GWAS_catalog_EDA_and_normalization

April 16, 2022

```
[1]: import pandas as pd
     import numpy as np
     import math
```

1 Exploratory analysis of CSV data retrieved from GWAS Catalog.

This specific data is for Schizophrenia, but other data in GWAS catalog can be retrieved in the same format.

1.1 Exploring raw data types, range of data, etc.

1.1.1 Overall

```
[2]: raw_schizo_df = pd.read_csv('schizophrenia_gwas_catalog_2022.csv')
     raw_schizo_df.head()
```

```
[2]:   Variant and risk allele    P-value P-value annotation   RAF     OR Beta  \
     0     rs11265461-<b>C</b>   2 x 10-7                NaN  0.41   1.45   '-
     1       rs230529-<b>T</b>   2 x 10-7                NaN  0.47   1.45   '-
     2      rs2237457-<b>T</b>   6 x 10-7  (Recessive model)  0.36   1.74   '-
     3      rs2269372-<b>A</b>   4 x 10-8                NaN    NR  1.313   '-
     4      rs7597593-<b>T</b>  9 x 10-11                NaN    NR  1.066   '-

                   CI    Mapped gene                       Reported trait  \
     0  [1.26-1.67]  SLAMF1, SETP9  Schizophrenia (treatment resistant)
     1  [1.26-1.66]          NFKB1  Schizophrenia (treatment resistant)
     2         [NR]          GRB10  Schizophrenia (treatment resistant)
     3         [NR]          RENBP                        Schizophrenia
     4  [1.05-1.09]        ZNF804A                        Schizophrenia

                                  Trait(s) Background trait(s) Study accession  \
     0  treatment refractory schizophrenia                  '-      GCST001458
     1  treatment refractory schizophrenia                  '-      GCST001458
     2  treatment refractory schizophrenia                  '-      GCST002604
```

```
3                       schizophrenia                  '-      GCST002190
4                       schizophrenia                  '-      GCST004946

      Location
0  1:160660353
1  4:102536261
2   7:50658447
3  X:153942092
4  2:184668853
```

[3]:

```python
total_rows = len(raw_schizo_df)
print(total_rows)
```

```
3849
```

1.1.2 Variants

[4]:

```python
num_unique_variants = len(raw_schizo_df['Variant and risk allele'].unique())
print(f"{num_unique_variants} unique variants out of {total_rows} records.")

variant_counts = raw_schizo_df['Variant and risk allele'].value_counts()
```

```
2739 unique variants out of 3849 records.
```

[5]:

```python
# Explore entries for one repeated variant to assess differences.
duplicates = raw_schizo_df.groupby('Variant and risk allele').filter(lambda x: len(x) > 1)
one_variant = duplicates.iloc[0]['Variant and risk allele']
duplicates[duplicates['Variant and risk allele'] == one_variant]
```

[5]:

```
     Variant and risk allele    P-value          P-value annotation   RAF  \
4         rs7597593-<b>T</b>  9 x 10-11                         NaN    NR
747       rs7597593-<b>T</b>  2 x 10-11                         NaN    NR
2878      rs7597593-<b>T</b>  3 x 10-12                         NaN  0.62
3625      rs7597593-<b>T</b>   8 x 10-6  (5 degree of freedom test)    NR

         OR Beta           CI Mapped gene  \
4     1.066   '-  [1.05-1.09]     ZNF804A
747   1.069   '-  [1.05-1.09]     ZNF804A
2878     '-   '-           '-     ZNF804A
3625  1.055   '-  [1.03-1.08]     ZNF804A

                                  Reported trait  \
4                                  Schizophrenia
747                                Schizophrenia
2878            Broad depression or schizophrenia
```

```
3625  Autism spectrum disorder, attention deficit-hy…

                                               Trait(s) Background trait(s)  \
4                                         schizophrenia                  '-
747                                       schizophrenia                  '-
2878                 unipolar depression, schizophrenia                  '-
3625  attention deficit hyperactivity disorder, unip…                  '-

     Study accession     Location
4         GCST004946  2:184668853
747       GCST007201  2:184668853
2878      GCST007257  2:184668853
3625      GCST001877  2:184668853
```

