Exploratory Data Analysis for dataset containing Lung Cancer associated SNPs from GWAS Catalog

Dataset can be found here - Link

Citations

1. Sollis E, Mosaku A, Abid A, Buniello A, Cerezo M, Gil L, Groza T, Güneş O, Hall P, Hayhurst J, Ibrahim A, Ji Y, John S, Lewis E, MacArthur JAL, McMahon A, Osumi-Sutherland D, Panoutsopoulou K, Pendlington Z, Ramachandran S, Stefancsik R, Stewart J, Whetzel P, Wilson R, Hindorff L, Cunningham F, Lambert SA, Inouye M, Parkinson H, Harris LW. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. Nucleic Acids Res. 2023 Jan 6;51(D1):D977-D985. doi: 10.1093/nar/gkac1010. PMID: 36350656; PMCID: PMC9825413.
2. Chen, C., Liu, Y., Luo, M., Yang, J., Chen, Y., Wang, R., Zhou, J., Zang, Y., Diao, L., & Han, L. (2024). PancanQTLv2.0: a comprehensive resource for expression quantitative trait loci across human cancers. Nucleic acids research, 52(D1), D1400–D1406. https://doi.org/10.1093/nar/gkad916

Appendix

Dataset Introduction

The dataset contains information about Single Nucleotide Polymorphisms (SNPs) associated with lung cancer sourced from the GWAS dataset. Each record within the dataset corresponds to a SNP reported in a published scientific literature and is reported to have a statistical link to lung cancer susceptibility or outcomes. The dataset includes key features such as SNP ID, the chromosome where the SNP is located at and its exact position in the chromosome, the significance of the SNP is captured through p-values, odds ratio which represents the probability of the said SNP to occur. Altogether, this dataset proves to be comprehensive resource for exploring genetic risk factors underlying lung cancer.

Dataset Justification

We chose this dataset as we acknowledge that lung cancer remains to be one of the leading cause of death in the world and understanding about the underlying risk factors that cause this fatal disease can help us understand the mechanism of the cancer and develop targeted therapies to prevent it. GWAS Catalog is a well-established knowledge resource for studying these risk factors that combines categorical columns such as SNP ID, Gene Name, ancestry as well as numerical variables such as p-values, odds ratio, and sample sizes. This makes the dataset a flexible option for numerical and categorical analysis. Its biomedical association, rich annotation, and potential to explore genetic risk factors make it a strong candidate for meaningful analysis.

Research Question

We plan to use the data from GWAS to perform eQTL analysis. eQTL or Expression Quantitative Trait Loci are genomic loci (positions) within the genome that influence gene expression levels. There are 2 types of eQTLs: cis-eQTLs are SNPs which occur within close proximity of a gene thereby controlling the gene expression by possibly altering the region where a transcription factor binds for gene expression. The second type is trans-eQTLs are SNPs which are located anywhere within the human genome, even on a different chromosome. These eQTLs are much harder to identify and often act together with other trans-eQTLs in coordinating gene expression. We would be using this dataset to identify potential eQTLs based on constraints such as p-value less than 5E-8, the proximity of the SNP to the gene etc. After identifying potential eQTLs for Lung Cancer, we would be comparing them against SNPs recorded in the established eQTL database for cancer PancanQTL

Data Pre-processing and cleanup

We found that the dataset contained a lot of NR values in the Risk Allele Frequency column which we plan to convert to NA values as they are not important because the column is actually numerical. We found dirty values within chromosome ID column which we plan to standardize to integers between 1 and 22, as well as X, Y, and MT. We found that certain numerical columns in the dataset that was misinterpreted to be string, which we would handling using as.integer or other similar functions for the other misinterpreted columns. We would also be checking for duplicate SNP IDs to maintain the uniqueness of each SNP.

Planned Statistical Methods

We could possibly implement classical model such as `linear regression` to understand the relationships between Expression which could be obtained from GEO (Gene Expression Omnibus) and Genotype which is the SNP that could be integrated using the existing R package `MatrixEQTL`. This will involve modeling gene expression as the dependent variable and SNP genotype (coded as 0, 1, or 2 based on minor allele count) as the independent variable, while controlling for potential confounders such as ancestry and study population. `Chi-square` tests will be employed to examine associations between categorical variables such as chromosome distribution and risk allele presence, while `correlation analysis` will assess relationships between continuous variables like p-values, odds ratios, and risk allele frequencies.

