GWAS_post_cleaning_analysis

April 18, 2022

1 Analysis of data after cleaning/normalization

1.1 Setup

[1]:

```
import pandas as pd
import numpy as np
import math
import seaborn as sns
```

[18]:

```
METADATA_FILE = 'gwas_trait_metadata.csv'
CLEANED_FILE_SUFFIX = '_cleaned.csv'
UNKNOWN_GENE = 'UNKNOWN'
CHILD_TRAIT_DELIMITER = ';'

metadata_df = pd.read_csv(METADATA_FILE)
all_traits = metadata_df['Trait'].tolist()
print(all_traits)
```

```
['attention deficit hyperactivity disorder', 'alzheimer disease', 'anxiety
disorder', 'autism spectrum disorder', 'bipolar disorder', 'drug dependence',
'eating disorder', 'personality disorder', 'schizophrenia', 'tourette syndrome',
'unipolar depression']
```

[3]:

```
def trait_to_cleaned_filename(trait):
    return trait.replace(" ", "_") + CLEANED_FILE_SUFFIX


trait_to_df = {
    trait: pd.read_csv(trait_to_cleaned_filename(trait)) for trait in all_traits
}
```

[4]:

```
trait_to_df['schizophrenia'].head()
```

[4]:

```
   Unnamed: 0    variant_and_allele       p_value          trait     gene
0        2214    rs3130820-<b>?</b>  2.000000e-44  schizophrenia   OR2U1P
1        2214    rs3130820-<b>?</b>  2.000000e-44  schizophrenia   OR2G1P
2          58  rs115329265-<b>A</b>  5.000000e-36  schizophrenia  NOP56P1
```

```
3          58  rs115329265-<b>A</b>  5.000000e-36  schizophrenia   RPSAP2
4         410    rs9257566-<b>?</b>  7.000000e-30  schizophrenia    OR2J2
```

1.2 Comparing summary stats of data for all traits

[5]:

```
trait_summaries = []
for trait in all_traits:
  trait_df = trait_to_df[trait]
  if trait == 'attention deficit hyperactivity disorder':
    # Shorten for plots
    trait = 'ADHD'
  trait_df['parent_trait'] = trait

  num_unknown_genes = len(trait_df.loc[trait_df['gene'] == UNKNOWN_GENE])
  trait_summary = {
      'parent_trait': trait,
      'num_variants': len(trait_df),
      'num_unique_genes': len(trait_df['gene'].unique()),
      'num_unknown_genes': num_unknown_genes,
      'min_pval': trait_df['p_value'].min(),
      'max_pval': trait_df['p_value'].max(),
  }
  trait_summaries.append(trait_summary)

summary_df = pd.DataFrame(trait_summaries)
summary_df
```

[5]:

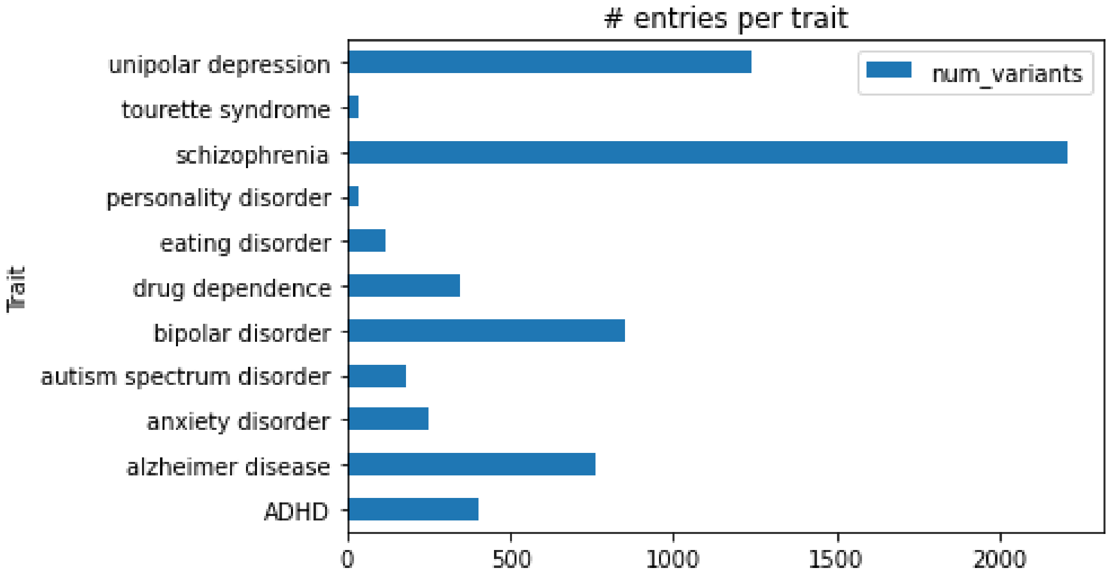
```
                parent_trait  num_variants  num_unique_genes  \
0                       ADHD           400               277
1          alzheimer disease           764               495
2           anxiety disorder           250               222
3   autism spectrum disorder           180               145
4           bipolar disorder           854               509
5            drug dependence           345               259
6            eating disorder           119               105
7       personality disorder            33                28
8              schizophrenia          2205              1122
9          tourette syndrome            39                38
10        unipolar depression          1238               810

    num_unknown_genes       min_pval  max_pval
0                  20   8.000000e-14  0.000009
1                  40  2.000000e-303  0.000009
2                  16   7.000000e-22  0.000009
3                  17   4.000000e-13  0.000009
4                 106   1.000000e-21  0.000009
```

```
5                    28   1.000000e-70  0.000009
6                    10   7.000000e-15  0.000009
7                     6   2.000000e-07  0.000009
8                   251   2.000000e-44  0.000009
9                     2   3.000000e-08  0.000009
10                   96   4.000000e-52  0.000009
```

