PHC 6088 - Final Project

Sumia Tahir

4/30/2020

BACKGROUND

Primary sclerosing cholangitis (PSC) is characterized by chronic inflammation and scarring of the bile ducts (Mayo Clinic). Due to the blocked ducts, bile may accumulate in the liver and lead to liver damage and cirrhosis. Symptoms progress very slowly and may include malaise, jaundice, itchy skin, pain in the upper right part of the abdomen, chills, night sweats, and enlarged liver. In advanced stages, it may lead to liver failure or cancers of the bile duct (cholangiocarcinoma) and liver. The only possible treatment for advanced primary sclerosing cholangitis is a liver transplant. In North America, the incidence of PSC ranges from 3.85 to 16.2 cases per 100,000 person-years.

PSC may be caused by autoimmune factors, and the risk for this disease has a strong genetic component. Inflammatory bowel disease (IBD), which includes ulcerative colitis and Crohn's disease, is also present in about 70% of people with primary sclerosing cholangitis. IBD has also been shown to have genetic predisposition. People with both PSC and IBD are at an increased risk for colon cancer. Crohn's disease most commonly affects the end of the small intestine (ileum) and colon, causing abdominal pain and diarrhea.

People with Celiac disease also have an increased risk for developing PSC. This association is a feature of autoimmunity. Celiac disease affects 1 in 100 people worldwide. It is characterized by inflammation of the small intestine due to ingestion of gluten, a protein found in wheat, rye and barley (Celiac Disease Foundation). The immune response damages the villi that line the small intestine, leading to malabsorption of nutrients. Celiac disease also has a strong hereditary component, with a 1 in 10 risk of developing the disease if a parent, child or sibling has it. If left untreated, celiac disease increases the risk for coronary artery disease and small bowel cancers, and can lead to the development of other autoimmune disorders such as Type I diabetes and multiple sclerosis.

Since PSC has been linked to Crohn's disease and Celiac disease, we looked at summary statistics from GWAS to find SNPs that were highly associated with these conditions. Some of these SNPs may have potential to be diagnostic markers for diseases that have slow progression and mild symptoms, or they may give an idea of the risk of susceptibility to a disease. Finding similar SNPs across diseases may increase our understanding of the underlying mechanisms and pathways and how one disease may increase risk for another.

INTRODUCTION TO DATASETS

The datasets used in this analysis were downloaded as "tsv" files from the Genome-Wide Association Study (GWAS) database at http://www.ebi.ac.uk/gwas. The phenotypes of interest were "sclerosing cholangitis", "Celiac disease" and "Crohn's disease". Undergoing a search in the GWAS database with the keywords "sclerosing cholangitis" yielded 308 associations/SNPs from 17 studies (Trait: EFO_0004268). The keyword search for "Celiac disease" gave 211 assocations from 15 studies (Trait: EFO_0001060), while the search for "Crohn's disease" provided 891 associations from 46 studies (Trait: EFO_0000384).

DATA ANALYSIS & RESULTS

```
#install.packages("qqman")
#install.packages("forestplot")
library(forestplot)
library(qqman)

psc0 <- read.table(file="gwas.tsv", sep='\t', header=TRUE)
crohns0 <- read.table("crohns.tsv",stringsAsFactors = FALSE,sep = "\t", fill = TRUE,
                      quote = "", header=TRUE)
celiac0 <- read.table("celiac.tsv",stringsAsFactors = FALSE,sep = "\t", fill = TRUE,
                      quote = "", header=TRUE)
```

```
dim(psc0)
```

```
## [1] 307  38
```

```
dim(celiac0)
```

```
## [1] 211  38
```

```
dim(crohns0)
```

```
## [1] 890  38
```

Note: One observation was removed from both the “sclerosing cholangitis” dataset and the “Crohns Disease” dataset because the p-value was on the order of 10e-341, and there was an error in its SNP ID.

All of the datasets featured 38 columns/variables as listed below:

```
colnames(psc0)
```

```
##  [1] "DATE.ADDED.TO.CATALOG"      "PUBMEDID"
##  [3] "FIRST.AUTHOR"               "DATE"
##  [5] "JOURNAL"                    "LINK"
##  [7] "STUDY"                      "DISEASE.TRAIT"
##  [9] "INITIAL.SAMPLE.SIZE"        "REPLICATION.SAMPLE.SIZE"
## [11] "REGION"                     "CHR_ID"
## [13] "CHR_POS"                    "REPORTED.GENE.S."
## [15] "MAPPED_GENE"                "UPSTREAM_GENE_ID"
## [17] "DOWNSTREAM_GENE_ID"         "SNP_GENE_IDS"
## [19] "UPSTREAM_GENE_DISTANCE"     "DOWNSTREAM_GENE_DISTANCE"
## [21] "STRONGEST.SNP.RISK.ALLELE"  "SNPS"
## [23] "MERGED"                     "SNP_ID_CURRENT"
## [25] "CONTEXT"                    "INTERGENIC"
## [27] "RISK.ALLELE.FREQUENCY"      "P.VALUE"
## [29] "PVALUE_MLOG"                "P.VALUE..TEXT."
## [31] "OR.or.BETA"                 "X95..CI..TEXT."
## [33] "PLATFORM..SNPS.PASSING.QC." "CNV"
## [35] "MAPPED_TRAIT"               "MAPPED_TRAIT_URI"
## [37] "STUDY.ACCESSION"            "GENOTYPING.TECHNOLOGY"
```

REMOVING OBSERVATIONS WITH MISSING SNP IDs

Some of the rows had missing values for the SNP identifier. These were removed from all of the datasets.

```
psc1 <- psc0[!is.na(psc0$SNP_ID_CURRENT),]
crohns1 <- crohns0[!is.na(crohns0$SNP_ID_CURRENT),]
celiac1 <- celiac0[!is.na(celiac0$SNP_ID_CURRENT),]
```

```
dim(psc1)
```

```
## [1] 300  38
```

```
dim(celiac1)
```

```
## [1] 208  38
```

```
dim(crohns1)
```

```
## [1] 885  38
```

COMBINING P-VALUES FROM MULTIPLE STUDIES (FISHER'S METHOD)

The datasets contain summary statistics from multiple studies. Therefore, some of the SNPs had multiple p-values. These were combined using Fisher's method, where -2 times the sum of the natural log of p-values from different studies follows a chi-squared distribution with degrees of freedom equal to twice the number of studies.

A new variable called "p_fish" was created to add p-values that have been adjusted for multiple studies.

```
psc1$p_fish <- psc1$P.VALUE
crohns1$p_fish <- crohns1$P.VALUE
celiac1$p_fish <- celiac1$P.VALUE
```

SNPs with multiple entries were identified and their p-values were combined.

Sclerosing Cholangitis dataset

```
# finding duplicate SNPs
dups_psc <- psc1[duplicated(psc1$SNP_ID_CURRENT)|duplicated(psc1$SNP_ID_CURRENT,
                                                            fromLast=TRUE),]
table(dups_psc$SNP_ID_CURRENT) #frequency of each duplicate SNP
```

```
##
##  1788097  1893592  2836883  3184504  3197999  3748816  4147359  7426056
##        2        2        2        3        4        2        2        3
##  7937682 11168249 13140464 56258221 60652743
##        2        2        2        2        2
```

```
ind_psc <- unique(dups_psc$SNP_ID_CURRENT) #ID numbers of duplicate SNPs
print(nrow_psc <- length(ind_psc)) #total number of SNPs with multiple entries
```

```
## [1] 13
```

```
# calculating Fisher's p-value for duplicate SNPs
for (i in 1:nrow_psc) {
  chisq <- (-2)*sum(log(psc1$P.VALUE[psc1$SNP_ID_CURRENT==ind_psc[i]]))
  df <- 2*length(which(psc1$SNP_ID_CURRENT==ind_psc[i]))
  pval <- pchisq(chisq, df, lower.tail=FALSE)
  psc1$p_fish[psc1$SNP_ID_CURRENT==ind_psc[i]] <- pval
}
```

The table shows the SNP ID numbers for 13 duplicate SNPs along with the number of entries for each SNP. Almost all duplicate SNPs have 2 entries, except for rs319799 that has 4 entries and rs3184504 that has 3 entries.

