GWAS analysis for alcohol frequency intake

Carla Casanova Suárez

Descriptive analysis of phenotype and covariates

The current work is focused on the analysis of genetic factors influencing alcohol intake. Alcohol consumption has been linked to over 200 diseases and it is responsible for over 5% of the global disease burden, however in 2012 roughly a quarter of all the deaths was in 20-39-year-age group (Clarke *et al.*, 2017). Most of the GWAS analysis performed are focused on the Alcohol consumption level and alcohol use disorder (AUD) diagnosis and the Alcohol Use Disorder Identification Test-Consumption (AUDIT-C) scores. Moreover, twin and adoption studies have shown that half of the risk of alcohol dependence, a subtype of AUD, is heritable (Kranzler *et al.*, 2019). Nevertheless, there is also a substantial genetic component to the variation in alcohol consumption since approximately 18% of the variance in alcohol consumption is attributable to common SNPs (Clarke *et al.*, 2017). Considering that the dataset analyzed does not have available data about AUD diagnosis or AUDIT-C scores, the current analysis will be focused on alcohol frequency in order to highlight the presence or absence of SNPs that could increase the risk of higher alcohol intake. However, this analysis cannot look for SNPs associated with alcohol consumption disorders, since there is no clinical data (diagnostic tests, etc.). Essentially, the analysis looks for SNPs that can explain variation in this frequency.

Nevertheless, alcohol consumption and alcohol disorders are related with other phenotypes too. Moreover, 17 phenotypes showed significant genetic correlation in alcohol dependence studies, including psychiatric (e.g., schizophrenia, depression), substance use (e.g., smoking and cannabis use), social (e.g., socio-economic deprivation), and behavioral (e.g., educational attainment) traits (Kranzler *et al.*, 2019). So, some covariates have been considered to adjust data such as people that ever smoked, presence of schizophrenia, alcohol intake in a normal alcohol day and the income score.

Methods

Alcohol frequency data was divided in 6 groups: *2 to 3 times per week*, *4 or more times per week*, *2 to 4 times per month, monthly*, *never* and *prefer not to answer*. Individuals that do not answer about alcohol frequency are discarded from the analysis. So, two main different approaches were assessed for dividing data in a binary variable in order to avoid population stratification. The first one was the consumption versus non-consumption of alcohol, assigning 0 to the individuals of the group *Never* and 1 to the rest. Nevertheless, without adjusting data the *lambda* value obtained in the *chi squared test* was 1.281085, but when considering the covariates ever smoke, schizophrenia, alcohol day intake and incomes, the *lambda* value was corrected to 1.059491. The second approach considered two groups, being high frequency those who drink more than twice or four times per week and low frequency the rest. Without adding any covariate, the *lambda* value is 1.0123312 and qqplot is represented in *Figure 3*. If considering other covariates, especially alcohol day intake and schizophrenia, the *lambda* value increases and the *qqplot* shows little inflation. Nevertheless, when analyzing high and low frequency with the covariate smoke the best lambda value is reached being 1.02225. To sum up, data has been divided following as criteria the closest *lambda* value to 1.

For performing the described data selection assessment, quality controls were performed to SNPs and individuals previously. In the case of SNPs, the quality control considered those with a call rate higher than 95%, minor allele frequency (MAF) higher to 5% and, in the case of the low frequency group, a Hardy-Weinberg equilibrium (HWE) higher than 3.3. The quality control of individuals discarded sex discrepancies, which

were obtained from the heterozygosity of the X chromosome (being higher than 0.2 for males and less than 0.2 for females), related individuals, a call rate lesser than 95% and heterozygosity values out of the upper and lower threshold. In this case, heterozygosity threshold was calculated as $\overline{x}$ of heterozygosity -+ $3*\sigma$. For related individuals an *identity-by-descent* (IBD) analysis was performed by using SNPRelate and SNPassoc pacakges, and the threshold of the kinship value was settled to greater than 0.1. Additionally, PCA analysis was performed to asses population stratification, which is displayed in *Figure 4*.

The qqman package was used for computing the Manhattan plot. Moreover, p-values from of the Additive model computed in the *rhs test* after quality controls were used. Should be considered that when applying standard Bonferroni correction of 5 x 10^{-8}, no significant SNPs were found for the high frequency of alcohol group. Hence, the threshold was adjusted to 5 x 10^{-5}.