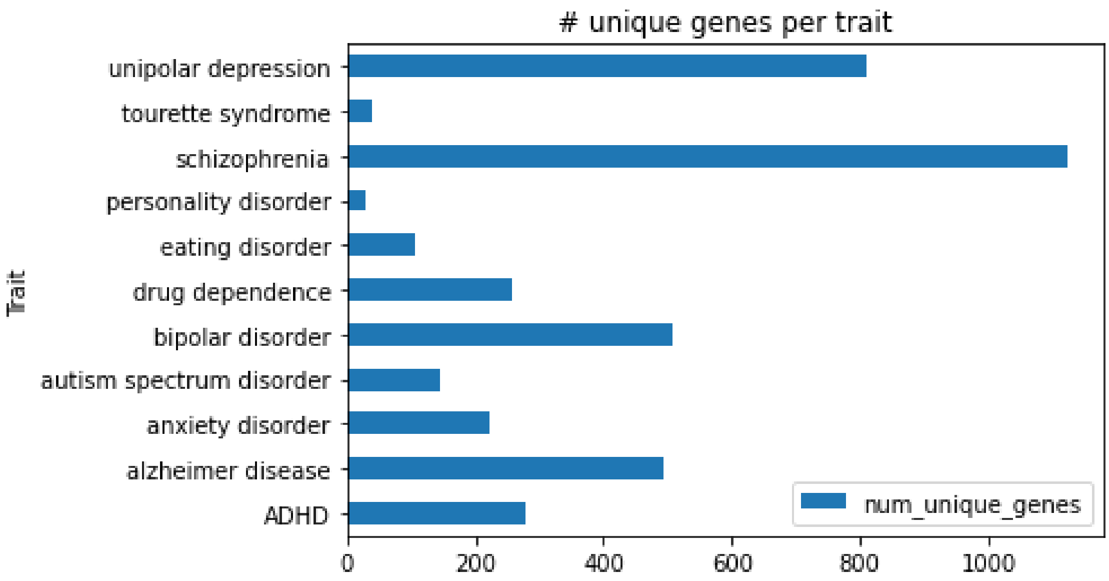
[6]:
```
summary_df.plot(kind='barh', title='# entries per trait',
                x='parent_trait', y='num_variants',
                xlabel='Trait')
```

[6]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f57fb5e8b90>
```



[7]:
```
summary_df.plot(kind='barh', title='# unique genes per trait',
                x='parent_trait', y='num_unique_genes',
                xlabel='Trait')
```

[7]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f57fb50d510>
```



