GWAS_post_cleaning_analysis

April 24, 2022

1 Analysis of data after cleaning/normalization

1.1 Setup

[1]:
```python
import pandas as pd
import matplotlib.pyplot as plt
from matplotlib.colors import LinearSegmentedColormap
import numpy as np
import math
import seaborn as sns
```

[2]:
```python
METADATA_FILE = 'gwas_trait_metadata.csv'
UNKNOWN_GENE = 'UNKNOWN'
CHILD_TRAIT_DELIMITER = ';'
TISSUE_DELIM = "&"

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]:
```python
def trait_to_cleaned_filename(trait):
    return trait.replace(" ", "_") + ".csv"


trait_to_df = {
    trait: pd.read_csv(trait_to_cleaned_filename(trait)) for trait in all_traits
}
# Shorten to ADHD for brevity in plots:
trait_to_df['ADHD'] = trait_to_df['attention deficit hyperactivity disorder']
trait_to_df.pop('attention deficit hyperactivity disorder')
all_traits.remove('attention deficit hyperactivity disorder')
all_traits.append('ADHD')
```

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

```
[4]:      variant_and_allele       p_value          trait     gene    location  \
     0    rs3130820-<b>?</b>  2.000000e-44  schizophrenia   OR2U1P  6:29238906
     1    rs3130820-<b>?</b>  2.000000e-44  schizophrenia   OR2G1P  6:29238906
     2  rs115329265-<b>A</b>  5.000000e-36  schizophrenia  NOP56P1  6:28744470
     3  rs115329265-<b>A</b>  5.000000e-36  schizophrenia   RPSAP2  6:28744470
     4    rs9257566-<b>?</b>  7.000000e-30  schizophrenia    OR2J2  6:29176755

         af                                            tissues
     0 -1.0                                                NaN
     1 -1.0                                                NaN
     2 -1.0  Brain_Caudate_basal_ganglia&Artery_Tibial&Brai…
     3 -1.0  Brain_Caudate_basal_ganglia&Artery_Tibial&Brai…
     4 -1.0  Artery_Coronary&Brain_Caudate_basal_ganglia&Ar…
```

1.2 Comparing summary stats of data for all traits

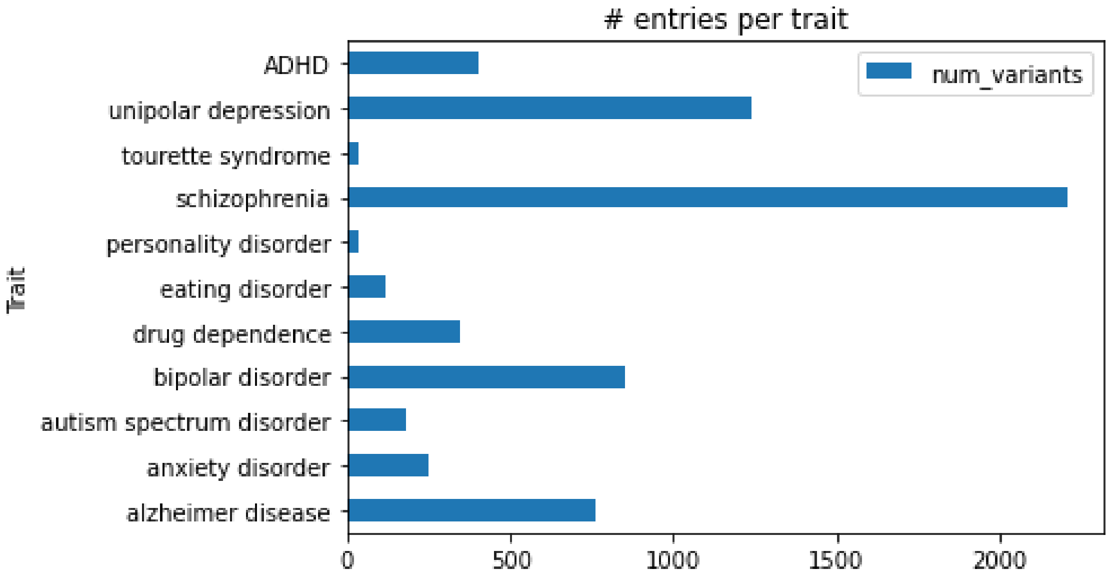
```
[5]: trait_summaries = []
     for trait in all_traits:
       trait_df = trait_to_df[trait]
       if trait == 'attention deficit hyperactivity disorder':
         # Shorten for plots
         trait = 'ADHD'
       trait_df['parent_trait'] = trait

       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         alzheimer disease           764               495
     1          anxiety disorder           250               222
     2  autism spectrum disorder           179               145
     3           bipolar disorder           854               509
     4            drug dependence           345               259
```

```
5            eating disorder           119               105
6        personality disorder           33                28
7               schizophrenia         2205              1122
8           tourette syndrome           39                38
9          unipolar depression        1237               810
10                       ADHD          400               277

    num_unknown_genes       min_pval  max_pval
0                  40  2.000000e-303  0.000009
1                  16   7.000000e-22  0.000009
2                  16   4.000000e-13  0.000009
3                 106   1.000000e-21  0.000009
4                  28   1.000000e-70  0.000009
5                  10   7.000000e-15  0.000009
6                   6   2.000000e-07  0.000009
7                 251   2.000000e-44  0.000009
8                   2   3.000000e-08  0.000009
9                  95   4.000000e-52  0.000009
10                 20   8.000000e-14  0.000009
```

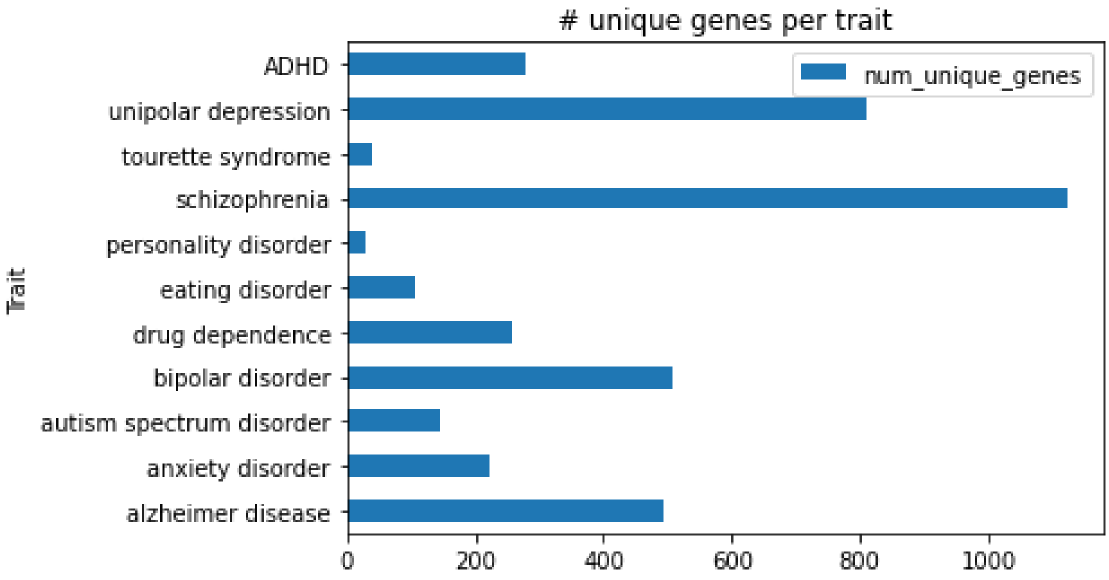
[6]:
```
summary_df.plot(kind='barh', title='# entries per trait',
                x='parent_trait', y='num_variants',
                xlabel='Trait')
```

[6]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f275046cc10>
```