Limitations

The limitation of the dataset is that there are a lot of NA values in a lot of fields that would affect the downstream analysis. We plan to analyse the dataset using the proposed method using only a select fiew fields as some of the other fields contain descriptions/annotations that are not useful for analysis. We also found that the data type of certain fields are being misinterpreted so we would explore the field before changing the data type of the field. Additionally, the presence of 337 duplicate SNP records suggests inconsistent data curation practices, requiring careful selection criteria to retain the most reliable entries without introducing selection bias. These missing values could introduce bias if they are not missing at random, potentially skewing our understanding of allele frequencies and effect sizes.

Libraries

```r
library(ggplot2)
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v lubridate 1.9.4     v tibble    3.3.0
v purrr     1.1.0     v tidyr     1.3.1
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```r
library(stringr)
```

Exploratory Data Analysis

Loading our dataset

```r
data <- read.table('data/gwas-association-downloaded_2025-09-25-MONDO_0008903-withChildTraits
```

Summary of our dataset

```r
column_names <- colnames(data)
cat(paste0('Column names of the GWAS dataset:\t', column_names, '\n'))
```

```
Column names of the GWAS dataset:   DATE.ADDED.TO.CATALOG
 Column names of the GWAS dataset:  PUBMEDID
 Column names of the GWAS dataset:  FIRST.AUTHOR
 Column names of the GWAS dataset:  DATE
 Column names of the GWAS dataset:  JOURNAL
 Column names of the GWAS dataset:  LINK
 Column names of the GWAS dataset:  STUDY
 Column names of the GWAS dataset:  DISEASE.TRAIT
 Column names of the GWAS dataset:  INITIAL.SAMPLE.SIZE
 Column names of the GWAS dataset:  REPLICATION.SAMPLE.SIZE
 Column names of the GWAS dataset:  REGION
 Column names of the GWAS dataset:  CHR_ID
 Column names of the GWAS dataset:  CHR_POS
 Column names of the GWAS dataset:  REPORTED.GENE.S.
 Column names of the GWAS dataset:  MAPPED_GENE
 Column names of the GWAS dataset:  UPSTREAM_GENE_ID
 Column names of the GWAS dataset:  DOWNSTREAM_GENE_ID
 Column names of the GWAS dataset:  SNP_GENE_IDS
 Column names of the GWAS dataset:  UPSTREAM_GENE_DISTANCE
 Column names of the GWAS dataset:  DOWNSTREAM_GENE_DISTANCE
 Column names of the GWAS dataset:  STRONGEST.SNP.RISK.ALLELE
 Column names of the GWAS dataset:  SNPS
 Column names of the GWAS dataset:  MERGED
 Column names of the GWAS dataset:  SNP_ID_CURRENT
 Column names of the GWAS dataset:  CONTEXT
 Column names of the GWAS dataset:  INTERGENIC
 Column names of the GWAS dataset:  RISK.ALLELE.FREQUENCY
 Column names of the GWAS dataset:  P.VALUE
 Column names of the GWAS dataset:  PVALUE_MLOG
 Column names of the GWAS dataset:  P.VALUE..TEXT.
 Column names of the GWAS dataset:  OR.or.BETA
 Column names of the GWAS dataset:  X95..CI..TEXT.
 Column names of the GWAS dataset:  PLATFORM..SNPS.PASSING.QC.
 Column names of the GWAS dataset:  CNV
 Column names of the GWAS dataset:  MAPPED_TRAIT
 Column names of the GWAS dataset:  MAPPED_TRAIT_URI
 Column names of the GWAS dataset:  STUDY.ACCESSION
 Column names of the GWAS dataset:  GENOTYPING.TECHNOLOGY
```

```
summary(data)
```

```
DATE.ADDED.TO.CATALOG    PUBMEDID        FIRST.AUTHOR           DATE
Length:1748           Min.   :18385676   Length:1748        Length:1748
Class :character      1st Qu.:28604730   Class :character   Class :character
Mode  :character      Median :31326317   Mode  :character   Mode  :character
                      Mean   :32965060
                      3rd Qu.:37689528
                      Max.   :40829600

  JOURNAL              LINK              STUDY          DISEASE.TRAIT
