Comparative Analysis of SNP Genomic Data in Patients with Anxiety Disorder and Major Depressive Disorder

2023-03-05

Ellie Kim (rk27863)

SDS 322E Elements of Data Science

I. Introduction

In this project, I will analyze two genomic meta datasets from two separate scientific papers. Both datasets were acquired from the Psychiatric Genomics Consortium (PGC) database. The data on patients with anxiety disorder (AD) came from the paper titled “Meta-analysis of genome-wide association studies of anxiety disorders” by Otowa, T et al., published in ‘Molecular Psychiatry’ in 2016. The data on patients with major depressive disorder (MDD) came from the paper titled “Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression” by the authors Wray, Naomi et al. This paper was published in Nature Genetics in 2018.

```
# load necessary packages
library(readr)
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✔ ggplot2 3.4.0     ✔ dplyr   1.1.0
## ✔ tibble  3.1.8     ✔ stringr 1.5.0
## ✔ tidyr   1.3.0     ✔ forcats 1.0.0
## ✔ purrr   1.0.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
```

```
# save datasets into variables

# anxiety disorder dataset
anx <- read_tsv('dataset/anxiety.meta.full.cc.tbl')
```

```
## Rows: 6330995 Columns: 10
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: "\t"
## chr (3): SNPID, Allele1, Allele2
## dbl (7): CHR, BP, Freq1, Effect, StdErr, P.value, TotalN
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# major depressive disorder dataset
mdd <- read_tsv('dataset/daner_pgc_mdd_meta_w2_no23andMe_rmUKBB')
```

```
## Rows: 9874289 Columns: 19
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: "\t"
## chr  (4): SNP, A1, A2, Direction
## dbl (15): CHR, BP, FRQ_A_45396, FRQ_U_97250, INFO, OR, SE, P, ngt, HetISqt, ...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

a. Dataset Description

Both dataset focuses on Single Nucleotide Polymorphism (SNP), which is a type of genomic mutation where one nucleotide in the DNA differs from other individuals in the population. In the study, the authors obtained the information of genomic SNP via comparing genome sequencing data from patient group and control group. I have a particular interest in this topic because I'm interested in the study of mental disorders, as well as genomic data analysis.

Each unique row in the datasets represent a single case of SNP. The columns display information about the SNP ID (categorical), chromosomal location (categorical), base pair (numeric), reference allele (categorical), alternative allele (categorical), allele frequency (numeric), number of subjects who display the SNP (numeric), and much more. I will be able to join the two datasets by using the SNP ID, which is a universal identifier assigned to each SNP.

b. Research Question

The potential trend I'm interested in investigating is how the patterns of SNP in the genome differs between those with anxiety disorder (AD) and major depressive disorder (MDD). Along with the differences, I expect some similarity between the two, due to the overlapping biological nature of the disorders. My research question for this project is: Are there any significant difference in SNP profiles between patients with AD or MDD, and can we identify any potential genetic markers associated with these disorders?

II. Joining/Merging

```
# total number of SNP in each dataset before joining
nrow(anx)
```

```
## [1] 6330995
```

```
nrow(mdd)
```

```
## [1] 9874289
```

```
# SNP in anx that is not in mdd
nrow(anti_join(anx, mdd, by = c("SNPID" = "SNP")))
```

```
## [1] 112304
```

```
# SNP in mdd that is not in anx
nrow(anti_join(mdd, anx, by = c("SNP" = "SNPID")))
```

```
## [1] 3655598
```

```
# SNP in common
nrow(inner_join(anx, mdd, by = c("SNPID" = "SNP")))
```

```
## [1] 6218691
```

```
# join and keep all rows in both dataset
anx_mdd <- full_join(anx, mdd, by = c("SNPID" = "SNP"),
                     suffix = c(".anx",".mdd"))

# number of rows in final dataset
nrow(anx_mdd)
```

```
## [1] 9986593
```

In this section, I join the SNP genomic dataset from AD and MDD based on SNP ID. Before joining, there are 6,330,995 SNP observed for anxiety disorder (AD) and 9,874,289 SNP observed for major depressive disorder (MDD). 112,304 SNP were characterized in AD but not in MDD, while 3,655,598 SNP were in MDD but not in AD. 6,218,691 SNP were commonly found in both disorders. It was interesting to find out that patients of AD and MDD shared so many similar SNP with each other.

I decided to join the two datasets while keeping all rows from both of them. The number of rows in the final data was 9,986,593. No observation was dropped, as all of the rows in both datasets were kept. A potential issue to keep in mind is distinguishing the SNP that were found in AD versus MDD when dealing with the merged dataset.

III. Tidying