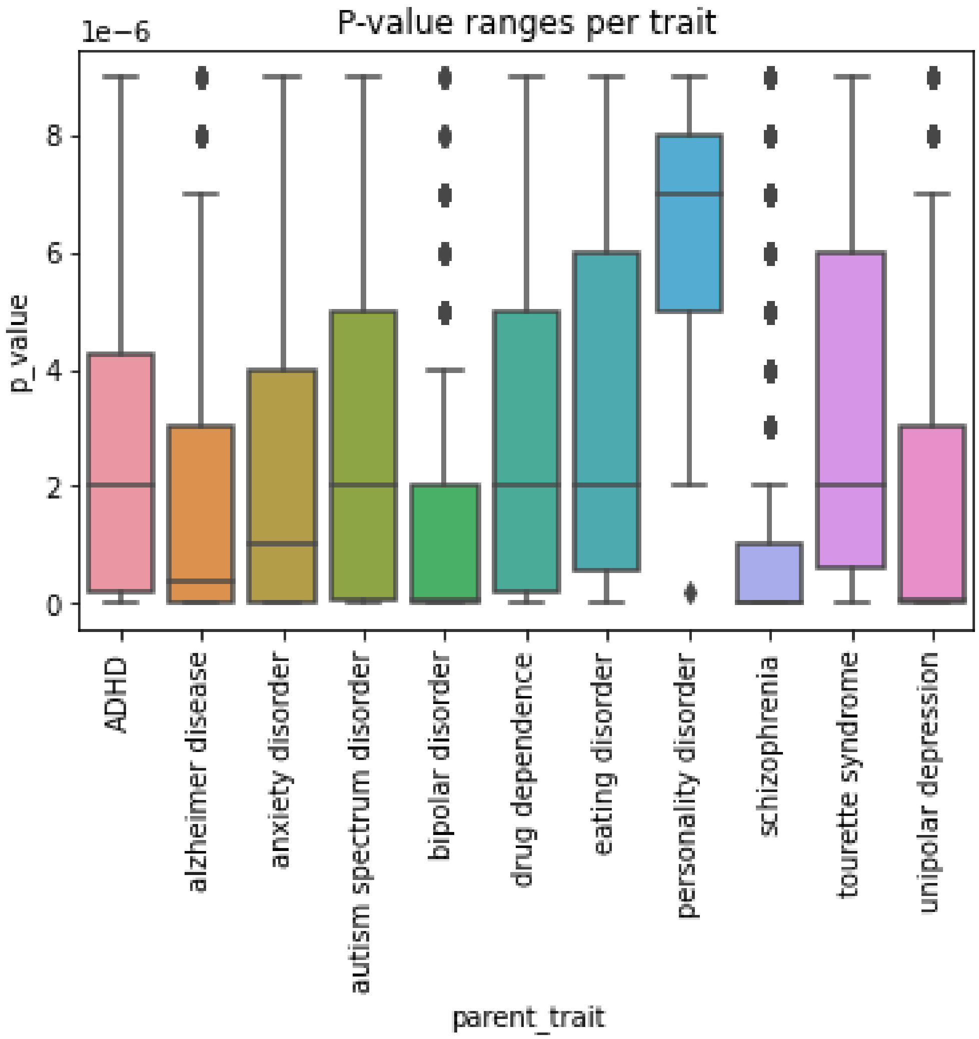
```
[8]: all_df = pd.concat(df for df in trait_to_df.values())
     all_df.head()
```

```
[8]:    Unnamed: 0   variant_and_allele       p_value  \
     0         265  rs12410155-<b>?</b>  8.000000e-14
     1         265  rs12410155-<b>?</b>  8.000000e-14
     2          91  rs12658032-<b>A</b>  1.000000e-13
     3          91  rs12658032-<b>A</b>  1.000000e-13
     4         280  rs11420276-<b>?</b>  2.000000e-13

                                           trait         gene parent_trait
     0  attention deficit hyperactivity disorder  ST3GAL3-AS1         ADHD
     1  attention deficit hyperactivity disorder      ST3GAL3         ADHD
     2  attention deficit hyperactivity disorder    LINC02163         ADHD
     3  attention deficit hyperactivity disorder    RNU6-334P         ADHD
     4  attention deficit hyperactivity disorder      ST3GAL3         ADHD
```

```
[10]: ax = sns.boxplot(x='parent_trait', y='p_value', data=all_df)
      ax.set_xticklabels(ax.get_xticklabels(), rotation=90)
      ax.set(title='P-value ranges per trait')
```

```
[10]: [Text(0.5, 1.0, 'P-value ranges per trait')]
```



The IQRs of all traits look roughly the same except for personality disorder (and schizophrenia has a much lower IQR and many outliers). It looks like filtering data points with p-values > 6e-6 may remove most of the outliers (and may need to exclude personality disorder; it also appears to have very little coverage based on the previous plots).