[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 0x7f2750368a90>
```

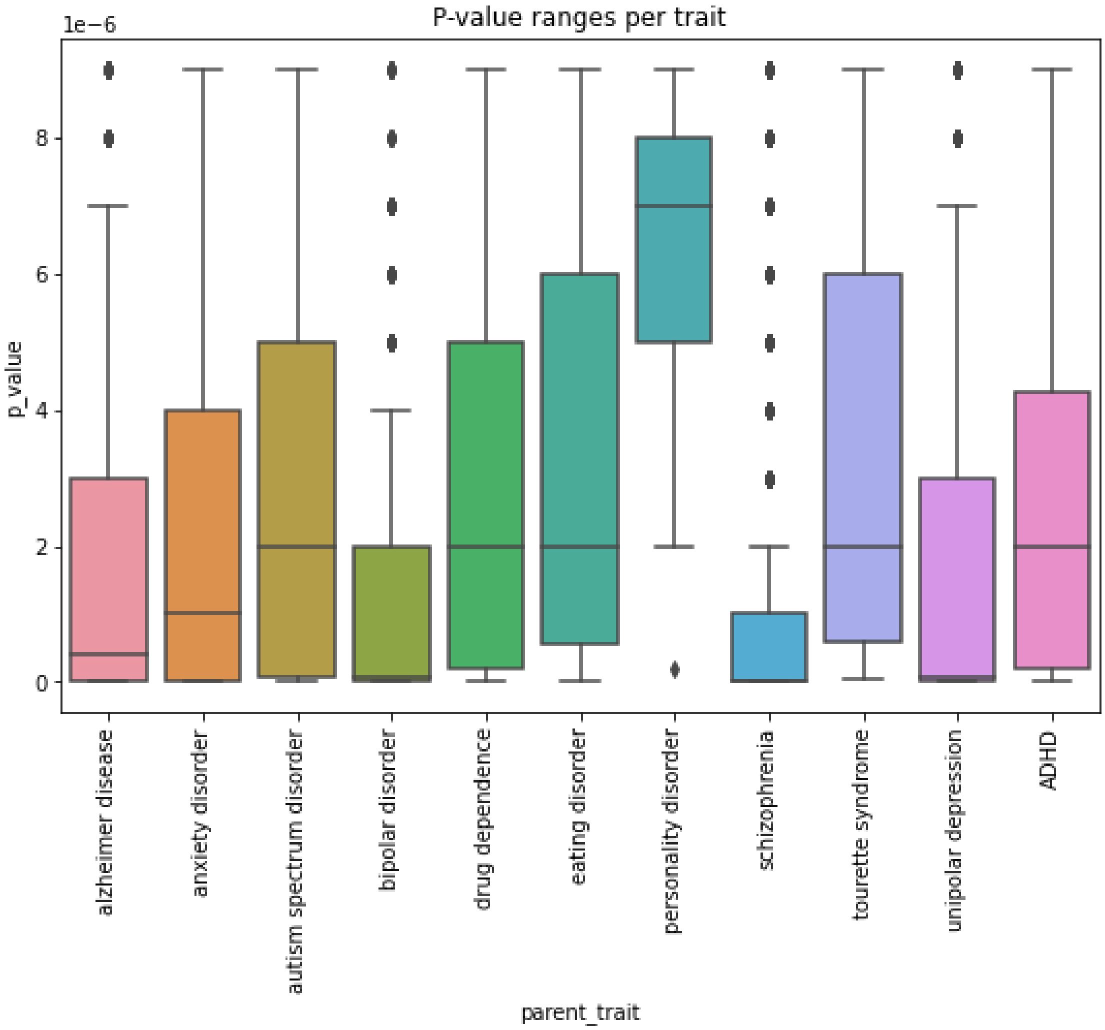


[8]:
```
all_df = pd.concat(df for df in trait_to_df.values())
all_df.head()
```

[8]:
```
     variant_and_allele        p_value              trait     gene  \
0    rs429358-<b>?</b>  2.000000e-303  alzheimer disease     APOE
1   rs2075650-<b>G</b>  1.000000e-295  alzheimer disease   TOMM40
2    rs429358-<b>C</b>  5.000000e-286  alzheimer disease     APOE
3  rs41289512-<b>G</b>  3.000000e-194  alzheimer disease  NECTIN2
4   rs2075650-<b>?</b>  2.000000e-157  alzheimer disease   TOMM40

      location        af  tissues       parent_trait
0  19:44908684 -1.000000      NaN  alzheimer disease
1  19:44892362  0.128939      NaN  alzheimer disease
2  19:44908684  0.074418      NaN  alzheimer disease
3  19:44848259  0.028322      NaN  alzheimer disease
4  19:44892362 -1.000000      NaN  alzheimer disease
```

[9]:
```
fig, ax = plt.subplots(figsize=(9, 6))
sns.boxplot(x='parent_trait', y='p_value', data=all_df, ax=ax)
ax.set_xticklabels(ax.get_xticklabels(), rotation=90)
ax.set(title='P-value ranges per trait')
```

[9]: [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.

[10]:
```
trait_to_genes = {}
for trait in all_traits:
  genes = set(trait_to_df[trait]['gene'])
  genes.remove(UNKNOWN_GENE)
  trait_to_genes[trait] = genes

# Stores dicts where keys are traits and values are # overlapping genes, to be
# used in creating a DF later.
overlapping_gene_count_dicts = []
for trait_a in all_traits:
  trait_a_dict = {}
  for trait_b in all_traits:
    if trait_a == trait_b:
      trait_a_dict[trait_a] = len(trait_to_genes[trait_a])
      continue

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

overlap_df = pd.DataFrame(overlapping_gene_count_dicts, index=all_traits)
```

```
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.
alzheimer disease and ADHD have 11 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.
anxiety disorder and ADHD have 9 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.
autism spectrum disorder and ADHD have 12 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.
bipolar disorder and ADHD have 15 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.
drug dependence and ADHD have 11 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.
eating disorder and ADHD have 8 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.
personality disorder and ADHD have 2 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.
schizophrenia and ADHD have 36 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.
tourette syndrome and ADHD have 1 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.
unipolar depression and ADHD have 34 overlapping genes.
ADHD and alzheimer disease have 11 overlapping genes.
ADHD and anxiety disorder have 9 overlapping genes.
ADHD and autism spectrum disorder have 12 overlapping genes.
ADHD and bipolar disorder have 15 overlapping genes.
ADHD and drug dependence have 11 overlapping genes.
ADHD and eating disorder have 8 overlapping genes.
ADHD and personality disorder have 2 overlapping genes.
ADHD and schizophrenia have 36 overlapping genes.
ADHD and tourette syndrome have 1 overlapping genes.
ADHD and unipolar depression have 34 overlapping genes.
```

[11]:

```
overlap_df.head()
```

[11]:

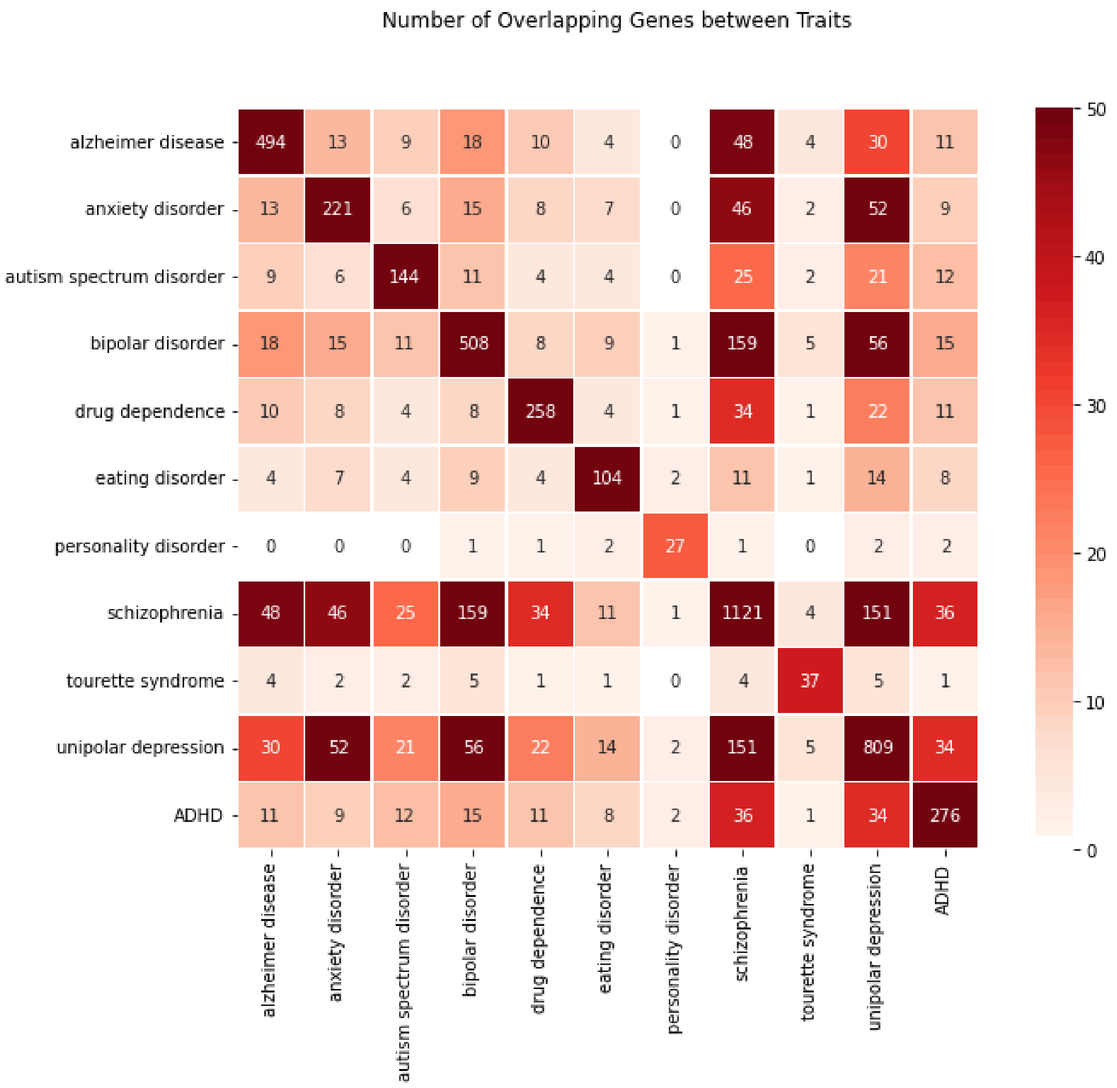
```
                          alzheimer disease  anxiety disorder  \
alzheimer disease                       494                13
anxiety disorder                         13               221
autism spectrum disorder                  9                 6
bipolar disorder                         18                15
drug dependence                          10                 8

