Using GWAS Summary Statistics to Uncover Genetic Relationships between Mental Illnesses

Jacob Stimes

Problem Statement

My project idea arose from three realizations I gleaned from examining literature on some genomics areas I'm interested in:

1. Mental health conditions are known to have some similarities with one another (Smail, 2021), but it's possible there are more associations we have not yet discovered.
2. Researchers have been compiling summary statistics from genome-wide association studies (GWAS), but research suggests that further analysis of this data is needed to derive clinical insight (Wu, 2021).
3. Clustering is a widely used technique in genomics for perceiving latent relationships in data (Rappoport, 2018).

Given that research suggests certain mental illnesses have genetic relationships and there is easy-access GWAS data that needs further study, I determined that applying clustering to this data could potentially help uncover these latent relationships.

Significance

Mental health disorders are prevalent in the United States, yet attitudes towards them can vary wildly amongst the population: according to the CDC, in 2020, about 1 in 4 US adults are afflicted by some form of mental illness (CDC, 2012). The CDC has invested significantly into understanding attitudes of Americans on mental disorders; in their latest study, they report that 35-67% of adults feel "people are caring and sympathetic to people with mental illness," suggesting stigmas on the subject (CDC, 2012). Furthermore, they also report that, although the majority of adults agree treatment for mental disorders is effective, the responses for certain demographics actually strongly disagree (CDC, 2012). With the ongoing COVID-19 pandemic, there has never been a better time to investigate causal factors in mental health conditions. The Mayo Clinic has shown, based on survey data, that the rate of mental health symptom reporting "show a major increase" compared to before the COVID-19 pandemic (Mayo Clinic, 2021), and other researchers estimate that the reduced number of people seeking care during the pandemic will lead to an increase of severe symptoms later on (Carr, 2021).

Background

Genome-wide Association Studies (GWAS)

Genome-wide association studies (GWAS) are a particular method that has been used recently to identify the genetic associations with diseases (Uffelmann, 2021). At a high-level, these studies are conducted by genotyping large sample sizes of cases and controls (people with and without a certain phenotype or condition) and analyzing the genetic variants which are only present in the cases. By nature of the "genome-wide" approach, it is possible to confirm the polygenic nature of certain mental health disorders, such as schizophrenia (Escudero, 2014). These studies provide summary statistics, which are usually lists of the statistically-significant SNPs found in cases but not in controls used in the study (Pasaniuc, 2016). When the results are made available in this way, they can be utilized in meta-analyses to generate further findings, as done by Landi et al in their research on cutaneous melanoma (Landi, 2020).

GWAS experiments are not without their limitations, though. Two prominent factors must be accounted for in all studies: population stratification and linkage disequilibrium (Tam, 2019). Population stratification refers to information encoded into a genome which reflects a person's ancestral origin and can lead to spurious associations when not accounted for in both the study design, for instance selection of study participants, and analysis, for example using the population stratification as a covariate in the derived statistical model (Marchini, 2004). This information is not only accurate enough to describe a person's continent of origin, but even their ancestral country, so choosing, for example, only Europeans in a study is not alone sufficient to account for this factor (Novembre, 2008). Linkage disequilibrium describes the tendency of adjacent alleles to be inherited together; in GWAS it must be considered when nearby SNPs are shown to be correlated with cases, as some SNPs may not be causal and simply correlated with a causal SNP due to linkage disequilibrium (Vilhjalmsson, 2015).

Clustering, a data analysis technique for grouping data points, is used frequently in the genomic space to uncover latent relationships about high-dimensional data (Han, 2017). Hierarchical clustering algorithms are especially effective in genomics due to their ability to give an overview of relationships between genes (or any dataset) at various granularities (Pagnuco, 2017). However, care must be used with these approaches since validating the results is non-trivial, but can still be done with tools such as the Silhouette and Dunn indices (Pagnuco, 2017). Combining GWAS results with eQTL gene expression data has been shown to be successful in discovering variant implications (Zhu, 2016). These results suggest there may be value in exploring the use of clustering in this area to confirm discoveries or identify new ones.

Mental Illness

As our world population continues to grow and life spans increase, it's expected more and more people will require treatment for some form of mental illness (Bailey, 2018). As mentioned above, researchers such as Escudero et al have succeeded in identifying the genetic basis of certain mental illnesses; these findings additionally allow them to recognize potentially related conditions based on shared affected genes, such as bipolar disorder and schizophrenia

(Escudero, 2014). While these associations aren't alone a means to clinically understand the relationships between conditions, they serve as a starting point for further exploring and quantifying connections that have long been debated, for instance the associations between depression and dementia (Bennett, 2014), delirium and cognitive decline (Popp, 2013), and obsessive-compulsive disorder and schizophrenia (Kayahan, 2005). More insight into these connections can help researchers better understand the conditions and allow doctors to better diagnose the conditions; it may even lead to more proactive screening of non-mental illnesses in those with a mental illness, as suggested by the correlation in coronary heart disease and mental disorders (Hert, 2018).

One challenge with classification and diagnosis of mental illnesses is they tend to be "defined solely by descriptive, usually behavioural, criteria" (Cross Disorder Phenotype Group of the Psychiatric GWAS Consortium, 2009). To elaborate further, symptoms of one condition may also be symptoms of another, and some conditions may be associated with many symptoms, even though only a subset of which need to be present to merit a diagnosis. Latest research into the genetic basis of psychiatric traits demonstrates that many are polygenic, which suggests that there is ample potential to analyze the genetic overlap of these conditions (Réthelyi, 2019). Multiple GWAS findings on a variety of mental illnesses have been published in recent years and cross-disorder studies are an area of interest emphasized by researchers in the field (Santoro, 2016).

Aims and Rationale

Understanding the genetic factors behind mental illnesses could potentially change social perspectives on the illness and treatments, provide insight for targeted drug discovery, and allow us to discover and quantify connections between different types of conditions. Therefore, for my project, I chose to examine genomic data to quantify relationships and characteristics of mental health disorders. Since there is publicly available GWAS summary data on more than 10 mental health conditions, I decided to download this data for further analysis and apply clustering techniques to attempt to determine similarities between conditions. I hypothesized that conditions that appeared to be associated together based on clustering results may be starting points for future study.

Methods

At a high-level, my methods consisted of the following steps:

1. Retrieving and exploring GWAS summary statistic data.
2. Cleaning the data based on the above explorations, and then analyzing cleaned data across conditions.
3. Performing clustering on the cleaned data, utilizing some insights from the above analyses.

Retrieving and Exploring GWAS Summary Statistics

I retrieved GWAS summary statistics from GWAS catalog (Buniello, 2019). This data can be found here and the specific download links are documented in a metadata file I created, located here. It's worth noting I only retrieved data for top-level mental health traits (deemed "parent traits" throughout). These parent traits include data for their child traits (documented in the metadata file linked above). This was to avoid redundancy in the analysis. The traits I downloaded were those in the child traits of "mental or behavioural disorder" which were not child traits of another mental or behavioral disorder (i.e. top-level) and had at least 30 reported associations. I was originally hoping to use the data hosted by the Psychiatric Genomics Consortium as well, until realizing this data is not reported in a standardized format as it is in GWAS catalog (Cross Disorder Phenotype Group of the Psychiatric GWAS Consortium).

That being said, the GWAS catalog data is not entirely standardized either. Through some exploratory analysis on data for a single condition, I identified some challenges and opportunities for normalization of this data. The entirety of this analysis can be found in the python notebook I used (and for each notebook, there is a generated PDF based on the notebook, although I personally found the notebook itself to be easier to read). For this analysis, mostly built-in python libraries were used, in addition to Pandas and NumPy (Reback, 2022; Harris, 2020).

I then retrieved more data for these variants by:

- Utilizing dbSNP to add allele frequency (AF) information if present/possible (Sherry, 1999).
- Parsing and joining "Single-Tissue cis-QTL Data" from GTEx portal to identify any tissues with which a variant is significantly associated with differential gene expression (Genotype-Tissue Expression Project). Detailed data setup steps described here.

Some exploration of this data for a single condition is performed in the same python notebook linked above.

Data Normalization / Cleaning

Based on my analysis of the raw GWAS summary statistic data, I developed a script to perform my desired normalization and cleaning of this data. I developed some additional scripts to add the allele frequency information and tissue association details.

Some more of the "why" will be discussed in the results section, but normalization consisted of:

- Filtering variants to just those directly associated with the parent trait or one of the defined child traits of the condition.
- Converting p-values from strings to numbers.
- Removing some bad data (some variants are reported as "undefined").
- Removing duplicate variants.
- Separating rows with multiple genes into multiple rows with one gene.

- Imputing variant location if not set and possible to obtain (sometimes variant ID was reported as the variant location instead of an RS ID).
- Dropping columns unused in subsequent analysis, renaming columns.

Analysis of Cleaned Data

With the data normalized and joined with extra information, I developed another python notebook to examine this data for all conditions. I utilized the Seaborn and Matplotlib libraries to assist in visualizing this data (Waskom, 2021; Hunter, 2022). Histograms, bar plots, box plots, heatmaps, and even pie charts are some tools used to help interpret this data and compare between conditions.

Encoding & Clustering

I developed a final notebook which loaded the clean data and utilized it in various data clustering experiments. In this notebook, I utilized the scikit-learn and scipy python libraries in addition to those used previously (Pedregosa, 2011; Virtanen, 2020).

I relied on vector-distance as a way to compare traits, and therefore needed to encode my data into vectors. To do this, I tried the following approaches:

Gene-based encoding

- I transformed data for a trait into an “implicated-gene” encoding, where each gene in my dataset is mapped to an index, and the vector V = (v1, v2…vN) for a trait is derived by setting vi to 1 if gene-i is has been implicated by a variant for that trait, and 0 if not. This concept was inspired by the Gene2vec work (Du, 2019).
- I iterated on that approach to filter my set of all genes implicated by any trait to only those genes which had been implicated by more than one trait (to reduce dimensionality of the data).
- I further iterated on this encoding to weight each element of the vector by how many times it was implicated by variants of a given trait, and then normalized the values by dividing by the max number of gene-implications. For example, suppose a trait had been mapped to geneA 1 times, geneB 2 times, and geneC 3 times: then, before normalizing, the vector would be (1, 2, 3). After normalizing, the vector would be (⅓, ⅔, 1). This ensures all vector elements for all traits are in the range [0,1].

Tissue-based encoding

I applied a similar approach as the gene-based encoding, except now mapped each implicated tissue to a vector index instead of each gene. I used the weighted, normalized vector with this experiment as it showed better results with the gene-based encoding (more details in results).

Combined gene & tissue encoding, Principal component analysis

I also experimented with concatenating the gene-implicated and tissue-implicated vectors as I thought it would provide more information into the clustering algorithm. Additionally, I experimented with performing principal component analysis (PCA) on this data and plotting the top two principal components to see if this exhibited any other types of patterns in the data.

Data filtering prior to encoding & clustering

To improve the clustering results, I employed the following methods to filter some potentially noisy or less significant data:

- I excluded certain traits with the least amount of data
- I attempted to use the same number of data points (variants) per trait by filtering to the top X variants where X is the number of data points in the trait with the fewest data points and “top” refers to lowest p-values for GWAS variant associations.
- I filtered data to certain allele frequency ranges
- I used the tissue-associations to limit the data to only variants known to be significantly associated with gene expression in a tissue.

Clustering

With encoded and filtered data, I explored using the ward hierarchical clustering algorithm and the k-means clustering algorithm (with various values for k). For hierarchical clustering, I plotted dendrograms to visualize results, whereas with k-means, I used PCA to reduce data to 2-dimensions to generate scatter plots to communicate the clustering outcomes.

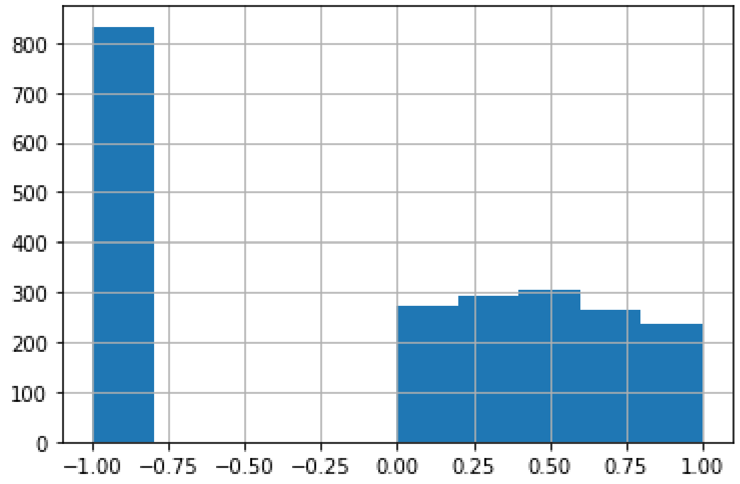
Results

Exploratory Analysis of GWAS Summary Statistics

Exploring the raw GWAS summary statistics revealed some limitations with the overall dataset that I felt needed to be addressed:

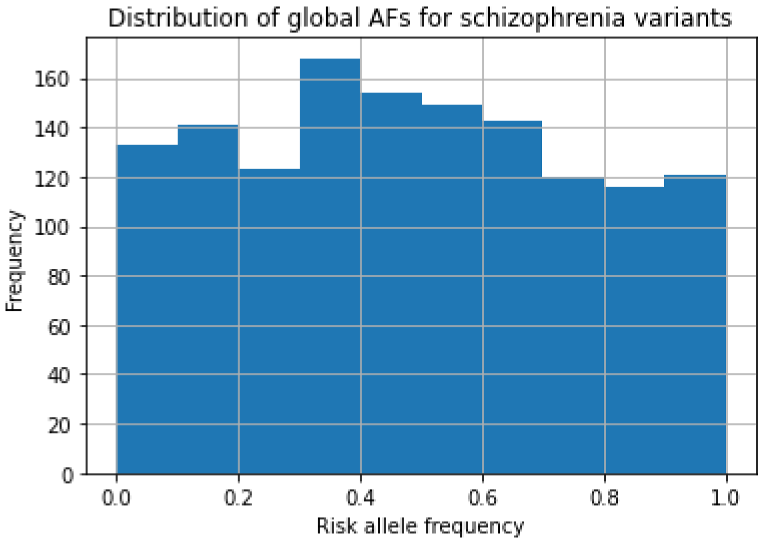
- Some variants were duplicated, since they were found in two or more studies. Since I didn’t consider the specific study origin in my subsequent analysis, I chose to filter duplicates and retain the copy with the smallest p-value.
- The “Trait(s)” column did not only include entries matching the trait of the given dataset. For example, for schizophrenia, some values for the ‘Trait’ column were: “*Schizophrenia*”, “*Schizophrenia or bipolar disorder*”, “*Schizophrenia vs bipolar disorder (ordinary least squares (OLS))*”, “*Anorexia nervosa, attention-deficit/hyperactivity disorder, autism spectrum disorder, bipolar disorder, major depression, obsessive-compulsive disorder, schizophrenia, or Tourette syndrome (pleiotropy)*”. Since I planned to compare traits with one another, I thought it would be best to avoid including variants that may belong to other traits in the dataset for a given trait. Therefore, I filtered the data to just variants that were associated with just the parent trait, or child traits of the parent trait.
- Not all variants had a mapped gene, which was important to notice to avoid encoding the “UNKNOWN” gene in the encoding prior to clustering.
- GWAS Catalog provides a “RAF” (risk allele frequency) column, but this is calculated based on the study samples (definition on this page), so I felt it wasn’t adequate for

comparing with other variants from other studies. Instead, I attempted to retrieve the global allele frequency (AF) from dbSNP if the variant had an RS ID and a known AF. I was unable to locate the global AF for a decent proportion of variants; the following chart depicts AF for variants associated with schizophrenia, where variants without a known AF are given the value -1 for visualization:



Global allele frequency for schizophrenia GWAS-identified variants, according to dbSNP. -1 represents unknown allele frequencies.