```
# create a new tidy dataset
anx_mdd_tidy <- anx_mdd %>%
  # pivot the two columns 'TotalN' and 'Nca'
  pivot_longer(cols = c('TotalN', 'Nca'),
               # make it so that the new 'disorder' column holds the names
               # this is because 'TotalN' and 'Nca' are number of patient cases in AD
and MDD respectively
               names_to = 'disorder',
               # store the values to a new column called 'cases'
               # this is the number of patients (AD or MDD) who display this SNP
               values_to = 'cases')
```

The merged dataset has columns named ‘TotalN’ and ‘Nca.’ The variable they each display are both ‘number of patients who display this SNP.’ However ‘TotalN’ column is from the AD dataset, while ‘Nca’ column is from the MDD dataset. As a result, technically, this column can be separated into two different variables: ‘disorder’ (whether this SNP is found in AD or MDD) and ‘cases’ (number of patients). I used ‘pivot_longer’ function to reshape the dataset so that each variable has its own column. Later in the ‘Wrangling’ portion, I will rename the new ‘disorder’ variable so that ‘TotalN’ is replaced with ‘AD’ and ‘Nca’ is replaced with ‘MDD,’ displaying disorder names.

In this new tidied datset, each row is now a single SNP observation in either AD or MDD. It is no longer a ‘unique’ SNP, but it is more tidy since every observation and every variable has their own row and column. Each row of SNP is now specific to a disorder, whose information is stored in the column called ‘disorder.’

IV. Wrangling

In order to manipulate the data to better suit the purposes of my research, I selected the necessary columns, filtered the data, and created new columns.

```
# create new variable to store the final dataset
anx_mdd_final <- anx_mdd_tidy %>%
  # recode the 'disorder' variable to actually store the name of the disorders
  mutate(disorder = recode(disorder, 'TotalN' = 'anx', 'Nca' = 'mdd')) %>%
  # create a new variable 'freq' and grab value from 'Freq1' if the disorder is 'an
x', and 'FRQ_A_45396' if the disorder is 'mdd'
  mutate(freq = ifelse(disorder == 'anx', Freq1, FRQ_A_45396)) %>% # freq is the alle
le frequency of the reference allele (allele frequency: proportion of individuals in
the population who have the allele)
  # select only the necessary columns and rename them
  select(SNP = SNPID,    # unique SNP ID
         chr = CHR.anx,  # chromosomal location of each SNP
         bp = BP.anx,    # base pair (bp) location of each SNP
         ref_allele = Allele1,   # reference allele (A, T, C, or G - standard allele
for this SNP locus)
         alt_allele = Allele2,   # alternative allele (A, T, C, or G - allele that re
placed the standard allele)
         freq = freq,  # allele frequency of the reference allele
         disorder,     # type of disorder (anx or mdd)
         cases) %>%    # number of (anx or mdd) cases
  # filter rows whose 'cases' variable is NA
  filter(!is.na(cases)) %>%     # NA means that there were no cases of SNP detected f
or that disorder (AD or MDD) so it makes sense to filter these rows out
  # create a new variable of 'allele_var' (allelic variation) that is a concatenated
combination of two other variables, 'ref_allele' and 'alt_allele'
  mutate(allele_var = paste(ref_allele, ">" ,alt_allele))
```

This is the structure of the final dataset I created from joining, tidying, and wrangling the initial datasets. For each SNP, I have the information on which chromosome it is in (chr), its exact location in base pairs (bp), the reference/alternative allele (ref_allele, alt_allele), allele frequency (freq), which disorder it was found in (disorder), how many patients of that disorder displayed this SNP (cases), and the allele substitution that happened in the SNP (allele_var).

```
str(anx_mdd_final)
```

```
## tibble [16,205,284 × 9] (S3: tbl_df/tbl/data.frame)
##  $ SNP       : chr [1:16205284] "rs1000033" "rs1000033" "rs1000050" "rs1000050"
...
##  $ chr       : num [1:16205284] 1 1 1 1 1 1 1 1 1 1 ...
##  $ bp        : num [1:16205284] 2.27e+08 2.27e+08 1.63e+08 1.63e+08 2.22e+08 ...
##  $ ref_allele: chr [1:16205284] "t" "t" "t" "t" ...
##  $ alt_allele: chr [1:16205284] "g" "g" "c" "c" ...
##  $ freq      : num [1:16205284] 0.827 0.832 0.855 0.856 0.336 ...
##  $ disorder  : chr [1:16205284] "anx" "mdd" "anx" "mdd" ...
##  $ cases     : num [1:16205284] 17310 45396 10864 45396 17310 ...
##  $ allele_var: chr [1:16205284] "t > g" "t > g" "t > c" "t > c" ...
```

a. Summary Statistics - Cases

Next, I computed the summary statistic of the ‘cases’ variable, which is a numeric variable that shows how many patients with the disorder displayed the SNP in their genotype. To distinguish between SNP that were found in AD and MDD, I grouped the data by disorder prior to calculating the summary statistic. The distributions of the numeric variable for both AD and MDD disorder were skewed, so I computed its median and IQR.