1.3 Comparing gene & variant overlap

Do pairwise comparison to see which traits share implicated genes.

```
[16]: trait_to_genes = {}
for trait in all_traits:
  genes = set(trait_to_df[trait]['gene'].unique())
  genes.remove(UNKNOWN_GENE)
  trait_to_genes[trait] = genes
```

```python
for trait_a in all_traits:
  for trait_b in all_traits:
    if trait_a == trait_b:
      continue

    overlapping_genes = trait_to_genes[trait_a].
↪intersection(trait_to_genes[trait_b])
    if len(overlapping_genes) > 0:
      print(f'{trait_a} and {trait_b} have {len(overlapping_genes)} overlapping␣
↪genes.')
```

```
attention deficit hyperactivity disorder and alzheimer disease have 11
overlapping genes.
attention deficit hyperactivity disorder and anxiety disorder have 9 overlapping
genes.
attention deficit hyperactivity disorder and autism spectrum disorder have 12
overlapping genes.
attention deficit hyperactivity disorder and bipolar disorder have 15
overlapping genes.
attention deficit hyperactivity disorder and drug dependence have 11 overlapping
genes.
attention deficit hyperactivity disorder and eating disorder have 8 overlapping
genes.
attention deficit hyperactivity disorder and personality disorder have 2
overlapping genes.
attention deficit hyperactivity disorder and schizophrenia have 36 overlapping
genes.
attention deficit hyperactivity disorder and tourette syndrome have 1
overlapping genes.
attention deficit hyperactivity disorder and unipolar depression have 34
overlapping genes.
alzheimer disease and attention deficit hyperactivity disorder have 11
overlapping genes.
alzheimer disease and anxiety disorder have 13 overlapping genes.
alzheimer disease and autism spectrum disorder have 9 overlapping genes.
alzheimer disease and bipolar disorder have 18 overlapping genes.
alzheimer disease and drug dependence have 10 overlapping genes.
alzheimer disease and eating disorder have 4 overlapping genes.
alzheimer disease and schizophrenia have 48 overlapping genes.
alzheimer disease and tourette syndrome have 4 overlapping genes.
alzheimer disease and unipolar depression have 30 overlapping genes.
anxiety disorder and attention deficit hyperactivity disorder have 9 overlapping
genes.
anxiety disorder and alzheimer disease have 13 overlapping genes.
anxiety disorder and autism spectrum disorder have 6 overlapping genes.
anxiety disorder and bipolar disorder have 15 overlapping genes.
anxiety disorder and drug dependence have 8 overlapping genes.
anxiety disorder and eating disorder have 7 overlapping genes.
anxiety disorder and schizophrenia have 46 overlapping genes.
anxiety disorder and tourette syndrome have 2 overlapping genes.
anxiety disorder and unipolar depression have 52 overlapping genes.
autism spectrum disorder and attention deficit hyperactivity disorder have 12
overlapping genes.
autism spectrum disorder and alzheimer disease have 9 overlapping genes.
autism spectrum disorder and anxiety disorder have 6 overlapping genes.
autism spectrum disorder and bipolar disorder have 11 overlapping genes.
autism spectrum disorder and drug dependence have 4 overlapping genes.
autism spectrum disorder and eating disorder have 4 overlapping genes.
autism spectrum disorder and schizophrenia have 25 overlapping genes.
autism spectrum disorder and tourette syndrome have 2 overlapping genes.
autism spectrum disorder and unipolar depression have 21 overlapping genes.
bipolar disorder and attention deficit hyperactivity disorder have 15
overlapping genes.
bipolar disorder and alzheimer disease have 18 overlapping genes.
bipolar disorder and anxiety disorder have 15 overlapping genes.
bipolar disorder and autism spectrum disorder have 11 overlapping genes.
bipolar disorder and drug dependence have 8 overlapping genes.
bipolar disorder and eating disorder have 9 overlapping genes.
bipolar disorder and personality disorder have 1 overlapping genes.