Removing those variants with unknown AFs for schizophrenia, the histogram looks like the following:

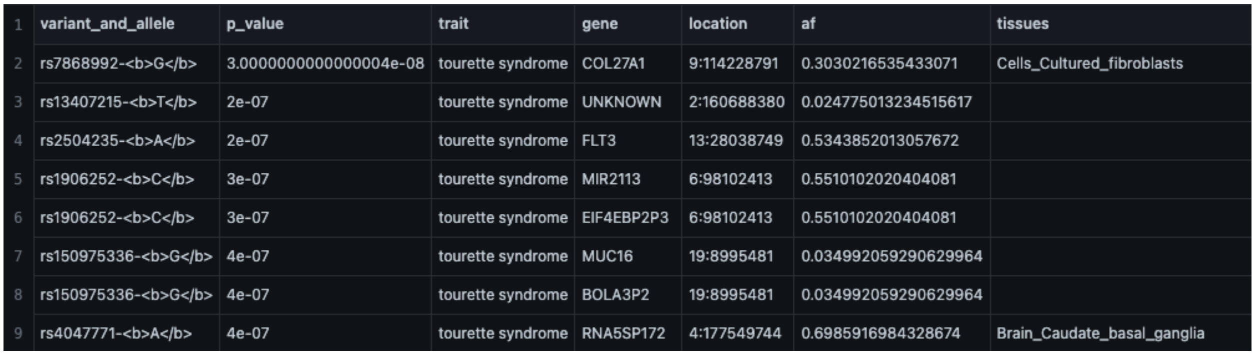


Sometimes the risk allele was unspecified. I initially thought to simply choose the minor allele reported by dbSNP for imputing the AF; however, I saw that a non-zero number of risk alleles reported in the GWAS data were actually major alleles, so I decided it may not be a safe assumption to always choose the minor allele if unspecified.

- The variant location was not always set in the ‘Location’ column from GWAS data; sometimes the variant was reported as a location, and so it was easy to impute the location based on the variant ID. Otherwise, I was unable to set the location.
- Significant variant-gene associations in tissues are identified by variant location. Fortunately, GWAS catalog and GTEx portal used the same human genome version so I was able to easily join this data if the variant location information was available in GWAS data.

Data Normalization / Cleaning

I cleaned the GWAS summary statistic data using the methods described above based on the observations listed above. The cleaned GWAS summary statistic data is available here. As mentioned, I also retrieved AF data and joined the data with tissue associations. The final, combined data is available here. The format for a given trait has a structure like this:



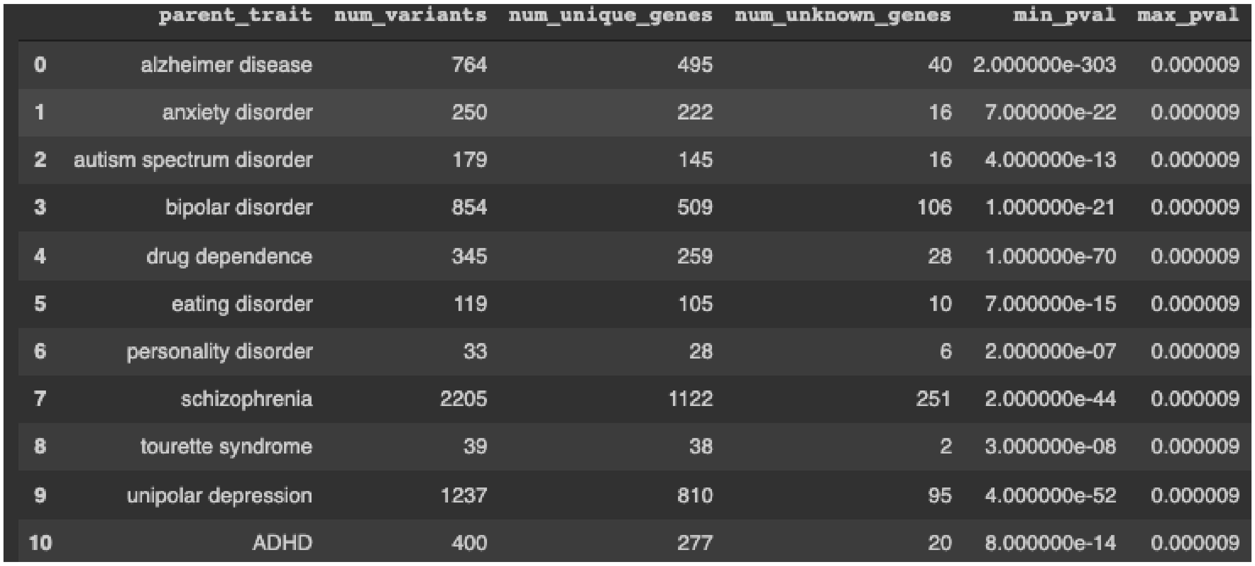
variant_and_allele	p_value	trait	gene	location	af	tissues
rs7868992-<b>G</b>	3.0000000000000004e-08	tourette syndrome	COL27A1	9:114228791	0.3030216535433071	Cells_Cultured_fibroblasts
rs13407215-<b>T</b>	2e-07	tourette syndrome	UNKNOWN	2:160688380	0.024775013234515617	
rs2504235-<b>A</b>	2e-07	tourette syndrome	FLT3	13:28038749	0.5343852013057672	
rs1906252-<b>C</b>	3e-07	tourette syndrome	MIR2113	6:98102413	0.5510102020404081	
rs1906252-<b>C</b>	3e-07	tourette syndrome	EIF4EBP2P3	6:98102413	0.5510102020404081	
rs150975336-<b>G</b>	4e-07	tourette syndrome	MUC16	19:8995481	0.034992059290629964	
rs150975336-<b>G</b>	4e-07	tourette syndrome	BOLA3P2	19:8995481	0.034992059290629964	
rs4047771-<b>A</b>	4e-07	tourette syndrome	RNA5SP172	4:177549744	0.6985916984328674	Brain_Caudate_basal_ganglia

The code for cleaning GWAS summary statistics, loading AF data, and joining with cis-eQTL data is also available for potential reuse in other similar experiments (e.g. to compare/contrast data for different cancers).

Analysis of Cleaned Data

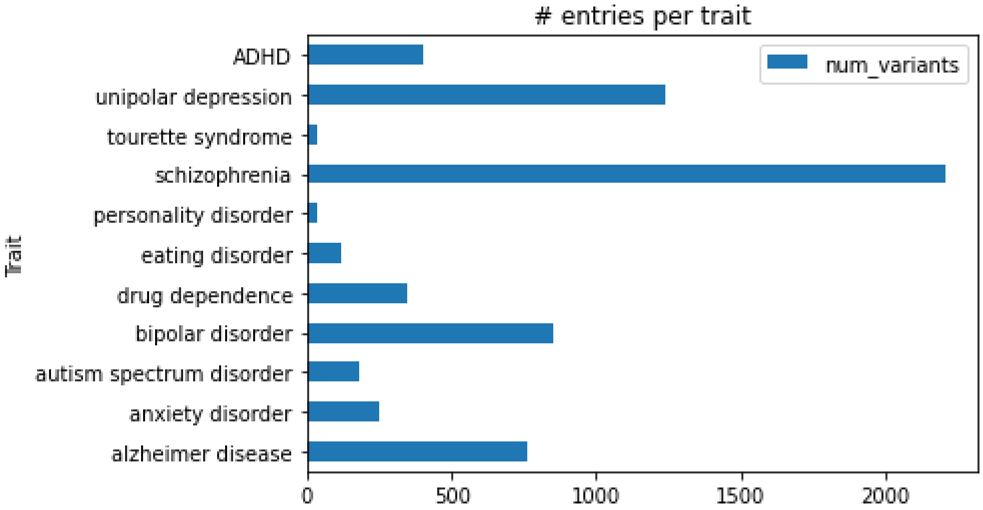
With the cleaned data for each trait, I performed some further analysis to compare traits prior to encoding and clustering.

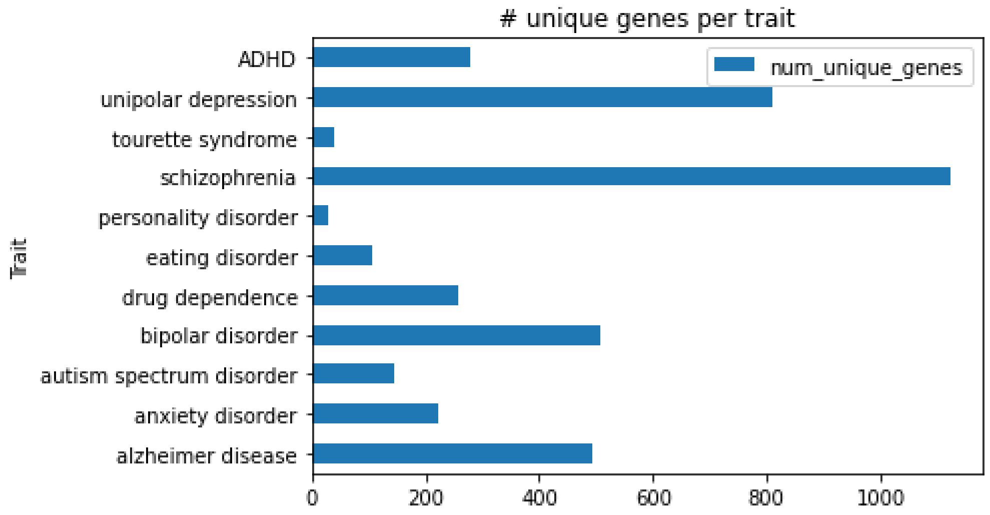
I generated some simple summary stats on each trait:



	parent_trait	num_variants	num_unique_genes	num_unknown_genes	min_pval	max_pval
0	alzheimer disease	764	495	40	2.000000e-303	0.000009
1	anxiety disorder	250	222	16	7.000000e-22	0.000009
2	autism spectrum disorder	179	145	16	4.000000e-13	0.000009
3	bipolar disorder	854	509	106	1.000000e-21	0.000009
4	drug dependence	345	259	28	1.000000e-70	0.000009
5	eating disorder	119	105	10	7.000000e-15	0.000009
6	personality disorder	33	28	6	2.000000e-07	0.000009
7	schizophrenia	2205	1122	251	2.000000e-44	0.000009
8	tourette syndrome	39	38	2	3.000000e-08	0.000009
9	unipolar depression	1237	810	95	4.000000e-52	0.000009
10	ADHD	400	277	20	8.000000e-14	0.000009

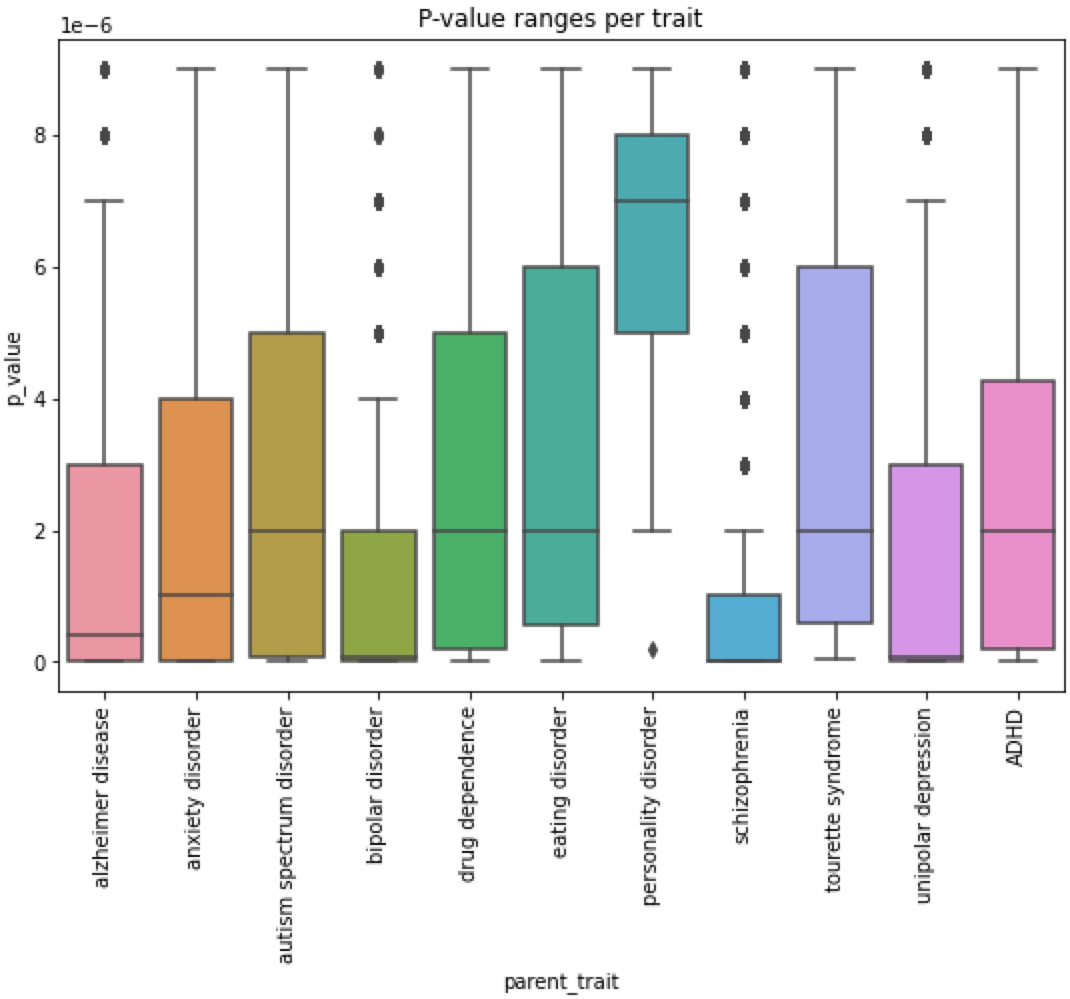
I generated bar plots of the number of entries per trait and number of unique genes per trait:



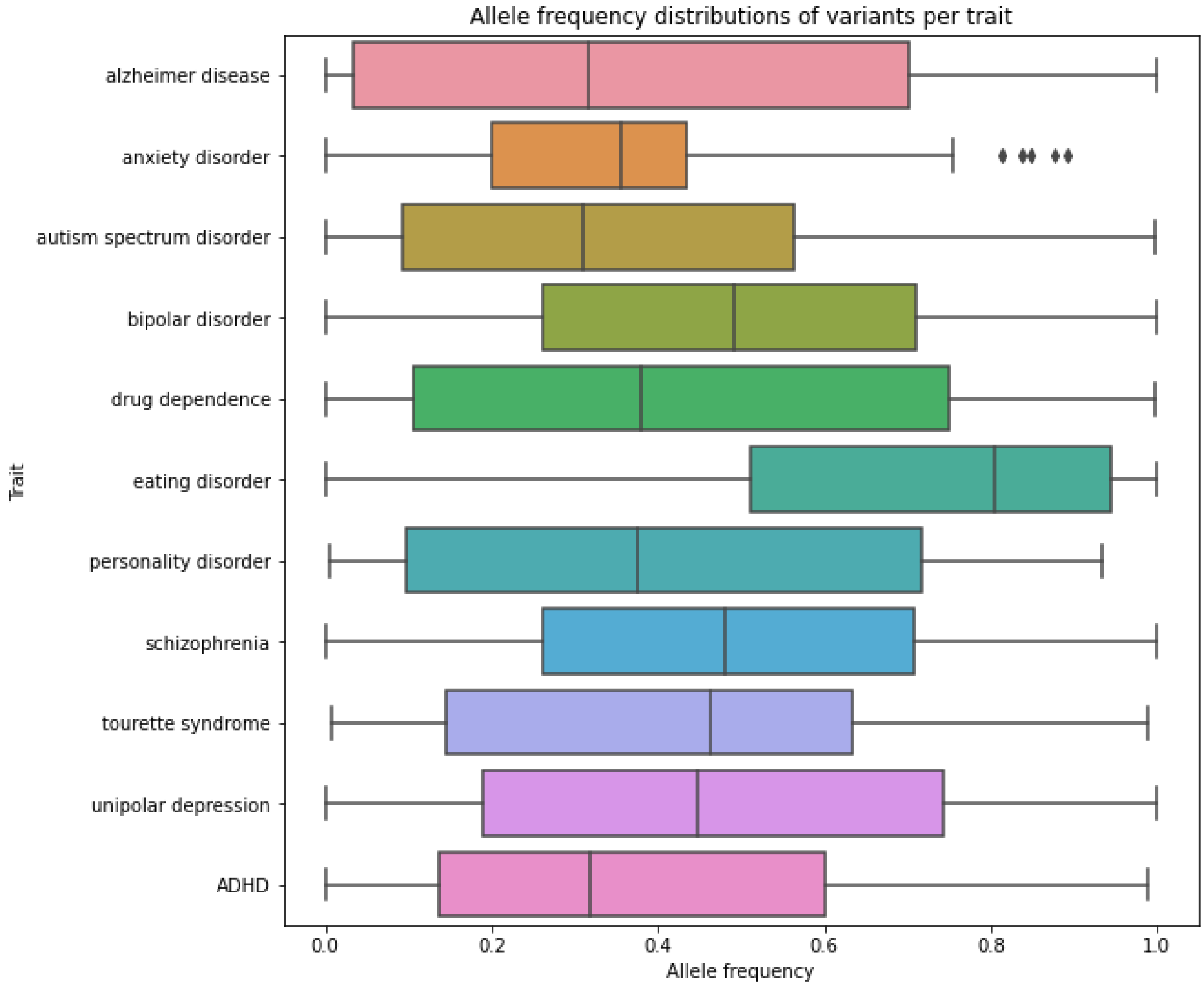


One very apparent observation is that there are significantly different amounts of data for each trait, with schizophrenia and unipolar depression appearing to be the most well-studied, and personality disorder and tourette syndrome being the least studied. Also, all traits have some associated variants which have no known mapped genes.