```
# summary statistic of numeric variable - patient cases
# median and IQR
# both distribution left skewed

anx_mdd_final %>%
  group_by(disorder) %>%  # group by disorder
  # display summary statistic of median and IQR
  summarize(median = median(cases),  # removing NA is not necessary, since it was alr
eady done in the 'wrangling' section
            IQR = IQR(cases))  # interquartile range
```

```
## # A tibble: 2 × 3
##   disorder median   IQR
##   <chr>     <dbl> <dbl>
## 1 anx       17310  4491
## 2 mdd       45396   951
```

For the SNP that were found in anxiety disorder (AD), the median cases of patients that displayed the SNP were 17,310 cases. The interquartile range was 4,491 cases. For the SNP that were found in major depressive disorder (MDD), the median cases of patients that displayed the SNP were 45,396 cases. The interquartile range was 951 cases. It is likely that the main reason why the median cases of patients were higher for SNP in MDD compared to SNP in AD is because the study on MDD was done with a much bigger sample size compared to the study on AD. Still, notably, the spread for SNP in MDD, as displayed with IQR, was much smaller (951 cases) than the spread for SNP in AD (4,491 cases).

b. Summary Statistics - Allele Frequency

In addition, I looked into the ‘freq’ variable, which is the allele frequency of the reference allele in each SNP. To distinguish between SNP that were found in AD and MDD, I grouped the data by disorder prior to calculating the summary statistic. The distribution was symmetrical, so I computed the mean and standard deviation.

```
# summary statistic of numeric variable - allele frequencies
anx_mdd_final %>%
  group_by(disorder) %>%  # group by disorder
  summarize(mean = mean(freq),  # removing NA not necessary
            sd = sd(freq))  # standard deviation
```

```
## # A tibble: 2 × 3
##   disorder  mean    sd
##   <chr>    <dbl> <dbl>
## 1 anx      0.473 0.304
## 2 mdd      0.457 0.366
```

The result is that the average allele frequency of SNP mutations in AD disorder is 0.4734, while those in MDD disorder is 0.4567. The standard deviation is 0.3037 and 0.3661 respectively. Reference allele frequency is a significant factor when conducting an association study that aims to characterize a link between genetic variants and a disorder. This is because SNP with inherently low reference allele frequency have higher likelihood to have occurred by chance, instead of actually being linked to the disorder. In general, allele frequency of 5% or higher is considered significant, since it means that the allele is not inherently rare in the population. The fact that the average allele frequency for SNP in both AD and MDD are around 40% to 50% is an indicator that the SNP studied were generally not from a rare reference allele. However, the relatively high standard deviation is a potential factor to keep in mind.

c. Summary Statistics - Allelic Variation

Lastly, I computed the summary statistic for the ‘allele_var’ variable. ‘allele_var’ (allelic variation) is a categorical variable that shows the reference allele and the alternative allele for each SNP. In this context, ‘allele’ refers to the specific nucleotide (A, T, C, or G) in that particular location. The variable ‘allele_var’ is organized so that ‘a > g,’ for instance, means that the standard allele for that location is A, but the SNP caused the allele to become G.

```
# summary statistic of categorical variable - allelic variation
anx_mdd_final %>%
  # filter out rows whose reference/alternative alleles were not specified
  filter(!str_detect(allele_var, "NA > NA")) %>%
  # find the distinct combinations of SNP IDs and allelic variations
  # this is because SNP IDs could be counted multiple times if they are present in bo
th AD and MDD
  distinct(SNP, allele_var) %>%
  # count the number of SNP in each allelic variation type
  count(allele_var) %>%
  # arrange by descending order
  arrange(desc(n)) %>%
  # display as data frame
  as.data.frame
```