bipolar disorder and schizophrenia have 159 overlapping genes.
bipolar disorder and tourette syndrome have 5 overlapping genes.
bipolar disorder and unipolar depression have 56 overlapping genes.
drug dependence and attention deficit hyperactivity disorder have 11 overlapping
genes.
drug dependence and alzheimer disease have 10 overlapping genes.
drug dependence and anxiety disorder have 8 overlapping genes.
drug dependence and autism spectrum disorder have 4 overlapping genes.
drug dependence and bipolar disorder have 8 overlapping genes.
drug dependence and eating disorder have 4 overlapping genes.
drug dependence and personality disorder have 1 overlapping genes.
drug dependence and schizophrenia have 34 overlapping genes.
drug dependence and tourette syndrome have 1 overlapping genes.
drug dependence and unipolar depression have 22 overlapping genes.
eating disorder and attention deficit hyperactivity disorder have 8 overlapping
genes.
eating disorder and alzheimer disease have 4 overlapping genes.
eating disorder and anxiety disorder have 7 overlapping genes.
eating disorder and autism spectrum disorder have 4 overlapping genes.
eating disorder and bipolar disorder have 9 overlapping genes.
eating disorder and drug dependence have 4 overlapping genes.
eating disorder and personality disorder have 2 overlapping genes.
eating disorder and schizophrenia have 11 overlapping genes.
eating disorder and tourette syndrome have 1 overlapping genes.
eating disorder and unipolar depression have 14 overlapping genes.
personality disorder and attention deficit hyperactivity disorder have 2
overlapping genes.
personality disorder and bipolar disorder have 1 overlapping genes.
personality disorder and drug dependence have 1 overlapping genes.
personality disorder and eating disorder have 2 overlapping genes.
personality disorder and schizophrenia have 1 overlapping genes.
personality disorder and unipolar depression have 2 overlapping genes.
schizophrenia and attention deficit hyperactivity disorder have 36 overlapping
genes.
schizophrenia and alzheimer disease have 48 overlapping genes.
schizophrenia and anxiety disorder have 46 overlapping genes.
schizophrenia and autism spectrum disorder have 25 overlapping genes.
schizophrenia and bipolar disorder have 159 overlapping genes.
schizophrenia and drug dependence have 34 overlapping genes.
schizophrenia and eating disorder have 11 overlapping genes.
schizophrenia and personality disorder have 1 overlapping genes.
schizophrenia and tourette syndrome have 4 overlapping genes.
schizophrenia and unipolar depression have 151 overlapping genes.
tourette syndrome and attention deficit hyperactivity disorder have 1
overlapping genes.
tourette syndrome and alzheimer disease have 4 overlapping genes.
tourette syndrome and anxiety disorder have 2 overlapping genes.
tourette syndrome and autism spectrum disorder have 2 overlapping genes.
tourette syndrome and bipolar disorder have 5 overlapping genes.
tourette syndrome and drug dependence have 1 overlapping genes.
tourette syndrome and eating disorder have 1 overlapping genes.
tourette syndrome and schizophrenia have 4 overlapping genes.
tourette syndrome and unipolar depression have 5 overlapping genes.
unipolar depression and attention deficit hyperactivity disorder have 34
overlapping genes.
unipolar depression and alzheimer disease have 30 overlapping genes.
unipolar depression and anxiety disorder have 52 overlapping genes.
unipolar depression and autism spectrum disorder have 21 overlapping genes.
unipolar depression and bipolar disorder have 56 overlapping genes.
unipolar depression and drug dependence have 22 overlapping genes.
unipolar depression and eating disorder have 14 overlapping genes.
unipolar depression and personality disorder have 2 overlapping genes.
unipolar depression and schizophrenia have 151 overlapping genes.
unipolar depression and tourette syndrome have 5 overlapping genes.
```