                          autism spectrum disorder  bipolar disorder  \
alzheimer disease                                9                18
anxiety disorder                                 6                15
autism spectrum disorder                       144                11
bipolar disorder                                11               508
drug dependence                                  4                 8

                          drug dependence  eating disorder  \
alzheimer disease                      10                4
anxiety disorder                        8                7
autism spectrum disorder                4                4
bipolar disorder                        8                9
drug dependence                       258                4

                          personality disorder  schizophrenia  \
alzheimer disease                            0             48
anxiety disorder                             0             46
autism spectrum disorder                     0             25
bipolar disorder                             1            159
drug dependence                              1             34

                          tourette syndrome  unipolar depression  ADHD
alzheimer disease                         4                   30    11
anxiety disorder                          2                   52     9
autism spectrum disorder                  2                   21    12
bipolar disorder                          5                   56    15
drug dependence                           1                   22    11
```

[12]:

```
# Custom color map where 0 is white to emphasize no gene overlap, and
# darkest color is at 50, so that traits with more data don't skew the chart
# too much.
cmap_reds = plt.get_cmap('Reds')
num_colors = 50
colors = ['white'] + [cmap_reds(i / num_colors) for i in range(1, num_colors)]
cmap = LinearSegmentedColormap.from_list('', colors, num_colors)

fig, ax = plt.subplots(figsize=(12, 8))
sns.heatmap(overlap_df,
            cmap=cmap,
            vmin=0,
            vmax=num_colors,
            annot=True,
            fmt="d",
            linewidths=0.5,
            square=True,
            ax=ax)
fig.suptitle('Number of Overlapping Genes between Traits')
```

[12]:

```
Text(0.5, 0.98, 'Number of Overlapping Genes between Traits')
```



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

[13]:

```
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

[14]:
```
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.')
```

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

Sanity-check a few of those.

[15]:

```
adhd_variants = trait_to_variants['ADHD']
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>'}
```

[16]:

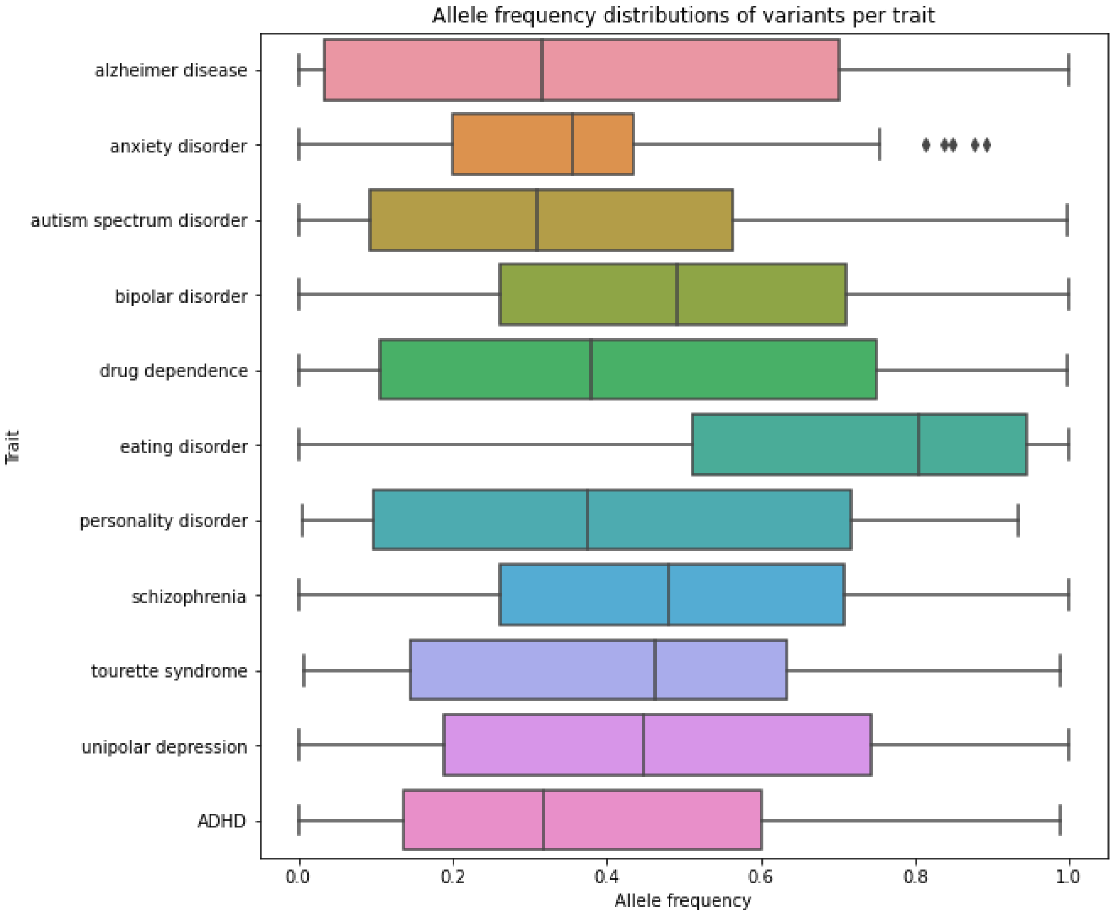
```
depression_variants = trait_to_variants['unipolar depression']
anxiety_variants = trait_to_variants['anxiety disorder']
overlapping_variants = depression_variants.intersection(anxiety_variants)
print(overlapping_variants)
```

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

1.4 Compare allele frequencies

[17]:
```
fig, ax = plt.subplots(figsize=(9, 9))
sns.boxplot(x='af', y='parent_trait', data=all_df[all_df['af'] >= 0.0], ax=ax)
# ax.set_xticklabels(ax.get_xticklabels(), rotation=90)
ax.set(title='Allele frequency distributions of variants per trait')
plt.xlabel('Allele frequency')
plt.ylabel('Trait')
```

[17]:
```
Text(0, 0.5, 'Trait')
```



Observations: * Eating disorder has a suspiciously high average AF. * All traits seems to be bounded by [0, 1] * With the exception of anxiety disorder, which appears to have the tightest IQR.

[18]:
```
missing_af_data_per_trait = all_df[all_df['af'] < 0.0].groupby('parent_trait').
 ↪size()
missing_af_data_per_trait
```

[18]:
```
parent_trait
ADHD                          175
alzheimer disease             313
anxiety disorder              131
autism spectrum disorder       54
bipolar disorder              279
drug dependence               112
eating disorder                14
personality disorder            8
schizophrenia                 833
unipolar depression           213
dtype: int64
```

1.5 Tissue associations

A variant has entries in the 'tissues' column if the variant was found to be significantly associated with gene expression in that tissue via cis-QTL analysis (+/- 1Mb windows around transcription start site, according to GTEx portal.