```
##   allele_var       n
## 1      a > g 2151071
## 2      t > c 2148617
## 3      c > g  547329
## 4      a > c  522402
## 5      t > g  520956
## 6      a > t  440620
```

The result showed that ‘a > g’ was the most common allelic variation among the SNP, followed by ‘t > c,’ and ‘c > g.’ The top two most common allelic variations, ‘a > g’ and ‘t > c,’ had far higher occurences compared to the rest.

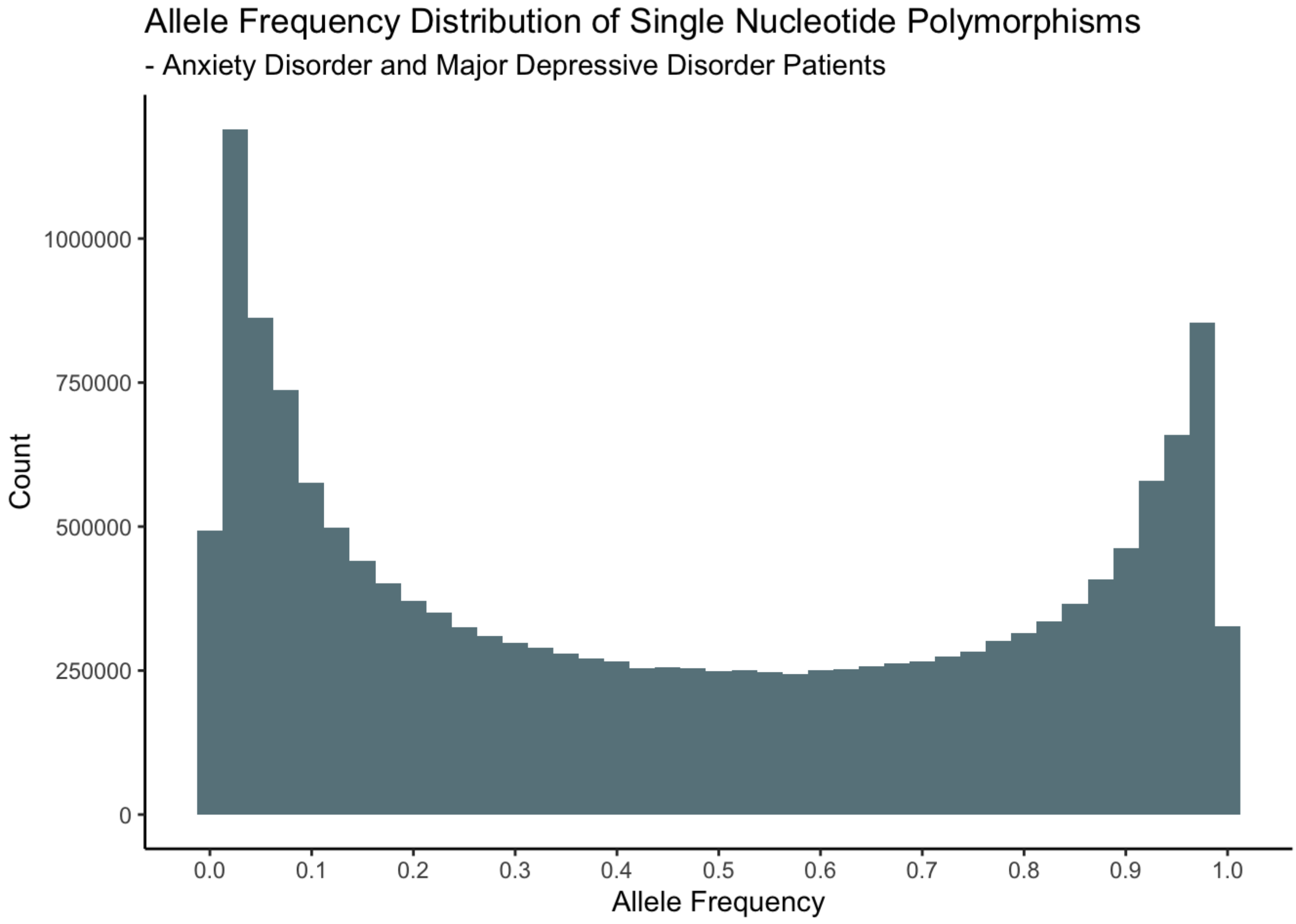
One thing to note when looking at this result is that not all 12 combinations of reference alleles and alternative alleles are shown in the data (12 because 4 (A, T, C, G) x 3 (A, T, C, G minus previous)). This is because of the complementary nature of the DNA, which is that A base pairs with T, and C base pairs with G. The information only displays half (12 / 2 = 6) of the combinations of allelic variation because the other half would just be the complementary version of the displayed.

V. Visualizing

a. Distribution of Allele Frequency

In the previous section ‘IV. Wrangling; b. Summary Statistics - Allele Frequency,’ I calculated the mean and standard deviation of the allele frequencies. Here, I visualized the same information into a histogram to better understand its distribution. This graph displays allele frequency on the x axis and the number of SNP that has the allele frequency on the y axis. Again, allele frequency refers to how prevalent the standard allele is in the general population. High allele frequency means that most people have the reference allele, while low frequency means that the proportion of those who have the reference allele is low.