1.1.3 P-values

[6]:
```
raw_schizo_df['P-value'].describe()
```

[6]:
```
count         3849
unique         163
top       2 x 10-8
freq           201
Name: P-value, dtype: object
```

[7]:
```
len(raw_schizo_df['Mapped gene'].unique())
```

[7]:
```
1427
```

1.1.4 Genes

[8]:
```
def has_multiple_genes(mapped_gene):
  return "," in mapped_gene


multi_gene_index = raw_schizo_df['Mapped gene'].apply(has_multiple_genes)
len(raw_schizo_df[multi_gene_index])
```

[8]:
```
913
```

1.1.5 Reported trait / Trait(s)

[9]:
```
raw_schizo_df['Reported trait'].unique()
```

```
[9]: array(['Schizophrenia (treatment resistant)', 'Schizophrenia',
            'Schizophrenia (MTAG)', 'Schizophrenia or bipolar disorder',
            'Schizophrenia (negative symptoms)', 'Methamphetamine dependence',
            'Early-onset schizophrenia',
            'Autism spectrum disorder or schizophrenia',
            'Gray matter volume (schizophrenia interaction)',
            'Schizophrenia (inflammation and infection response interaction)',
            'Broad depression or schizophrenia',
            'Dentate gyrus volume x schizophrenia interaction',
            'Schizophrenia vs type 2 diabetes',
            'Schizophrenia and type 2 diabetes',
            'Autism and schizophrenia (MTAG)',
            'Left superior temporal gyrus thickness (schizophrenia interaction)',
            'Bipolar disorder and schizophrenia',
            'Schizophrenia (cytomegalovirus infection interaction)',
            'Schizophrenia (age at onset)',
            'Schizophrenia or schizoaffective disorder',
            'Schizophrenia vs autism spectrum disorder (ordinary least squares
     (OLS))',
            'Schizophrenia vs bipolar disorder (ordinary least squares (OLS))',
            'Schizophrenia vs anorexia nervosa (ordinary least squares (OLS))',
            'Schizophrenia vs ADHD (ordinary least squares (OLS))',
            'Schizophrenia vs major depressive disorder (ordinary least squares
     (OLS))',
            "Schizophrenia vs Tourette's syndrome and other tic disorders (ordinary
     least squares (OLS))",
            'Schizophrenia x sex interaction',
            'Bipolar disorder lithium response (continuous) or schizophrenia',
            'Bipolar disorder lithium response (categorical) or schizophrenia',
            'Cognitive ability, years of educational attainment or schizophrenia
     (pleiotropy)',
            'Brain imaging in schizophrenia (dorsolateral prefrontal cortex
     interaction)',
            'Schizophrenia, schizoaffective disorder or bipolar disorder',
            'Schizophrenia, bipolar disorder or recurrent major depressive disorder x
     sex interaction (3df)',
            'Schizophrenia, bipolar disorder or recurrent major depressive disorder',
            'Schizophrenia, bipolar disorder or major depressive disorder x sex
     interaction',
            'Schizophrenia, bipolar disorder or major depressive disorder',
            'Schizophrenia, bipolar disorder or major depressive disorder x sex
     interaction (3df)',
            'Neuropsychiatric disorders',
            'Autism spectrum disorder, attention deficit-hyperactivity disorder,
     bipolar disorder, major depressive disorder, and schizophrenia (combined)',
            'Psychiatric diseases (pleiotropy) (HIPO component 1)',
            'Schizophrenia, bipolar disorder or recurrent major depressive disorder x
     sex interaction',
            'Anorexia nervosa, attention-deficit/hyperactivity disorder, autism
     spectrum disorder, bipolar disorder, major depression, obsessive-compulsive
     disorder, schizophrenia, or Tourette syndrome (pleiotropy)'],
           dtype=object)
```

```
[10]: raw_schizo_df['Trait(s)'].unique()
```

```
[10]: array(['treatment refractory schizophrenia', 'schizophrenia',
       'autism spectrum disorder, schizophrenia',
       'schizophrenia, grey matter volume measurement',
       'schizophrenia, cytomegalovirus seropositivity',
       'schizophrenia, HSV1 seropositivity',
       'schizophrenia, Toxoplasma gondii seropositivity',
       'unipolar depression, schizophrenia',
       'dentate gyrus volume measurement, schizophrenia',
       'schizophrenia, type 2 diabetes mellitus',
       'schizophrenia, bipolar disorder',
       'schizophrenia, left superior temporal gyrus thickness measurement',
       'schizophrenia, cytomegalovirus infection',
       'schizophrenia, age at onset',
       'schizophrenia, schizoaffective disorder',
       'anorexia nervosa, schizophrenia',
       'attention deficit hyperactivity disorder, schizophrenia',
       'Tourette syndrome, schizophrenia',
       'schizophrenia, sex interaction measurement',
       'schizophrenia, bipolar disorder, response to lithium ion',
       'schizophrenia, intelligence, self reported educational attainment',
       'schizophrenia, dorsolateral prefrontal cortex functional measurement,
brain measurement',
       'schizophrenia, bipolar disorder, schizoaffective disorder',
       'unipolar depression, schizophrenia, sex interaction measurement, bipolar
disorder',
       'disease recurrence, unipolar depression, schizophrenia, bipolar
disorder',
       'unipolar depression, schizophrenia, bipolar disorder',
       'attention deficit hyperactivity disorder, autism spectrum disorder,
schizophrenia, bipolar disorder, major depressive disorder',
       'attention deficit hyperactivity disorder, unipolar depression, autism
spectrum disorder, schizophrenia, bipolar disorder',
       'disease recurrence, unipolar depression, schizophrenia, sex interaction
measurement, bipolar disorder',
       'anorexia nervosa, obsessive-compulsive disorder, attention deficit
hyperactivity disorder, Tourette syndrome, unipolar depression, autism spectrum
disorder, schizophrenia, bipolar disorder'],
      dtype=object)
```