Celiac Disease dataset

```
dups_cel <- celiac1[duplicated(celiac1$SNP_ID_CURRENT)|duplicated(celiac1$SNP_ID_CURRENT,
                                                                  fromLast=TRUE),]
table(dups_cel$SNP_ID_CURRENT) #frequency of duplicate SNPs
```

```
##
##   653178  1250552  1464510  1738074  1893592  1980422  2187668  2816316
##        2        2        2        2        2        2        2        2
##  4821124  6679677  6691768  6822844 13003464 13151961 17264332 17810546
##        2        2        2        2        2        2        2        2
```

```
ind_cel <- unique(dups_cel$SNP_ID_CURRENT) #ID numbers of duplicate SNPs
print(nrow_cel <- length(ind_cel)) #total number of SNPs with multiple entries
```

```
## [1] 16
```

```
# calculating Fisher's p-value for duplicate SNPs
for (i in 1:nrow_cel) {
  chisq <- (-2)*sum(log(celiac1$P.VALUE[celiac1$SNP_ID_CURRENT==ind_cel[i]]))
  df <- 2*length(which(celiac1$SNP_ID_CURRENT==ind_cel[i]))
  pval <- pchisq(chisq, df, lower.tail=FALSE)
  celiac1$p_fish[celiac1$SNP_ID_CURRENT==ind_cel[i]] <- pval
}
```

The "Celiac disease" dataset contained 16 SNPs with multiple entries from different studies. The table shows that each duplicate SNP has two entries.

Crohn's Disease dataset

```
dups_cro <- crohns1[duplicated(crohns1$SNP_ID_CURRENT)|duplicated(crohns1$SNP_ID_CURRENT,
                                                                  fromLast=TRUE),]
table(dups_cro$SNP_ID_CURRENT) #frequency of duplicate SNPs
```

```
##
##    6596   17119   26528  212388  224136  259964  395157  516246
##       3       2       2       5       2       3       3       3
##  559928  568617  653178  724016  921720  925255 1042058 1049526
##       3       2       2       2       2       3       2       3
## 1142287 1250550 1260326 1292053 1363907 1456896 1569328 1748195
##       2       2       2       2       2       3       3       2
## 1819333 1819658 1847472 1893217 2024092 2062305 2066847 2076756
##       2       2       4       2       4       2       3       5
## 2188962 2227551 2241880 2284553 2301436 2413583 2476601 2538470
##       3       2       3       4       3       3       4       2
## 2542151 2581828 2823286 2836878 2872507 2930047 2945412 3024505
##       3       2       2       2       2       2       2       3
## 3091315 3091316 3197999 3749171 3764147 3766606 3792109 3853824
##       2       2       5       2       4       3       2       3
## 4077515 4243971 4246905 4409764 4656958 4703855 4802307 4845604
##       2       2       2       4       2       2       4       3
##  5743289  5763767  6062496  6425143  6561151  6651252  6679677  6716753
##        2        2        2        2        2        5        2        2
##  6738825  6863411  6908425  7015630  7097656  7236492  7282490  7517810
##        2        3        3        2        2        2        2        2
##  7517847  7554511  7555082  7556897  7608910  7702331  7746082  7954567
##        3        3        2        2        3        2        3        2
##  8005161  9264942  9271366  9286879  9292777  9297145  9491697  9491891
##        3        3        2        2        2        2        2        3
##  9858542 10045431 10065637 10486483 10495903 10758669 10761659 10775412
##        2        2        2        2        3        3        5        2
## 10781499 10865331 10883365 10995271 11195128 11209026 11229555 11230563
##        2        2        2        2        2        5        2        2
## 11236797 11465804 11681525 11741861 11742570 11879191 11924265 12718244
##        2        2        2        2        4        2        2        2
## 12720356 12942547 12946510 13126505 13333062 13407913 16967103 17293632
##        3        3        2        2        3        3        3        4
## 17391694 17622378 17694108 34687326 34779708 34804116 35320439 56116661
##        2        2        2        2        2        3        2        2
## 56167332 61839660 71559680 71624119 75900472 76418789
##        4        2        3        2        2        2
```

```
ind_cro <- unique(dups_cro$SNP_ID_CURRENT) #ID numbers of duplicate SNPs
print(nrow_cro <- length(ind_cro)) #total number of SNPs with multiple entries
```

```
## [1] 142
```

```
# calculating Fisher's p-value for duplicate SNPs
for (i in 1:nrow_cro) {
  chisq <- (-2)*sum(log(crohns1$P.VALUE[crohns1$SNP_ID_CURRENT==ind_cro[i]]))
  df <- 2*length(which(crohns1$SNP_ID_CURRENT==ind_cro[i]))
  pval <- pchisq(chisq, df, lower.tail=FALSE)
  crohns1$p_fish[crohns1$SNP_ID_CURRENT==ind_cro[i]] <- pval
}
```

The dataset for “Crohn’s disease” contained 142 SNPs with multiple entries from different studies. The number of entries varied from 2 to 5.

COMBINING P-VALUES USING FIXED EFFECTS METHOD

An alternative way to compute meta p-values is to use the “Fixed effects” meta-analysis model. This model assumes that the different “betas” (effect sizes) from each study are approximations of a single common “beta”, and that variation arises from the sampling variation of each study.

In the case where different studies have differing “true” betas, indicating inhomogeneity of studies, then the “random effects” model for meta-analysis can be implemented. The homogeneity of the samples is tested using the Cochran’s Q test.

For our datasets, the odds ratio was reported as the “effect size”. However, not all of the studies reported the effect size.

```
sum(is.na(psc1$OR.or.BETA))
```

```
## [1] 251
```

```
sum(is.na(celiac1$OR.or.BETA))
```

```
## [1] 63
```

```r
sum(is.na(crohns1$OR.or.BETA))
```

```
## [1] 463
```

The "sclerosing cholangitis" dataset had 251 missing values for the Odds ratio, the "Celiac" dataset had 63 missing values, and the "Crohns" dataset had 463 missing values. Therefore, the fixed effects method was not used to get an estimate of the combined p-value for all the SNPs.

To get a sampling of the fixed effects method, we will compute p-values for a couple of select SNPs and compare these to the p-values computed by Fisher's method.

The following function in R can be used to calculate the fixed effects p-values. (This function was taken from the "Week 13" notes of the Statistical Analysis of Genetics Data course.)

```r
meta=function(betahat,se){
  S=length(betahat)
  # Below considers fixed effects
  w=1/se^2
  betahat.fixed=sum(w*betahat)/sum(w)
  se.betahat.fixed=1/sqrt(sum(w))
  z.betahat.fixed=betahat.fixed/se.betahat.fixed
  Q=sum(w*(betahat-betahat.fixed)^2)
  pval=1-pchisq(Q,S-1)
  # Below considers random effects
  wbar=mean(w)
  sw2=var(w)
  U=(S-1)*(wbar-sw2/sum(w))
  if (Q<=S-1){sigmabeta2=0}
  else {sigmabeta2=(Q-(S-1))/U}
  wstar=1/(sigmabeta2+1/w)
  muhat.random=sum(wstar*betahat)/sum(wstar)
  se.muhat.random=1/sqrt(sum(wstar))
  z.random=muhat.random/se.muhat.random
  return(list(betahat.fixed=betahat.fixed,se.betahat.fixed=se.betahat.fixed,
              z.betahat.fixed=z.betahat.fixed,Q=Q,pval=pval,
              muhat.random=muhat.random,se.muhat.random=se.muhat.random,
              z.random=z.random))
}
```

SNP rs3197999 (Sclerosing Cholangitis)

```r
# extract odds ratios and CIs
# print(or <- psc1$OR.or.BETA[psc1$SNP_ID_CURRENT==3197999])
# print(CI <- psc1$X95..CI..TEXT.[psc1$SNP_ID_CURRENT==3197999])

betahat.psc <- log(c(1.39, 1.33, 1.33))  #taking log of odds ratios to get betas
upper.psc <- log(c(1.56, 1.40, 1.40))
lower.psc <- log(c(1.24, 1.26, 1.26))
se.psc <- (upper.psc - lower.psc)/(2*1.96) #getting standard error from CIs

result.psc <- meta(betahat.psc,se.psc)
print(result.psc)
```

```
## $betahat.fixed
## [1] 0.2893831
##
```

```
## $se.betahat.fixed
## [1] 0.01807733
##
## $z.betahat.fixed
## [1] 16.00806
##
## $Q
## [1] 0.5135771
##
## $pval
## [1] 0.7735317
##
## $muhat.random
## [1] 0.2893831
##
## $se.muhat.random
## [1] 0.01807733
##
## $z.random
## [1] 16.00806
```

```
OR.fixed=exp(result.psc$betahat.fixed)
OR.CI=exp(c(result.psc$betahat.fixed-1.96*result.psc$se.betahat.fixed,
            result.psc$betahat.fixed+1.96*result.psc$se.betahat.fixed))

print(OR.fixed)
```

```
## [1] 1.335603
```

```
print(OR.CI)
```

```
## [1] 1.289109 1.383774
```

```
2*pnorm(-abs(result.psc$z.betahat.fixed))
```

```
## [1] 1.1225e-57
```

```
2*pnorm(-abs(result.psc$z.random))
```

```
## [1] 1.1225e-57
```

For the rs3197999 SNP, the combined odds-ratio is 1.336 with a 95% CI of [1.29-1.38]. The Cochran Q test for homogeneity of the samples gave a pval » 0.05, indicating that the samples are homogeneous; therefore, the fixed effects model is suitable. The fixed effects p-value is 1.1225e-57.