I additionally generated some box plots to examine the distributions of different properties of traits:

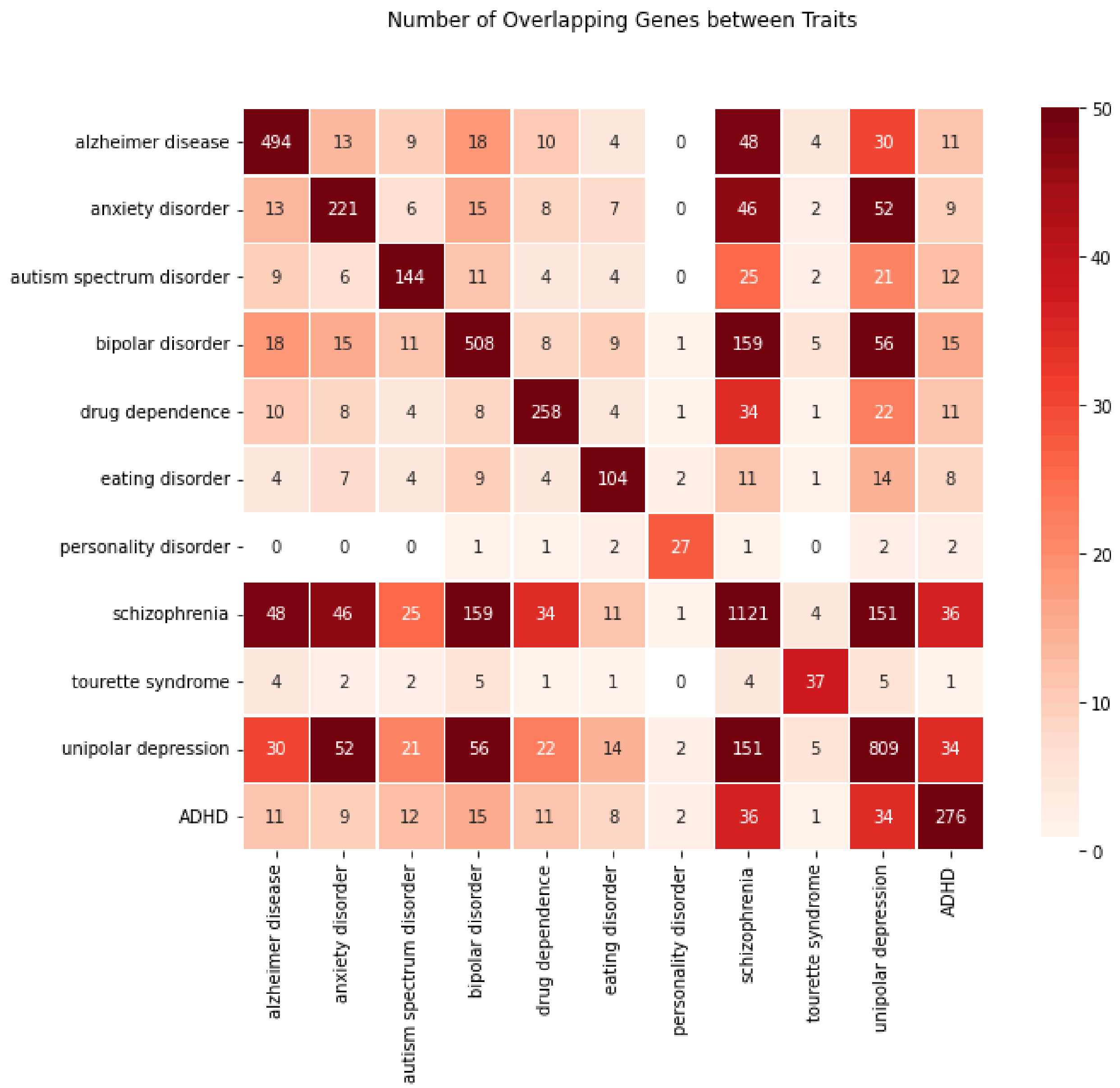


Even though schizophrenia has the most data points, it has a noticeably smaller interquartile range (IQR) of p-values for the reported variant associations, whereas personality disorder sticks out as having a higher IQR of p-values. Otherwise, the other traits have roughly the same p-value ranges.



Similarly, the AF distribution of each trait is roughly similar. An obvious distinction is that eating disorder's IQR is actually from 0.5 to 0.9, suggesting that eating disorders may be caused by common rather than minor variants. Of course, as mentioned previously, I was not able to obtain this data for all variants so the true distribution is not depicted. Another distinction is that anxiety disorder has a more-constrained AF distribution from ~0.2-0.4. This could suggest that minor but not especially rare variants make up the genetic basis of anxiety disorder.

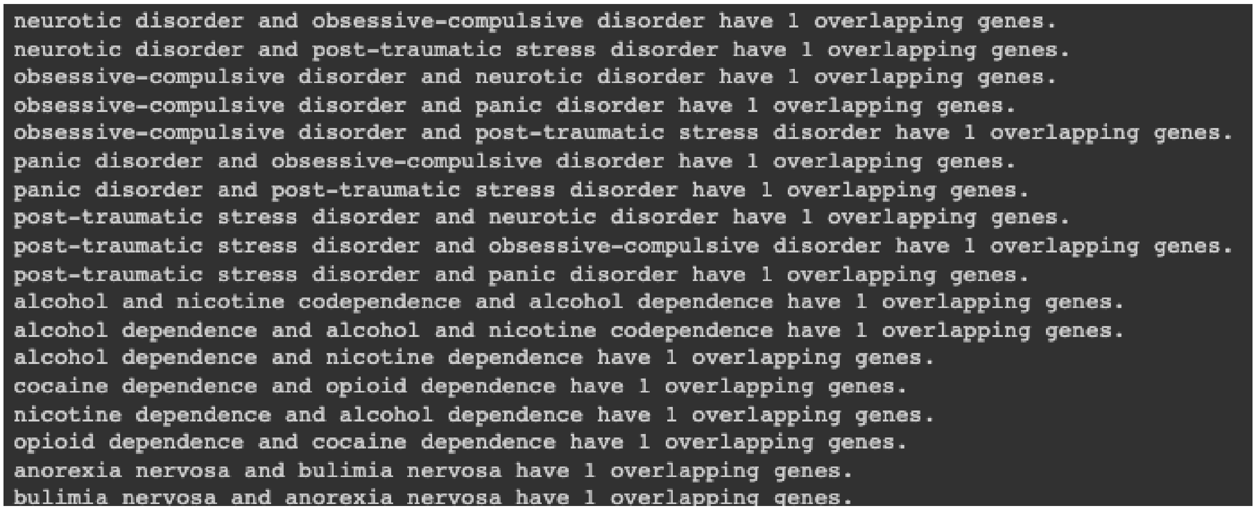
I additionally computed the number of overlapping genes for each trait (those that are implicated by a variant of each), and plotted this via a heatmap:



	alzheimer disease	anxiety disorder	autism spectrum disorder	bipolar disorder	drug dependence	eating disorder	personality disorder	schizophrenia	tourette syndrome	unipolar depression	ADHD
alzheimer disease	494	13	9	18	10	4	0	48	4	30	11
anxiety disorder	13	221	6	15	8	7	0	46	2	52	9
autism spectrum disorder	9	6	144	11	4	4	0	25	2	21	12
bipolar disorder	18	15	11	508	8	9	1	159	5	56	15
drug dependence	10	8	4	8	258	4	1	34	1	22	11
eating disorder	4	7	4	9	4	104	2	11	1	14	8
personality disorder	0	0	0	1	1	2	27	1	0	2	2
schizophrenia	48	46	25	159	34	11	1	1121	4	151	36
tourette syndrome	4	2	2	5	1	1	0	4	37	5	1
unipolar depression	30	52	21	56	22	14	2	151	5	809	34
ADHD	11	9	12	15	11	8	2	36	1	34	276

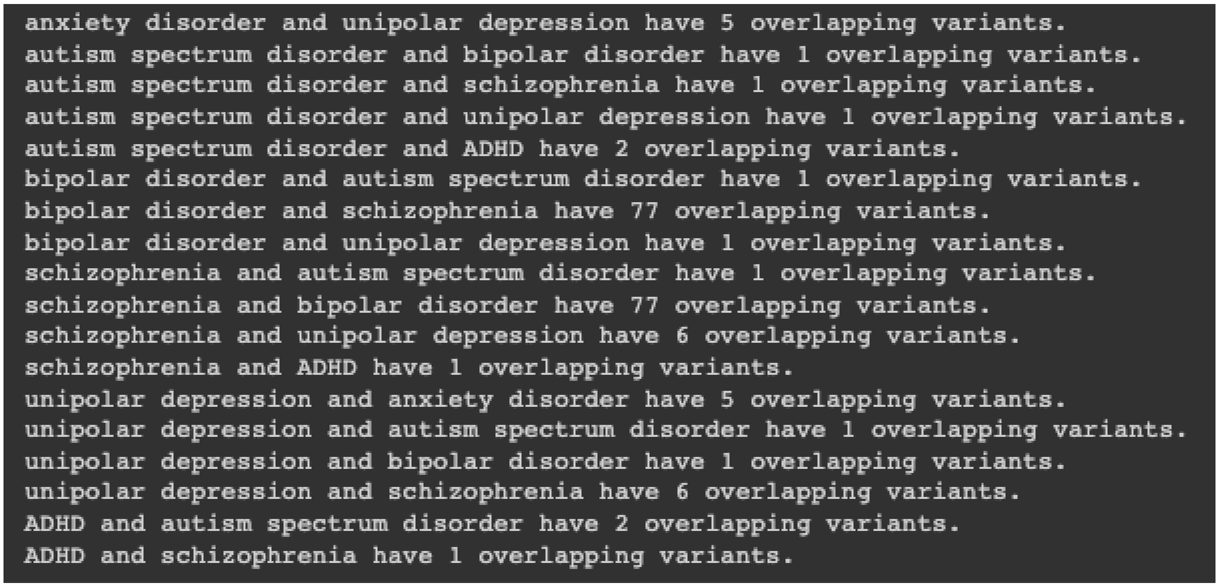
Again, an observation here is that some traits have much more data which skews the graph a bit; I chose the max coloring-cutoff to be 50 to account for it slightly, but you can still notice schizophrenia and unipolar depression appear to be related to most other traits. However, I believe this is simply because they are more studied than other traits and therefore have more known implicated genes, and thus are more likely to have overlap with other traits. Nonetheless, a takeaway is that all traits have variants that affect at least one gene in common with another trait. To further examine the significance of this, it may have been useful to include data from traits which are not mental health conditions.

I similarly compared gene overlap for child traits of a given trait. However, since no two child traits had more than one gene in common, I did not visualize the data. The results are the following (note that I neglected to reduce duplication of each pair):



```
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.
```

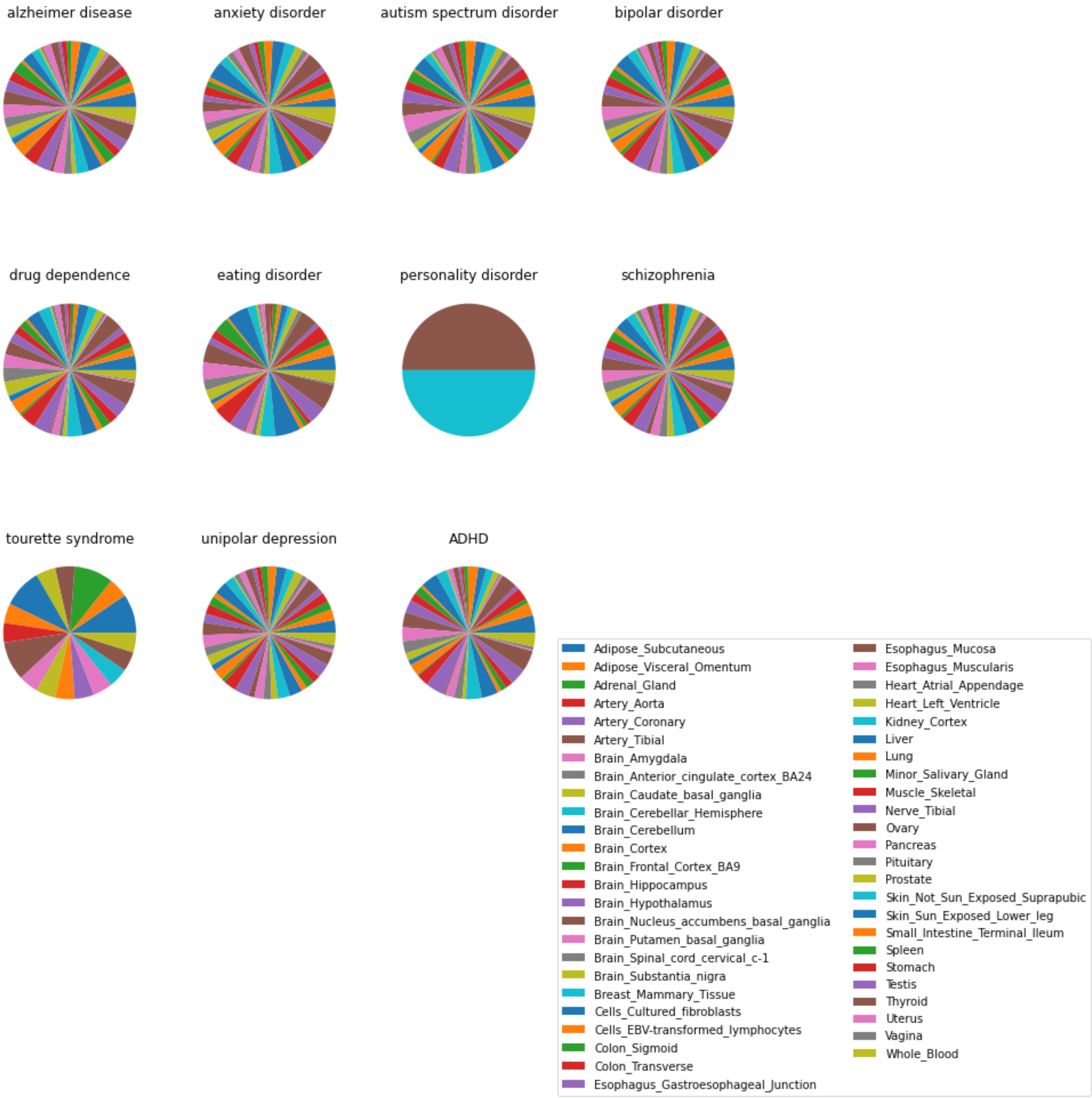
I similarly checked for overlap of specific variants rather than genes, and my results were the following:



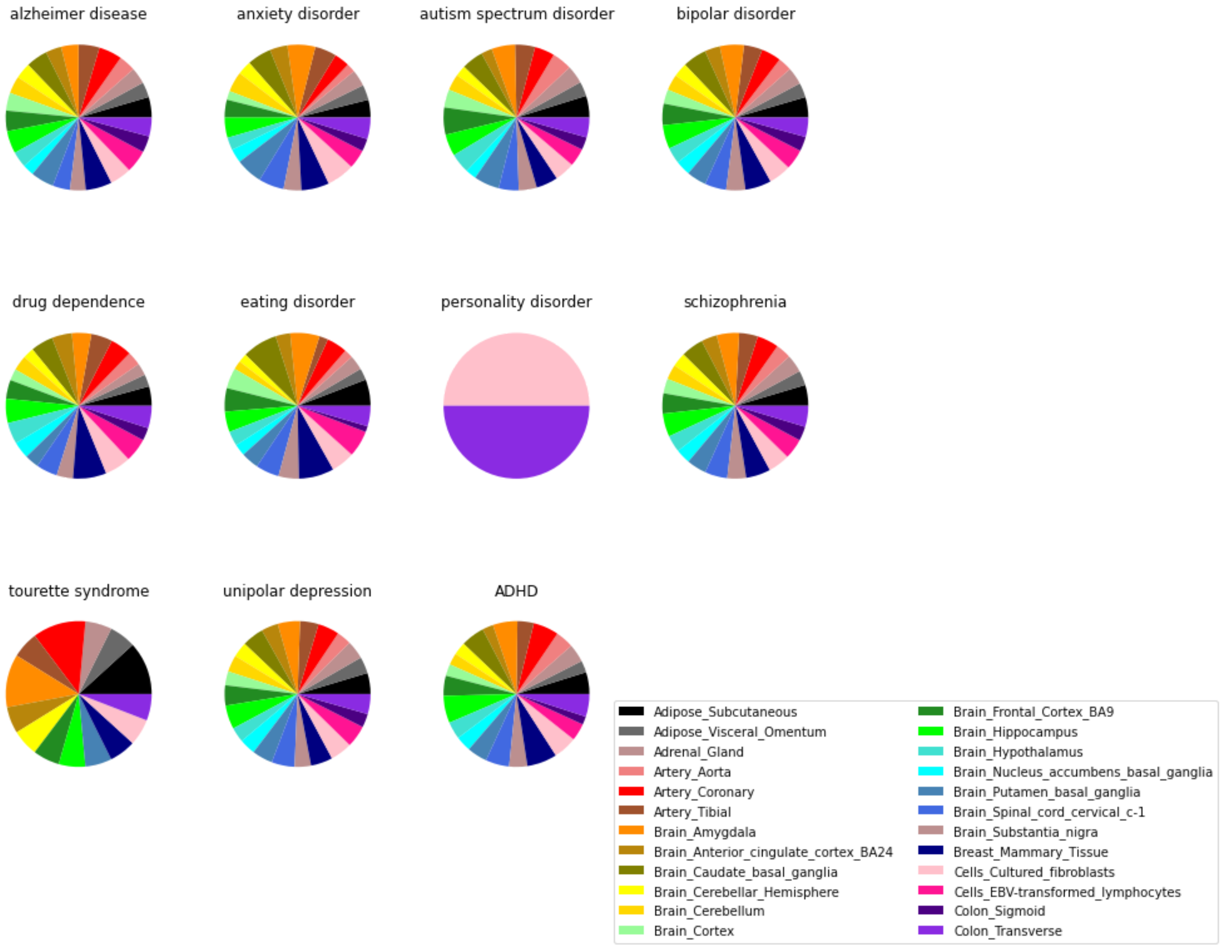
```
anxiety disorder and unipolar depression have 5 overlapping variants.
autism spectrum disorder and bipolar disorder have 1 overlapping variants.
autism spectrum disorder and schizophrenia have 1 overlapping variants.
autism spectrum disorder and unipolar depression have 1 overlapping variants.
autism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

A noteworthy result is the 77 overlapping variants between schizophrenia and bipolar disorder. This reflects results reported by Escudero (2018).

Finally, I also examined the various tissue associations of each variant. A negative result here is that I had a hard time visualizing and therefore interpreting this data. Along with the prior theme of my results, one issue was that tissue associations are also biased by how many variants are reported to be associated with a trait in GWAS catalog. To account for this, I thought this may be a good use of pie charts:

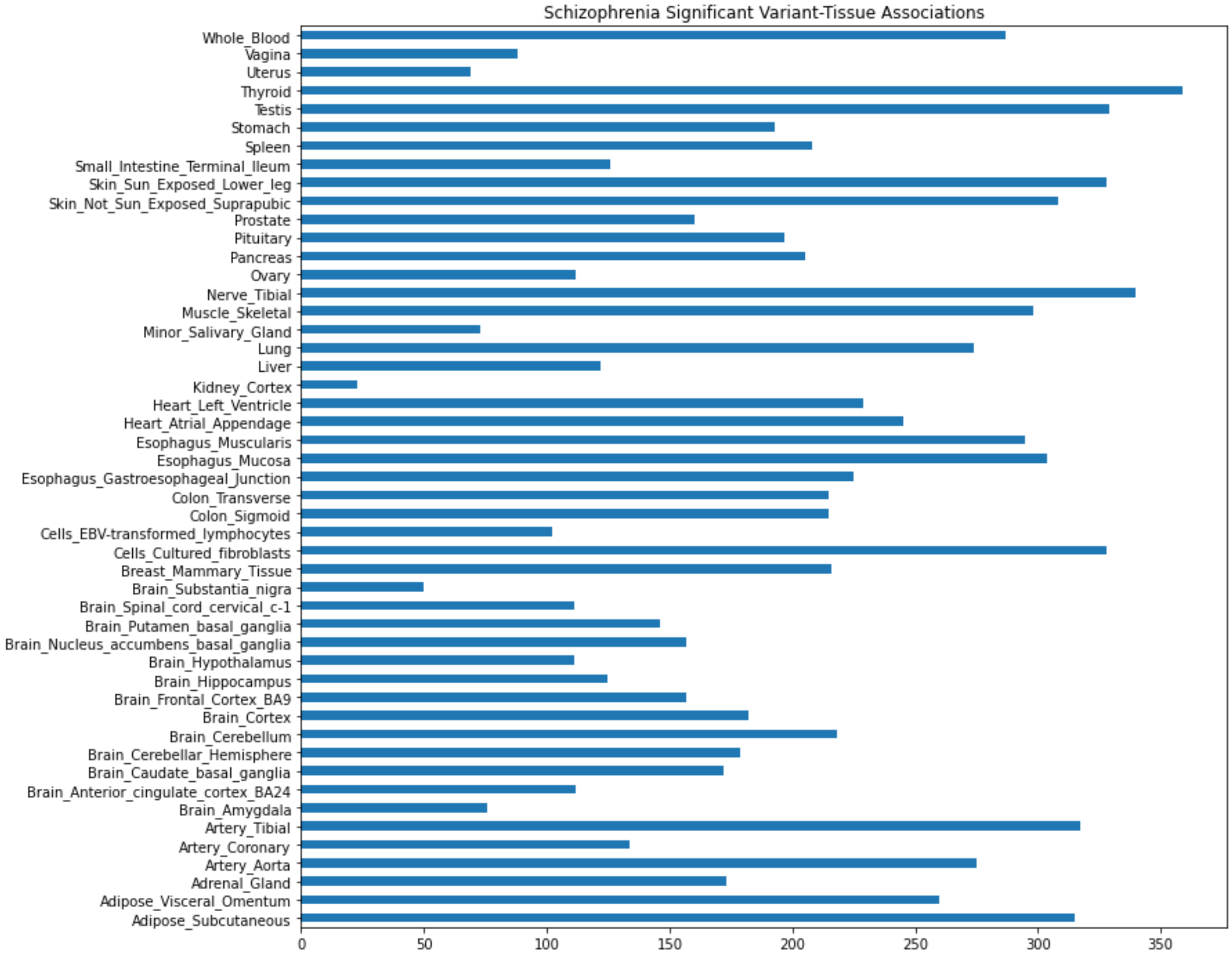


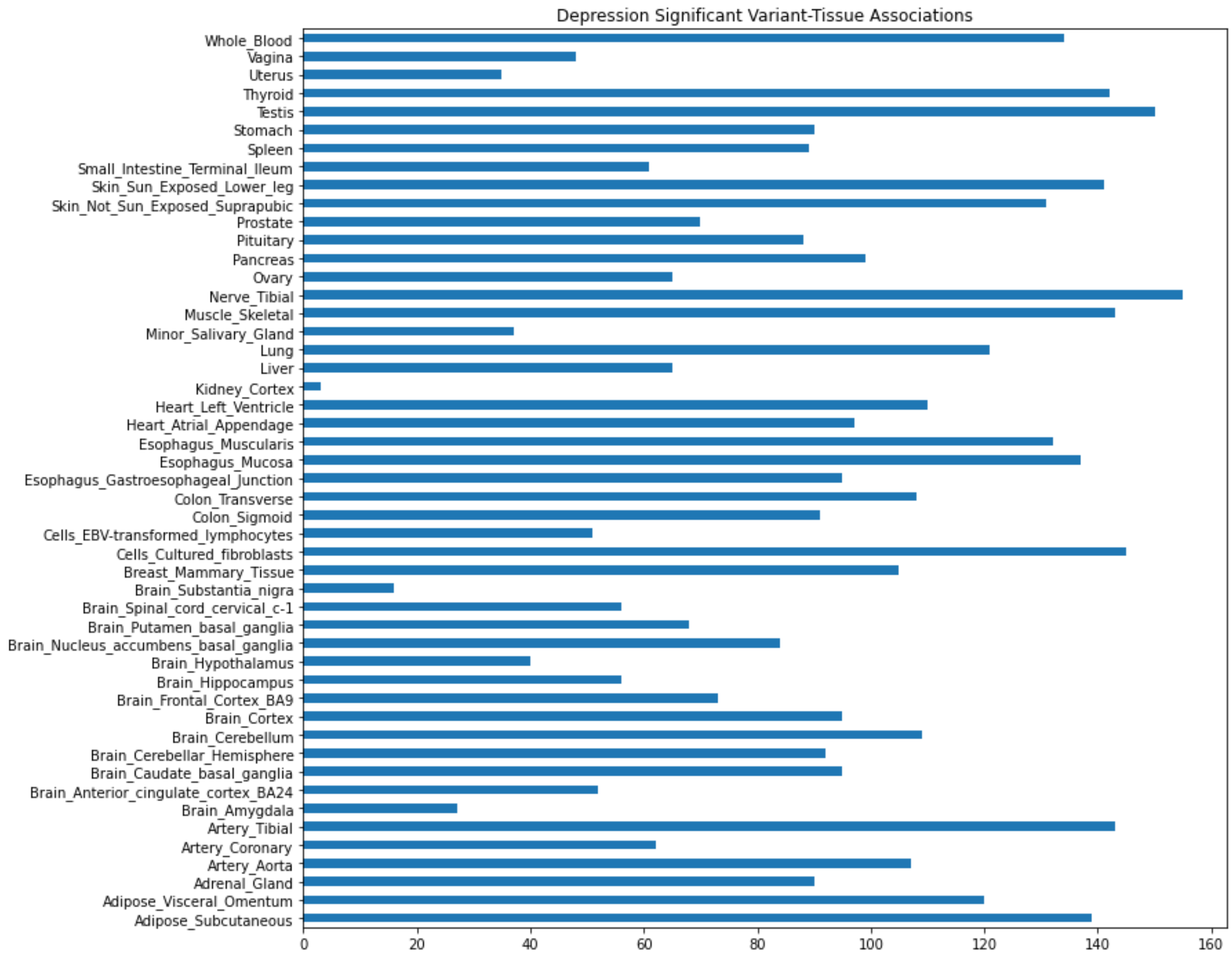
However, there are more tissues in the dataset than values in typical colors palettes. So I tried reducing it to the top 25 tissues across all traits and using more distinct colors:



I still found it pretty hard to decipher, and the traits with low amounts of data stick out the most. However, if you look closely, you may notice some differences: the brown slice, Artery_Tibial, is significantly smaller in eating disorder than the other traits.

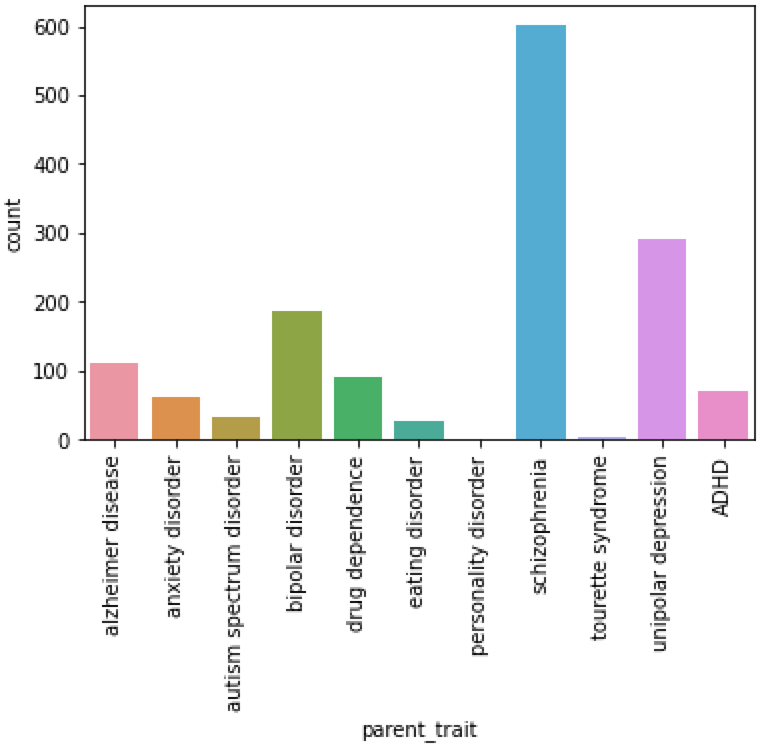
I instead then just tried comparing bar plots of tissue associations for two specific traits with the most data:


Schizophrenia 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
50
100
150
200
250
300
350

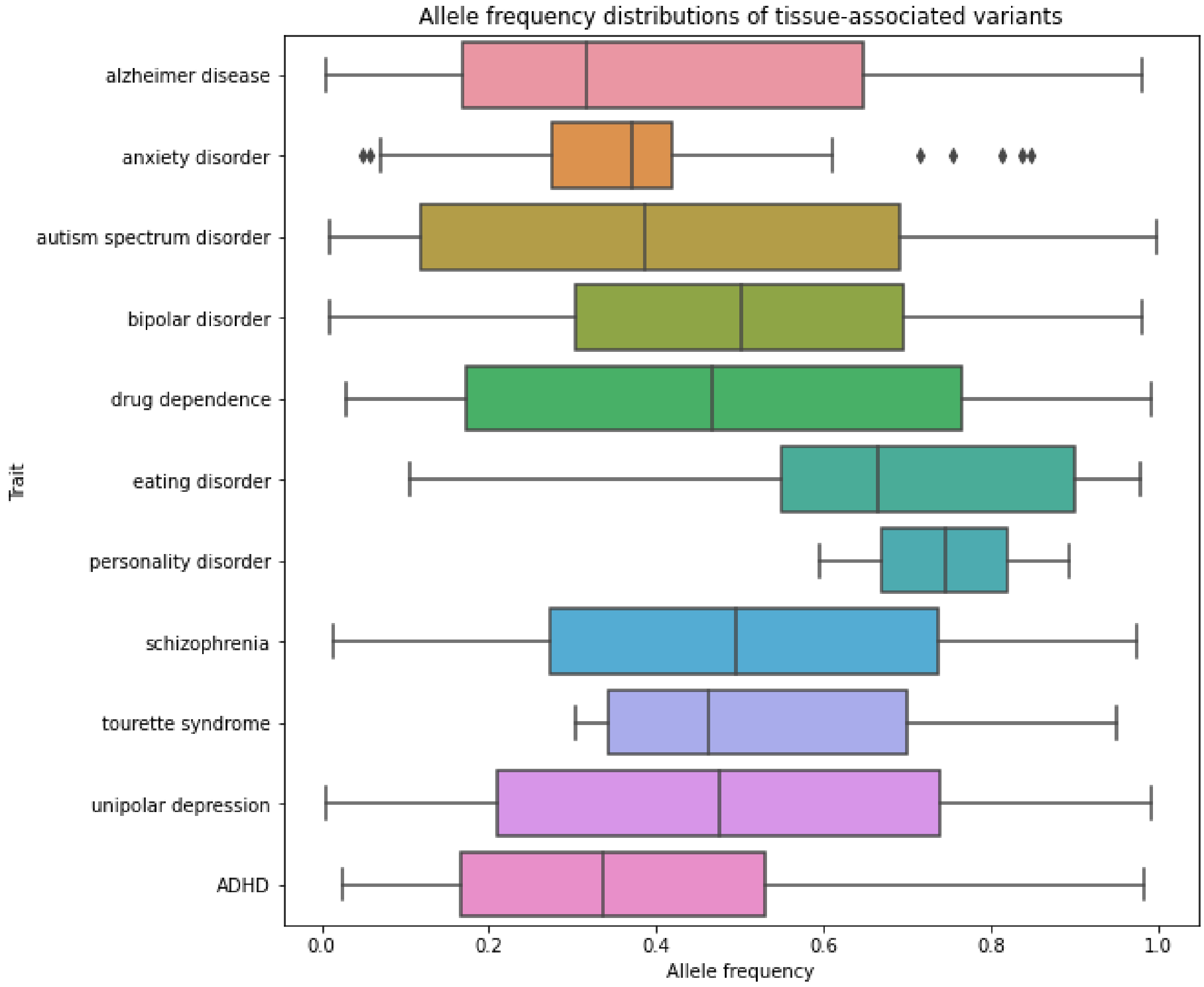


Again, I found it hard to glean insights by comparing these charts. However, some observations are that the most common tissue differs between the traits (Nerve_Tibial for depression, Thyroid for schizophrenia).

I also filtered my dataset to just variants associated with tissues before generating some previous plots, although there weren't many additional insights from this. Numbers of variants per-trait associated with tissues still reflects trait data-size:



And AF distributions are roughly the same, albeit slightly more constricted:

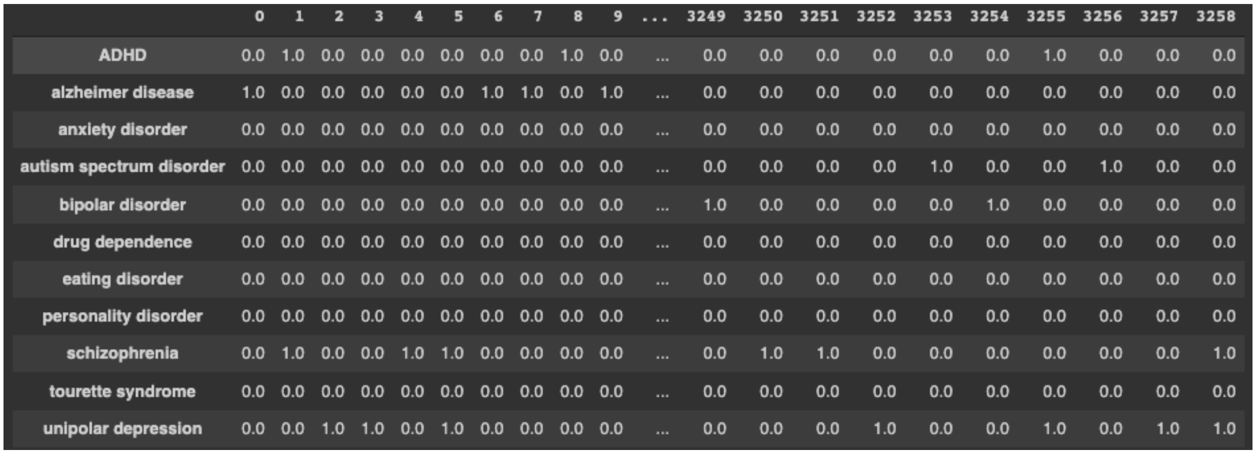


Personality disorder now also appears to consist of more common variants, but this may be spurious due to the low data size.

Encoding and Clustering

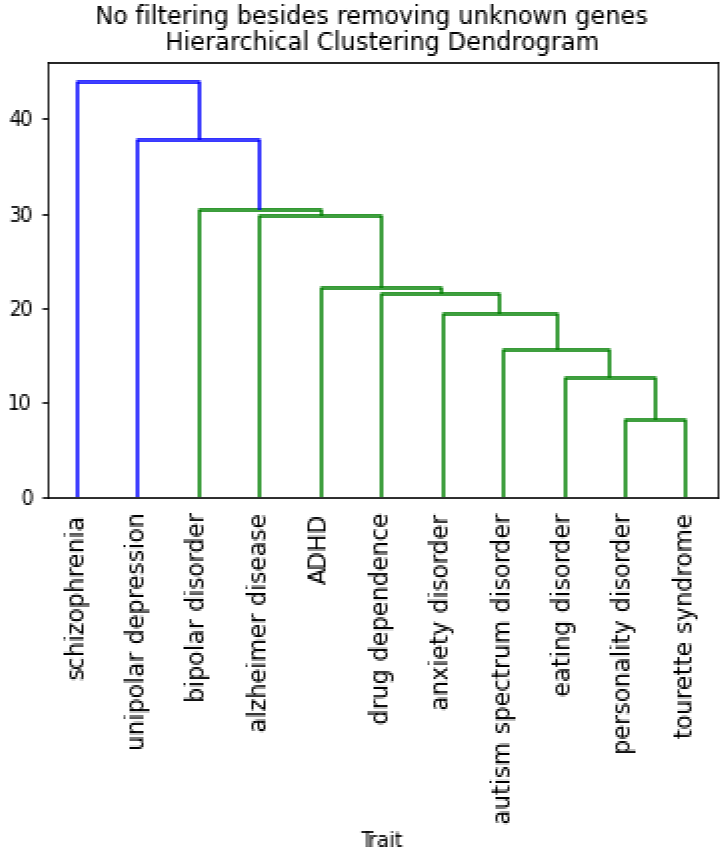
Hierarchical Clustering

My gene-based encoding described above yielded a data structure depicted by the following, where numbers represent unique genes:



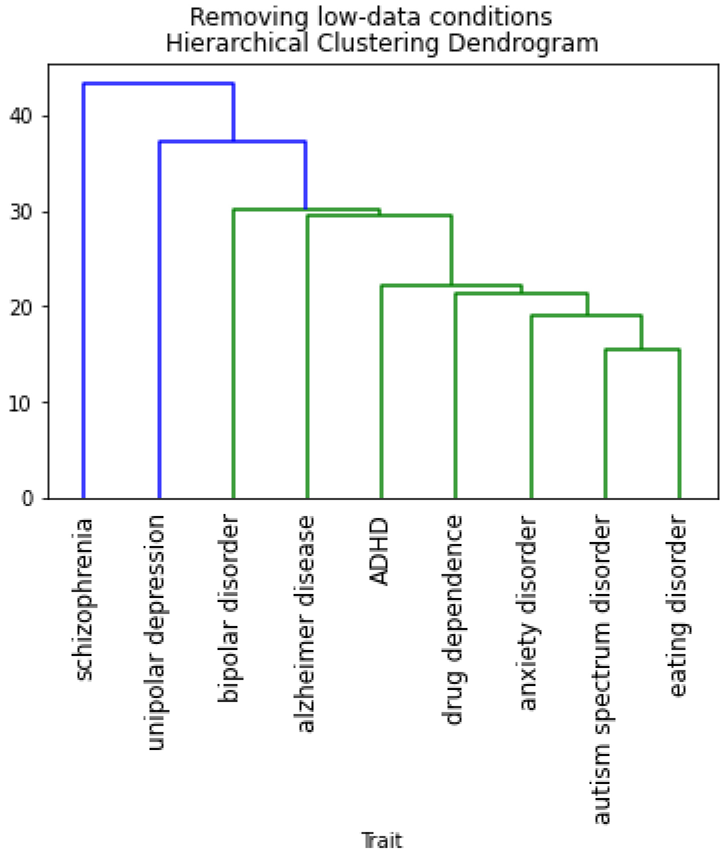
