main

August 3, 2021

1 Extract male bias genes on the X chromosome

[1]:

```
import pandas as pd
```

[2]:

```
def get_deg(tissue):
    fn = "../../../../differential_expression/%s/" % tissue +\
    "metrics_summary/_m/chrom_annotation_genes.txt"
    return pd.read_csv(fn, sep='\t').loc[:, ["gene_id", "seqname", "Symbol",
                                                         "t", "adj.P.Val"]]
```

[3]:

```
df = pd.DataFrame()
for tissue in ["caudate", "dlpfc", "hippocampus"]:
    dt = get_deg(tissue)
    dt["Tissue"] = tissue
    df = pd.concat([df, dt], axis=0)
df["ensemblID"] = df.gene_id.str.replace("\\..*", "", regex=True)
df.shape
```

[3]:

```
(1058, 7)
```

[4]:

```
xci = pd.read_csv("../../_h/xci_status_hg19.txt", sep='\t')
xci["ensemblID"] = xci["Gene ID"].str.replace("\\..*", "", regex=True)
xci.head(2)
```

[4]:

```
  Gene name            Gene ID Chr  Start position  End position  \
0    PLCXD1  ENSG00000182378.8   X          192989        220023
1    GTPBP6  ENSG00000178605.8   X          220025        230886

  Transcript type Combined XCI status        ensemblID
0  protein_coding              escape  ENSG00000182378
1  protein_coding              escape  ENSG00000178605
```

[5]:

```
xci.groupby("Combined XCI status").size()
```

[5]:

```
Combined XCI status
escape       99
inactive    431
variable    101
```

```
dtype: int64
```

```
[6]: tt = df.merge(xci[(xci["Combined XCI status"] == "escape")], on="ensemblID")
     tt[(tt['t'] > 0)]
```

```
[6]:                    gene_id seqname  Symbol         t     adj.P.Val  \
     55  ENSG00000182378.13_PAR_Y    chrY  PLCXD1  6.988151  4.239261e-09
     56  ENSG00000182378.13_PAR_Y    chrY  PLCXD1  6.107645  9.826634e-07
     57  ENSG00000182378.13_PAR_Y    chrY  PLCXD1  4.772588  7.966110e-04
     72  ENSG00000002586.18_PAR_Y    chrY    CD99  4.264277  3.645847e-03
     73  ENSG00000002586.18_PAR_Y    chrY    CD99  3.901722  1.477514e-02
     74  ENSG00000002586.18_PAR_Y    chrY    CD99  4.073952  1.481474e-02
     80  ENSG00000169093.15_PAR_Y    chrY   ASMTL  3.831986  1.701108e-02
     83  ENSG00000178605.13_PAR_Y    chrY  GTPBP6  3.599197  2.847852e-02

             Tissue        ensemblID Gene name             Gene ID Chr  \
     55     caudate  ENSG00000182378    PLCXD1   ENSG00000182378.8   X
     56       dlpfc  ENSG00000182378    PLCXD1   ENSG00000182378.8   X
     57  hippocampus  ENSG00000182378    PLCXD1   ENSG00000182378.8   X
     72     caudate  ENSG00000002586      CD99  ENSG00000002586.13   X
     73       dlpfc  ENSG00000002586      CD99  ENSG00000002586.13   X
     74  hippocampus  ENSG00000002586      CD99  ENSG00000002586.13   X
     80       dlpfc  ENSG00000169093     ASMTL  ENSG00000169093.10   X
     83       dlpfc  ENSG00000178605    GTPBP6   ENSG00000178605.8   X

         Start position  End position Transcript type Combined XCI status
     55          192989        220023  protein_coding              escape
     56          192989        220023  protein_coding              escape
     57          192989        220023  protein_coding              escape
     72         2609220       2659350  protein_coding              escape
     73         2609220       2659350  protein_coding              escape
     74         2609220       2659350  protein_coding              escape
     80         1522032       1572655  protein_coding              escape
     83          220025        230886  protein_coding              escape
```