Nevertheless, in order to filter SNPs with stronger associations to alcohol consumption, the genetic score was computed with PredictABEL and MASS packages. A SNP matrix was created with the top SNPs which showed p-values from the Additive model smaller to the changed Bonferroni threshold. A multivariate GLM following the Akaike information criteria (AIC) was computed with the selected SNPs in order to reduce the number of significative SNPs for alcohol frequency. Additionally, when the score was computed, another GLM was performed between scores and frequencies.

Finally, in order to analyze possible functionality of the genes affected by the top SNPs selection, *dbSNP* from NCBI and *GWAS Catalog* were consulted.

Results

After quality controls were performed, a total of 511336 SNPs and 2401 individuals were analyzed. From the original dataset a total of 58498 SNPs and 103 individuals were removed, however 64 individuals were found to be related. Additionally, 21 sex discrepancies were found, which is represented in the *Figure 1*, and few individuals were found out of the upper and lower limit of the Heterozigosity threshold, which is also represented in *Figure 2*.

When the *chi squared test* and the *snp rhs test* analysis was performed with high and low frequencies intake adding the ever smoke covariate, a *lambda* value of 1.02225 was obtained, so data did not showed stratification. However, as shown in *Figure 5*, most of the SNPs are not significant for the high frequency of alcohol consumption. The Manhattan plot, which is represented in the *Figure 6*, shows that significant SNPs are just found when Bonferroni significant level is equal to 5 x 10^{-5} (which correspond to 5 in the logarithmic scale). However, just 14 SNPs out of 511336 were significant. Nevertheless, when the risk score was computed, only 12 SNPs were strongly significant to perform the analysis. Additionally, just one of them showed a p-value smaller than 5 x 10^{-6} (rs1835844) and a couple of them also presented values of 10^{-6} magnitude (rs1588481 and rs308605).

The risk score analysis showed that higher alcohol frequency intake increases up to 29% per each risk allele compared to others. Additionally, distribution of scores across individuals and *Receiver Operator Characteristic* (ROC) plot was also computed as shown in *Figure 7* and *Figure 8* respectively. The ROC plot showed a moderately displaced curve to the top left corner, indicating good accuracy for the test.

In addition, in *Table 1* is gathered the OR for Recessive, Dominant and Additive models alongside other features, such as gene annotation and chromosome among others, for the 12 top SNPs obtained from the whole analysis. As spotted, ORs are moderately higher than 1, so no strong associations are expected. The highest ones, in the case of Additive model, are from the SNPs rs487616, rs12662272 and rs1835844 being only the last two associated to the genes SUPT3H and CCDC26 respectively according *dbSNP* from NCBI. Both genes were searched in *GWAS catalog* and none of them are associated to alcohol dependence, nevertheless CCDC26 seems to be associated with bipolar depression and nicotine dependence. Interestingly, rs1835844, which is associated with this gene, was the only one from the analysis which passed an adjusted Bonferroni correction of 5 x 10^{-6} as mentioned previously. Moreover, Locuszoom plots have been generated for rs487616, rs12662272 and rs1835844 as spotted in *Figure 9*, *Figure 10* and *Figure 11* respectively. The rs487616 and rs1835844 SNPs are both in regions with high recombination rates and, as can be spotted, rs1835844 is located near to FAM49B gene, GSDMC and LINC00977, and rs12662272 is associated to SUPT3H.

Finally, one strong negative association would be spotted for rs56403353 which showed an OR from the Additive model of 0.57. Nevertheless, no genes are associated with this SNP. Finally, the highest ORs were found for the Dominant model, nevertheless Additive model was preferred for analyzing results.

Discussion

As previously mentioned, several GWAS have been focused on variants for the AUD diagnosis or AUDIT-C scores. However, associated genes are mostly related with alcohol metabolizing enzymes (ADH1B/ADH1C/ADH5), genes implicated in the neurobiology of substance use (DRD2, PDE4B) among others (GCKR, CADM2, FAM69C and KLB) (Clarke *et al.*, 2017). Nevertheless, should be highlighted that these analyses are performed to spot pathologic consumption of alcohol since are based on diagnosis data supported by manuals of mental disorders (such as DSM-IV). As spotted in the present analysis, standard Bonferroni correction did not show any significant association between the set of 511336 SNPs and the individuals which alcohol intake was from 2 times per week to 4 or more. Nevertheless, the most significant SNP of the current study is associated with the gene CCDC26 which showed in *GWAS Catalog* to be associated with bipolar depression and nicotine dependence.