We can compare this to the p-value obtained by Fisher's method.

```
psc1$p_fish[psc1$SNP_ID_CURRENT==3197999]
```

```
## [1] 2.560327e-115 2.560327e-115 2.560327e-115 2.560327e-115
```

The Fisher's p-value is much smaller. However, for the Fisher's method, we used 4 data points whereas for the Fixed effects method, we only had 3 data points. This may partly account for the difference.

SNP rs6651252 (Crohn's Disease)

```
# get odds ratios and CIs
print(or <- crohns1$OR.or.BETA[crohns1$SNP_ID_CURRENT==6651252])
```

```
## [1]       NA 1.160706 1.230000 1.185000       NA
```

```r
print(CI <- crohns1$X95..CI..TEXT.[crohns1$SNP_ID_CURRENT==6651252])
```

```
## [1] ""              "[1.12-1.2]"    "[1.17-1.30]"   "[1.128-1.246]"
## [5] ""
```

```r
betahat.cro1 <- log(c(1.16, 1.23, 1.185))  #taking log of odds ratios to get betas
upper.cro1 <- log(c(1.2, 1.3, 1.246))
lower.cro1 <- log(c(1.12, 1.17, 1.128))
se.cro1 <- (upper.psc - lower.psc)/(2*1.96) #getting standard error from CIs

result.cro1 <- meta(betahat.cro1,se.cro1)
print(result.cro1)
```

```
## $betahat.fixed
## [1] 0.1845713
##
## $se.betahat.fixed
## [1] 0.01807733
##
## $z.betahat.fixed
## [1] 10.2101
##
## $Q
## [1] 1.382645
##
## $pval
## [1] 0.5009133
##
## $muhat.random
## [1] 0.1845713
##
## $se.muhat.random
## [1] 0.01807733
##
## $z.random
## [1] 10.2101
```

```r
OR.fixed=exp(result.cro1$betahat.fixed)
OR.CI=exp(c(result.cro1$betahat.fixed-1.96*result.cro1$se.betahat.fixed,
            result.cro1$betahat.fixed+1.96*result.cro1$se.betahat.fixed))

print(OR.fixed)
```

```
## [1] 1.202703
```

```r
print(OR.CI)
```

```
## [1] 1.160835 1.246080
```

```r
2*pnorm(-abs(result.cro1$z.betahat.fixed))
```

```
## [1] 1.786878e-24
```

```r
2*pnorm(-abs(result.cro1$z.random))
```

```
## [1] 1.786878e-24
```

```
crohns1$p_fish[crohns1$SNP_ID_CURRENT==6651252]
```

```
## [1] 1.124114e-68 1.124114e-68 1.124114e-68 1.124114e-68 1.124114e-68
```

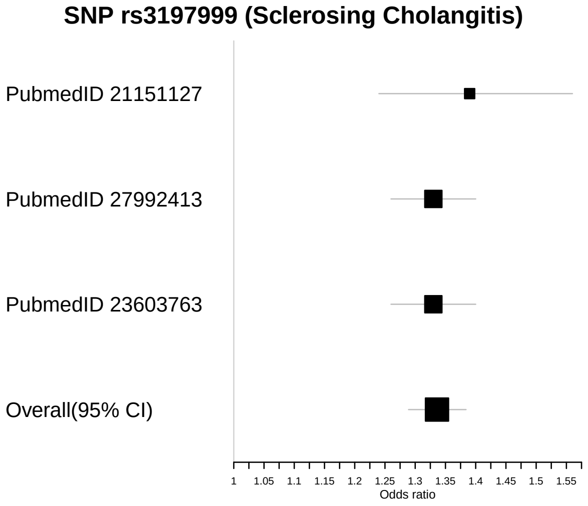
For the rs6651252 SNP from the Crohn's dataset, the combined odds-ratio was 1.184 with a 95% CI of [1.16-1.25]. The Cochran Q test gave a pval » 0.05, therefore, the samples were homogeneous and the fixed effects estimation holds. The fixed effects p-value is 1.787e-24, compared to the Fisher's p-value of 1.124e-68.

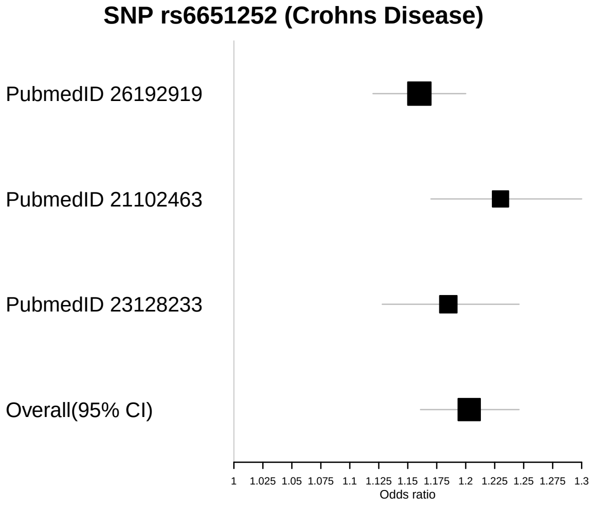
One factor that may account for this difference is that to calculate the Fisher's p-value, 5 data points were combined. However, for the Fixed effects method, we had some missing values so only 3 data points were used.

FOREST PLOTS

As an example of forest plots, we can look at the two SNPs for which the Fixed effect p-value calculations were done. These plots were generated using the "forestplot" package from CRAN. The mean odds ratio and overall 95% confidence interval is displayed at the bottom. The y-axis shows the Pubmed article IDs for the different studies from which the odds ratios were extracted.

```
pubmed <- psc1$PUBMEDID[psc1$SNP_ID_CURRENT==3197999]
row_names <- list(c(paste("PubmedID", pubmed[-4]), "Overall(95% CI)"))
point_psc <- c(1.39, 1.33, 1.33, 1.336)
high_psc <- c(1.56, 1.40, 1.40, 1.384)
low_psc <- c(1.24, 1.26, 1.26, 1.289)

grid.newpage()
pushViewport(viewport(layout = grid.layout(1, 2)))
pushViewport(viewport(layout.pos.col = 1))

forestplot(row_names, point_psc,low_psc, high_psc, zero = 1, cex = 2, lineheight = "auto",
xlab = "Odds ratio", title="SNP rs3197999 (Sclerosing Cholangitis)", new_page=FALSE)

pubmed2 <- crohns1$PUBMEDID[crohns1$SNP_ID_CURRENT==6651252]
row_names2 <- list(c(paste("PubmedID", pubmed2[2:4]), "Overall(95% CI)"))
or.cro1 <- c(1.16, 1.23, 1.185, 1.203)
high.cro1 <- c(1.2, 1.3, 1.246, 1.246)
low.cro1 <- c(1.12, 1.17, 1.128, 1.161)

popViewport()
pushViewport(viewport(layout.pos.col = 2))
forestplot(row_names2, or.cro1, low.cro1, high.cro1, zero = 1,cex = 2, lineheight = "auto",
xlab = "Odds ratio", title="SNP rs6651252 (Crohns Disease)", new_page = FALSE)
popViewport(2)
```


SNP rs3197999 (Sclerosing Cholangitis)


SNP rs6651252 (Crohns Disease)

The plots show that all of the odds ratios reported in the different studies are significant because the 95% confidence intervals do not include one.