Escaped genes are also located on the PAR regions of the Y chromosome.

```
[7]: xlinked = df[(df['seqname'] == 'chrX')].copy()
     xx_male = df[(df['seqname'].isin(["chrX", "chrY"])) & (df["t"] > 0)].copy()
     xlinked_male = xlinked[(xlinked["t"] > 0)].copy()
     xlinked_female = xlinked[(xlinked["t"] < 0)].copy()
```

```
[8]: xlinked.groupby("Tissue").size()
```

```
[8]: Tissue
     caudate        45
     dlpfc          60
     hippocampus    31
```

```
dtype: int64
```

```
[9]: xlinked_male.groupby("Tissue").size()
```

```
[9]: Tissue
     caudate         3
     dlpfc          18
     hippocampus     1
     dtype: int64
```

```
[10]: xlinked_female.groupby("Tissue").size()
```

```
[10]: Tissue
      caudate        42
      dlpfc          42
      hippocampus    30
      dtype: int64
```

```
[11]: xlinked_male
```

```
[11]:
                gene_id seqname     Symbol         t  adj.P.Val       Tissue  \
126   ENSG00000213468.4    chrX      FIRRE  4.689886   0.000705      caudate
132   ENSG00000186675.6    chrX     MAGEE2  4.613375   0.000956      caudate
133  ENSG00000102001.12    chrX    CACNA1F  4.608831   0.000963      caudate
97   ENSG00000172465.13    chrX     TCEAL1  4.402338   0.003347        dlpfc
99    ENSG00000236064.1    chrX        NaN  4.364580   0.003866        dlpfc
125   ENSG00000277883.1    chrX    NLRP3P1  4.155232   0.007473        dlpfc
130  ENSG00000184515.10    chrX       BEX5  4.126387   0.008105        dlpfc
199  ENSG00000204071.10    chrX     TCEAL6  3.836015   0.017011        dlpfc
232  ENSG00000147155.10    chrX        EBP  3.751150   0.020360        dlpfc
277   ENSG00000232119.7    chrX      MCTS1  3.632172   0.026693        dlpfc
291  ENSG00000198932.12    chrX    GPRASP1  3.611593   0.027517        dlpfc
314   ENSG00000186675.6    chrX     MAGEE2  3.555431   0.031388        dlpfc
330  ENSG00000102054.17    chrX      RBBP7  3.530376   0.032470        dlpfc
361   ENSG00000278530.4    chrX   CHMP1B2P  3.485690   0.035151        dlpfc
384   ENSG00000133169.5    chrX       BEX1  3.454265   0.037068        dlpfc
399   ENSG00000184905.8    chrX     TCEAL2  3.428241   0.039058        dlpfc
408  ENSG00000184867.13    chrX     ARMCX2  3.417176   0.039832        dlpfc
483  ENSG00000102401.19    chrX     ARMCX3  3.328593   0.045811        dlpfc
486  ENSG00000133134.11    chrX       BEX2  3.325992   0.045811        dlpfc
505   ENSG00000224204.1    chrX   PHEX-AS1  3.311985   0.046573        dlpfc
528   ENSG00000178947.8    chrX  SMIM10L2A  3.297549   0.047002        dlpfc
85   ENSG00000147124.12    chrX      ZNF41  4.105295   0.013322  hippocampus

           ensemblID
126  ENSG00000213468
132  ENSG00000186675
```

```
133   ENSG00000102001
97    ENSG00000172465
99    ENSG00000236064
125   ENSG00000277883
130   ENSG00000184515
199   ENSG00000204071
232   ENSG00000147155
277   ENSG00000232119
291   ENSG00000198932
314   ENSG00000186675
330   ENSG00000102054
361   ENSG00000278530
384   ENSG00000133169
399   ENSG00000184905
408   ENSG00000184867
483   ENSG00000102401
486   ENSG00000133134
505   ENSG00000224204
528   ENSG00000178947
85    ENSG00000147124
```