As previously described, alcohol consumption is influenced by several phenotypes since it is related with mental disorders, socioeconomic conditions, behavioral traits, and substance abuse. So, it is not strange that little association could be found for SNPs related with those phenotypes. However, it is difficult to quantify the genetic association with alcohol consumption phenotype since Kranzler *et al.* (2019) concluded in their GWAS analysis about AUD and AUDIT-C scores that heavy drinking is a key risk factor for AUD, but it is not a sufficient cause of the disorder. So, alcohol consumption phenotype varies widely even without reaching pathologic level. Hence, it is a possibility that common SNPs that are not directly related with alcohol consumption can be also influencing variation alongside the environmental factors.

Nevertheless, the present analysis has several limitations since a binary variable has been used to analyze different groups of frequencies. It is probably that GWAS power to detect SNPs related with higher alcohol intake is lost since individuals can vary widely between the classification of alcohol frequency. However, when selecting individuals from 6 to 10 or more drinks in a typical alcohol day as covariate, *lambda* value was slightly inflated. It is probably, because this group is not present in the two groups that were included in the high frequency group.

In conclusion, alcohol frequency does not show enough power to detect associations between SNPs and individuals with higher alcohol consumption levels. Additionally, should be highlighted that having high alcohol intake, even without reaching pathologic levels, is a risk factor to develop disorders of alcohol consumption. So, further investigation would be needed on variants influencing alcohol intake phenotype, even if they are not directly related with alcohol.

References

1. Clarke, TK., Adams, M., Davies, G. et al. Genome-wide association study of alcohol consumption and genetic overlap with other healthrelated traits in UK Biobank (N=112 · 117). Mol Psychiatry 22, 1376-1384 (2017). https://doi.org/10.1038/mp.2017.153

2. Kranzler, H.R., Zhou, H., Kember, R.L. et al. Genome-wide association study of alcohol consumption and use disorder in 274,424 individuals from multiple populations. Nat Commun 10, 1499 (2019). https://doi.org/10.1038/s41467-019-09480-8

Appendix

Appendix 0 Loading libraries

```
library(snpStats)
library(SNPRelate)
library(SNPassoc)
library(tidyverse)
library(ggrepel)
library(ggplot2)
library(qqman)
library(MASS)
library(PredictABEL)
library(knitr)
library(tinytex)
library(kableExtra)
```

Appendix 1 Creating variates and covariates

```
temp <- "/Users/carla/Documents/R directory/Final_task_GWAS/ega_synthetic_phenotypes.tsv"
pheno <- read.delim(temp)
table(pheno$alcohol_frequency)
```

```
##
##    2_to_3_times_a_week   2_to_4_times_a_month 4_or_more_times_a_week
##                    427                    439                    384
##        Monthly_or_less                  Never   Prefer_not_to_answer
##                    423                    415                    414
```

```
# Filter the groups of people who drink twice or more per
# week (high frequency) and people that drink 2-4 times per
# month/monthly/never (low frequency). Assign NA values to
# those who prefer not to answer.
pheno$HighLow <- 0
pheno$HighLow[pheno$alcohol_frequency == "4_or_more_times_a_week"] <- 1
pheno$HighLow[pheno$alcohol_frequency == "2_to_3_times_a_week"] <- 1
pheno$HighLow[pheno$alcohol_frequency %in% "Prefer_not_to_answer"] <- NA
table(pheno$HighLow)
```

```
##
##    0    1
## 1279  811
```

```
# Possible covariates Smoke
pheno$Smoke <- 0
pheno$Smoke[pheno$ever_smoked == "Yes"] <- 1

# Alcohol_day
pheno$alcDay <- 0
pheno$alcDay[pheno$alcohol_day == "10_or_more"] <- 1
pheno$alcDay[pheno$alcohol_day == "7_8_or_9"] <- 1
pheno$alcDay[pheno$alcohol_day %in% "Prefer_not_to_answer"] <- NA

# Mental disorder (schizo)
pheno$schizo <- 0
pheno$schizo[pheno$mental_disorders == "schizophrenia"] <- 1
```

Appendix 2 Creating objects from PLINK files with data

```
# Open plink files
path <- "/Users/carla/Documents/R directory/Final_task_GWAS/ega_synthetic_snps_600k/"
al.plink <- read.plink(file.path(path, "ega_synthetic_snps_600k"))
names(al.plink)
```

```
## [1] "genotypes" "fam"       "map"
```

```
# Assign annotations, genotypic information and family
# individuals to variables
al.geno <- al.plink$genotypes
annotation <- al.plink$map
family <- al.plink$fam
```