[19]:
```
all_df['tissues'] = all_df['tissues'].fillna('')
```

[20]:
```
# In previous cleaning, if a variant mapped to multiple genes, I created
# separate rows for it. For tissue association I just blindly match the variant
# location and don't care about the gene though... so drop duplicate locations
# as they may misrepresent total counts in subsequent tissue analysis here.
tissues_df = all_df.copy()
tissues_df.drop_duplicates(subset=['location', 'parent_trait'], inplace=True)
total_variants = len(tissues_df)
tissues_df['has_tissues'] = tissues_df['tissues'].map(lambda tissues:␣
 ↪len(tissues) > 0)
tissues_df = tissues_df[tissues_df['has_tissues'] == True]
num_variants_with_tissues = len(tissues_df)
print(f'{num_variants_with_tissues} variants significantly associated with␣
 ↪tissues of {total_variants} total.')
```

```
1480 variants significantly associated with tissues of 4615 total.
```

[21]:
```
tissue_metadata_file = 'tissue_metadata.txt'
all_tissues = []
with open(tissue_metadata_file) as tissue_file:
  for tissue in tissue_file.readlines():
    all_tissues.append(tissue.replace("\n", ""))

print(all_tissues)
```

```
['Adipose_Subcutaneous', 'Adipose_Visceral_Omentum', 'Adrenal_Gland',
'Artery_Aorta', 'Artery_Coronary', 'Artery_Tibial', 'Brain_Amygdala',
'Brain_Anterior_cingulate_cortex_BA24', 'Brain_Caudate_basal_ganglia',
'Brain_Cerebellar_Hemisphere', 'Brain_Cerebellum', 'Brain_Cortex',
'Brain_Frontal_Cortex_BA9', 'Brain_Hippocampus', 'Brain_Hypothalamus',
'Brain_Nucleus_accumbens_basal_ganglia', 'Brain_Putamen_basal_ganglia',
'Brain_Spinal_cord_cervical_c-1', 'Brain_Substantia_nigra',
'Breast_Mammary_Tissue', 'Cells_Cultured_fibroblasts', 'Cells_EBV-
transformed_lymphocytes', 'Colon_Sigmoid', 'Colon_Transverse',
'Esophagus_Gastroesophageal_Junction', 'Esophagus_Mucosa',
'Esophagus_Muscularis', 'Heart_Atrial_Appendage', 'Heart_Left_Ventricle',
'Kidney_Cortex', 'Liver', 'Lung', 'Minor_Salivary_Gland', 'Muscle_Skeletal',
'Nerve_Tibial', 'Ovary', 'Pancreas', 'Pituitary', 'Prostate',
'Skin_Not_Sun_Exposed_Suprapubic', 'Skin_Sun_Exposed_Lower_leg',
'Small_Intestine_Terminal_Ileum', 'Spleen', 'Stomach', 'Testis', 'Thyroid',
'Uterus', 'Vagina', 'Whole_Blood']
```

1.5.1 Some unsuccessful attempts to visualize tissue association per-trait…

[22]:

```
# Aggregate number of per-tissue associations per trait.
# This is a dict from trait to Dict[tissue, tissue association count]
trait_to_tissue_to_count = {}
for trait in all_traits:
  trait_to_tissue_to_count[trait] = {tissue: 0 for tissue in all_tissues}

for _, row in tissues_df.iterrows():
  trait = row['parent_trait']
  variant_tissues = row['tissues'].split(TISSUE_DELIM)
  for tissue in variant_tissues:
    trait_to_tissue_to_count[trait][tissue] += 1

rows = []
index = []
for trait, tissue_dict in trait_to_tissue_to_count.items():
  index.append(trait)
  rows.append(tissue_dict)

trait_to_tissue_count_df = pd.DataFrame(rows, index=index)
trait_to_tissue_count_df.head()
```

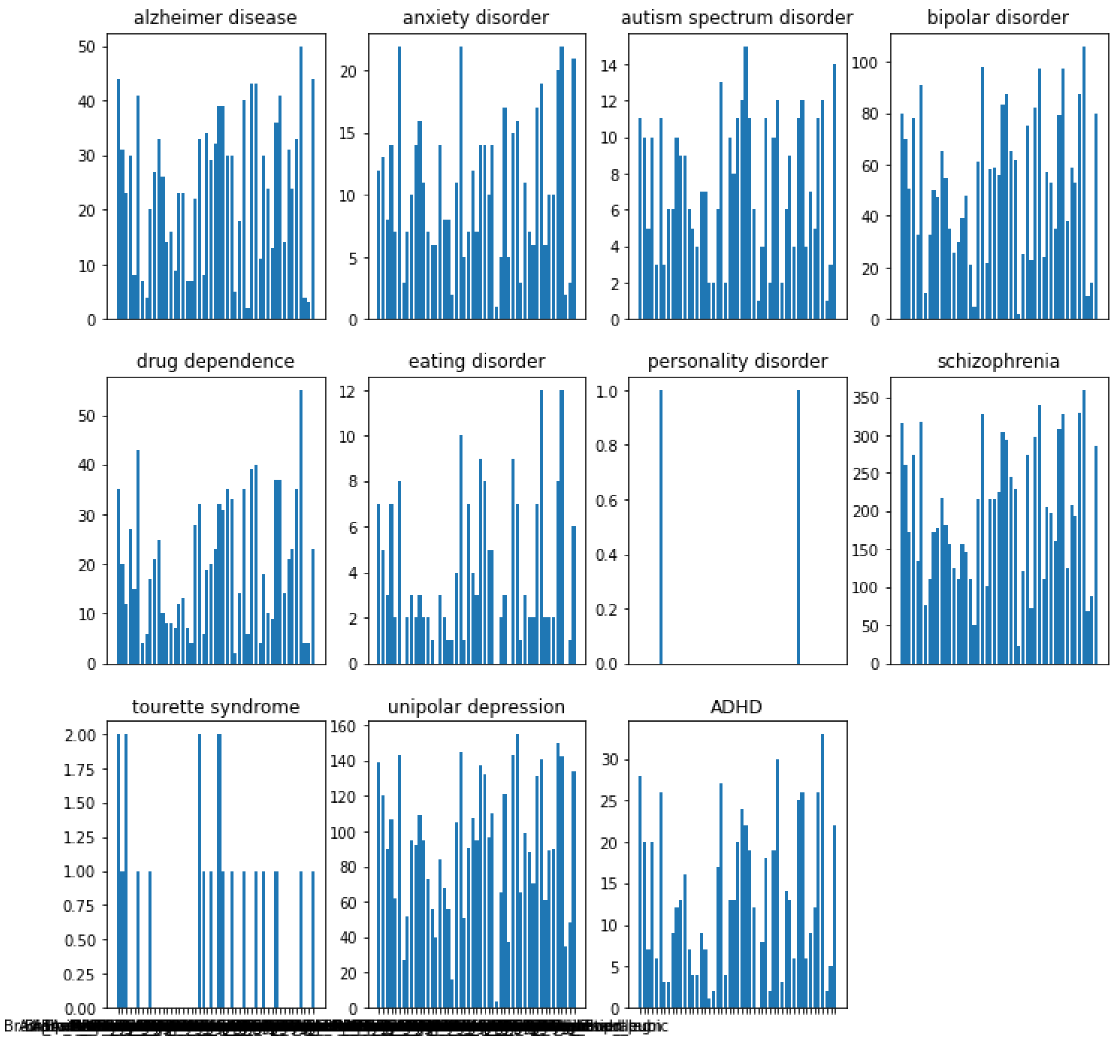
[22]:

```
                          Adipose_Subcutaneous  Adipose_Visceral_Omentum  \
alzheimer disease                           44                        31
anxiety disorder                            12                        13
autism spectrum disorder                    11                        10
bipolar disorder                            80                        70
drug dependence                             35                        20

                          Adrenal_Gland  Artery_Aorta  Artery_Coronary  \
alzheimer disease                          23           30                 8
anxiety disorder                            8           14                 7
autism spectrum disorder                    5           10                 3
bipolar disorder                           51           78                33
drug dependence                            12           27                15

                          Artery_Tibial  Brain_Amygdala  \
alzheimer disease                    41               7
anxiety disorder                     22               3
autism spectrum disorder             11               3
bipolar disorder                     91              10
drug dependence                      43               4

                          Brain_Anterior_cingulate_cortex_BA24  \
alzheimer disease                                            4
anxiety disorder                                             7
autism spectrum disorder                                     6
bipolar disorder                                            33
drug dependence                                              6

                          Brain_Caudate_basal_ganglia  \
alzheimer disease                                  20
anxiety disorder                                   10
autism spectrum disorder                            6
bipolar disorder                                   50
drug dependence                                    17

                          Brain_Cerebellar_Hemisphere  …  \
alzheimer disease                                  27  …
anxiety disorder                                   14  …
autism spectrum disorder                           10  …
bipolar disorder                                   47  …
drug dependence                                    21  …

                          Skin_Not_Sun_Exposed_Suprapubic  \
alzheimer disease                                      36
anxiety disorder                                       17
autism spectrum disorder                               11
bipolar disorder                                       79
drug dependence                                        37

                          Skin_Sun_Exposed_Lower_leg  \
alzheimer disease                                 41
anxiety disorder                                  19
autism spectrum disorder                          12
bipolar disorder                                  97
drug dependence                                   37
```

```
                          Small_Intestine_Terminal_Ileum  Spleen  Stomach  \
alzheimer disease                                     14      31       24
anxiety disorder                                       6      10       10
autism spectrum disorder                               4       7        5
bipolar disorder                                      38      59       53
drug dependence                                       14      21       23

                          Testis  Thyroid  Uterus  Vagina  Whole_Blood
alzheimer disease             33       50       4       3           44
anxiety disorder              20       22       2       3           21
autism spectrum disorder      11       12       1       3           14
bipolar disorder              87      106       9      14           80
drug dependence               35       55       4       4           23

[5 rows x 49 columns]
```