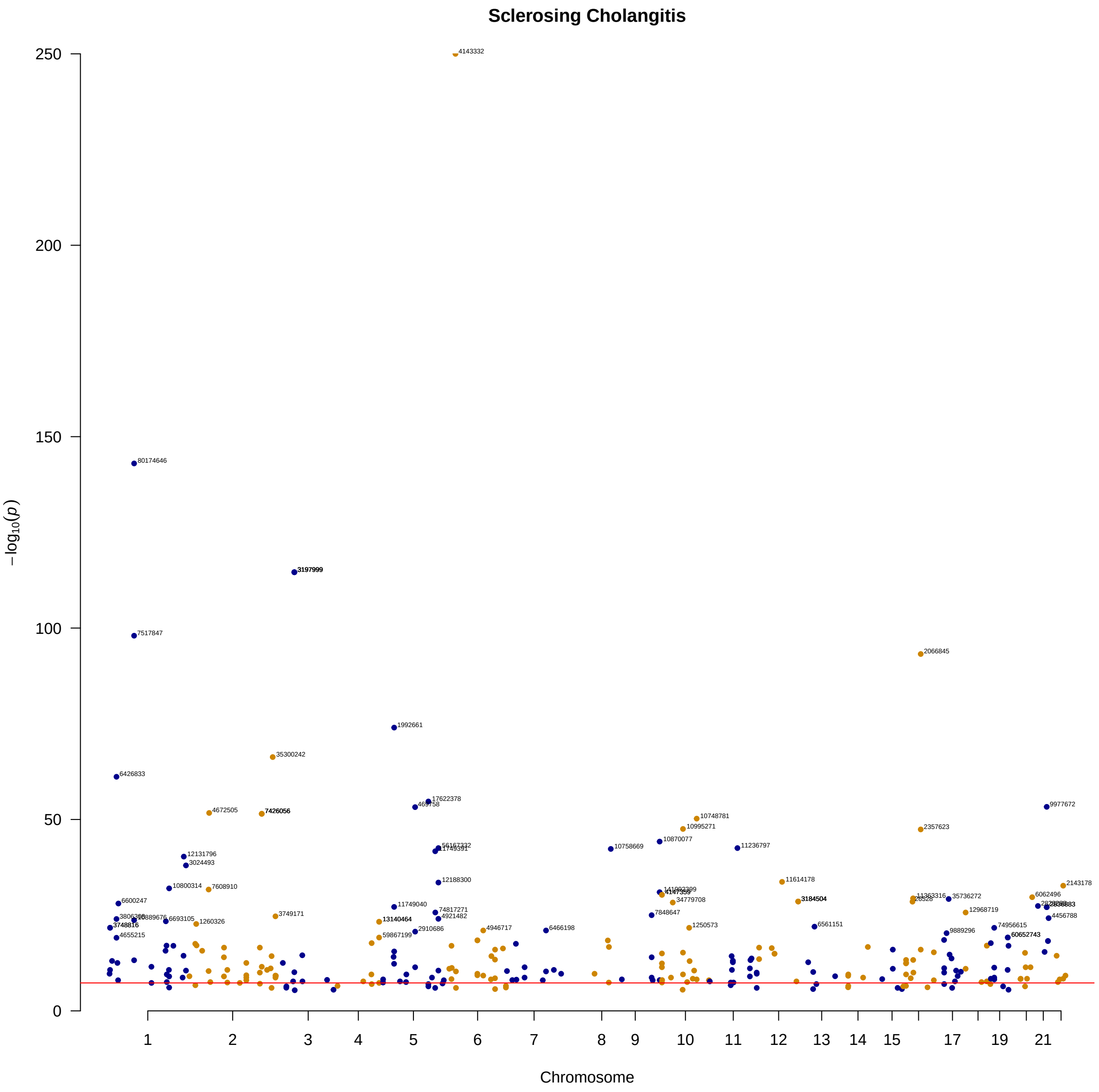
MANHATTAN PLOTS

Before creating Manhattan plots, we need to remove any rows that have missing values for chromosome ID or chromosome position (base pair).

```
psc2 <- psc1[!is.na(psc1$CHR_ID),]
psc2 <- psc1[!is.na(psc1$CHR_POS),]
celiac2 <- celiac1[!is.na(celiac1$CHR_ID),]
celiac2 <- celiac1[!is.na(celiac1$CHR_POS),]
crohns2 <- crohns1[!is.na(crohns1$CHR_ID),]
crohns2 <- crohns1[!is.na(crohns1$CHR_POS),]
```

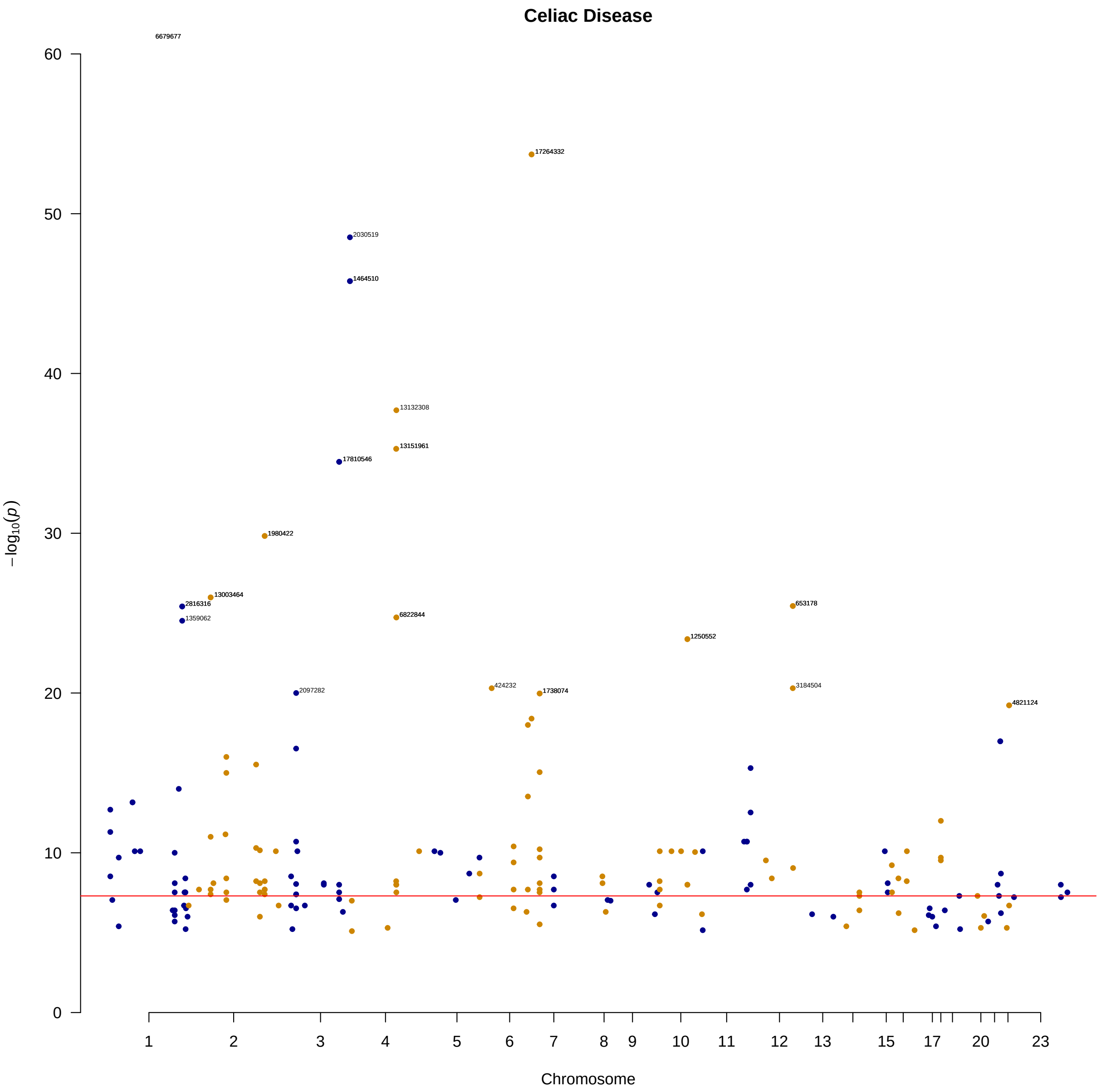
The Manhattan plots were generated by the "qqman" package from CRAN. In the plots below, the red line indicates the significance p-value threshold for GWAS, which is 5 × 10—8. All SNPs that have a p-value < 10e-20 have been annotated by their SNP ID number. This plot displays the SNPs with reference to their position on the chromosomes (along the x-axis). The y-axis indicates the p-value in -log base 10, therefore, smaller p-values appear larger. The plot also shows correlations between SNPs located in the same regions. If there is linkage disequilibrium between a pair of SNPs, then if one of them is statistically significant, the other will also likely be significant.

```
manhattan(psc2, main="Sclerosing Cholangitis", chr="CHR_ID", bp="CHR_POS",
          snp="SNP_ID_CURRENT", p="p_fish",
          col = c("blue4", "orange3"), suggestiveline=FALSE,
          annotatePval = 10e-20, ylim=c(0,250), annotateTop = FALSE)
```



