-16 -1.276590

        slope_se_m  pval_nominal     slope  slope_se
297207     0.07247  9.917080e-60 -0.591873  0.031047
338770     0.10469  4.896640e-51 -1.193670  0.056659

[2 rows x 22 columns]
```

```
[17]: gtex_sig[(gtex_sig['ensembl_gene_id'].isin(bs.gene_id))]\
          .to_csv("siEQTL_gtex_comparison.csv", index=False)
```

2 Session information

```
[18]: session_info.show()
```

```
[18]: <IPython.core.display.HTML object>
```