	0	1	2	3	4	5	6	7	8	9	...	3249	3250	3251	3252	3253	3254	3255	3256	3257	3258
ADHD	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
alzheimer disease	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
anxiety disorder	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
autism spectrum disorder	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0
bipolar disorder	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
drug dependence	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
eating disorder	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
personality disorder	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
schizophrenia	0.0	1.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	...	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
tourette syndrome	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
unipolar depression	0.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	1.0	1.0

Clustering these vectors with ward hierarchical clustering yielded an odd looking dendrogram:

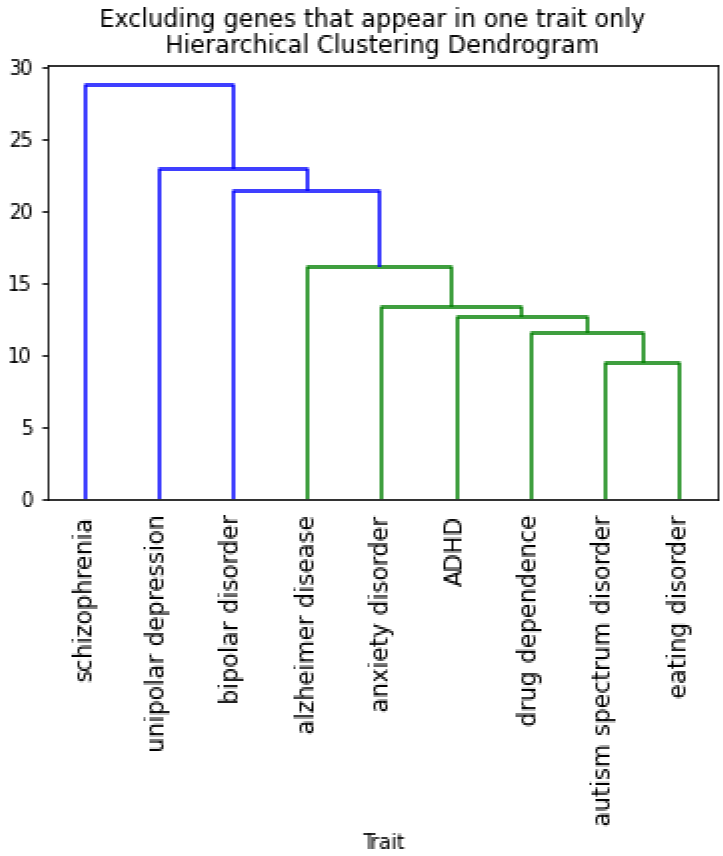


I noticed this ordering of trait similarity reflects data size per trait, and realized it should have been expected: a vector (1, 0, 0, 0, 0, 0) is much closer to a vector (0, 1, 0, 0, 0, 0) than it is to a vector (1, 1, 1, 1, 1, 1) simply based on magnitude (the sparse vectors here represent the low-data traits, personality disorder and tourette syndrome). With this observation, I tried making some improvements to compensate for the data size discrepancies.

First, I simply decided to remove the two lowest-data traits from further analysis as I did not think they had enough data to provide meaningful outcomes (they had order of 10's data points, whereas all other traits had on the order of hundreds or thousands). This still didn't change the clustering representing trait data size:

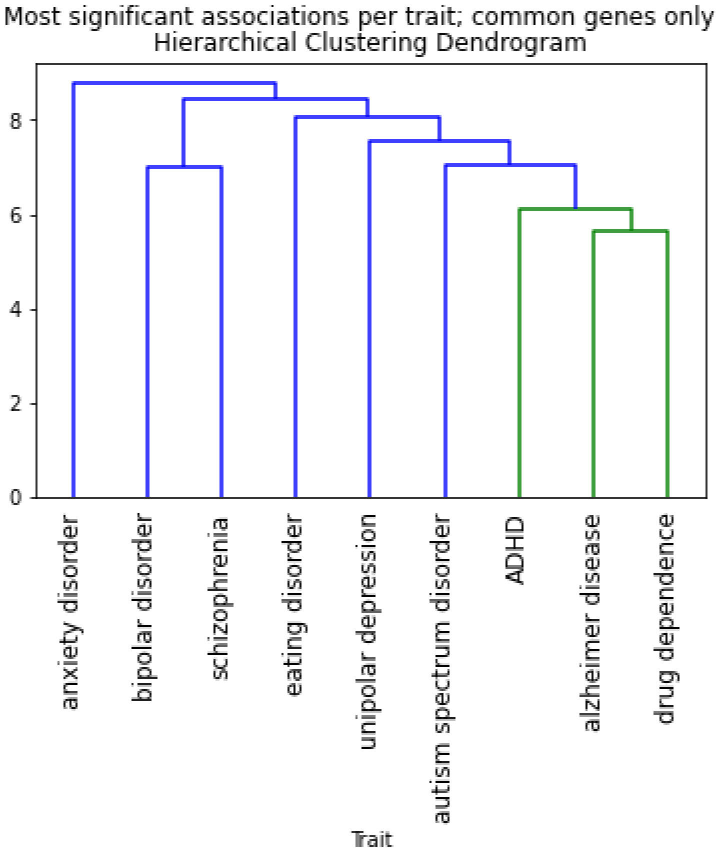


I considered whether genes associated with a single trait were useful in clustering. If each trait had roughly the same number of associated genes, I thought they may be useful in expressing differences in traits, but due to the data size differences, I decided those genes were not adding information about similarity across traits. Therefore I then excluded genes that only appear in a single trait and repeated the clustering:



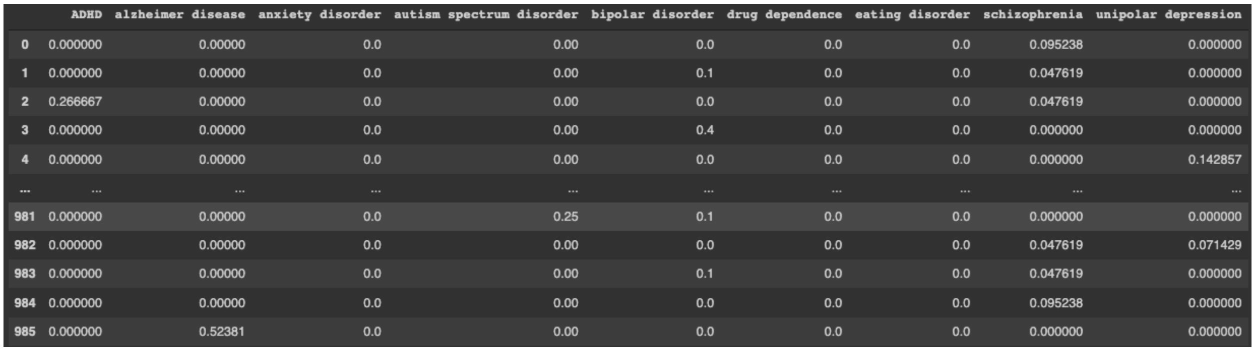
Alas the ordering of traits was different, but I thought the “telescoping” effect of sub-clusters still depicted some sort of data size bias.

Next I tried removing that bias by using the same number of variants per trait. For this I took the most significant X variants per trait, where significance is based on the reported p-value for the variant association, and X is the number of variants in the smallest trait dataset. Finally, this yielded a slightly different dedrogram:



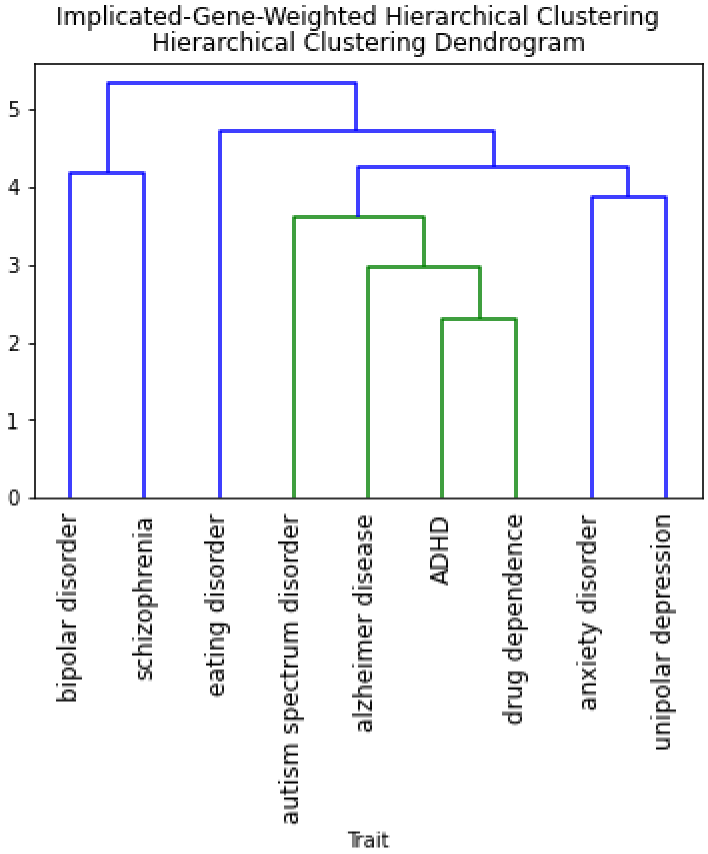
This clustering shows bipolar disorder and schizophrenia being closely related, which matches some associations reported in other literature (Smail, 2021). Additionally, it identified drug dependence and Alzheimer's as being closely related. This prompted some further literature review for me and I identified some research which examined how drug use can play a role in the cause of dementia (Hulse, 2005). This of course does not imply that the conditions may otherwise be related outside of that potential causal relationship.

As described in the methods section, I extended the gene-based encoding to count gene-associations for a trait and normalize the counts by the max count, generating vectors as depicted by the following image:



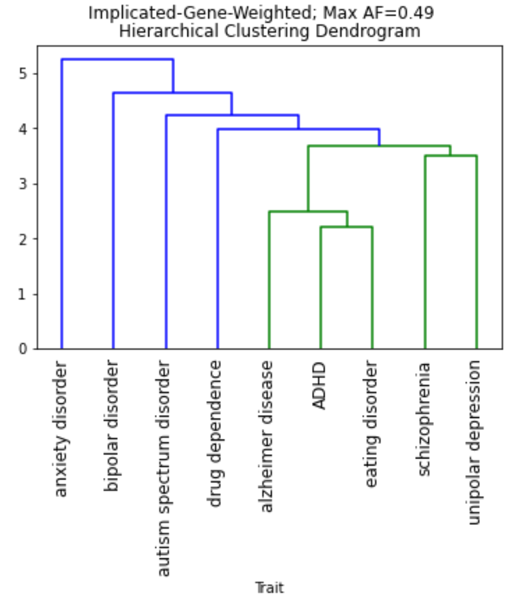
	ADHD	alzheimer disease	anxiety disorder	autism spectrum disorder	bipolar disorder	drug dependence	eating disorder	schizophrenia	unipolar depression
0	0.000000	0.00000	0.0	0.00	0.0	0.0	0.0	0.095238	0.000000
1	0.000000	0.00000	0.0	0.00	0.1	0.0	0.0	0.047619	0.000000
2	0.266667	0.00000	0.0	0.00	0.0	0.0	0.0	0.047619	0.000000
3	0.000000	0.00000	0.0	0.00	0.4	0.0	0.0	0.000000	0.000000
4	0.000000	0.00000	0.0	0.00	0.0	0.0	0.0	0.000000	0.142857
...	...	...	...	...	...	...	...	...	...
981	0.000000	0.00000	0.0	0.25	0.1	0.0	0.0	0.000000	0.000000
982	0.000000	0.00000	0.0	0.00	0.0	0.0	0.0	0.047619	0.071429
983	0.000000	0.00000	0.0	0.00	0.1	0.0	0.0	0.047619	0.000000
984	0.000000	0.00000	0.0	0.00	0.0	0.0	0.0	0.095238	0.000000
985	0.000000	0.52381	0.0	0.00	0.0	0.0	0.0	0.000000	0.000000

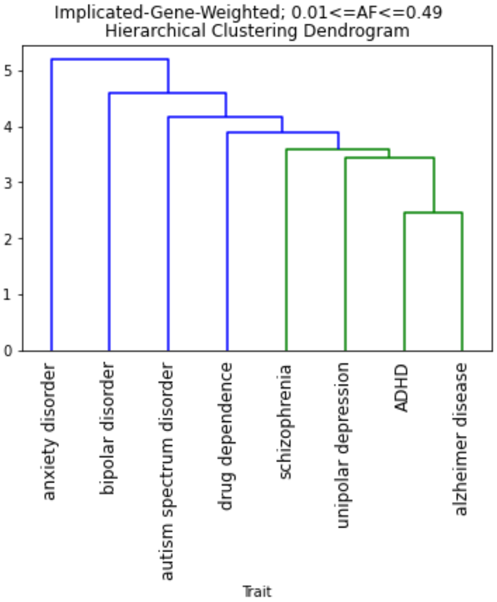
This encoding led to a slightly different hierarchical clustering outcome:



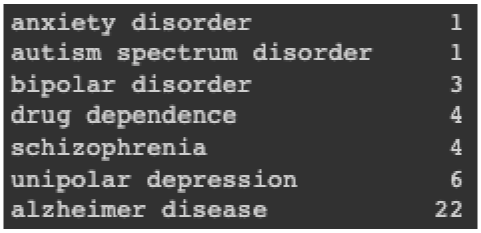
Again, bipolar disorder and schizophrenia are close together; Alzheimer's and drug dependence are close, although ADHD is closer to drug dependence. Also, a new association between anxiety and depression appeared. I had previously heard that anecdotally, these conditions tend to co-occur, but upon some further exploration I discovered that there is "tentative" support for a genetic association between the conditions (Hettema, 2008).

I further explored hierarchical clustering by first filtering to various ranges of allele frequencies. Unfortunately, it seemed each iteration led to differing results:





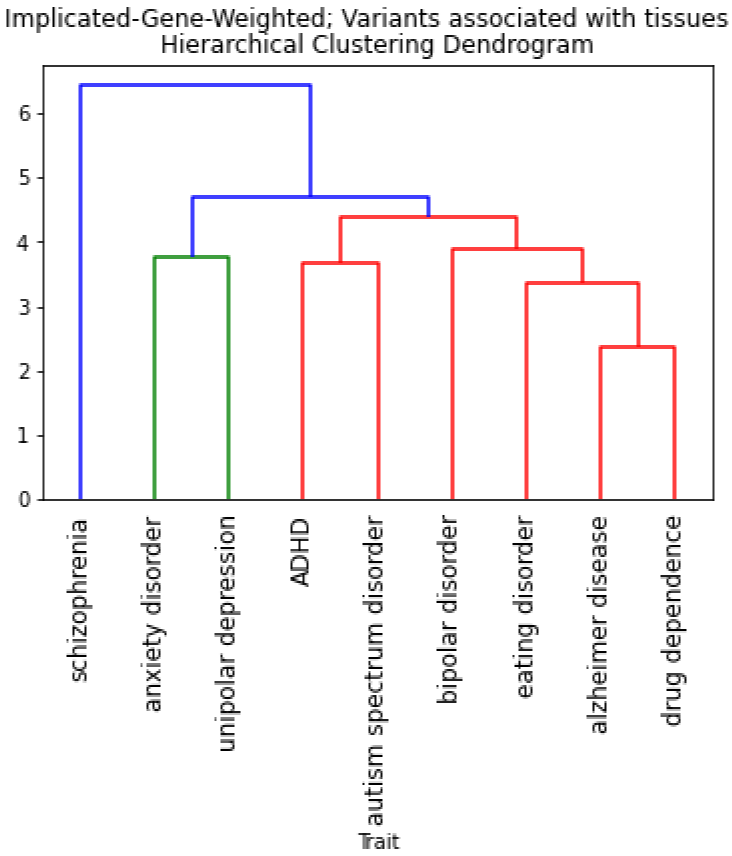
A paper I read said they excluded variants with AF < 0.01 in their GWAS meta-analysis, so I tried the same above (Eichler, 2019). That didn't actually remove many variants, although I found it interesting the Alzheimer's had the most rare variants (0.01) even though it was not among the traits with the most total variants:



```
anxiety disorder              1
autism spectrum disorder      1
bipolar disorder              3
drug dependence               4
schizophrenia                 4
unipolar depression           6
alzheimer disease            22
```

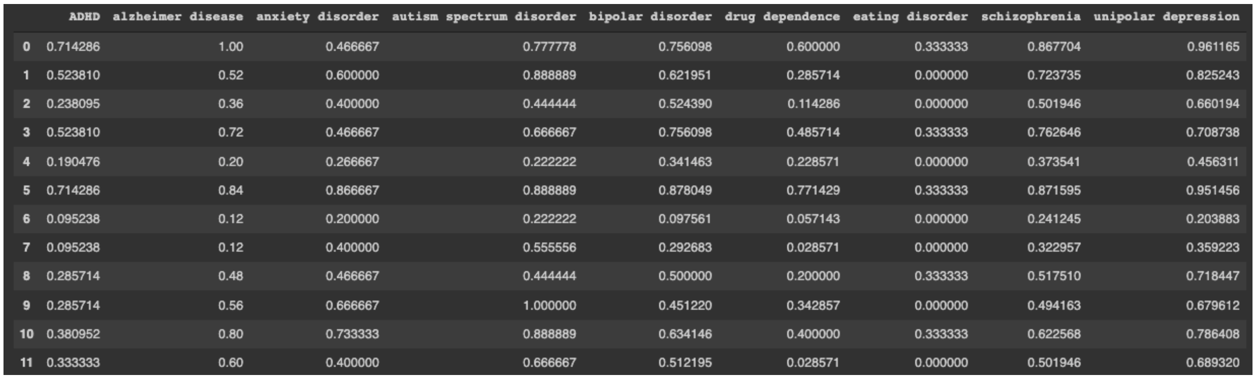
Rare (AF < 0.01) variants per trait.

My final gene-based encoding, hierarchical clustering experiment was to first filter the data to only variants with known tissue associations:



Interestingly, this still shows the Alzheimer's and drug dependence similarity, the depression and anxiety similarity, although now schizophrenia is the least similar to all other traits.

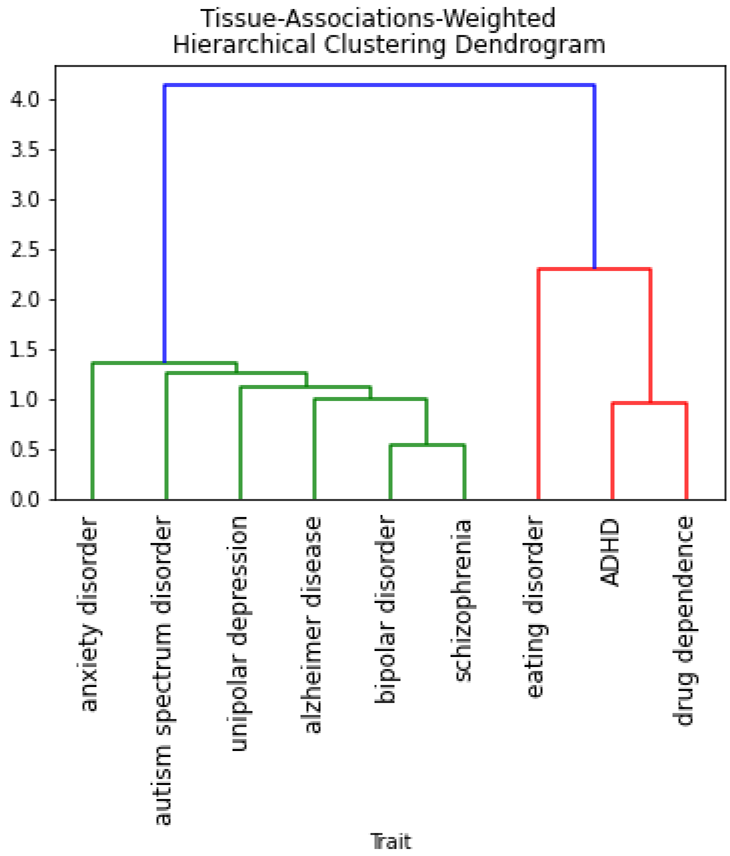
I then encoded the data based on tissue associations as described in the methods. This generated data in the following format:



	ADHD	alzheimer disease	anxiety disorder	autism spectrum disorder	bipolar disorder	drug dependence	eating disorder	schizophrenia	unipolar depression
0	0.714286	1.00	0.466667	0.777778	0.756098	0.600000	0.333333	0.867704	0.961165
1	0.523810	0.52	0.600000	0.888889	0.621951	0.285714	0.000000	0.723735	0.825243
2	0.238095	0.36	0.400000	0.444444	0.524390	0.114286	0.000000	0.501946	0.660194
3	0.523810	0.72	0.466667	0.666667	0.756098	0.485714	0.333333	0.762646	0.708738
4	0.190476	0.20	0.266667	0.222222	0.341463	0.228571	0.000000	0.373541	0.456311
5	0.714286	0.84	0.866667	0.888889	0.878049	0.771429	0.333333	0.871595	0.951456
6	0.095238	0.12	0.200000	0.222222	0.097561	0.057143	0.000000	0.241245	0.203883
7	0.095238	0.12	0.400000	0.555556	0.292683	0.028571	0.000000	0.322957	0.359223
8	0.285714	0.48	0.466667	0.444444	0.500000	0.200000	0.333333	0.517510	0.718447
9	0.285714	0.56	0.666667	1.000000	0.451220	0.342857	0.000000	0.494163	0.679612
10	0.380952	0.80	0.733333	0.888889	0.634146	0.400000	0.333333	0.622568	0.786408
11	0.333333	0.60	0.400000	0.666667	0.512195	0.028571	0.000000	0.501946	0.689320

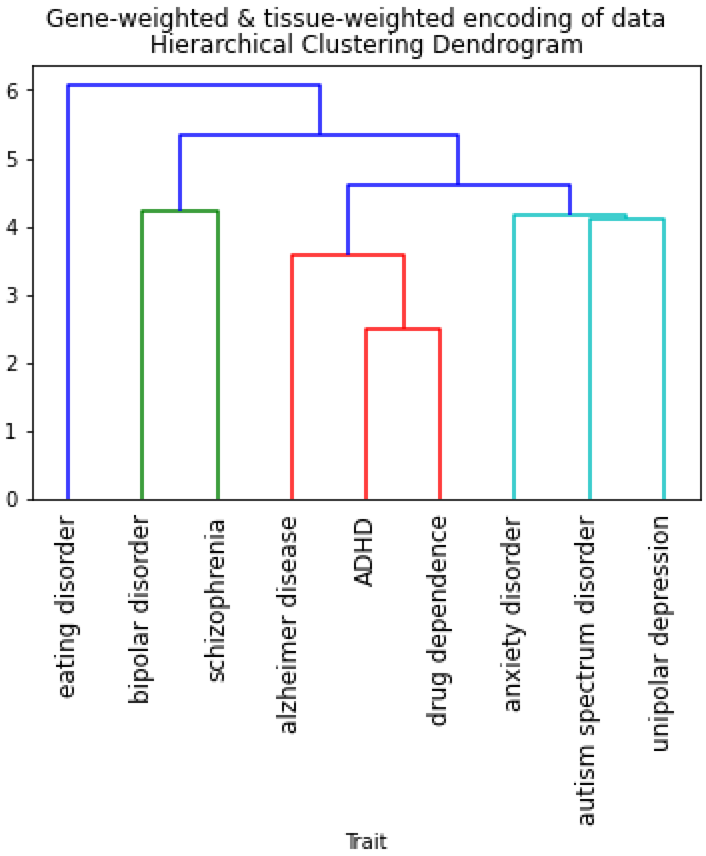
where the numbers represent distinct tissues and go up to 48 (total number of tissues).

Applying hierarchical clustering on this encoding actually generated a fairly distinct dendrogram from the others, but still retained some previous patterns:



ADHD and drug dependence appear together again, although now without Alzheimer's nearby. Schizophrenia and bipolar disorder are close together again, but are a part of a telescoping effect on the left half of the dendrogram.

My final hierarchical clustering trial was to combine both the gene-encoding vectors and tissues-encoding vectors for each trait, attempting to provide as much information as possible to the clustering algorithm. This provided the following dedrogram, which is slightly different than previous versions:

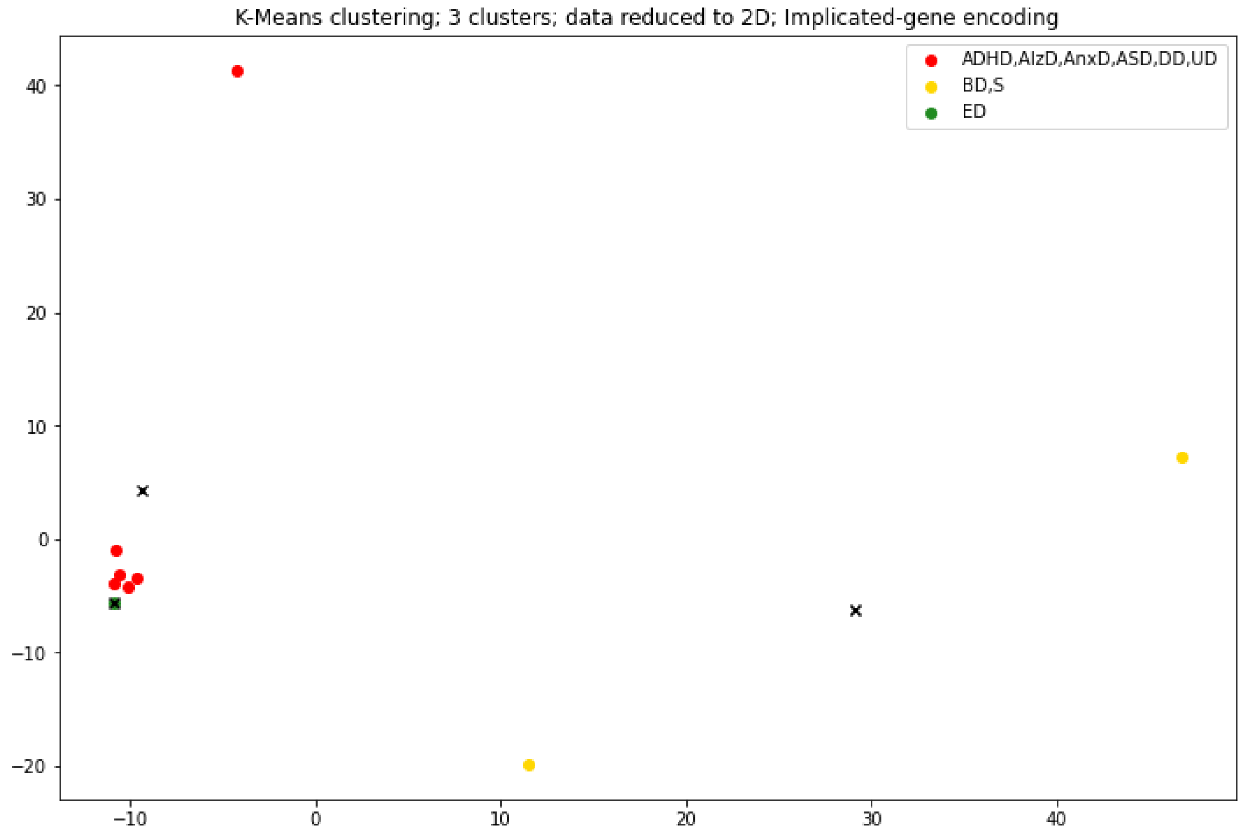


Notably, the bipolar disorder and schizophrenia associations continue to appear, anxiety and depression are still close together, as are Alzheimer's and drug dependence.

A negative result of these hierarchical clusterings is the frequently different associations that appear, as it’s not possible to confirm the encoding is the most representative and the results carry significance. Hierarchical clustering always associates all points in some way, which is one limitation of these approaches (Pagnuco, 2017).

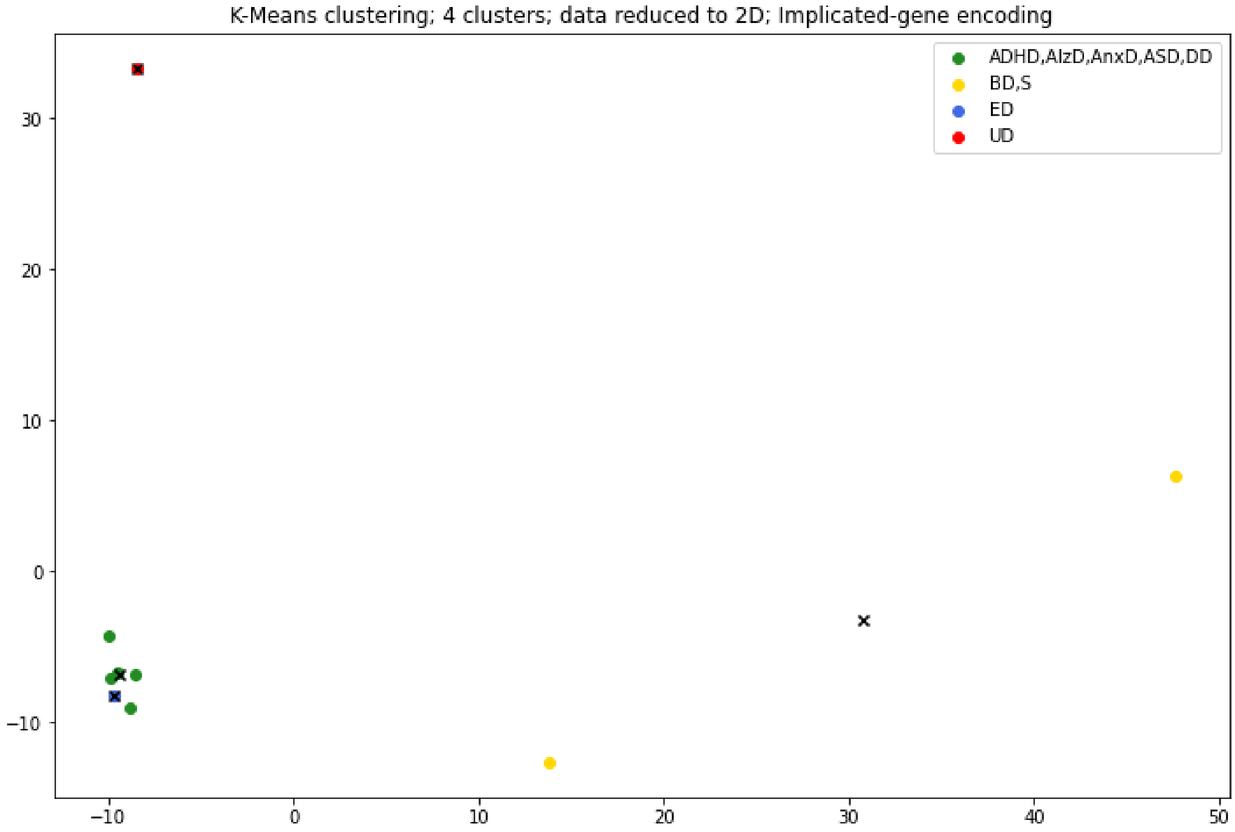
K-means Clustering

I then experimented by running the encodings generated above into a k-means clustering algorithm, trying out some different numbers of desired clusters as this is a required input to this algorithm. K-means clustering also does not provide output in the form of a dendrogram, but rather the cluster labels for the inputs and the determined cluster means (vectors representing the perceived centers of each cluster). Therefore to help visualize these results, I tried running PCA on the vectors (input vectors and output mean-vectors) to be able to plot the top two components as a 2D scatter plot. This is one such outcome with no prior data filtering:



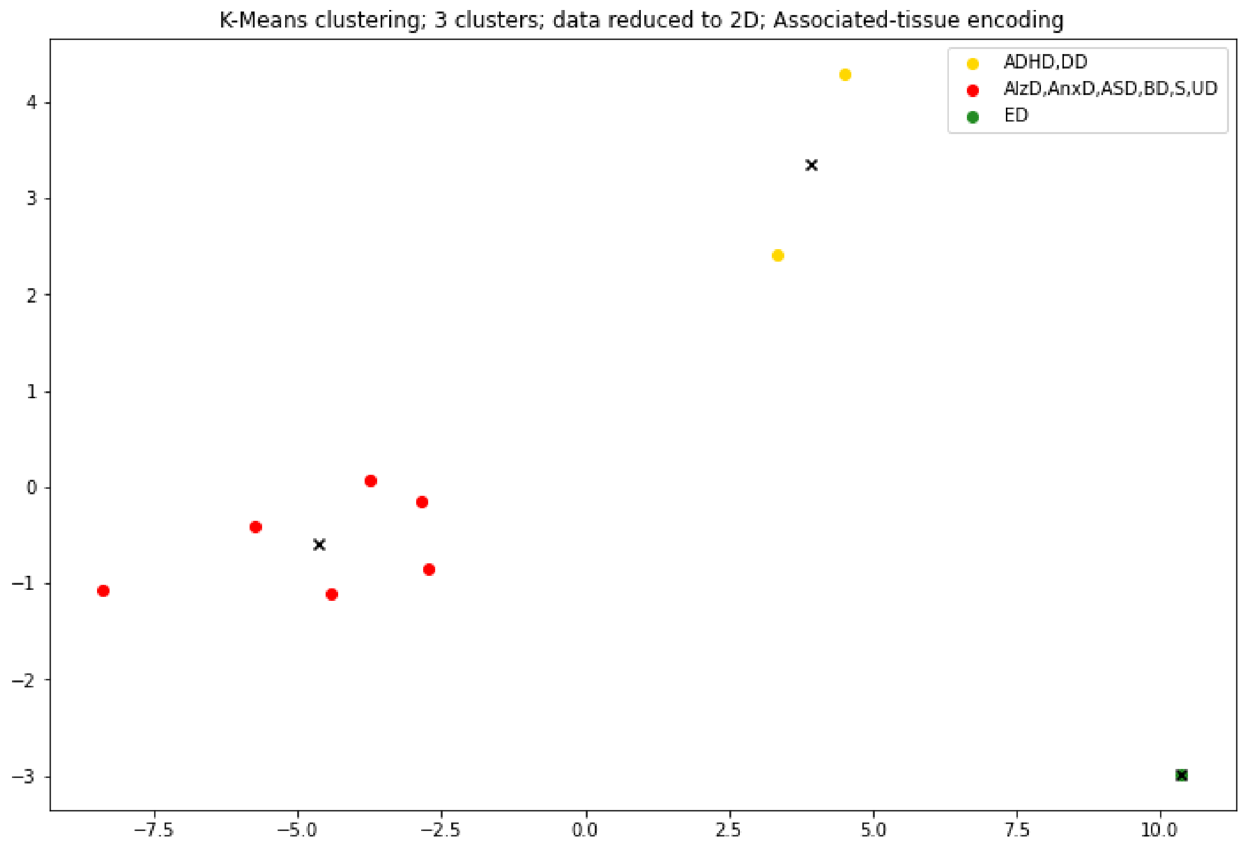
Eating disorder (ED in the legend) appears by itself, bipolar disorder (BD) and schizophrenia are in another cluster, and the rest are assigned the remaining cluster. I thought this was not strictly grouping on data size since BD has nearly half as much data as unipolar depression (UD). However, ED is the smallest trait by data size after removing personality disorder and tourette syndrome. I think it appears closer to the red cluster than it really is based on some imperfection in PCA.

Trying 4 desired clusters yielded this outcome:



Unsurprisingly, the point in the previous red group which was furthest away, UD, is now in its own cluster.

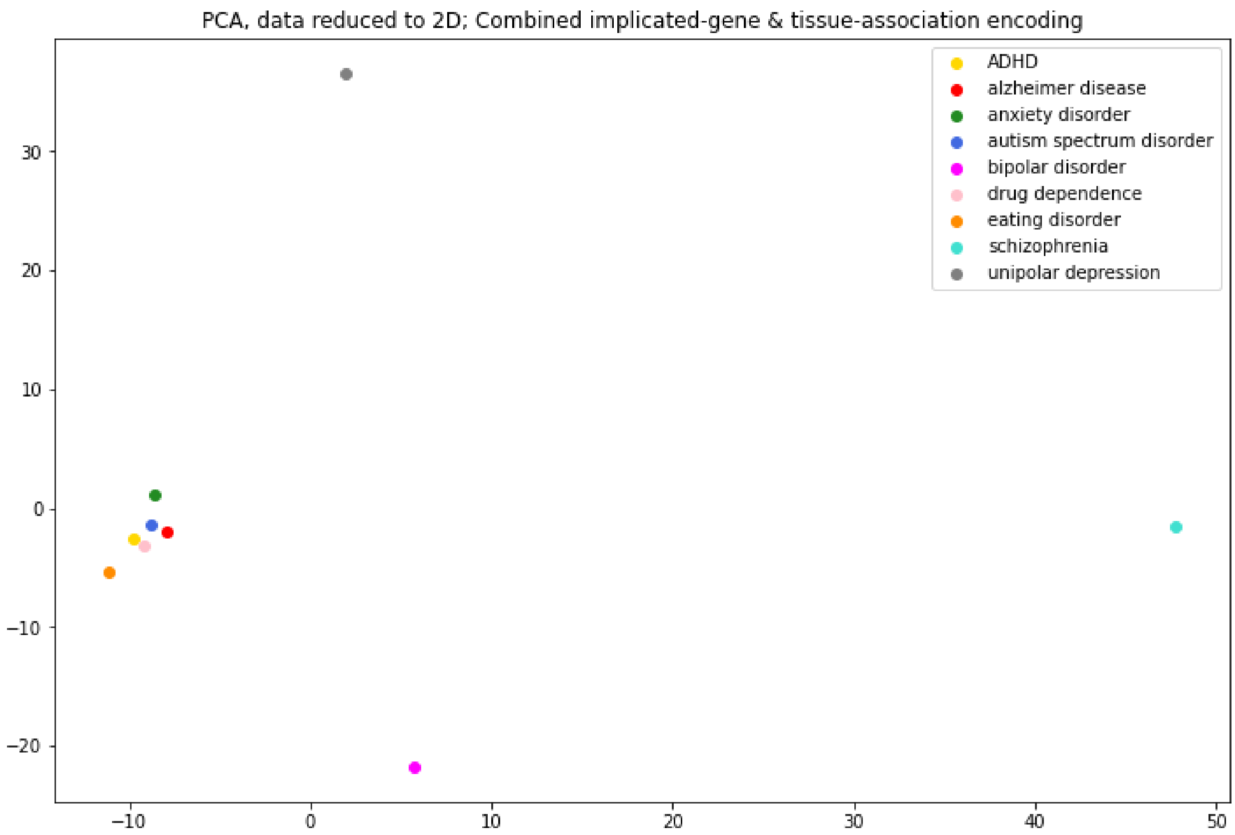
I tried some other filtering on the gene-based encoding before using k-means clustering, but I thought the most seemingly “correct” clustering based on the PCA plot was when I used the tissue-encoding:



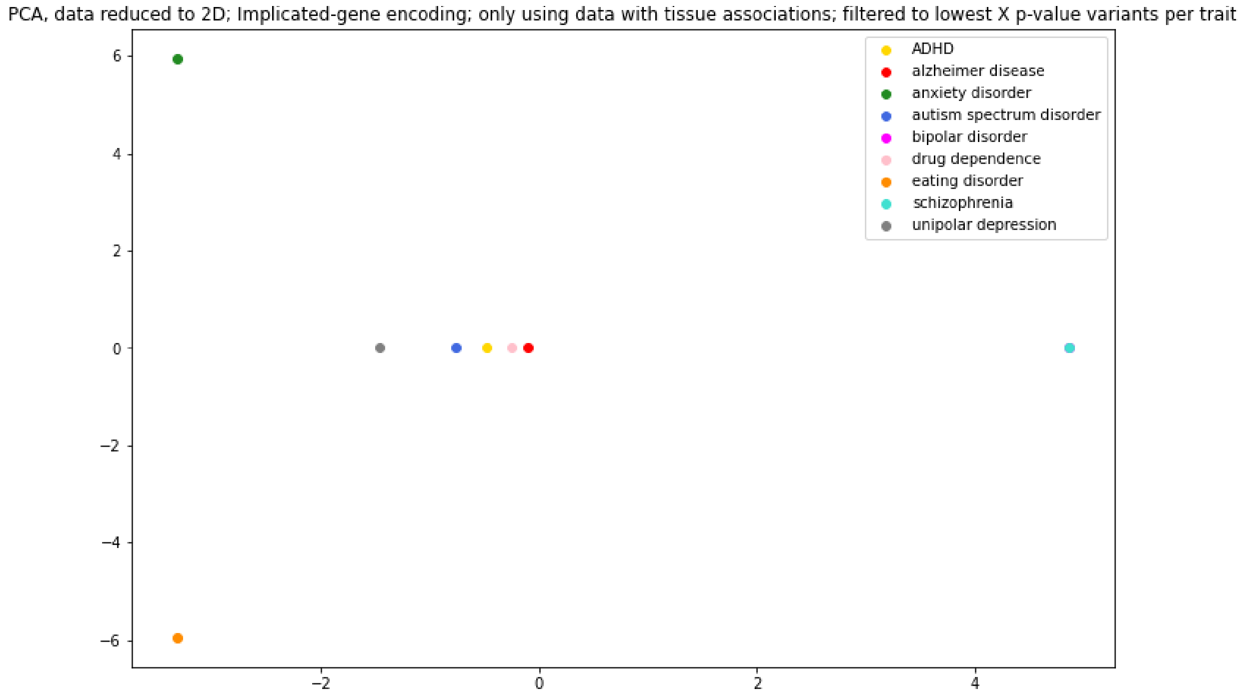
Here the clusters are visually distinct. Eating disorder is almost always by itself when employing k-means, and I struggled to determine whether this is simply due to its low data size or whether it really does have some distinctive characteristics based on implicated genes and tissues.

PCA Alone

Finally, I ran PCA directly on the combined gene and tissue encodings and just plotted the top two components to determine if PCA could reveal latent relationships between the trait data:



I felt this was still depicting data-size per trait in some way, so limited the input to the most significant variants by p-value again, to get the following, fairly strange looking plot:



It’s very peculiar that most points fall on a straight line, but we do see Alzheimer's and drug dependence appear to be closest together of any two traits, reinforcing previous results with other methods.

Limitations

As noted throughout, I felt that a significant limitation of this comparison across conditions was the very different sizes of data available for each condition. I attempted to account for it in a variety of ways, but it was difficult to verify what was the correct way to handle it as my experiments did not have a “right” answer. I attempted to do some initial exploration of the available data before committing to this project. I had not realized at the time of the limitations that would require me to filter out a sizable amount of the raw data (for example the non-standard reported traits). In hindsight, perhaps choosing a more well-studied class of conditions, such as cancers, would have provided more ample data for this analysis. I had originally hoped to test out my methods on other data in this way, but underestimated the initial data organization and cleaning costs.

Another limitation was that clustering algorithm output can not be interpreted as fact. If I had more time, I would have explored the verification techniques such as the Dunn index, discussed by Pagnuco (2017). Likely I should have narrowed the scope of my project to focus on more specific comparisons (for example between just 2 conditions) or to propose well-defined statistical tests so that a null hypothesis could be rejected or accepted at the end. In general, a limitation with analyses like these is that they can only identify areas for future study; they do not by themselves identify clinically actionable insights.

Conclusions

Some specific outcomes I obtained from my analysis are that clustering techniques can reveal genetic relationships suggested via other methods - for example the schizophrenia and bipolar associations. Additionally, my analysis showed that Alzheimer’s disease and drug dependence may have a genetic connection, as may anxiety disorder and unipolar depression. However, I also realized that analyses like mine can be easily influenced by confounding factors such as data size, and therefore my results need to be viewed skeptically. Similarly, I learned firsthand how in these high-dimensional bioinformatics studies, it’s easy to tweak the parameters and inputs to iterate until you achieve results that you want to see. This was apparent with the many different clustering outcomes based on different encodings and data filtering. Conducting research in this way allows for biases to creep into the experiment. Going forward I should specify more constrained and specific analyses before beginning and spend more time accounting for the potential skew caused by imperfect data, perhaps through study design or more extensive literature review.

References