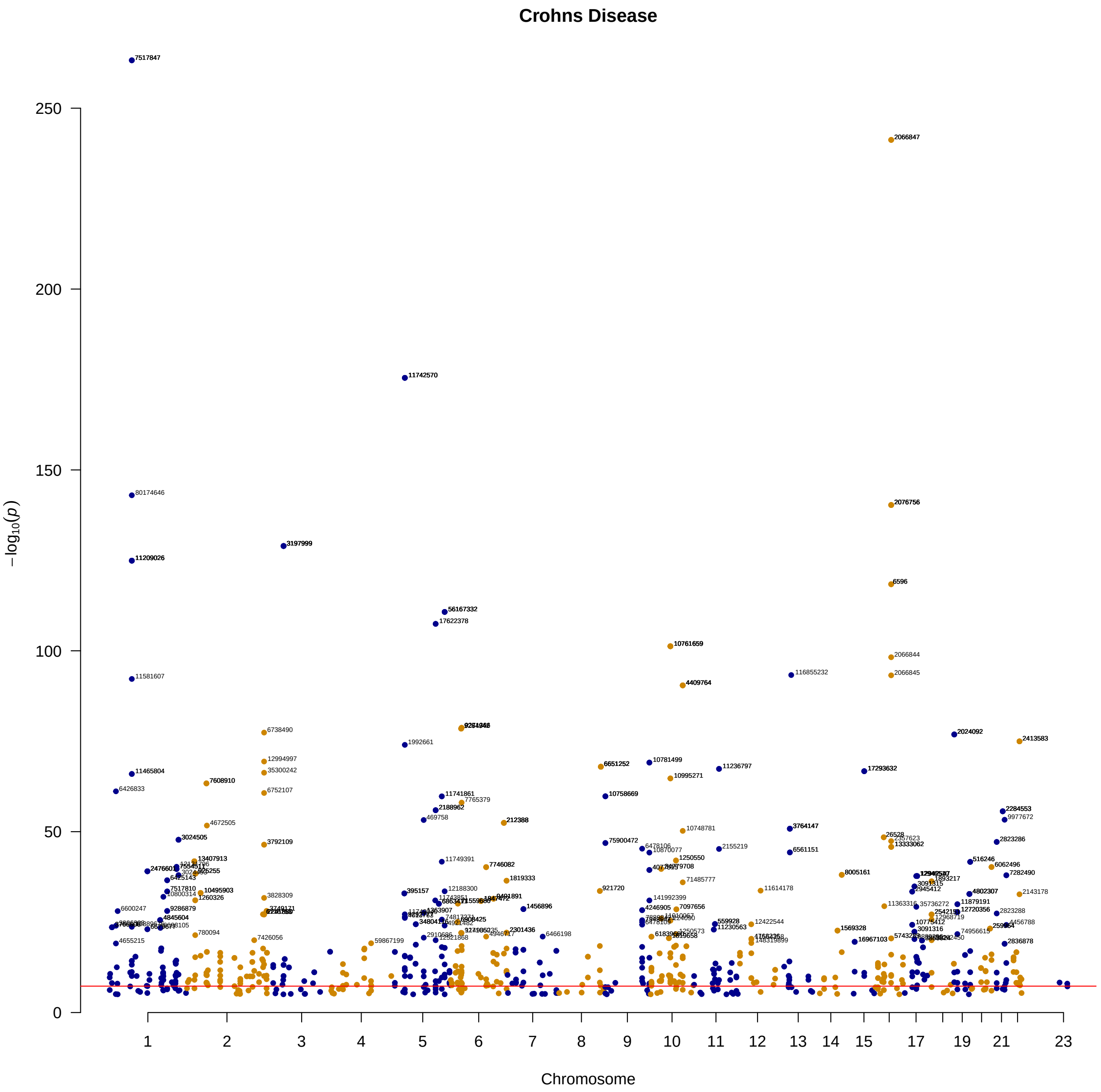
Almost all of the SNPs crossed the GWAS threshold of 5*10e-8, and their p-values were very small, indicating a strong association between variation of that SNP and sclerosing cholangitis. Chromosomes 1, 3, 5, 10 and 17 showed regions of correlated SNPs.

```
celiac2$CHR_ID[celiac2$CHR_ID=="X"] <- 23
celiac2$CHR_ID <- as.numeric(celiac2$CHR_ID)
manhattan(celiac2, main="Celiac Disease", chr="CHR_ID", bp="CHR_POS", snp="SNP_ID_CURRENT",
          p="p_fish", col = c("blue4", "orange3"),suggestiveline=FALSE,
          annotatePval = 10e-20, ylim = c(0, 60), annotateTop = FALSE)
```



Most of the p-values were very small. Chromosomes 2, 5, and 7 showed correlated SNPs.

```r
crohns2$CHR_ID[crohns2$CHR_ID=="X"] <- 23
crohns2$CHR_ID <- as.numeric(crohns2$CHR_ID)
crohns2$CHR_POS <- as.numeric(crohns2$CHR_POS)
crohns2 <- crohns2[-642,]  #missing value
manhattan(crohns2, main="Crohns Disease", chr="CHR_ID", bp="CHR_POS", snp="SNP_ID_CURRENT",
          p="p_fish", col = c("blue4", "orange3"), ylim=c(0,265), suggestiveline=FALSE,
          annotatePval = 10e-20, annotateTop = FALSE)
```

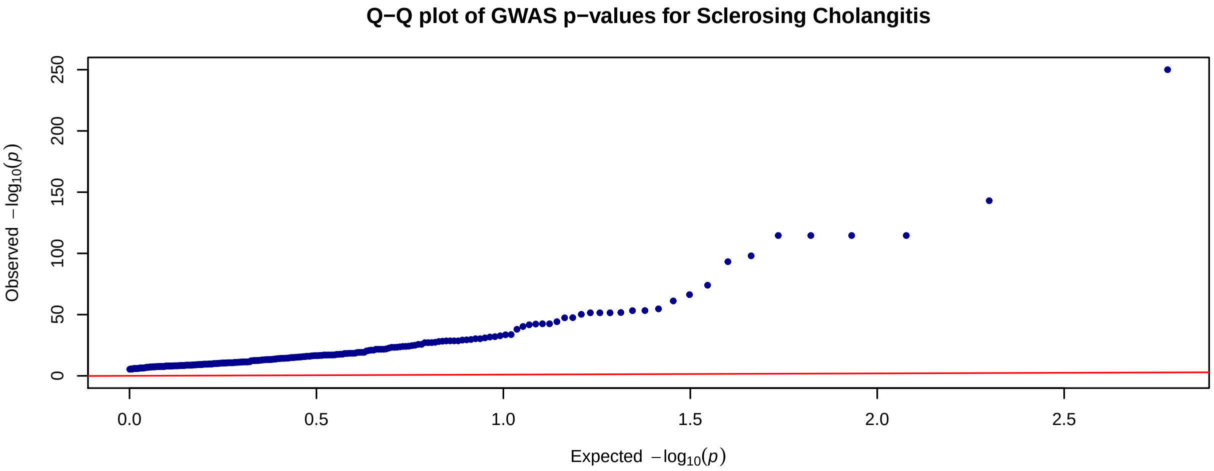


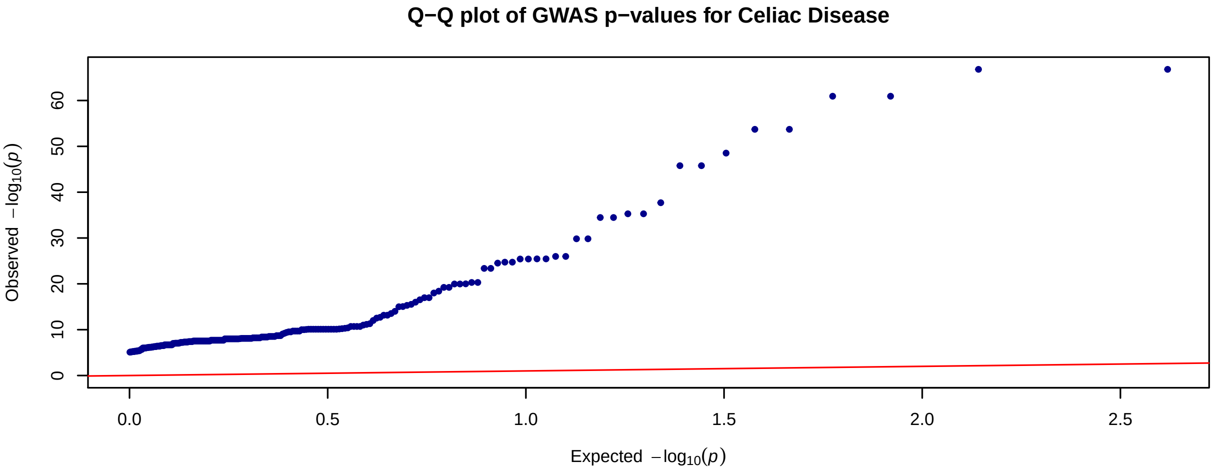
The Crohn's dataset showed many "highrises", indicating correlated SNPs. Most of the SNPs are highly significant.