Length:1748        Length:1748        Length:1748        Length:1748
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character




INITIAL.SAMPLE.SIZE REPLICATION.SAMPLE.SIZE    REGION
Length:1748         Length:1748             Length:1748
Class :character    Class :character        Class :character
Mode  :character    Mode  :character        Mode  :character




   CHR_ID            CHR_POS          REPORTED.GENE.S.   MAPPED_GENE
```

```
Length:1748        Length:1748        Length:1748        Length:1748      
Class :character   Class :character   Class :character   Class :character  
Mode  :character   Mode  :character   Mode  :character   Mode  :character  




UPSTREAM_GENE_ID   DOWNSTREAM_GENE_ID SNP_GENE_IDS      
Length:1748        Length:1748        Length:1748       
Class :character   Class :character   Class :character  
Mode  :character   Mode  :character   Mode  :character  




UPSTREAM_GENE_DISTANCE DOWNSTREAM_GENE_DISTANCE STRONGEST.SNP.RISK.ALLELE
Min.   :      1        Min.   :     30          Length:1748              
1st Qu.:  10198        1st Qu.:   8771          Class :character         
Median :  29873        Median :  25382          Mode  :character         
Mean   :  94663        Mean   :  95607                                   
3rd Qu.:  96508        3rd Qu.:  86828                                   
Max.   :3249767        Max.   :4177110                                   
NA's   :1199           NA's   :1199                                      
    SNPS               MERGED        SNP_ID_CURRENT        CONTEXT         
Length:1748        Min.   :0.00000   Min.   :6.569e+03   Length:1748       
Class :character   1st Qu.:0.00000   1st Qu.:4.400e+06   Class :character  
Mode  :character   Median :0.00000   Median :1.234e+07   Mode  :character  
                   Mean   :0.05492   Mean   :5.399e+07                     
                   3rd Qu.:0.00000   3rd Qu.:7.744e+07                     
                   Max.   :1.00000   Max.   :1.827e+09                     
                                     NA's   :114                           
  INTERGENIC     RISK.ALLELE.FREQUENCY    P.VALUE           PVALUE_MLOG    
Min.   :0.0000   Length:1748           Min.   :0.000e+00   Min.   :  5.000  
1st Qu.:0.0000   Class :character      1st Qu.:6.000e-10   1st Qu.:  5.398  
Median :0.0000   Mode  :character      Median :4.000e-07   Median :  6.398  
Mean   :0.3754                         Mean   :2.115e-06   Mean   :  9.308  
3rd Qu.:1.0000                         3rd Qu.:4.000e-06   3rd Qu.:  9.222  
Max.   :1.0000                         Max.   :1.000e-05   Max.   :178.097  
NA's   :27                                                                  
P.VALUE..TEXT.       OR.or.BETA      X95..CI..TEXT.    
Length:1748        Min.   :  0.010   Length:1748       
Class :character   1st Qu.:  1.070   Class :character  
Mode  :character   Median :  1.176   Mode  :character  
```

```
                    Mean   :  1.973
                    3rd Qu.:  1.645
                    Max.   :101.639
                    NA's   :121
 PLATFORM..SNPS.PASSING.QC.     CNV            MAPPED_TRAIT
 Length:1748                Length:1748        Length:1748
 Class :character           Class :character   Class :character
 Mode  :character           Mode  :character   Mode  :character



 MAPPED_TRAIT_URI   STUDY.ACCESSION    GENOTYPING.TECHNOLOGY
 Length:1748        Length:1748        Length:1748
 Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character
```

We see that some variables such as Chromosome position and Pubmed ID which misinterpreted as different data types by R. So in the next section, we change the data type of the column to match their true quality.