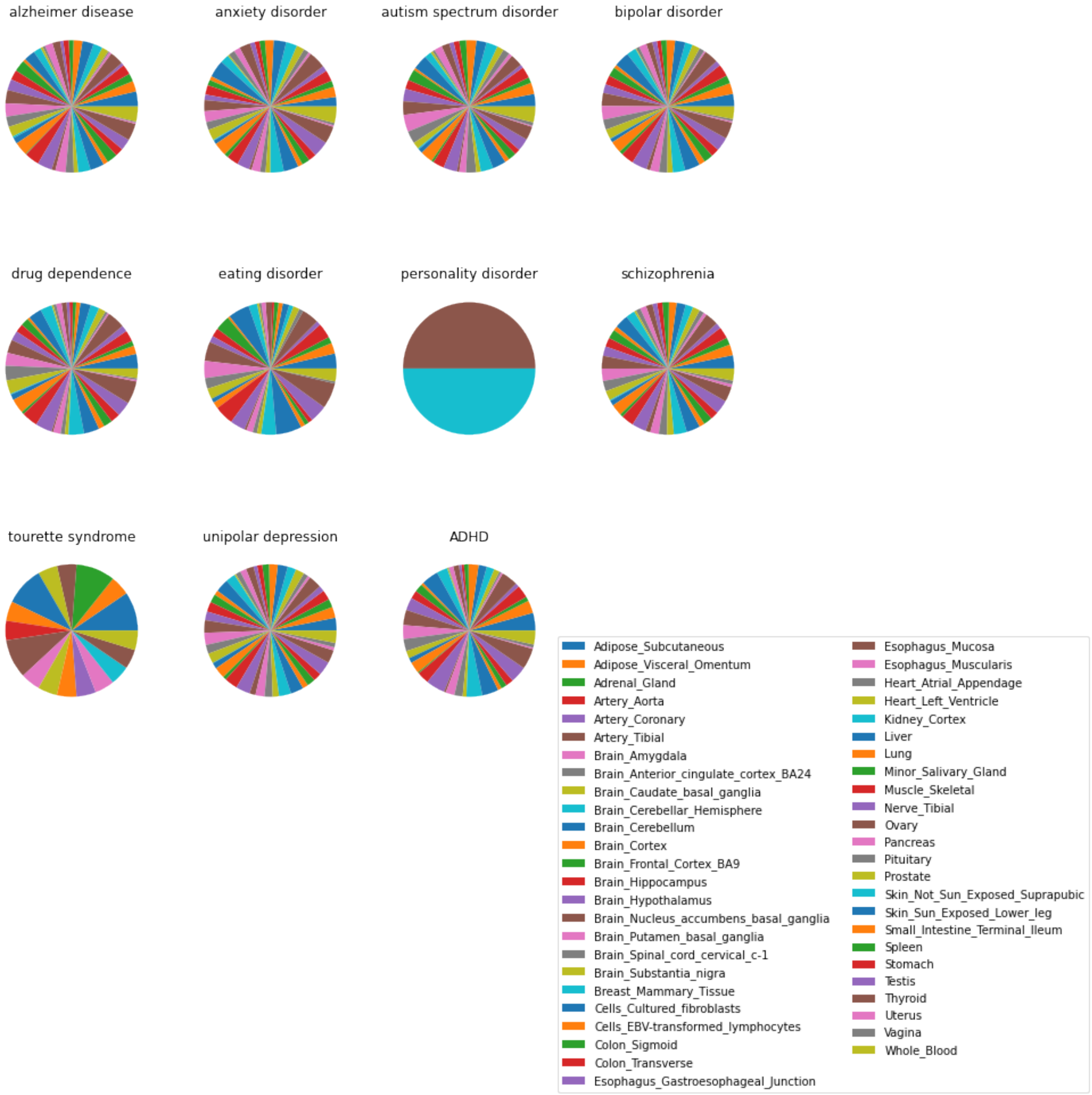
[23]:

```
fig_width = 12
fig_height = 12
rows = 3
cols = 4
fig, axs = plt.subplots(figsize=(fig_width, fig_height),
                        nrows=rows,
                        ncols=cols)
subplot_height = fig_height / rows
subplot_width = fig_width / cols
axs_flat = axs.flatten()
for idx, trait in enumerate(all_traits):
  trait_data = trait_to_tissue_count_df.loc[trait]
  ax = axs_flat[idx]
  barplot = ax.bar(trait_to_tissue_count_df.columns, trait_data.values)
  if math.floor(idx / cols) < rows - 1:
    ax.axes.xaxis.set_visible(False)
  ax.set(title=trait)

fig.delaxes(axs_flat[-1])
ax.set_xticklabels(ax.get_xticklabels(), rotation = 90)
plt.show()
```

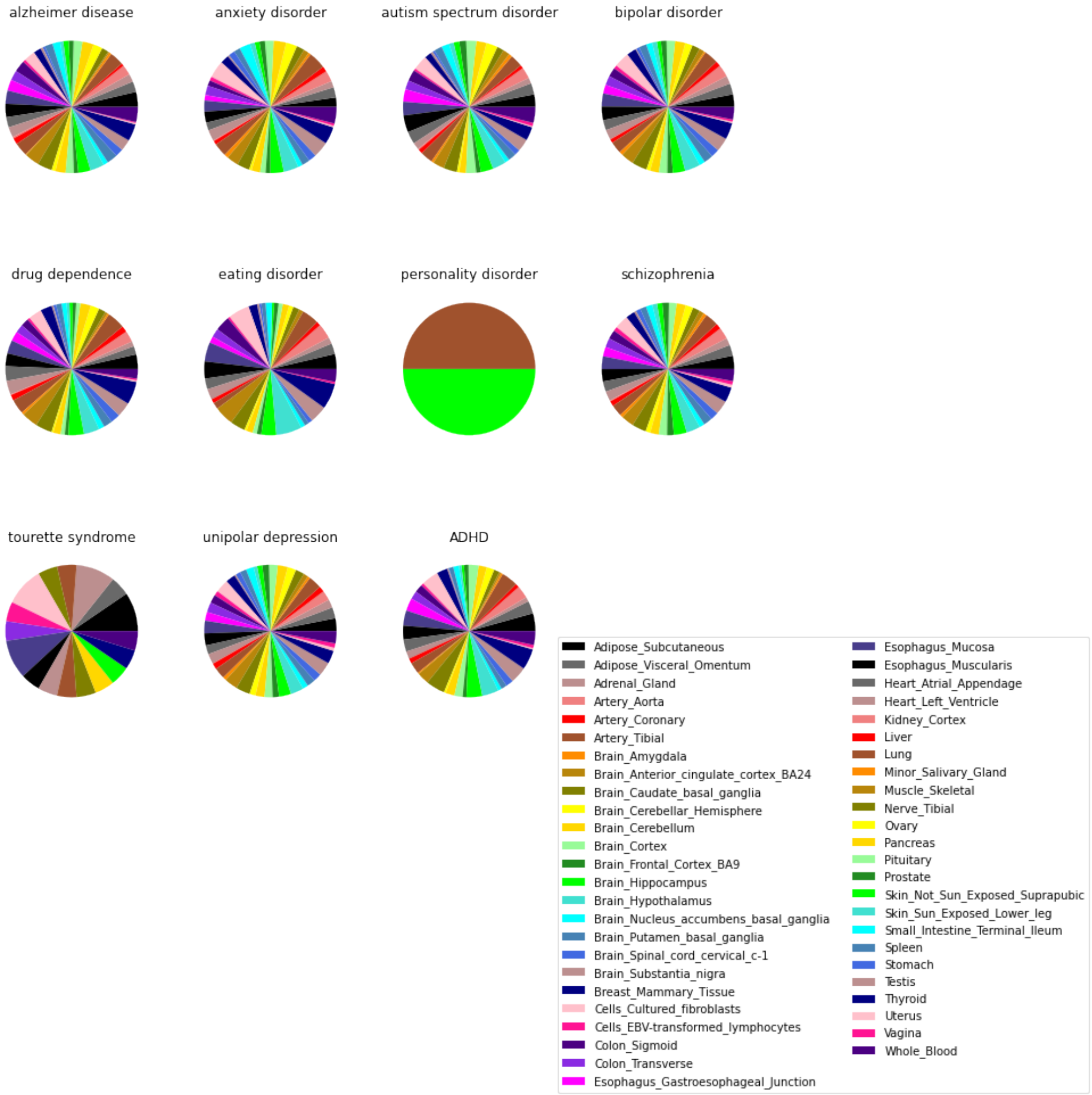


[24]:

```
fig_width = 12
fig_height = 12
rows = 3
cols = 4
fig, axs = plt.subplots(figsize=(fig_width, fig_height),
                        nrows=rows,
                        ncols=cols)

axs_flat = axs.flatten()
for idx, trait in enumerate(all_traits):
  trait_data = trait_to_tissue_count_df.loc[trait]
  ax = axs_flat[idx]
  if math.floor(idx / cols) < rows - 1:
    ax.axes.xaxis.set_visible(False)
  ax.pie(trait_data)
  ax.set(title=trait)

fig.delaxes(axs_flat[-1])
plt.xticks(rotation = 90)
plt.legend(trait_to_tissue_count_df.columns, loc='best', ncol=2,␣
 ↪bbox_to_anchor=(1, 0., 0.5, 0.5))
plt.show()
```