[11]:

```python
num_just_schizo = len(raw_schizo_df[raw_schizo_df['Trait(s)'] == 'schizophrenia'])
print(f"{num_just_schizo} / {total_rows} rows are for the trait schizophrenia only.")
```

```
2564 / 3849 rows are for the trait schizophrenia only.
```

1.1.6 Initial observations:

- 3849 records total
- P-values are currently objects/strings
- A lot of genes - 1427 unique values, although some normalization seems to be required (e.g. to fix ”SLAMF1, SETP9”). After normalizing it may be good to analyze counts per gene - maybe genes only implicated once are less signficant than others which appear multiple times.
- Many records have multiple traits in addition to schizophrenia (e.g. one trait value is ”anorexia nervosa, obsessive-compulsive disorder, attention deficit hyperactivity disorder, Tourette syndrome, unipolar depression, autism spectrum disorder, schizophrenia, bipolar disorder”). I assume these studies examined patients with either condition, but it’s not entirely clear without checking the studies themselves. To make this a scalable approach, it may be best to omit records that are for more than just schizophrenia to avoid any potential biases in the future similarity analysis.
- A fair amount of the variants in the dataset appear multiple times (e.g. reported by different studies). It’s worth noting this, although at the moment it’s unclear what the best way to handle this is. Maybe subsequent analysis should only focus on variants identified multiple time; maybe for each repeated variant, only the lowest p-value should be retained. However, some care should be applied given the above point about traits (maybe want the lowest p-value among records for just the trait schizophrenia).