Changing the some data type of some columns as they were declared as str but were int

```r
data$CHR_POS <- as.integer(data$CHR_POS)
```

```
Warning: NAs introduced by coercion
```

```r
data$PUBMEDID <- as.character(data$PUBMEDID)
```

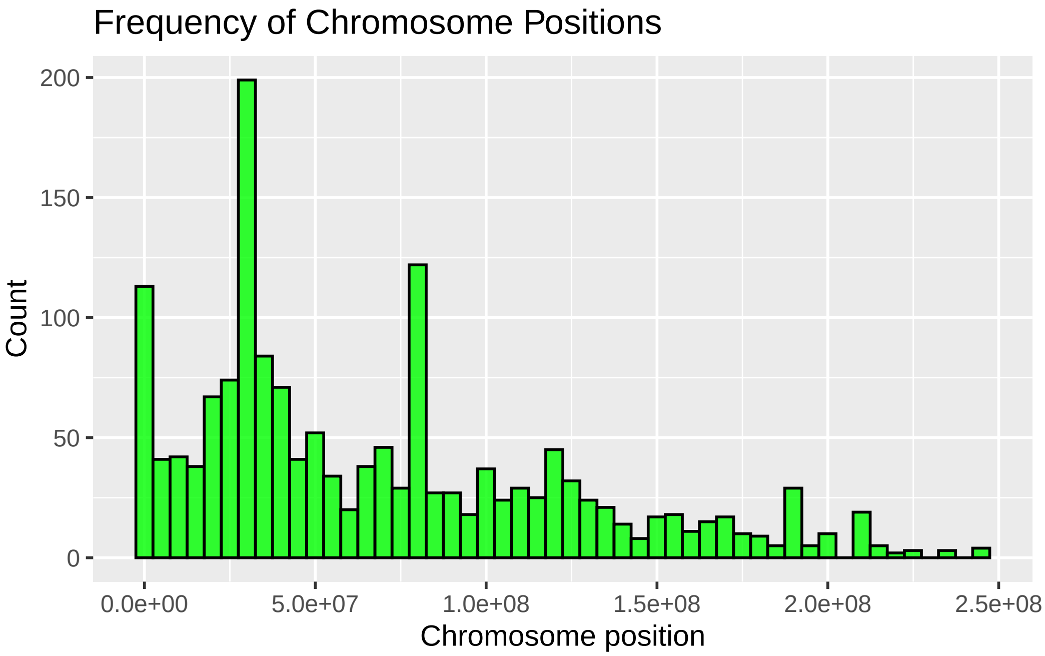
Variable Description

Histogram of the Positions of the SNPs

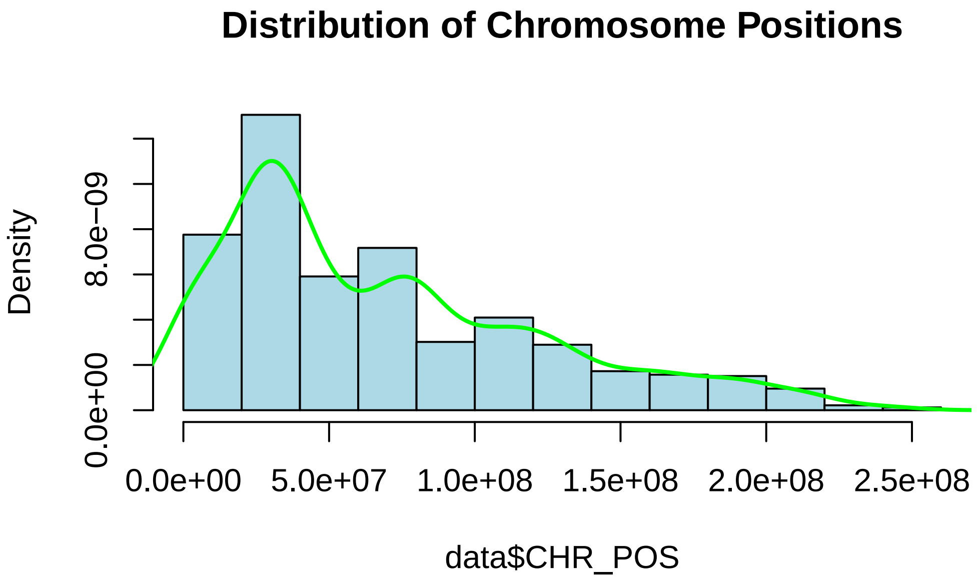
We are looking at the distribution of the positions where these SNPs occur within the genome to get an idea if most of the mutations occur upstream (towards the 5' end) or downstream (towards the 3' end) of the genome.

```r
ggplot(data, aes(CHR_POS, fill=CHR_ID)) +
    geom_histogram(bins = 50, fill = 'green', color = 'black', alpha = 0.8) +
    labs(
        title = 'Frequency of Chromosome Positions',
        x = 'Chromosome position',
        y = 'Count'
    )
```

```
Warning: Removed 124 rows containing non-finite outside the scale range
(`stat_bin()`).
```



```r
h <-hist(data$CHR_POS, probability =  TRUE, col = 'lightblue', main = 'Distribution of Chrom
lines(density(na.omit(data$CHR_POS)), col = 'green', lwd = 2)
```



```
summary(data$CHR_POS)
```