Do a similar check, except for child traits of each given parent trait

[21]:

```
for parent_trait in all_traits:
  trait_row = metadata_df.loc[metadata_df['Trait'] == parent_trait]
  child_trait_entry = trait_row['Child traits'].astype(str)
  if len(child_trait_entry) == 0:
    continue
```

```
child_traits = child_trait_entry.tolist()[0].split(CHILD_TRAIT_DELIMITER)
child_traits = [c_trait.strip().lower() for c_trait in child_traits]
trait_to_genes = {}
for child_trait in child_traits:
  parent_df = trait_to_df[parent_trait]
  child_trait_df = parent_df.loc[parent_df['trait'] == child_trait]
  child_trait_genes = set(child_trait_df['gene'].unique())
  if UNKNOWN_GENE in child_trait_genes:
    child_trait_genes.remove(UNKNOWN_GENE)
  trait_to_genes[child_trait] = child_trait_genes

for child_trait_a in child_traits:
  for child_trait_b in child_traits:
    if child_trait_a == child_trait_b:
      continue

    genes_a = trait_to_genes[child_trait_a]
    genes_b = trait_to_genes[child_trait_b]
    overlapping_genes = genes_a.intersection(genes_b)
    if len(overlapping_genes) > 0:
      print(f'{child_trait_a} and {child_trait_b} have␣
↪{len(overlapping_genes)} overlapping genes.')
```

```
neurotic disorder and obsessive-compulsive disorder have 1 overlapping genes.
neurotic disorder and post-traumatic stress disorder have 1 overlapping genes.
obsessive-compulsive disorder and neurotic disorder have 1 overlapping genes.
obsessive-compulsive disorder and panic disorder have 1 overlapping genes.
obsessive-compulsive disorder and post-traumatic stress disorder have 1
overlapping genes.
panic disorder and obsessive-compulsive disorder have 1 overlapping genes.
panic disorder and post-traumatic stress disorder have 1 overlapping genes.
post-traumatic stress disorder and neurotic disorder have 1 overlapping genes.
post-traumatic stress disorder and obsessive-compulsive disorder have 1
overlapping genes.
post-traumatic stress disorder and panic disorder have 1 overlapping genes.
alcohol and nicotine codependence and alcohol dependence have 1 overlapping
genes.
alcohol dependence and alcohol and nicotine codependence have 1 overlapping
genes.
alcohol dependence and nicotine dependence have 1 overlapping genes.
cocaine dependence and opioid dependence have 1 overlapping genes.
nicotine dependence and alcohol dependence have 1 overlapping genes.
opioid dependence and cocaine dependence have 1 overlapping genes.
anorexia nervosa and bulimia nervosa have 1 overlapping genes.
bulimia nervosa and anorexia nervosa have 1 overlapping genes.
```

Finally check if any variants are implicated in multiple (parent) traits

[23]:

```
trait_to_variants = {}
for trait in all_traits:
  variants = set(trait_to_df[trait]['variant_and_allele'].unique())
  trait_to_variants[trait] = variants


for trait_a in all_traits:
  for trait_b in all_traits:
    if trait_a == trait_b:
      continue

    overlapping_variants = trait_to_variants[trait_a].
 ↪intersection(trait_to_variants[trait_b])
    if len(overlapping_variants) > 0:
      print(f'{trait_a} and {trait_b} have {len(overlapping_variants)}␣
 ↪overlapping variants.')
```

```
attention deficit hyperactivity disorder and autism spectrum disorder have 2
overlapping variants.
attention deficit hyperactivity disorder and schizophrenia have 1 overlapping
variants.
anxiety disorder and unipolar depression have 5 overlapping variants.
autism spectrum disorder and attention deficit hyperactivity disorder have 2
overlapping variants.
autism spectrum disorder and bipolar disorder have 1 overlapping variants.
autism spectrum disorder and schizophrenia have 1 overlapping variants.
autism spectrum disorder and unipolar depression have 2 overlapping variants.
bipolar disorder and autism spectrum disorder have 1 overlapping variants.
bipolar disorder and schizophrenia have 77 overlapping variants.
bipolar disorder and unipolar depression have 1 overlapping variants.
schizophrenia and attention deficit hyperactivity disorder have 1 overlapping
variants.
schizophrenia and autism spectrum disorder have 1 overlapping variants.
schizophrenia and bipolar disorder have 77 overlapping variants.
schizophrenia and unipolar depression have 6 overlapping variants.
unipolar depression and anxiety disorder have 5 overlapping variants.
unipolar depression and autism spectrum disorder have 2 overlapping variants.
unipolar depression and bipolar disorder have 1 overlapping variants.
unipolar depression and schizophrenia have 6 overlapping variants.
```

Sanity-check a few of those.

[25]:

```
adhd_variants = trait_to_variants['attention deficit hyperactivity disorder']
autism_variants = trait_to_variants['autism spectrum disorder']
overlapping_variants = adhd_variants.intersection(autism_variants)
print(overlapping_variants)
```

```
{'rs6584649-<b>?</b>', 'rs4916723-<b>C</b>'}
```

[26]:

```
depression_variants = trait_to_variants['unipolar depression']
anxiety_variants = trait_to_variants['anxiety disorder']
overlapping_variants = depression_variants.intersection(anxiety_variants)
print(overlapping_variants)
```

```
{'rs3135296-<b>?</b>', 'rs4543289-<b>?</b>', 'rs1002656-<b>C</b>',
'rs30266-<b>A</b>', 'rs3807866-<b>A</b>'}
```