1.2 Cleaning/normalizing data

[12]:

```python
# Create copy of DF to hold normalized data and leave raw DF untouched.
schizo_df = raw_schizo_df.copy()
```

1.2.1 P-values

[13]:

```python
def pval_to_num(pval):
  parts = pval.split(" x 10-")
  return float(parts[0]) * pow(10, -float(parts[1]))


print(pval_to_num("2 x 10-7"))
```

```
2e-07
```

[14]:

```
schizo_df['P-value_norm'] = raw_schizo_df['P-value'].apply(pval_to_num)
```

[15]:

```
schizo_df['P-value_norm'].describe()
```

[15]:

```
count    3.849000e+03
mean     1.072234e-06
std      2.255918e-06
min      2.000000e-44
25%      3.000000e-10
50%      2.000000e-08
75%      6.000000e-07
max      1.000000e-05
Name: P-value_norm, dtype: float64
```

1.2.2 Traits

[16]:

```
# As mentioned above, it may be best to use the subset of data which focused
# solely on the trait of interest (schizophrenia).
# There are some others that are probably fine to include (e.g. treatment
# refractory schizophrenia), but for the sake of simplicity and
# generalizability, we'll assume there is one canonical GWAS catalog trait of
# interest for each condition to be analyzed.
canonical_trait = 'schizophrenia'
filtered_df = schizo_df[schizo_df['Trait(s)'] == canonical_trait]
print(f"Filtered from {len(schizo_df)} rows to {len(filtered_df)} rows.")
```

```
Filtered from 3849 rows to 2564 rows.
```

[17]:

```
# The majority of the data is retained, so we'll use just this subset.
schizo_df = filtered_df
```

1.2.3 Variants

[18]:

```
# Sanity-check that all duplicated variants are reported to map to same gene(s)
# before we split multi-gene associations into separate rows.
# If all repeated variants map to same gene, we can just retain the entry with
# lowest p-value (or any really, since subsequent analysis just cares about
# variant ID and implicated genes).
duplicate_variants = schizo_df.groupby('Variant and risk allele').filter(lambda␣
 ↪x: len(x) > 1)['Variant and risk allele'].unique()
all_good = True
for variant in duplicate_variants:
  all_mapped_genes = schizo_df[schizo_df['Variant and risk allele'] ==␣
 ↪variant]['Mapped gene'].unique()
  if len(all_mapped_genes) > 1:
    print(f"Found variant, {variant}, with differing mapped gene values.")
    all_good = False


if all_good:
  print("No repeated variants with differing mapped gene values.")
```

```
No repeated variants with differing mapped gene values.
```

[19]:

```
# Proceed with just choosing the record with the lowest p-value.
# It may later be useful to revisit this step and retain these duplicates -
# maybe only focusing on those associations that have been found in multiple
# independent studies will lead to better results in the subsequent analysis.
min_indices = schizo_df.groupby('Variant and risk allele')['P-value_norm'].
 ↪idxmin()
schizo_df = schizo_df.loc[min_indices]
```

[20]:

```
# Sanity-check duplicates are gone:
num_unique_variants = len(schizo_df['Variant and risk allele'].unique())
num_total = len(schizo_df)
print(f"{num_unique_variants} unique variants of {num_total} records")
```

```
1822 unique variants of 1822 records
```

1.2.4 Genes

[21]:

```
# Genes are comma-separated so `explode` can be used to create a new row for
# each gene (with all other columns identical).
# https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.explode.html
schizo_df['gene_norm'] = raw_schizo_df['Mapped gene'].apply(lambda val: val.
 ↪split(", "))
exploded_schizo_df = schizo_df.explode('gene_norm')
len(exploded_schizo_df)
```

[21]:

```
2201
```

[22]:

```
# Sanity check that the final number of rows is expected:
schizo_df['gene_norm'].apply(lambda x: len(x)).value_counts()
```

[22]:

```
1    1444
2     377
3       1
Name: gene_norm, dtype: int64
```

[23]:

```
# 1444 entries with one gene + 2 * 377 entries with two + 3 * 1 entries with␣
 ↪three
```

```python
assert len(exploded_schizo_df) == 1444 + 2 * 377 + 3 * 1
```

[24]:

```python
# Sanity-check passes so set schizo_df to the exploded version.
schizo_df = exploded_schizo_df
```