[25]:
```
fig_width = 12
fig_height = 12
rows = 3
cols = 4
fig, axs = plt.subplots(figsize=(fig_width, fig_height),
                        nrows=rows,
                        ncols=cols)
colors = ['black', 'dimgray', 'rosybrown', 'lightcoral', 'red',
            'sienna', 'darkorange', 'darkgoldenrod', 'olive', 'yellow',
            'gold', 'palegreen', 'forestgreen', 'lime', 'turquoise',
            'aqua', 'steelblue', 'royalblue', 'rosybrown', 'navy',
            'pink', 'deeppink', 'indigo', 'blueviolet', 'magenta',
            'darkslateblue']

axs_flat = axs.flatten()
for idx, trait in enumerate(all_traits):
  trait_data = trait_to_tissue_count_df.loc[trait]
  ax = axs_flat[idx]
  if math.floor(idx / cols) < rows - 1:
    ax.axes.xaxis.set_visible(False)
  ax.pie(trait_data, colors=colors)
  ax.set(title=trait)

fig.delaxes(axs_flat[-1])
plt.xticks(rotation = 90)
plt.legend(trait_to_tissue_count_df.columns, loc='best', ncol=2,␣
 ↪bbox_to_anchor=(1, 0., 0.5, 0.5))
plt.show()
```



alzheimer disease

anxiety disorder

autism spectrum disorder

bipolar disorder

drug dependence

eating disorder

personality disorder

schizophrenia

tourette syndrome

unipolar depression

ADHD

[26]:
```
len(all_tissues)
```

[26]:
```
49
```

Too many values for unique colors...

[27]:
```
sums = trait_to_tissue_count_df.sum(axis=0)
sums.sort_values(inplace=True)
sums[0:25].index.tolist()
```

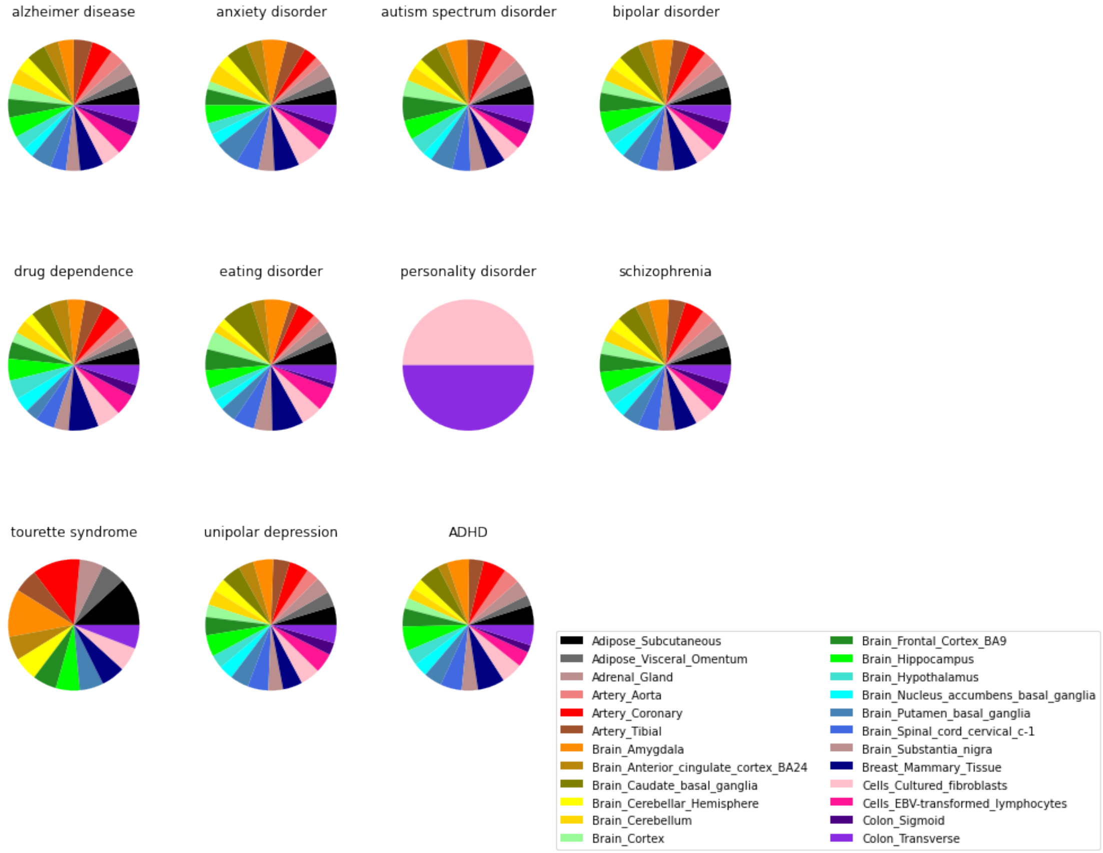
[27]:
```
['Kidney_Cortex',
 'Brain_Substantia_nigra',
 'Uterus',
 'Brain_Amygdala',
 'Minor_Salivary_Gland',
 'Vagina',
 'Cells_EBV-transformed_lymphocytes',
 'Brain_Hypothalamus',
 'Brain_Spinal_cord_cervical_c-1',
 'Brain_Anterior_cingulate_cortex_BA24',
 'Ovary',
 'Brain_Hippocampus',
 'Liver',
 'Artery_Coronary',
 'Small_Intestine_Terminal_Ileum',
 'Prostate',
 'Brain_Frontal_Cortex_BA9',
 'Brain_Putamen_basal_ganglia',
 'Brain_Nucleus_accumbens_basal_ganglia',
 'Adrenal_Gland',
 'Brain_Caudate_basal_ganglia',
 'Pituitary',
 'Brain_Cerebellar_Hemisphere',
 'Brain_Cortex',
 'Stomach']
```

[28]:

```
tissue_plotting_df = trait_to_tissue_count_df.copy()
columns_to_retain = set(tissue_plotting_df.columns.tolist()).
 ↪difference(set(sums[0:25].index.tolist()))
columns_to_retain_list = [col for col in columns_to_retain]
tissue_plotting_df = tissue_plotting_df[columns_to_retain_list]
```

[29]:

```
fig_width = 12
fig_height = 12
rows = 3
cols = 4
fig, axs = plt.subplots(figsize=(fig_width, fig_height),
                        nrows=rows,
                        ncols=cols)
colors = ['black', 'dimgray', 'rosybrown', 'lightcoral', 'red',
          'sienna', 'darkorange', 'darkgoldenrod', 'olive', 'yellow',
          'gold', 'palegreen', 'forestgreen', 'lime', 'turquoise',
          'aqua', 'steelblue', 'royalblue', 'rosybrown', 'navy',
          'pink', 'deeppink', 'indigo', 'blueviolet', 'magenta',
          'darkslateblue']

axs_flat = axs.flatten()
for idx, trait in enumerate(all_traits):
  trait_data = tissue_plotting_df.loc[trait]
  ax = axs_flat[idx]
  if math.floor(idx / cols) < rows - 1:
    ax.axes.xaxis.set_visible(False)
  ax.pie(trait_data, colors=colors)
  ax.set(title=trait)

fig.delaxes(axs_flat[-1])
plt.xticks(rotation = 90)
plt.legend(trait_to_tissue_count_df.columns, loc='best', ncol=2,␣
 ↪bbox_to_anchor=(1, 0., 0.5, 0.5))
plt.show()
```



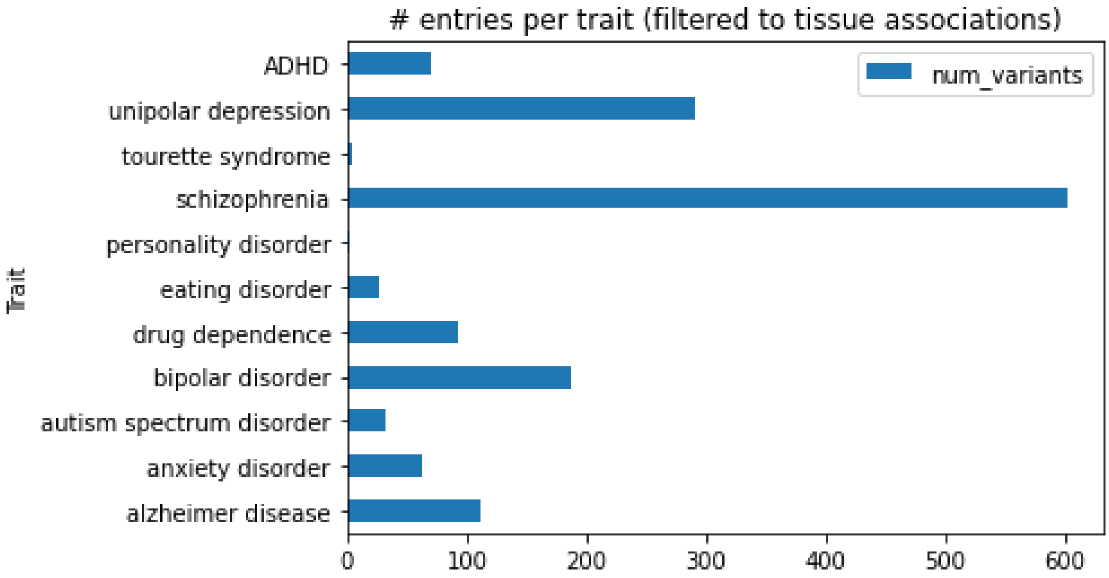
1.5.2 Re-analyzing other variables with only the tissue-associated variants