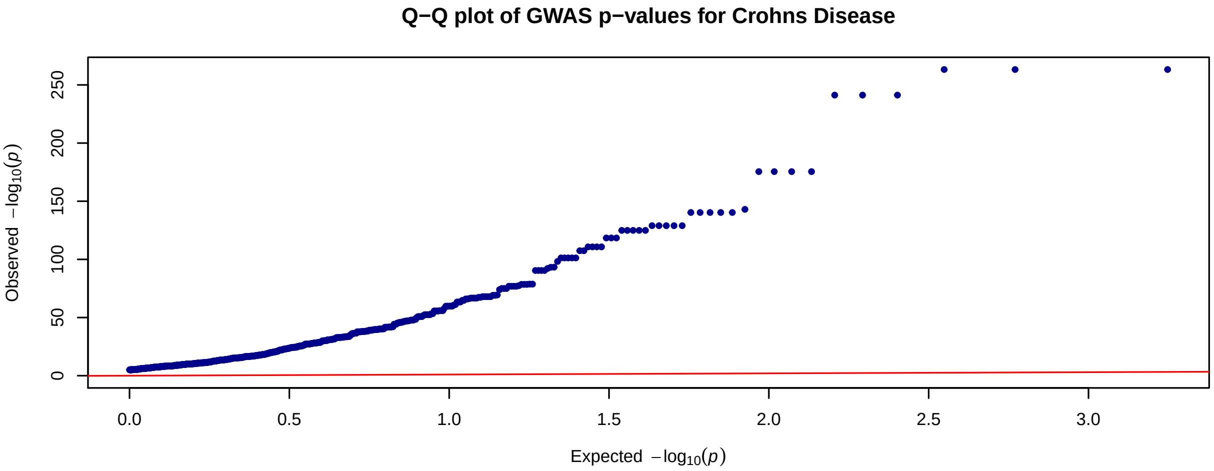
Q-Q PLOTS

The QQ plot graphically depicts the deviation of the observed p values from the null hypothesis. These were also generated by the "qqman" package.

```r
par(mfrow=c(3,1))
qq(psc2$p_fish, main = "Q-Q plot of GWAS p-values for Sclerosing Cholangitis", col = "blue4")
qq(celiac2$p_fish, main = "Q-Q plot of GWAS p-values for Celiac Disease", col = "blue4")
qq(crohns2$p_fish, main = "Q-Q plot of GWAS p-values for Crohns Disease", col = "blue4")
```







In the QQ plots above, we see that there is great deviation from the expected line. This indicates that most of the p values are highly significant. Since there is a noticeable separation of the expected line and the observed values, this can also mean that many of the p values are inflated (much smaller than expected) due to allele frequencies being systematically different between subpopulations of the total sample.

TOP 30 SNPs

We can rank the SNPs based on the p-values (Fisher's p-value for multiple entries), and take a look at some of the genes associated with the most significant SNPs.

Sclerosing Cholangitis

```r
top_psc <- psc2[order(psc2$p_fish),]
top_psc[1:30, c("SNP_ID_CURRENT", "P.VALUE", "p_fish", "MAPPED_GENE")]
```

```
##     SNP_ID_CURRENT P.VALUE        p_fish             MAPPED_GENE
## 58         4143332  1e-250 1.000000e-250               ZDHHC20P2
## 94        80174646  1e-143 1.000000e-143                   IL23R
## 3          3197999   1e-16 2.560327e-115                    MST1
## 12         3197999   5e-26 2.560327e-115                    MST1
## 46         3197999   2e-26 2.560327e-115                    MST1
## 198        3197999   7e-55 2.560327e-115                    MST1
## 121        7517847   1e-98  1.000000e-98         IL23R, C1orf141
## 273        2066845   6e-94  6.000000e-94                    NOD2
## 209        1992661   1e-74  1.000000e-74 AC108105.1 - AC093277.1
## 191       35300242   5e-67  5.000000e-67                 ATG16L1
## 90         6426833   7e-62  7.000000e-62      AL391883.1 - OTUD3
## 69        17622378   2e-55  2.000000e-55     C5orf56, AC116366.3
## 244        9977672   5e-54  5.000000e-54  AF064858.1 - RPL23AP12
## 66          469758   6e-54  6.000000e-54                   ERAP1
## 98         4672505   2e-52  2.000000e-52   RN7SL51P - AC093159.2
## 11         7426056   2e-16  3.305656e-52         CD28 - KRT18P39
## 45         7426056   2e-20  3.305656e-52         CD28 - KRT18P39
## 186        7426056   1e-20  3.305656e-52         CD28 - KRT18P39
## 159       10748781   6e-51  6.000000e-51  AL391684.1 - LINC01475
## 152       10995271   3e-48  3.000000e-48 AC024598.1 - AC067751.1
## 274        2357623   4e-48  4.000000e-48       NKD1 - AC007608.3
## 142       10870077   6e-45  6.000000e-45                   CARD9
## 75        56167332   3e-43  3.000000e-43              AC008691.1
## 286       11236797   3e-43  3.000000e-43       EMSY - AP001189.2
## 136       10758669   5e-43  5.000000e-43       HNRNPA1P41 - JAK2
## 73        11749391   2e-42  2.000000e-42                    IRGM
## 113       12131796   5e-41  5.000000e-41                   INAVA
## 114        3024493   1e-38  1.000000e-38                    IL10
## 297       11614178   2e-34  2.000000e-34                IFNG-AS1
## 76        12188300   3e-34  3.000000e-34              AC008691.1
```

rs4143332

This SNP is located on the "zinc finger DHHC-type containing 20 pseudogene 2" (ZDHHC20P2) gene, which is also associated with type 2 diabetes (GWAS catalog).

rs80174646

The SNP rs80174646 is part of the gene that encodes for the interleukin 23 (IL-23) receptor, which is found on the outer cell membranes of several types of immune system cells, such as T cells and natural killer cells. Upon binding of interleukin 23 to the IL-23 receptor, a cascade of signals in the inflammatory response pathway are triggered. Therefore, the rs80174646 is associated with immune reponse (NIH, 2017).

rs3197999

This SNP is part of the “macrophage stimulating 1” (MST1) gene, and is a known variant for primary sclerosing cholangitis. It was found that the [AA] genotype of this SNP increased the genetic risk of sporadic extrahepatic cholangiocarcinoma (Krawczyk et al, 2013).

rs2066845

This SNP is located on the nucleotide-binding oligomerization domain 2 (NOD2) gene. The protein encoded by this gene is an intracellular receptor for bacterial products. In the normal type, when this receptor is activated, it inhibits the signalling from another receptor in the inflammation pathway. If this gene carries a mutation, then that ultimately leads uncontrolled inflammation of the gut. Therefore, the NOD2 gene is known to be associated with Crohn’s Disease (Rhodes, 2006).