[25]:

```python
schizo_df['gene_norm'].value_counts()
```

[25]:

```
'-            251
LINC01470      21
CACNA1C        15
Y_RNA          15
VRK2           11
             ...
ARHGAP31        1
ADAMTS6         1
NLRC5           1
VN1R18P         1
NRIP1           1
Name: gene_norm, Length: 1118, dtype: int64
```

[26]:

```python
# 491 / 4764 entries have "'-" for their gene; I'm assuming this indicates an
# unknown/unconfirmed gene association.
UNKNOWN_GENE = "UNKNOWN"

def replace_unknown_gene(gene):
  return UNKNOWN_GENE if gene == "'-" else gene


schizo_df['gene_norm'] = schizo_df['gene_norm'].apply(replace_unknown_gene)
schizo_df['gene_norm'].value_counts()
```

[26]:

```
UNKNOWN       251
LINC01470      21
CACNA1C        15
Y_RNA          15
VRK2           11
             ...
ARHGAP31        1
ADAMTS6         1
NLRC5           1
VN1R18P         1
NRIP1           1
Name: gene_norm, Length: 1118, dtype: int64
```

1.3 Output

Finally, write out the normalized version of the data for use in further analysis.

[27]:
```
schizo_df.head()
```

[27]:
```
     Variant and risk allele   P-value P-value annotation     RAF         OR  \
2388  chr6:55564517-<b>?</b>  3 x 10-6           (female)  0.5665         '-
1176      rs1001780-<b>G</b>  8 x 10-6                NaN      NR  1.0752687
2036     rs10043984-<b>?</b>  5 x 10-8                NaN      '-         '-
615      rs10043984-<b>T</b>  4 x 10-8                NaN  0.2614         '-
236      rs10046758-<b>?</b>  9 x 10-8                NaN      NR         '-

                           Beta            CI Mapped gene  \
2388       0.1622 unit increase  [0.094-0.23]          '-
1176                         '-          [NR]     DLX2-DT
2036                         '-            '-       KDM3B
615   0.067151085 unit increase [0.043-0.091]       KDM3B
236                          '-            '-       CSMD1

            Reported trait       Trait(s) Background trait(s) Study accession  \
2388         Schizophrenia  schizophrenia                  '-      GCST012309
1176         Schizophrenia  schizophrenia                  '-      GCST003048
2036  Schizophrenia (MTAG)  schizophrenia                  '-      GCST010640
615   Schizophrenia (MTAG)  schizophrenia                  '-      GCST012089
236          Schizophrenia  schizophrenia                  '-      GCST008459

                   Location  P-value_norm gene_norm
2388  Mapping not available  3.000000e-06   UNKNOWN
1176            2:172107630  8.000000e-06   DLX2-DT
2036            5:138376432  5.000000e-08     KDM3B
615             5:138376432  4.000000e-08     KDM3B
236               8:4326648  9.000000e-08     CSMD1
```

[28]:
```
# Keep only the relevant, normalized columns for brevity. This can always be
# updated later to retain more if there's a use for it.

out_df = schizo_df[['Variant and risk allele', 'P-value_norm', 'Trait(s)',␣
 ↪'gene_norm']]
column_remapping = {
    'Variant and risk allele': 'variant_and_allele',
    'P-value_norm': 'p_value',
    'Trait(s)': 'trait',
    'gene_norm': 'gene',
}
out_df = out_df.rename(columns=column_remapping)
out_df.head()
```

```
[28]:            variant_and_allele       p_value          trait     gene
      2388  chr6:55564517-<b>?</b>  3.000000e-06  schizophrenia  UNKNOWN
      1176      rs1001780-<b>G</b>  8.000000e-06  schizophrenia  DLX2-DT
      2036     rs10043984-<b>?</b>  5.000000e-08  schizophrenia    KDM3B
      615      rs10043984-<b>T</b>  4.000000e-08  schizophrenia    KDM3B
      236      rs10046758-<b>?</b>  9.000000e-08  schizophrenia    CSMD1
```

```
[29]: out_df.to_csv('schizophrenia_gwas_catalog_2022_cleaned.csv')
```