```
# visualizing allele frequency
anx_mdd_final %>%
  distinct(SNP, freq) %>%  # plot only the frequencies for distinct SNP
  ggplot(aes(x = freq)) +  # plot frequencies in x axis
  geom_histogram(binwidth = 0.025, fill = 'lightblue4') +   # plot a histogram
  scale_x_continuous(breaks = seq(0, 1, by = 0.1)) +   # modify scales in x axis
  # specify title and axes labels
  labs(x = 'Allele Frequency', y = 'Count', title = 'Allele Frequency Distribution of
Single Nucleotide Polymorphisms', subtitle = '- Anxiety Disorder and Major Depressive
Disorder Patients') +
  theme_classic()   # modify theme
```

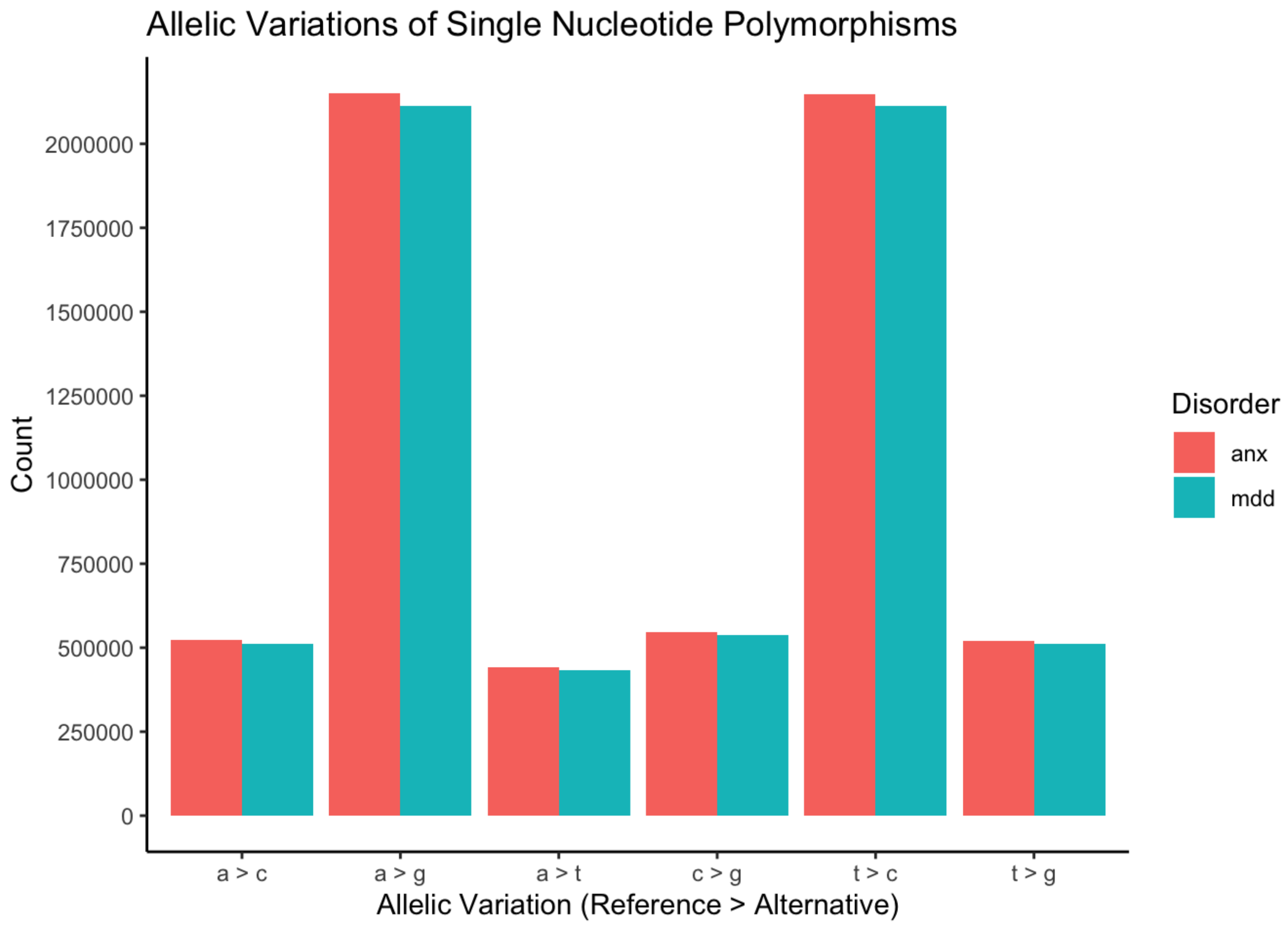


The general trend in the visualization is a u-shaped histogram. There are high proportions of SNP that have extreme values of allele frequency, both on the lowest end and the highest end. This indicates that many of the SNP mutations detected were from a very rare mutation or a very common mutation. This result colors the conclusions we can derive from other analyses, because common mutations are difficult to link with specific disorders.

b. Number of Allelic Variations by Disorder

Previously, in ‘IV. Wrangling; c. Summary Statistics - Allelic Variation,’ I calculated the number of SNP in each allelic variation type. Here, I created a bar graph to visualize the number of allelic variations by disorder. The x axis displays 6 different types of allelic variations, which are combinations of reference alleles and alternative alleles. The y axis displays the number of SNP cases in each category. Different colors were used to distinguish between the SNP in AD and MDD disorders.