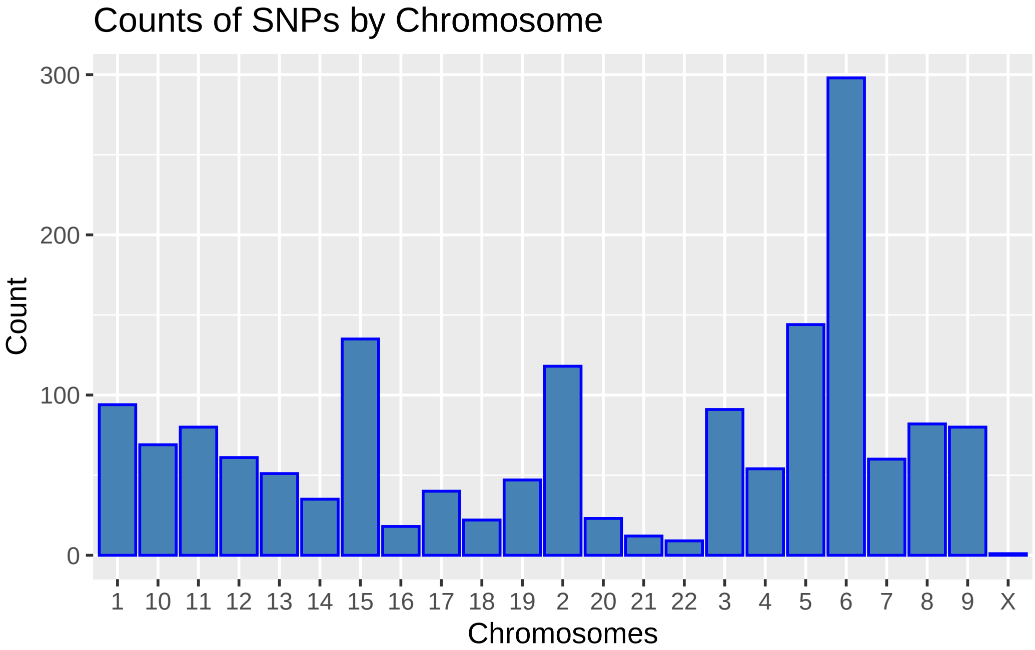
```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's 
   242581  28645826  51549004  68685302 101493781 245121489       124 
```

The SNP chromosome positions nearly span the entirety of the human genome with values ranging from 242kb (kilobases) to 245Mb (Megabases). The median SNP position is approximately 51.5Mb while the mean which is being skewed by larger values is 68.6Mb. SNPs are distributed across the genome, while clusters of SNPs occurring in certain regions may indicate the possibility of potential trans-eQTLs.

Barplot of Chromosomes

These chromosomes are condensed forms of the DNA where the SNPs occur and we are visualizng which of the chromosomes in the body seems to be associated with frequent mutations that leads to Lung cancer.

```
valid_chr <- as.character(c(1:22, "X", "Y", "MT"))
data <- data %>%      # Cleaning the CHR_ID as it contained dirty mixed values
    mutate(
        CHR_ID_clean = str_trim(CHR_ID),
        CHR_ID_clean = str_extract(CHR_ID_clean, "^[0-9]+$|^X$|^Y$|^MT$"),
        CHR_ID_clean = ifelse(CHR_ID_clean %in% valid_chr, CHR_ID_clean, NA)
) %>%
    filter(!is.na(CHR_ID_clean))
ggplot(data, aes(x = CHR_ID_clean)) +
    geom_bar(fill = 'steelblue',color = 'blue', na.rm=TRUE) +
    labs(
        title = 'Counts of SNPs by Chromosome',
        x = 'Chromosomes',
        y = 'Count'
    )
```