Celiac Disease

```
top_cel <- celiac2[order(celiac2$p_fish),]
top_cel[1:30, c("SNP_ID_CURRENT","P.VALUE", "p_fish", "MAPPED_GENE")]
```

```
##     SNP_ID_CURRENT P.VALUE       p_fish        MAPPED_GENE
## 47         2187668   1e-19 1.598784e-67           HLA-DQA1
## 117        2187668   1e-50 1.598784e-67           HLA-DQA1
## 79         6679677   1e-53 1.170288e-61      PHTF1 - RSBN1
## 167        6679677   8e-11 1.170288e-61      PHTF1 - RSBN1
## 70        17264332   3e-27 1.943090e-54         AL356234.2
## 146       17264332   5e-30 1.943090e-54         AL356234.2
## 141        2030519   3e-49 3.000000e-49                LPP
## 52         1464510   5e-09 1.666779e-46                LPP
## 85         1464510   3e-40 1.666779e-46                LPP
## 142       13132308   2e-38 2.000000e-38           IL21-AS1
## 21        13151961   3e-11 5.202388e-36           KIAA1109
## 116       13151961   2e-27 5.202388e-36           KIAA1109
## 50        17810546   1e-09 3.392374e-35          IL12A-AS1
## 115       17810546   4e-28 3.392374e-35          IL12A-AS1
## 68         1980422   2e-17 1.479792e-30    CD28 - KRT18P39
## 134        1980422   1e-15 1.479792e-30    CD28 - KRT18P39
## 100       13003464   4e-13 1.040038e-26              PUS10
## 129       13003464   4e-16 1.040038e-26              PUS10
## 55          653178   8e-08 3.569978e-26              ATXN2
## 118         653178   7e-21 3.569978e-26              ATXN2
## 49         2816316   3e-11 3.820837e-26         AL390957.1
## 93         2816316   2e-17 3.820837e-26         AL390957.1
## 48         6822844   1e-14 1.862136e-25         IL2 - IL21
## 56         6822844   3e-13 1.862136e-25         IL2 - IL21
## 127        1359062   3e-25 3.000000e-25         AL390957.1
## 107        1250552   9e-10 4.240305e-24              ZMIZ1
## 150        1250552   8e-17 4.240305e-24              ZMIZ1
## 23          424232   5e-21 5.000000e-21 NOTCH4 - TSBP1-AS1
## 154        3184504   5e-21 5.000000e-21     "ATXN2, SH2B3"
## 136        2097282   1e-20 1.000000e-20    UQCRC2P1 - CCR2
```

rs2187668

This SNP is located on the HLA-DQA1 gene, which is part of a family of genes called human leukocyte antigen (HLA) complex. The gene encodes for proteins on the outer membranes of certain immune cells that help the immune system distinguish the body's own proteins from foreign proteins (NIH, 2003).

rs2030519, rs1464510

These SNPs encode for the "lipoma preferred partner (LPP) gene". Polymorphisms of this gene are associated with celiac disease (Huang at al, 2017).

Crohn's Disease

```
top_cro <- crohns2[order(crohns2$p_fish),]
top_cro[1:30, c("SNP_ID_CURRENT", "P.VALUE", "p_fish", "MAPPED_GENE")]
```

```
##     SNP_ID_CURRENT P.VALUE        p_fish             MAPPED_GENE
## 20         7517847   3e-12 5.752942e-264       "IL23R, C1orf141"
## 447        7517847  1e-159 5.752942e-264       "IL23R, C1orf141"
## 699        7517847   1e-98 5.752942e-264       "IL23R, C1orf141"
## 34         2066847   2e-15 5.816610e-242      "AC007728.2, NOD2"
## 301        2066847   3e-24 5.816610e-242      "AC007728.2, NOD2"
## 588        2066847  6e-209 5.816610e-242      "AC007728.2, NOD2"
## 141       11742570   1e-06 3.490783e-176 AC108105.1 - AC093277.1
## 201       11742570   1e-55 3.490783e-176 AC108105.1 - AC093277.1
## 448       11742570   4e-87 3.490783e-176 AC108105.1 - AC093277.1
## 540       11742570   7e-36 3.490783e-176 AC108105.1 - AC093277.1
## 698       80174646  1e-143 1.000000e-143                   IL23R
## 18         2076756   7e-14 4.916251e-141                    NOD2
## 115        2076756   1e-37 4.916251e-141                    NOD2
## 123        2076756   1e-21 4.916251e-141                    NOD2
## 138        2076756   3e-10 4.916251e-141                    NOD2
## 543        2076756   4e-69 4.916251e-141                    NOD2
## 209        3197999   3e-23 1.071540e-129                    MST1
## 299        3197999   1e-12 1.071540e-129                    MST1
## 368        3197999   2e-33 1.071540e-129                    MST1
## 534        3197999   6e-17 1.071540e-129                    MST1
## 820        3197999   7e-55 1.071540e-129                    MST1
## 1         11209026   4e-21 1.205950e-125                   IL23R
## 44        11209026   2e-18 1.205950e-125                   IL23R
## 104       11209026   1e-18 1.205950e-125                   IL23R
## 139       11209026   4e-14 1.205950e-125                   IL23R
## 517       11209026   1e-64 1.205950e-125                   IL23R
## 158           6596   2e-54 3.878595e-119                   SNX20
## 161           6596   6e-26 3.878595e-119                   SNX20
## 164           6596   8e-45 3.878595e-119                   SNX20
## 129       56167332   9e-08 1.791669e-111              AC008691.1
```

The SNPs rs7517847/80174646, rs2066847/2076756, and rs3197999 correspond to the genes IL23R, NOD2, and MST1, respectively which have also been mentioned in the sclerosing cholangitis section.

Sclerosing Cholangitis & Celiac Disease

The significant p-value threshold for GWAS studies is 5*10e-8. Therefore, we looked at all the SNPs that were statistically significant at this level, and determined the SNPs that were in common between the traits.

```
sig_psc <- psc2[psc2$p_fish < 5*10e-8,]
sig_cel <- celiac2[celiac2$p_fish < 5*10e-8, ]
```

```r
match_psc_cel <- intersect(sig_psc$SNP_ID_CURRENT, sig_cel$SNP_ID_CURRENT)
length(match_psc_cel)
```

```
## [1] 8
```

```r
match_psc_cel
```

```
## [1]  4676410  3748816  3184504  1893592 72928038  6651252 13132308 11221332
```

There were 8 SNPs that were highly significant in both the sclerosing cholangitis and celiac disease datasets.

rs4676410

The SNP rs4676410 is part of region 16p11 near the cytokine gene IL27 which is associated with susceptibility to early-onset inflammatory bowel disease such as Crohn's disease (Imielinski et al, 2009).

rs1893592

The rs1893592 SNP is found on the "Ubiquitin-associated and SH3 domain-containing protein A" (UBASH3A) gene. Variants of this gene have been associated with increased susceptibility to rheumatoid arthritis, a complex autoimmune disorder, in the Han Chinese population (Liu et al, 2017).

Sclerosing Cholangitis & Crohn's Disease

```r
sig_cro <- crohns2[crohns2$p_fish < 5*10e-8, ]
match_psc_cro <- intersect(sig_psc$SNP_ID_CURRENT, sig_cro$SNP_ID_CURRENT)
length(match_psc_cro)
```

```
## [1] 240
```

```r
match_psc_cro
```

```
##   [1]   4676410   3197999   7426056   3184504   1893592  11168249  11749040
##   [8]   9687958    353339  71624119   4703855  34804116    469758   2910686
##  [15]   2549803  17622378  17622517   1004234   6863411  11749391  74817271
##  [22]  56167332  12188300   4921482   6556411  72812861   1267499   2328530
##  [29]    714830  71559680  72928038  34920465   2816958   6697886   2234161
##  [36]   3766606   6426833   3806308   4655215   1260326  80174646  77981966
##  [43]  10889676    702872   4672505  11675538   4845604   6693105   4129267
##  [50]   4971079  35667974   3747517  72871627  17229679   6434978   6425143
##  [57]  16841904   7552167  12131796   3024493  12075255   2666218  13407913
##  [64] 201014116   6600247    925255   7517847   7608910 183686347   2476601
##  [71] 114202211   4851529  12987977    871656  11691685   2111485  78973538
##  [78]   1333062  10800314  61802846   6651252  10758669   2812378   7848647
##  [85]    726657   7468800   4986790  10870077 141992399   3124998  61839660
##  [92]   3118471  76913543   2104286   2236379   2050392  34779708  10995271
##  [99]   7915475   2227551   1250573   7097656   1800682   2497318  10748781
## [106]   1847472   4946717  28701841   9491891    582757    928722   9494840
## [113]   2451258 111305875   1182188   1525735  28550029    860262   4917129
## [120]  12718244   9297145   6466198   7805114   4728142   2538470  10094579
## [127]   1551399   2042011   1405108   5837881  11676348   7556897  12694846
## [134]  35300242   3749171   4676406  35320439  73178598  10510607   1001007
## [141] 116046827   6781808  11098964  13107612   3774937  59867199  13132308
## [148]  11750385   3776414    395157   1992661  28998802   9797244   2779255
## [155]   9889296  35736272  12942547  12943464   3853824   1292035    196941
## [162]  17780256   7236492  12968719  62097857  66504140    587259   2024092
## [169] 72977586 74956615 35018800 12720356 35074907  4802307   679574
## [176]  4243971  6058869  4812833 79493594  1883832  1328454   259964
## [183]  6062496  2823288  2284553  9977672  4456788  2266961   140135
## [190]  2143178  5757584  1569414 10761648 10775412  2026029  9554587
## [197]  2145623  8006884 12879003  1569328 11624293 16967103 17293632
## [204] 35874463   367569 11649613  8061882  7195296  7404095    26528
## [211] 11363316 11574938  1870293  2066845  2357623 72796367 11117431
## [218] 12932970 11190133 10743181 11229555 10750899 11230563   174535
## [225]   559928   568617 11236797  7115956  4561177   661054  7933433
## [232] 11221332  1860545 11616188  7313895 11614178 12369214 17085007
## [239]   941823  6561151
```

There was a large overlap of significant SNPs between the sclerosing cholangitis dataset and Crohn's disease dataset. This may be due to the same studies being included under both traits on the GWAS database.