```
# visualizing allele variation by disorder
anx_mdd_final %>%
  filter(!str_detect(allele_var, "NA > NA")) %>%   # remove values with NA
  ggplot(aes(x = allele_var, fill = disorder)) +   # plot allelic variation on x axis
and distinguish the 'disorder' variable by color
  geom_bar(stat = 'count', position = 'dodge') +   # use stat function to specify the
display of counts
  scale_y_continuous(breaks = seq(0, 2000000, by = 250000)) +   # adjust y axis scale
s
  # specify title and axes labels
  labs(x = 'Allelic Variation (Reference > Alternative)', y = 'Count', title = 'Allel
ic Variations of Single Nucleotide Polymorphisms', fill = 'Disorder') +
  theme_classic()   # modify theme
```



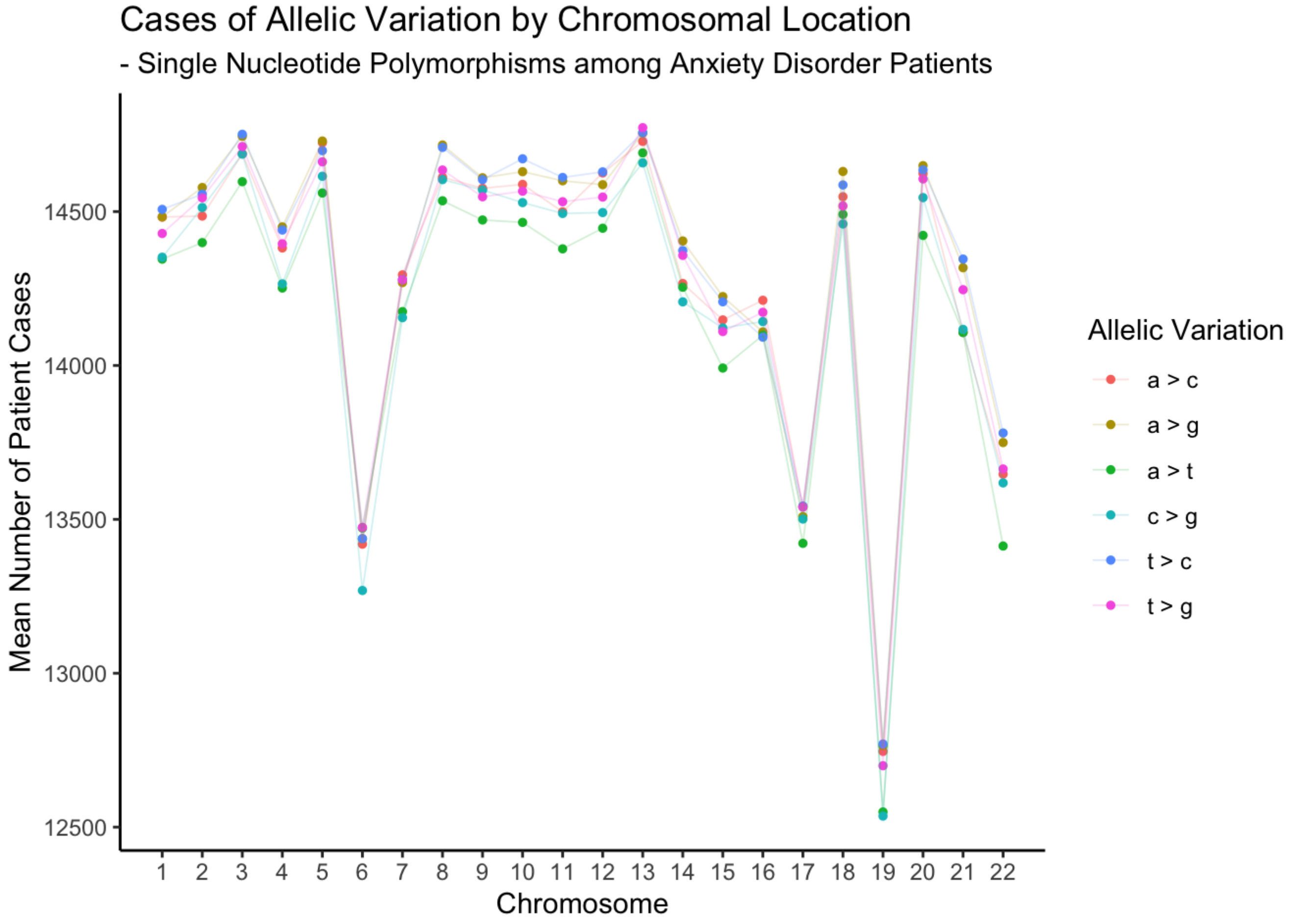
The apparent trend in this graph is that the variations A > G and T > C are highly prevalent compared to all other combinations. Additionally, it seems like there is no significant difference between A > G and T > C count. Same goes for the rest, since A > C, A > T, C > G, and T > G all have similar amount of SNP in their category. Furthermore, there are no apparent differences in the SNP pattern between patients with anxiety disorder and patients with major depressive disorder.

c. Cases of Allelic Variation by Chromosomal Location

In this section, I plotted two different scatterplots that each display information on three variables: the chromosomal location of the SNP, allelic variation types, and the mean number of patients that have the SNP. The two different figures are for separating SNP data from anxiety disorder and major depressive disorder. In both graphs, the x axis represent chromosomal location, from chromosome 1 to 22. The y axis is the mean number of patients that displayed the type of SNP. The SNP are grouped by chromosomal location and allelic variation before the mean numbers of cases were computed. Color distinguishes between the allelic variation types. Lines were added to help visualize the groups of points based on allelic variation.

