main

March 14, 2023

1 Extract male bias genes on the X chromosome

[1]:
```
import session_info
import pandas as pd
from pyhere import here
```

[2]:
```
def get_deg():
    fn = here("differential_expression/tissue_comparison/summary_table",
              "_m/differential_expression_analysis_4features_sex.txt.gz")
    df = pd.read_csv(fn, sep='\t').loc[:, ["Tissue", "Feature", "ensemblID",␣
 ↪"Symbol",
                                           "seqnames", "Type", "t",␣
 ↪"Chrom_Type"]]
    return df[(df["Type"] == "Gene")].copy()
```

[3]:
```
df = get_deg()
```

[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]:

	Tissue	Feature	ensemblID	Symbol	seqnames	\
27	Caudate	CD99|ENSG00000002586.20	ENSG00000002586	CD99	chrX	
28	DLPFC	CD99|ENSG00000002586.20	ENSG00000002586	CD99	chrX	
29	Hippocampus	CD99|ENSG00000002586.20	ENSG00000002586	CD99	chrX	
36	Caudate	ZBED1|ENSG00000214717.12	ENSG00000214717	ZBED1	chrX	
37	DLPFC	ZBED1|ENSG00000214717.12	ENSG00000214717	ZBED1	chrX	
..	…	…	…	…	…	
213	Caudate	DNAAF6|ENSG00000080572.14	ENSG00000080572	DNAAF6	chrX	
216	DLPFC	IQSEC2|ENSG00000124313.18	ENSG00000124313	IQSEC2	chrX	
218	Caudate	NXF5|ENSG00000126952.18	ENSG00000126952	NXF5	chrX	
219	DLPFC	NXF5|ENSG00000126952.18	ENSG00000126952	NXF5	chrX	
220	Hippocampus	NXF5|ENSG00000126952.18	ENSG00000126952	NXF5	chrX	

	Type	t	Chrom_Type	Gene name	Gene ID	Chr	\
27	Gene	20.718194	Allosome	CD99	ENSG00000002586.13	X	
28	Gene	10.690853	Allosome	CD99	ENSG00000002586.13	X	
29	Gene	10.917016	Allosome	CD99	ENSG00000002586.13	X	
36	Gene	14.600959	Allosome	ZBED1	ENSG00000214717.5	X	
37	Gene	6.980985	Allosome	ZBED1	ENSG00000214717.5	X	
..	…	…	…	…	…	..	
213	Gene	0.315468	Allosome	PIH1D3	ENSG00000080572.8	X	
216	Gene	0.353653	Allosome	IQSEC2	ENSG00000124313.8	X	
218	Gene	0.137540	Allosome	NXF5	ENSG00000126952.12	X	
219	Gene	0.122630	Allosome	NXF5	ENSG00000126952.12	X	
220	Gene	0.196594	Allosome	NXF5	ENSG00000126952.12	X	

	Start position	End position	Transcript type	Combined XCI status
27	2609220	2659350	protein_coding	escape
28	2609220	2659350	protein_coding	escape
29	2609220	2659350	protein_coding	escape
36	2404455	2419008	protein_coding	escape
37	2404455	2419008	protein_coding	escape
..	…	…	…	…
213	106449862	106487473	protein_coding	escape
216	53262058	53350522	protein_coding	escape
218	101087085	101112549	protein_coding	escape
219	101087085	101112549	protein_coding	escape
220	101087085	101112549	protein_coding	escape

[78 rows x 15 columns]

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

[7]:
```
xlinked = df[(df['seqnames'] == 'chrX')].copy()
xx_male = df[(df['seqnames'].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        871
DLPFC          858
Hippocampus    876
dtype: int64
```

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

[9]:
```
Tissue
Caudate        492
DLPFC          462
Hippocampus    392
dtype: int64
```

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

[10]:
```
Tissue
Caudate        379
DLPFC          396
Hippocampus    484
dtype: int64
```

[11]:
```
xlinked_male
```

[11]:

	Tissue	Feature	ensemblID	\
53	Caudate	CD99|ENSG00000002586.20	ENSG00000002586	
56	Caudate	PRKCIP1|ENSG00000237682.2	ENSG00000237682	
62	Caudate	ZBED1|ENSG00000214717.12	ENSG00000214717	
65	Caudate	DHRSX|ENSG00000169084.15	ENSG00000169084	
66	Caudate	ENSG00000289007|ENSG00000289007.2	ENSG00000289007	
…	…	…	…	
1359849	Hippocampus	ZC4H2|ENSG00000126970.17	ENSG00000126970	
1359856	Hippocampus	ZDHHC15|ENSG00000102383.14	ENSG00000102383	
1359895	Hippocampus	AMOT|ENSG00000126016.17	ENSG00000126016	
1360040	Hippocampus	DANT2|ENSG00000235244.6	ENSG00000235244	
1360062	Hippocampus	NUS1P1|ENSG00000235636.1	ENSG00000235636	

	Symbol	seqnames	Type	t	Chrom_Type
53	CD99	chrX	Gene	20.718194	Allosome
56	PRKCIP1	chrX	Gene	17.897618	Allosome

```
62                ZBED1     chrX  Gene  14.600959   Allosome
65                DHRSX     chrX  Gene  13.689327   Allosome
66       ENSG00000289007     chrX  Gene  12.947296   Allosome
…                    …      …    …        …          …
1359849           ZC4H2     chrX  Gene   0.011004   Allosome
1359856         ZDHHC15     chrX  Gene   0.010772   Allosome
1359895            AMOT     chrX  Gene   0.009385   Allosome
1360040           DANT2     chrX  Gene   0.002949   Allosome
1360062           NUS1P1    chrX  Gene   0.001518   Allosome

[1346 rows x 8 columns]
```

[12]:
```
xlinked_male.merge(xci[["ensemblID", "Combined XCI status"]], on="ensemblID",␣
↪how="left").fillna("unknown")
```

[12]:
```
           Tissue                              Feature        ensemblID  \
0         Caudate           CD99|ENSG00000002586.20  ENSG00000002586
1         Caudate         PRKCIP1|ENSG00000237682.2  ENSG00000237682
2         Caudate          ZBED1|ENSG00000214717.12  ENSG00000214717
3         Caudate          DHRSX|ENSG00000169084.15  ENSG00000169084
4         Caudate  ENSG00000289007|ENSG00000289007.2  ENSG00000289007
…            …                                  …              …
1341  Hippocampus          ZC4H2|ENSG00000126970.17  ENSG00000126970
1342  Hippocampus        ZDHHC15|ENSG00000102383.14  ENSG00000102383
1343  Hippocampus           AMOT|ENSG00000126016.17  ENSG00000126016
1344  Hippocampus          DANT2|ENSG00000235244.6   ENSG00000235244
1345  Hippocampus          NUS1P1|ENSG00000235636.1  ENSG00000235636

               Symbol seqnames  Type          t Chrom_Type Combined XCI status
0                CD99     chrX  Gene  20.718194   Allosome               escape
1             PRKCIP1     chrX  Gene  17.897618   Allosome              unknown
2               ZBED1     chrX  Gene  14.600959   Allosome               escape
3               DHRSX     chrX  Gene  13.689327   Allosome               escape
4     ENSG00000289007     chrX  Gene  12.947296   Allosome              unknown
…                   …      …    …        …          …                      …
1341            ZC4H2     chrX  Gene   0.011004   Allosome             inactive
1342          ZDHHC15     chrX  Gene   0.010772   Allosome             inactive
1343             AMOT     chrX  Gene   0.009385   Allosome             inactive
1344            DANT2     chrX  Gene   0.002949   Allosome              unknown
1345            NUS1P1    chrX  Gene   0.001518   Allosome              unknown

[1346 rows x 9 columns]
```

[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", "seqnames"], ascending=True)\
  .to_csv("BrainSeq_male_biased_genes_XCI_status.tsv", sep='\t', index=False)
```

1.1 Session information

[15]:
```
session_info.show()
```

[15]:
```
<IPython.core.display.HTML object>
```