The roles of the SNPs rs4676410, rs3197999, rs1893592 have been mentioned before.

Celiac Disease & Crohn's Disease

```
match_cel_cro <- intersect(sig_cel$SNP_ID_CURRENT, sig_cro$SNP_ID_CURRENT)
length(match_cel_cro)
```

```
## [1] 51
```

```
match_cel_cro
```

```
##  [1]    653178  10188217    212388  72928038   6651252   1893592   6679677
##  [8]   1893217  13003464  11221332  13132308   3184504  11580078  34884278
## [15]   6689858   2075184  36001488   4676410      4625  62324212   7725052
## [22]   7731626   4869313  11741255    755374  36051895   4246905  11145763
## [29]    706778  10822050   1250563   1332099  17885785  17466626   1689510
## [36]  72743477  12598357 117372389  12232497  62131887    602662   2836882
## [43]   2066363 114846446   7672495   7660520   7042370   7100025  77150043
## [50]   2807264  12863738
```

There were 51 significant SNPs that were common to Celiac disease and Crohn's disease.

rs72928038

The SNP rs72928038 is located on the "BTB Domain And CNC Homolog 2" (BACH2) gene, which is a transcription factor expressed in B and T lymphocytes. Many autoimmune disorders are associated with genetic variants in the BACH2 gene, including multiple sclerosis, rheumatoid arthritis, inflammatory bowel disease and type I diabetes (Yang et al, 2019).

CONCLUSIONS

We looked at the summary statistics from GWAS to study SNPs that were significantly associated with Primary Sclerosing Cholangitis, Celiac Disease and Crohn's Disease. Since the datasets included results from different studies, some SNPs had multiple entries. For duplicate SNPs, the meta p-values were calculated using the Fisher's method instead of the Fixed effects meta-analysis model because most studies did not report a value for the "effect size" (odds ratio/beta).

The Manhattan plots and QQ plots indicated that most of the SNPs crossed the p-value significance threhold for GWAS which is 5*10e-8. In fact, the overall p-values were very small, probably inflated due to some population stratification, with allele frequencies being different between subpopulations. The Manhattan plots showed regions with correlated SNPs.

For the sclerosing cholangitis dataset, we investigated a few of the highly significant SNPs: rs4143332, rs80174646, rs3197999, and rs2066845. These SNPs were located on genes associated with type 2 diabetes (ZDHHC20P2 gene), the inflammatory response pathway (IL23R gene), primary sclerosing cholangitis and cholangiocarcinoma (MST1 gene), and Crohn's Disease (NOD2 gene), respectively.

A few of the highly significant SNPs for Celiac disease were rs2187668 (located on the HLA-DQA1 gene, involved in immune response), and rs2030519/rs1464510 (located on the LPP gene, associated with Celiac disease).

Some of the notable SNPs for Crohn's disease were rs7517847/rs80174646 (IL23R gene, inflammatory response), rs2066847/rs2076756 (NOD2 gene, Crohn's disease), and rs3197999 (MST1 gene, PSC). These were the same SNPs as in sclerosing cholangitis.

For each disease, we subsetted all the SNPs that were significant at the GWAS threshold of 5*10e-8, and we examined how many of these SNPs overlapped between the diseases.

There were 8 significant SNPs that were in common between PSC and Celiac disease. These included the SNP rs4676410, which is located near the IL27 gene and is associated with susceptibility to early-onset inflammatory bowel disease, and the SNP rs1893592, located on the UBASH3A gene, associated with rheumatoid arthritis.

Between the sclerosing cholangitis dataset and the Crohn's disease dataset, 240 significant SNPs overlapped. One reason for this large number may be that the same studies were included under both traits from the GWAS database. Some of these SNPs (rs4676410, rs3197999, rs1893592) were associated with Celiac disease, PSC, and rheumatoid arthritis.

There were 51 SNPs that were in common between Celiac disease and Crohn's disease. One of the SNPs was rs72928038, which is located on the BACH2 gene and is associated with a myriad of autoimmune disorders including multiple sclerosis, rheumatoid arthritis, and inflammatory bowel disease.

From this analysis, we concluded that since the traits of sclerosing cholangitis, Celiac disease and Crohn's disease have a genetic component, it is understandable that we would find many SNPs that are signficantly associated with these conditions. Also, investigating some of the specific SNPs and their mapped genes, we saw that all of SNPs/genes were associated with the general inflammatory response or with other autoimmune disorders. Therefore, there is an underlying genetic link between these diseases.

Since most people with PSC have inflammatory bowel disease, we can look at SNPs such as rs4676410 (susceptibility to early-onset IBD) and rs2066845 (associated with Crohn's disease) to assess the risk for developing PSC. Likewise, the risk for cholangiocarcinoma (bile duct cancer) may be associated with the SNP rs3197999, located in the MSTI gene. Cholangiocarcinoma is often diagnosed in the final stages and has very poor prognosis and life-expectancy. Therefore, timely screening may be made available to those with the risk variant at this SNP.

One of the primary limitations of this study was that the risk alleles for each SNP were not reported for the vast majority of the observations. Therefore, we can only state that a certain SNP is associated with a disease but cannot state to which allele the risk is attributed.

REFERENCES

Celiac Disease Foundation. What is Celiac Disease? https://celiac.org/about-celiac-disease/what-is-celiac-disease/

"forestplot" package vignette: https://cran.r-project.org/web/packages/forestplot/forestplot.pdf

GWAS catalog: ZDHHC20P2. https://www.ebi.ac.uk/gwas/genes/ZDHHC20P2.

Huang S, Zhang N, Zhou Z, et al. Association of LPP and TAGAP Polymorphisms with Celiac Disease Risk: A Meta-Analysis. *Int J Environ Res Public Health.* 2017 Feb; 14(2): 171.

Imielinski M, Baldassano RN, Griffiths A, et al. Common variants at five new loci associated with early-onset inflammatory bowel disease. *Nat Genet.* 2009 Dec; 41(12): 1335–1340.

Krawczyk M, Höblinger A, Mihalache F, et al. Macrophage stimulating protein variation enhances the risk of sporadic extrahepatic cholangiocarcinoma. *Dig Liver Dis.* 2013 Jul;45(7):612-5.

Liu D, Liu J, Cui G, et al. Evaluation of the association of UBASH3A and SYNGR1 with rheumatoid arthritis and disease activity and severity in Han Chinese. *Oncotarget.* 2017; 8:103385-103392.

Mayo Clinic. Primary Sclerosing Cholangitis. https://www.mayoclinic.org/diseases-conditions/primary-sclerosing-cholangitis/symptoms-causes/syc-20355797.

NIH: Genetics Home Reference. HLA-DQA1 gene. https://ghr.nlm.nih.gov/gene/HLA-DQA1. March 2013.

NIH: Genetics Home Reference. IL23R gene. https://ghr.nlm.nih.gov/gene/IL23R. December 2017.

"qqman" package vignette: https://cran.r-project.org/web/packages/qqman/vignettes/qqman.html

Rhodes D. Primary Sclerosing Cholangitis Literature: Inflammatory Bowel Disease Genetics (Part 1). http://www.psc-literature.org/IBDarticle1.htm. 05/16/06.

Yang L, Chen S, Zhao Q, et al. The Critical Role of Bach2 in Shaping the Balance between CD4+ T Cell Subsets in Immune-Mediated Diseases. *Mediators Inflamm.* 2019 Dec 30;2019.