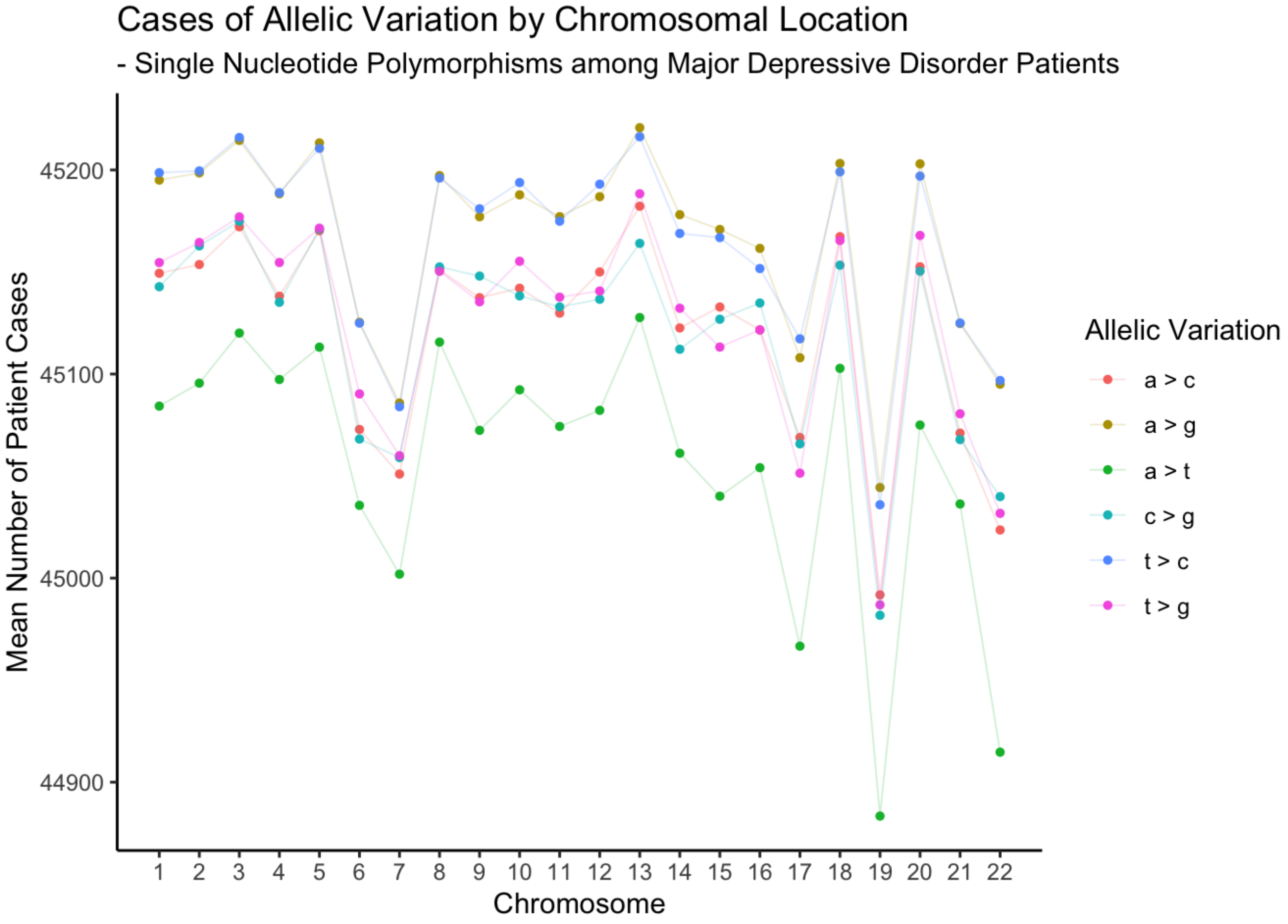
```
anx_mdd_final %>%
  filter(disorder == 'anx') %>%   # display data on Anxiety patients only
  filter(!str_detect(allele_var, "NA > NA")) %>%   # remove values with NA
  group_by(chr, allele_var) %>%   # group by chromosome and allelic variation to compute mean cases of patients
  # compute mean cases of patients
  summarize(mean_cases = mean(cases, na.rm = T), .groups = 'drop') %>%
  # plot chromosome location on x axis, mean number of cases on y axis, and use color to distinguish allelic variation types
  ggplot(aes(x = chr, y = mean_cases, color = allele_var)) +
  geom_point(size = 1) +   # scatter plot with size 1
  geom_line(alpha = 0.2, size = 0.3) +   # connect the allelic variation types
  scale_x_continuous(breaks = seq(1, 22, by = 1)) +   # adjust x axis scales
  # specify titles and axes labels
  labs(x = 'Chromosome', y = 'Mean Number of Patient Cases', color = 'Allelic Variation', title = 'Cases of Allelic Variation by Chromosomal Location', subtitle = '- Single Nucleotide Polymorphisms among Anxiety Disorder Patients') +
  theme_classic()   # modify theme
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
```



This graph summarizes the SNP information from anxiety disorder (AD) patients. AD patients had specifically low mutations in chromosome 6 and 19, followed by 17 and 22. The differences of patient cases between chromosome 6 and 19 compared to the rest were very big. Chromosome 18 and 20 had specifically high cases of SNP despite its relatively shorter length. Allelic variations T > C and A > G were the two most prevalent variations in patients across all chromosomes, while C > G and A > T were the two most uncommon.