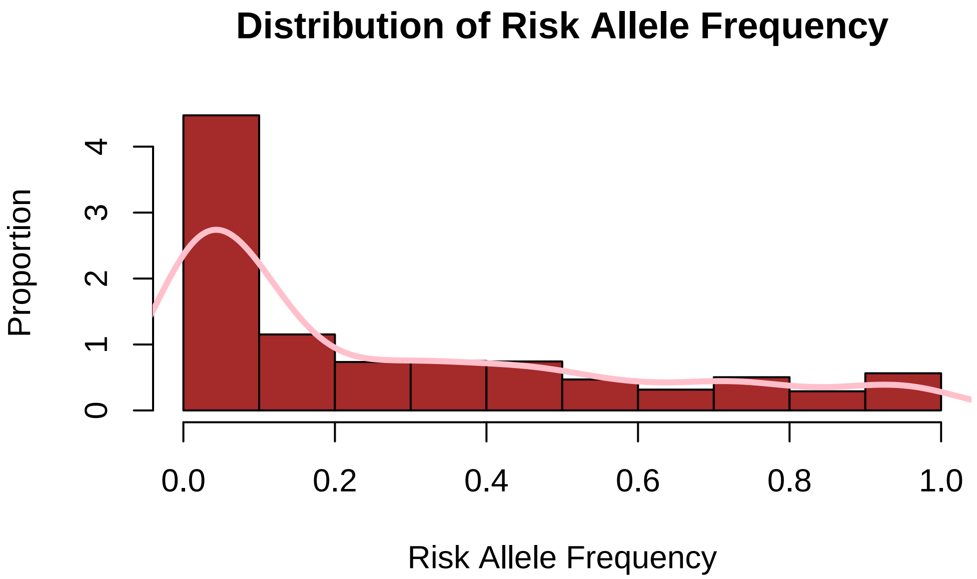
We find that most of the SNPs are found on Chromosome 6, followed by chromosomes 15 and 2. From this, we can infer that mutations associated with chromosome 6 may play a particularly important role in lung cancer susceptibility. At a glance, these findings indicate hotspots for targeted therapy that may decrease the risk of lung cancer susceptibility.

Histogram of Risk Allele frequency

Risk Allele Frequency is the proportion of chromosomes in a population that carry this risk allele as these alleles increase the likelihood of developing the disease.

```r
data <- data %>%
    mutate(
        RAF = ifelse(RISK.ALLELE.FREQUENCY == "NR", NA, RISK.ALLELE.FREQUENCY),
        RAF = as.numeric(RAF)
    )

h <- hist(data$RAF, probability = TRUE,col = 'brown', main = 'Distribution of Risk Allele Fr
lines(density(na.omit(data$RAF)), col = 'pink', lwd = 3)
```



```r
summary(data$RAF)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
 0.0000  0.0240  0.1230  0.2728  0.4576  0.9998     453 
```

The Risk Allele Frequency (RAF) values range from 0 (rare alleles) to 0.9998 (nearly fixed alleles). The median RAF 0.1230, which is more useful in this feature of the dataset as it is right-skewed, indicating that half of the risk alleles occur in less than 12% of the population. The mean RAF 0.2728 though higher is affected by some common alleles. The IQR which lies between 0.0240 and 0.4576 captures majority of the moderately frequent risk alleles. Also 453 alleles lacked RAF values and were marked `NR` which were converted to `NA` values.

Checking for duplicated SNPs

```
sum(duplicated(data$SNPS))
```

```
[1] 337
```

We found that there are around 337 duplicate SNPs that need to be filtered out. These duplicated records may arise from multiple studies reporting on the same SNP under different experimental conditions, sample populations, and statistical models. We plan to do the filtering based on biologically meaningful properties such as Risk Allele Frequency, p-value, odds ratio. This ensures that for each unique SNP, only the most reliable and informative record is being retained.