[12]:

```
xci["ensemblID"] = xci["Gene ID"].str.replace("\\..*", "", regex=True)
xlinked_male["ensemblID"] = xlinked_male.gene_id.str.replace("\\..*", "",␣
 ↪regex=True)
xlinked_male.merge(xci[["ensemblID", "Combined XCI status"]], on="ensemblID",␣
 ↪how="left").fillna("unknown")
```

[12]:

```
               gene_id seqname    Symbol         t  adj.P.Val     Tissue  \
0    ENSG00000213468.4    chrX     FIRRE  4.689886   0.000705    caudate
1    ENSG00000186675.6    chrX    MAGEE2  4.613375   0.000956    caudate
2   ENSG00000102001.12    chrX   CACNA1F  4.608831   0.000963    caudate
3   ENSG00000172465.13    chrX    TCEAL1  4.402338   0.003347      dlpfc
4    ENSG00000236064.1    chrX   unknown  4.364580   0.003866      dlpfc
5    ENSG00000277883.1    chrX   NLRP3P1  4.155232   0.007473      dlpfc
6   ENSG00000184515.10    chrX      BEX5  4.126387   0.008105      dlpfc
7   ENSG00000204071.10    chrX    TCEAL6  3.836015   0.017011      dlpfc
8   ENSG00000147155.10    chrX       EBP  3.751150   0.020360      dlpfc
9    ENSG00000232119.7    chrX     MCTS1  3.632172   0.026693      dlpfc
10  ENSG00000198932.12    chrX   GPRASP1  3.611593   0.027517      dlpfc
11   ENSG00000186675.6    chrX    MAGEE2  3.555431   0.031388      dlpfc
12  ENSG00000102054.17    chrX     RBBP7  3.530376   0.032470      dlpfc
13   ENSG00000278530.4    chrX  CHMP1B2P  3.485690   0.035151      dlpfc
14   ENSG00000133169.5    chrX      BEX1  3.454265   0.037068      dlpfc
15   ENSG00000184905.8    chrX    TCEAL2  3.428241   0.039058      dlpfc
16  ENSG00000184867.13    chrX    ARMCX2  3.417176   0.039832      dlpfc
17  ENSG00000102401.19    chrX    ARMCX3  3.328593   0.045811      dlpfc
18  ENSG00000133134.11    chrX      BEX2  3.325992   0.045811      dlpfc
```

```
19   ENSG00000224204.1    chrX   PHEX-AS1  3.311985  0.046573        dlpfc
20   ENSG00000178947.8    chrX  SMIM10L2A  3.297549  0.047002        dlpfc
21  ENSG00000147124.12    chrX      ZNF41  4.105295  0.013322  hippocampus

          ensemblID Combined XCI status
0   ENSG00000213468            variable
1   ENSG00000186675             unknown
2   ENSG00000102001            inactive
3   ENSG00000172465            inactive
4   ENSG00000236064            inactive
5   ENSG00000277883             unknown
6   ENSG00000184515             unknown
7   ENSG00000204071             unknown
8   ENSG00000147155             unknown
9   ENSG00000232119            inactive
10  ENSG00000198932            inactive
11  ENSG00000186675             unknown
12  ENSG00000102054            variable
13  ENSG00000278530             unknown
14  ENSG00000133169             unknown
15  ENSG00000184905            inactive
16  ENSG00000184867            inactive
17  ENSG00000102401            inactive
18  ENSG00000133134            inactive
19  ENSG00000224204            inactive
20  ENSG00000178947             unknown
21  ENSG00000147124            inactive
```

[13]:

```
dx = xlinked_male.merge(xci[["ensemblID", "Combined XCI status"]],␣
 ↪on="ensemblID", how="left").fillna("unknown")
dx = dx[(dx["Combined XCI status"] == "unknown")].copy()
```

[14]:

```
pd.concat([xx_male.merge(xci[["ensemblID", "Combined XCI status"]],␣
 ↪on="ensemblID"), dx], axis=0)\
  .sort_values(["Tissue",  "Combined XCI status", "seqname"], ascending=True)\
  .to_csv("BrainSeq_male_biased_genes_XCI_status.tsv", sep='\t', index=False)
```