```
anx_mdd_final %>%
  filter(disorder == 'mdd') %>%   # display data on Major Depressive Disorder patient
s only
  filter(!str_detect(allele_var, "NA > NA")) %>%   # remove values with NA
  group_by(chr, allele_var) %>%   # group by chromosome and allelic variation to comp
ute mean cases of patients
  # compute mean cases of patients
  summarize(mean_cases = mean(cases, na.rm = T), .groups = 'drop') %>%
  # plot chromosome location on x axis, mean number of cases on y axis, and use color
to distinguish allelic variation types
  ggplot(aes(x = chr, y = mean_cases, color = allele_var)) +
  geom_point(size = 1) +   # scatter plot with size 1
  geom_line(alpha = 0.2, size = 0.3) +   # connect the allelic variation types
  scale_x_continuous(breaks = seq(1, 22, by = 1)) +   # adjust x axis scales
  # specify titles and axes labels
  labs(x = 'Chromosome', y = 'Mean Number of Patient Cases', color = 'Allelic Variati
on', title = 'Cases of Allelic Variation by Chromosomal Location', subtitle = '- Sing
le Nucleotide Polymorphisms among Major Depressive Disorder Patients') +
  theme_classic()   # modify theme
```



This graph summarizes the SNP information from major depressive disorder (MDD) patients. MDD patients had specifically low mutations in chromosome 7, 17, 19, and 22. Compared to AD disorder, the difference of number of patient SNP cases across all chromosome was smaller, meaning it has less spread in the y axis. Unlike AD, chromosome 7 was one of the chromosomes with the lowest SNP cases. Similar to AD, allelic variations T > C and A > G were the two most prevalent variations in patients across all chromosomes. Here, A >T was the variation type with the single lowest patient cases in every chromosome.

VI. Discussion

a. Results

Throughout the project, I analyzed two genomic datasets obtained from studies on Single Nucleotide Polymorphisms (SNP) in patients that have anxiety disorder (AD) and major depressive disorder (MDD). The research question I investigated was: Are there any significant difference in SNP profiles between patients with AD or MDD, and can we identify any potential genetic markers associated with these disorders?

Firstly, it was observed from initial data analysis in the 'II. Joining/Merging' section that the two SNP datasets share a lot of common single nucleotide polymorphisms. In addition, there were more SNP that were characterized in MDD patients but not in AD, compared to vice versa.

The visualization of allele frequency in 'V. Visualizing; a. Distribution of Allele Frequency' section revealed a u-shaped histogram. This indicates that the disorders are potentially associated with rare mutations in the population. Still, additional analysis on how the patient and control groups were sampled is necessary to validate this conclusion.

In the analysis of the allelic variations in section ‘V. Visualizing; b. Number of Allelic Variations by Disorder,’ both disorders had high instances of A > G and T > C mutations compared to any other allelic combinations. No significant differences were observed when comparing the prevalence of allelic variation between AD patients and MDD patients.

Finally, in the section ‘V. Visualizing; c. Cases of Allelic Variation by Chromosomal Location,’ comprehensive analysis on SNP profiles for each disorder was possible. Both disorders had common chromosomes with low SNP mutations, such as chromosome 6 and 19. Notably, chromosome 6 had low SNP cases in both disorders despite its relatively long length. On the other hand, chromosome 7 had significantly low SNP prevalence in MDD patients, but not as much in AD patients. This could be a potential genetic marker differentiating those with MDD from AD, but further analysis on the actual mutations within chromosome 7 is necessary to verify the indication of this result. Furthermore, analysis on AD patients showed a bigger difference in number of cases of SNP across chromosomes, but it is possible that this effect is due to differing sample size, as discovered in section ‘IV. Wrangling; a. Summary Statistics - Cases.’

b. Reflection

Conducting this project was challenging, especially in understanding the information stored in the genomic data. There was a lot of extra studying required to comprehend and apply the biological knowledge to conduct data analysis. However, I learned a lot about how genomic data are organized, and the significance of each of the variables. Through this project, I gained a deeper understanding in the field of genomics and its potential of interdisciplinary research using data science.

c. Acknowledgements

I want to express gratitude to the authors of the two papers and the Psychiatric Genomics Consortium (PGC) for making this data publicly available for researchers.

Special thanks to Dr. Goheun Kim and Dr. Kadir Akdemir who gave me the genetics and computational biology knowledge that was crucial for conducting this project.

Finally, I would like to acknowledge and appreciate Dr. Layla Guyot for her efforts to provide us with this opportunity of conducting our own data science project.

VII. References

Otowa, T et al. “Meta-analysis of genome-wide association studies of anxiety disorders.” *Molecular psychiatry* vol. 21,10 (2016): 1391-9. doi:10.1038/mp.2015.197 (doi:10.1038/mp.2015.197)

Wray, Naomi R et al. “Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression.” *Nature genetics* vol. 50,5 (2018): 668-681. doi:10.1038/s41588-018-0090-3 (doi:10.1038/s41588-018-0090-3)