Bailey, R., Sharpe, D., Kwiatkowski, T., Watson, S., Dexter Samuels, A. ., & Hall, J. (2018). Mental Health Care Disparities Now and in the Future. *Journal of Racial and Ethnic Health Disparities*, *5*(2), 351–356. https://doi.org/10.1007/s40615-017-0377-6

Bennett, S., & Thomas, A. J. (2014). Depression and dementia: Cause, consequence or coincidence? *Maturitas*, *79*(2), 184–190. https://doi.org/10.1016/j.maturitas.2014.05.009

Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E, Suveges D, Vrousgou O, Whetzel PL, Amode R, Guillen JA, Riat HS, Trevanion SJ, Hall P, Junkins H, Flicek P, Burdett T, Hindorff LA, Cunningham F and Parkinson H. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Research, 2019, Vol. 47 (Database issue): D1005-D1012.

Carr, M. J., Steeg, S., Webb, R. T., Kapur, N., Chew-Graham, C. A., Abel, K. M., Hope, H., Pierce, M., & Ashcroft, D. M. (2021). Effects of the COVID-19 pandemic on primary care-recorded mental illness and self-harm episodes in the UK: a population-based cohort study. *The Lancet Public Health*, *6*(2), e124–e135. https://doi.org/10.1016/S2468-2667(20)30288-7

CDC. (n.d.). *Attitudes Toward Mental Illness: Results from the Behavioral Risk Factor Surveillance System*.

Cross Disorder Phenotype Group of the Psychiatric GWAS Consortium. (2009). Dissecting the phenotype in genome-wide association studies of psychiatric illness. *British Journal of Psychiatry*, *195*(2), 97–99. https://doi.org/10.1192/bjp.bp.108.063156

Du, J., Jia, P., Dai, Y., Tao, C., Zhao, Z., & Zhi, D. (2019). Gene2vec: distributed representation of genes based on co-expression. *BMC Genomics*, *20*(1), 82. https://doi.org/10.1186/s12864-018-5370-x

Eichler, E. E. (2019). Genetic Variation, Comparative Genomics, and the Diagnosis of Disease. *New England Journal of Medicine*, *381*(1), 64–74. https://doi.org/10.1056/NEJMra1809315

Escudero, I., & Johnstone, M. (2014). Genetics of Schizophrenia. *Current Psychiatry Reports*, *16*(11), 502. https://doi.org/10.1007/s11920-014-0502-8

Falola, O., Osamor, V. C., Adebiyi, M., & Adebiyi, E. (2017). Analyzing a single nucleotide polymorphism in schizophrenia: a meta-analysis approach. *Neuropsychiatric Disease and Treatment*, *13*, 2243–2250. https://doi.org/10.2147/NDT.S111900

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from the GTEx Portal on 04/21/22

Han, E., Carbonetto, P., Curtis, R. E., Wang, Y., Granka, J. M., Byrnes, J., Noto, K., Kermany, A. R., Myres, N. M., Barber, M. J., Rand, K. A., Song, S., Roman, T., Battat, E., Elyashiv, E., Guturu, H., Hong, E. L., Chahine, K. G., & Ball, C. A. (2017). Clustering of 770,000 genomes reveals post-colonial population structure of North America. *Nature Communications*, *8*(1), 14238. https://doi.org/10.1038/ncomms14238

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hert, M. D., Detraux, J., & Vancampfort, D. (2018). The intriguing relationship between coronary heart disease and mental disorders. *Dialogues in Clinical Neuroscience*, *20*(1), 31–40. https://doi.org/10.31887/DCNS.2018.20.1/mdehert

Hettema, J. M. (2008). What is the genetic relationship between anxiety and depression? *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, *148C*(2), 140–146. https://doi.org/10.1002/ajmg.c.30171

Hulse, G. K., Lautenschlager, N. T., Tait, R. J., & Almeida, O. P. (2005). Dementia associated with alcohol and other drug use. *International Psychogeriatrics*, *17*(s1), S109–S127. https://doi.org/10.1017/S1041610205001985

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Kayahan, B., Oztürk, O., & Veznedaroğlu, B. (2005). [Obsessive-compulsive symptoms in schizophrenia]. *Turk Psikiyatri Dergisi = Turkish Journal of Psychiatry*, *16*(3), 205–215.

Landi, M. T., Bishop, D. T., MacGregor, S., Machiela, M. J., Stratigos, A. J., Ghiorzo, P., Brossard, M., Calista, D., Choi, J., Fargnoli, M. C., Zhang, T., Rodolfo, M., Trower, A. J., Menin, C., Martinez, J., Hadjisavvas, A., Song, L., Stefanaki, I., Scolyer, R., … Law, M. H. (2020). Genome-wide association meta-analyses combining multiple risk phenotypes provide insights into the genetic architecture of cutaneous melanoma susceptibility. *Nature Genetics*, *52*(5), 494–504. https://doi.org/10.1038/s41588-020-0611-8

Marchini, J., Cardon, L. R., Phillips, M. S., & Donnelly, P. (2004). The effects of human population structure on large genetic association studies. *Nature Genetics*, *36*(5), 512–517. https://doi.org/10.1038/ng1337

Mayo Foundation for Medical Education and Research. (n.d.). *COVID-19: How to manage your mental health during the pandemic*. Mayo Clinic. Retrieved February 28, 2022, from https://www.mayoclinic.org/diseases-conditions/coronavirus/in-depth/mental-health-covid-19/art-20482731

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., Nelson, M. R., Stephens, M., & Bustamante, C. D. (2008). Genes mirror geography within Europe. *Nature*, *456*(7218), 98–101. https://doi.org/10.1038/nature07331

Pagnuco, I. A., Pastore, J. I., Abras, G., Brun, M., & Ballarin, V. L. (2017). Analysis of genetic association using hierarchical clustering and cluster validation indices. *Genomics*, *109*(5–6), 438–445. https://doi.org/10.1016/j.ygeno.2017.06.009

Pasaniuc, B., & Price, A. L. (2017). Dissecting the genetics of complex traits using summary association statistics. *Nature Reviews Genetics*, *18*(2), 117–127. https://doi.org/10.1038/nrg.2016.142

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html

Popp, J. (2013). Delirium and cognitive decline: more than a coincidence. *Current Opinion in Neurology*, *26*(6), 634–639. https://doi.org/10.1097/WCO.0000000000000030

Psychiatric Genomics Consortium. (n.d.). Psychiatric Genomics Consortium. Retrieved February 28, 2022, from https://www.med.unc.edu/pgc/

Rappoport, N., & Shamir, R. (2018). Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Research*, *46*(20), 10546–10562. https://doi.org/10.1093/nar/gky889

Reback, J., Jbrockmendel, McKinney, W., Van Den Bossche, J., Augspurger, T., Roeschke, M., Hawkins, S., Cloud, P., Gfyoung, Sinhrks, Hoefler, P., Klein, A., Terji Petersen, Tratner, J., She, C., Ayd, W., Naveh, S., JHM Darbyshire, Garcia, M., … Battiston, P. (2022). *pandas-dev/pandas: Pandas 1.4.2* (v1.4.2) [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.3509134

Réthelyi, J., Pulay, A., Balogh, L., & Nemoda, Z. (2019). [New directions in psychiatric genetics: Genome-wide association studies, polygenic risk score and cross-disorder analysis].

Psychiatria Hungarica: A Magyar Pszichiatriai Tarsasag Tudomanyos Folyoirata, *34*(4), 411–418.

Santoro, M. L., Moretti, P. N., Pellegrino, R., Gadelha, A., Abílio, V. C., Hayashi, M. A. F., Belangero, S. I., & Hakonarson, H. (2016). A current snapshot of common genomic variants contribution in psychiatric disorders. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *171*(8), 997–1005. https://doi.org/10.1002/ajmg.b.32475

Sherry, S. T., Ward, M. and Sirotkin, K. (1999) dbSNP—Database for Single Nucleotide Polymorphisms and Other Classes of Minor Genetic Variation. Genome Res., 9, 677–679.

Smail, M. A., Wu, X., Henkel, N. D., Eby, H. M., Herman, J. P., McCullumsmith, R. E., & Shukla, R. (2021). Similarities and dissimilarities between psychiatric cluster disorders. *Molecular Psychiatry*, *26*(9), 4853–4863. https://doi.org/10.1038/s41380-021-01030-3

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., & Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, *20*(8), 467–484. https://doi.org/10.1038/s41576-019-0127-1

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., & Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, *1*(1), 1–21. https://doi.org/10.1038/s43586-021-00056-9

Vilhjálmsson, B. J., Yang, J., Finucane, H. K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P.-R., Bhatia, G., Do, R., Hayeck, T., Won, H.-H., Kathiresan, S., Pato, M., Pato, C., Tamimi, R., Stahl, E., Zaitlen, N., Pasaniuc, B., … Zheng, W. (2015). Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *The American Journal of Human Genetics*, *97*(4), 576–592. https://doi.org/10.1016/j.ajhg.2015.09.001

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … Vázquez-Baeza, Y.

(2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Waskom, M. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Wu, P., Wang, B., Lubitz, S. A., Benjamin, E. J., Meigs, J. B., & Dupuis, J. (2021). Approximate conditional phenotype analysis based on genome wide association summary statistics. *Scientific Reports*, *11*(1), 2518. https://doi.org/10.1038/s41598-021-82000-1

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., Montgomery, G. W., Goddard, M. E., Wray, N. R., Visscher, P. M., & Yang, J. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics*, *48*(5), 481–487. https://doi.org/10.1038/ng.3538