[30]:

```
trait_summaries = []
for trait in all_traits:
  trait_df = tissues_df[tissues_df['parent_trait'] == trait]
  trait_summary = {
      'parent_trait': trait,
      'num_variants': len(trait_df),
  }
```

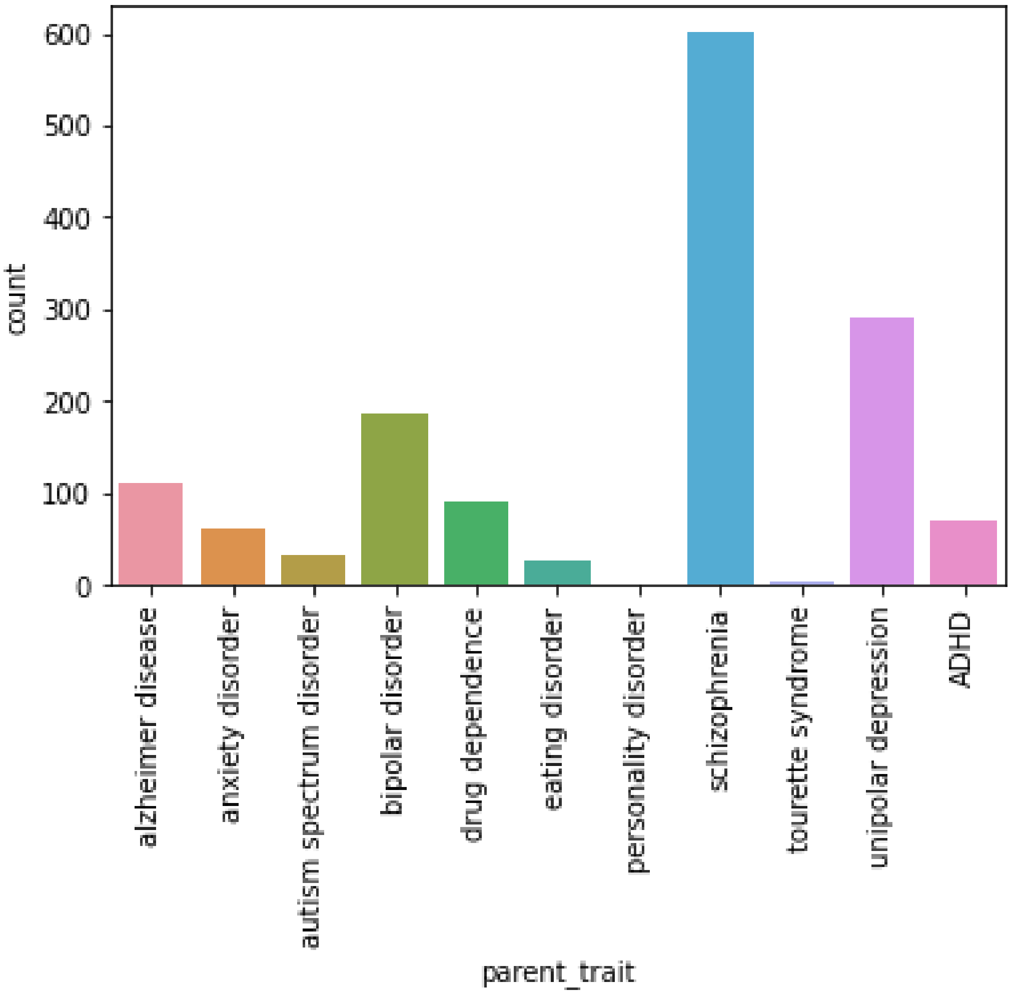
```
    trait_summaries.append(trait_summary)

summary_df = pd.DataFrame(trait_summaries)
summary_df.plot(kind='barh', title='# entries per trait (filtered to tissue␣
↪associations)',
                x='parent_trait', y='num_variants',
                xlabel='Trait')
```

```
[30]: <matplotlib.axes._subplots.AxesSubplot at 0x7f274a366ad0>
```



```
[31]: ax = sns.countplot(x='parent_trait', data=tissues_df)
      ax.set_xticklabels(ax.get_xticklabels(), rotation = 90)
```

```
[31]: [Text(0, 0, 'alzheimer disease'),
       Text(0, 0, 'anxiety disorder'),
       Text(0, 0, 'autism spectrum disorder'),
       Text(0, 0, 'bipolar disorder'),
       Text(0, 0, 'drug dependence'),
       Text(0, 0, 'eating disorder'),
       Text(0, 0, 'personality disorder'),
       Text(0, 0, 'schizophrenia'),
       Text(0, 0, 'tourette syndrome'),
       Text(0, 0, 'unipolar depression'),
       Text(0, 0, 'ADHD')]
```

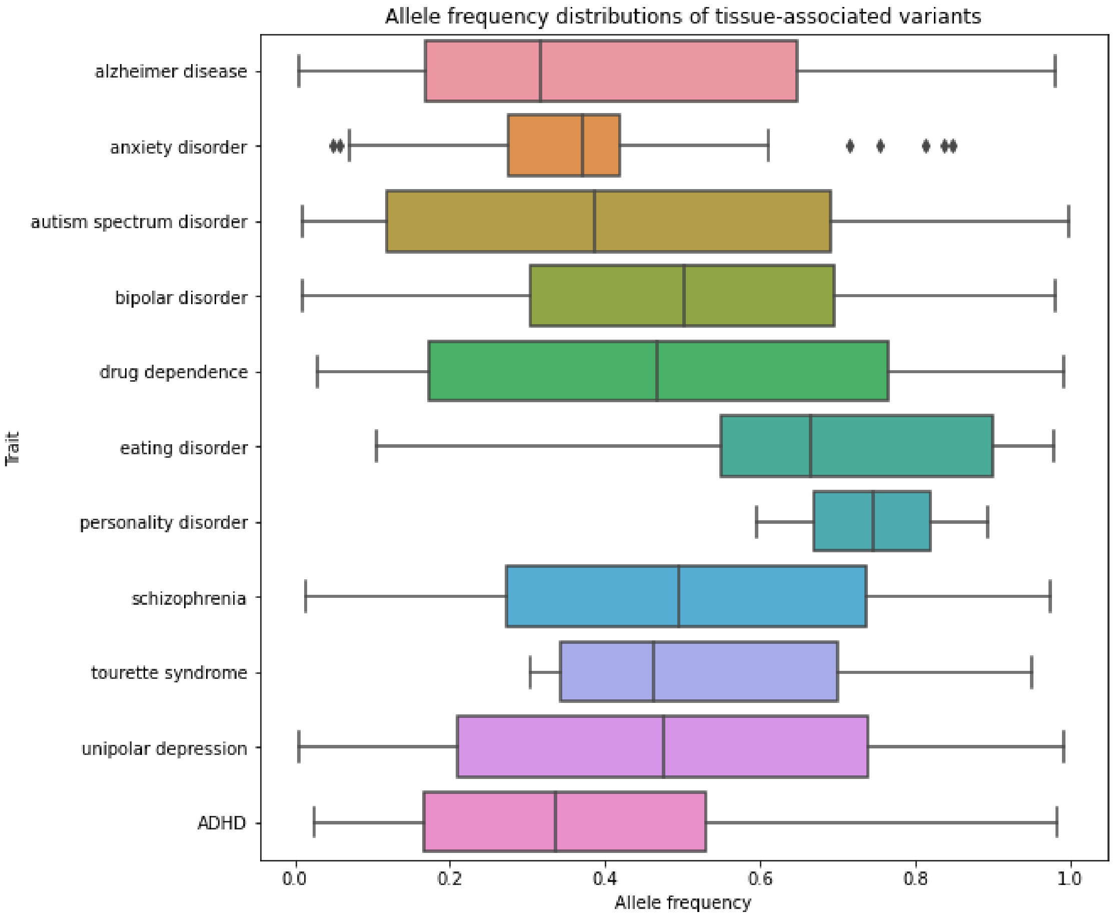


[32]:

```
fig, ax = plt.subplots(figsize=(9, 9))
sns.boxplot(x='af', y='parent_trait', data=tissues_df[tissues_df['af'] >= 0.0],␣
 ↪ax=ax)
# ax.set_xticklabels(ax.get_xticklabels(), rotation=90)
ax.set(title='Allele frequency distributions of tissue-associated variants')
plt.xlabel('Allele frequency')
plt.ylabel('Trait')
```

[32]:

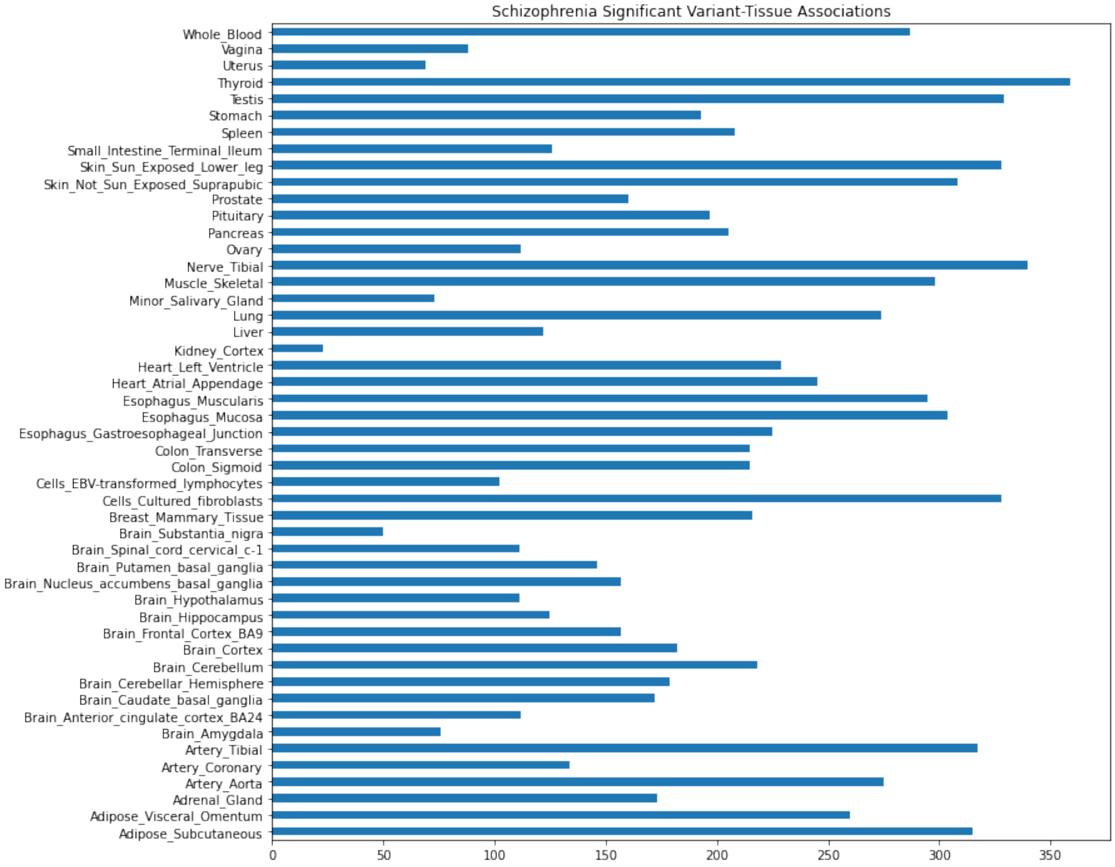
```
Text(0, 0.5, 'Trait')
```



1.5.3 Focusing on tissue associations of one trait

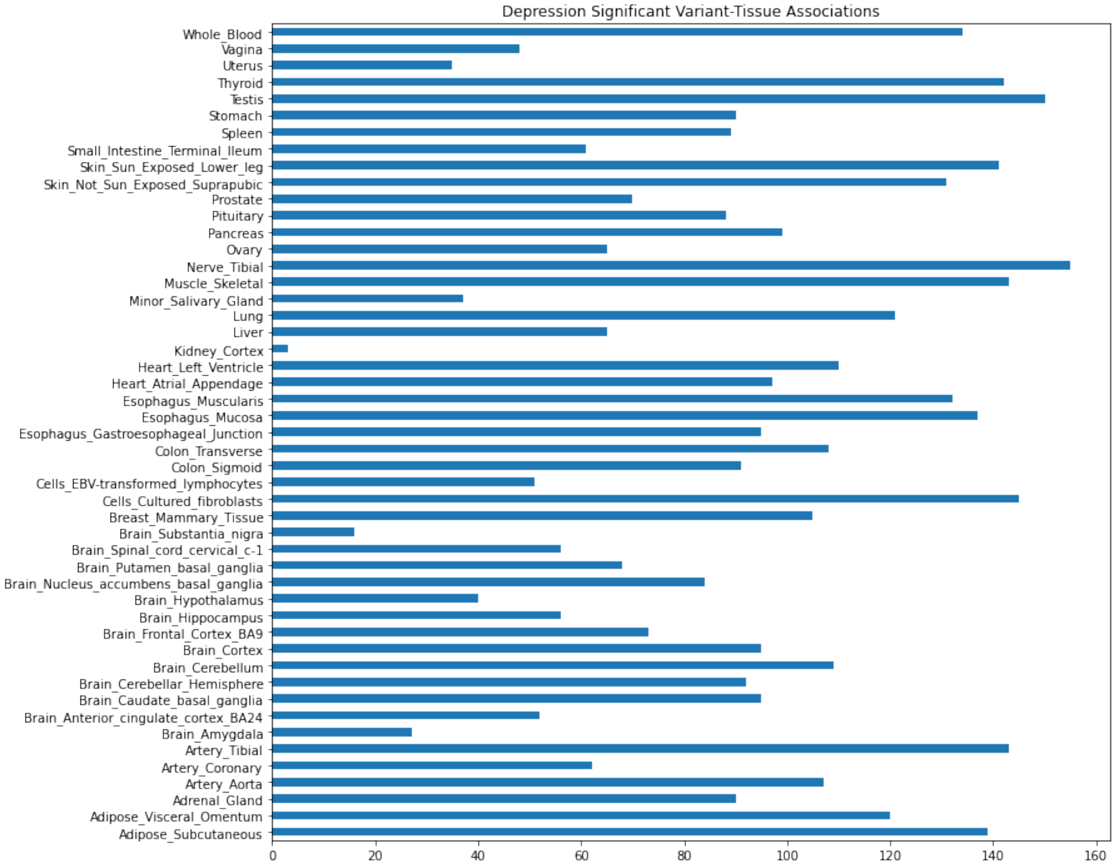
[33]:
```
fig, ax = plt.subplots(figsize=(12,12))
trait_to_tissue_count_df.loc['schizophrenia'].plot(
    kind='barh',
    ax=ax,
    title='Schizophrenia Significant Variant-Tissue Associations')
```

[33]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f274a652e10>
```



```
[34]: fig, ax = plt.subplots(figsize=(12,12))
      trait_to_tissue_count_df.loc['unipolar depression'].plot(
          kind='barh',
          ax=ax,
          title='Depression Significant Variant-Tissue Associations')
```

```
[34]: <matplotlib.axes._subplots.AxesSubplot at 0x7f274b11bf90>
```


Depression Significant Variant-Tissue Associations
Whole_Blood
Vagina
Uterus
Thyroid
Testis
Stomach
Spleen
Small_Intestine_Terminal_Ileum
Skin_Sun_Exposed_Lower_leg
Skin_Not_Sun_Exposed_Suprapubic
Prostate
Pituitary
Pancreas
Ovary
Nerve_Tibial
Muscle_Skeletal
Minor_Salivary_Gland
Lung
Liver
Kidney_Cortex
Heart_Left_Ventricle
Heart_Atrial_Appendage
Esophagus_Muscularis
Esophagus_Mucosa
Esophagus_Gastroesophageal_Junction
Colon_Transverse
Colon_Sigmoid
Cells_EBV-transformed_lymphocytes
Cells_Cultured_fibroblasts
Breast_Mammary_Tissue
Brain_Substantia_nigra
Brain_Spinal_cord_cervical_c-1
Brain_Putamen_basal_ganglia
Brain_Nucleus_accumbens_basal_ganglia
Brain_Hypothalamus
Brain_Hippocampus
Brain_Frontal_Cortex_BA9
Brain_Cortex
Brain_Cerebellum
Brain_Cerebellar_Hemisphere
Brain_Caudate_basal_ganglia
Brain_Anterior_cingulate_cortex_BA24
Brain_Amygdala
Artery_Tibial
Artery_Coronary
Artery_Aorta
Adrenal_Gland
Adipose_Visceral_Omentum
Adipose_Subcutaneous
0
20
40
60
80
100
120
140
160