Appendix 3 Checking rows from tsv file and PLINK data

```
# Rename rows of pheno file with IDs
rownames(pheno) <- pheno$subject_id
# Check if rows of geno and pheno have the same info
identical(rownames(pheno), rownames(al.geno))
```

```
## [1] FALSE
```

```
# Fix row info of both variables
ids <- intersect(rownames(pheno), rownames(al.geno))
geno <- al.geno[ids, ]
al <- pheno[ids, ]
identical(rownames(al), rownames(geno))
```

```
## [1] TRUE
```

```
family <- family[ids, ]
```

Appendix 4 QC of SNPs

```
info.snps <- col.summary(geno)
head(info.snps)
```

```
##             Calls Call.rate Certain.calls       RAF          MAF       P.AA
## rs572818783  2487 0.9932109             1 0.9997990 0.0002010454 0.00000000
## rs571093408  2488 0.9936102             1 0.9917605 0.0082395498 0.00000000
## rs540466151  2489 0.9940096             1 0.9993973 0.0006026517 0.00000000
## rs554008981  2487 0.9932109             1 0.9965822 0.0034177724 0.00000000
## rs577106641  2487 0.9932109             1 0.9997990 0.0002010454 0.00000000
## rs531646671  2488 0.9936102             1 0.8522910 0.1477090032 0.00522508
##                      P.AB      P.BB      z.HWE
## rs572818783 0.0004020909 0.9995979 0.01002812
## rs571093408 0.0164790997 0.9835209 0.41440204
## rs540466151 0.0012053033 0.9987947 0.03008435
## rs554008981 0.0068355448 0.9931645 0.17102827
## rs577106641 0.0004020909 0.9995979 0.01002812
## rs531646671 0.2849678457 0.7098071 6.57433471
```

```
# Get HW of controls, in this case, group of non-drinkers
controls <- al$HighLow == 0 & !is.na(al$HighLow)
geno.controls <- geno[controls, ]
info.controls <- col.summary(geno.controls)

# Compute QC to SNPs
use <- info.snps$Call.rate > 0.95 & info.snps$MAF > 0.05 & abs(info.controls$z.HWE <
    3.3)
mask.snps <- use & !is.na(use)

geno.qc.snps <- geno[, mask.snps]

annotation <- annotation[mask.snps, ]
```

Appendix 5 QC of individuals

```
info.indv <- row.summary(geno.qc.snps)
head(info.indv)
```

```
##         Call.rate Certain.calls Heterozygosity
## HG02583 0.9999961             1      0.2958907
## HG00419 1.0000000             1      0.2672020
## HG00531 1.0000000             1      0.2632164
## NA20849 1.0000000             1      0.2869131
## NA18592 1.0000000             1      0.2608129
## NA19375 0.9999980             1      0.2912005
```

```
# Stats for gender discrepancies (use chromosome X)
geno.X <- geno.qc.snps[, annotation$chromosome == "23" & !is.na(annotation$chromosome)]
info.X <- row.summary(geno.X)

# Select the sex discrepancies (males with non-zero
# heterozygosity for X chromosomes)
sex.discrep <- (al$sex == "male" & info.X$Heterozygosity > 0.2) |
    (al$sex == "female" & info.X$Heterozygosity < 0.2)

# Compute the expected threshold for heterozygosity. Use x
# +- 3*sigma as the confidence band of individuals with a
# correct heterozygosity.
x.heterozygosity <- mean(info.indv$Heterozygosity)
s.heterozygosity <- sd(info.indv$Heterozygosity)
upper.limit <- x.heterozygosity + (3 * s.heterozygosity)
lower.limit <- x.heterozygosity - (3 * s.heterozygosity)
het.col <- ifelse(info.indv$Heterozygosity < upper.limit & info.indv$Heterozygosity >
    lower.limit, "grey", "red")
```

Figure 1 Heterozygosity for chromosome X in males and females

```
# Graph gender discrepancies
mycol <- ifelse(al$sex == "male", "blue", "red")
plot(info.X$Heterozygosity, col = mycol, pch = 16, xlab = "Individuals",
    ylab = "Heterozygosity in chromosome X")
```

```
legend("topright", c("Females", "Males"), col = c("red", "blue"),
    pch = 16)
```

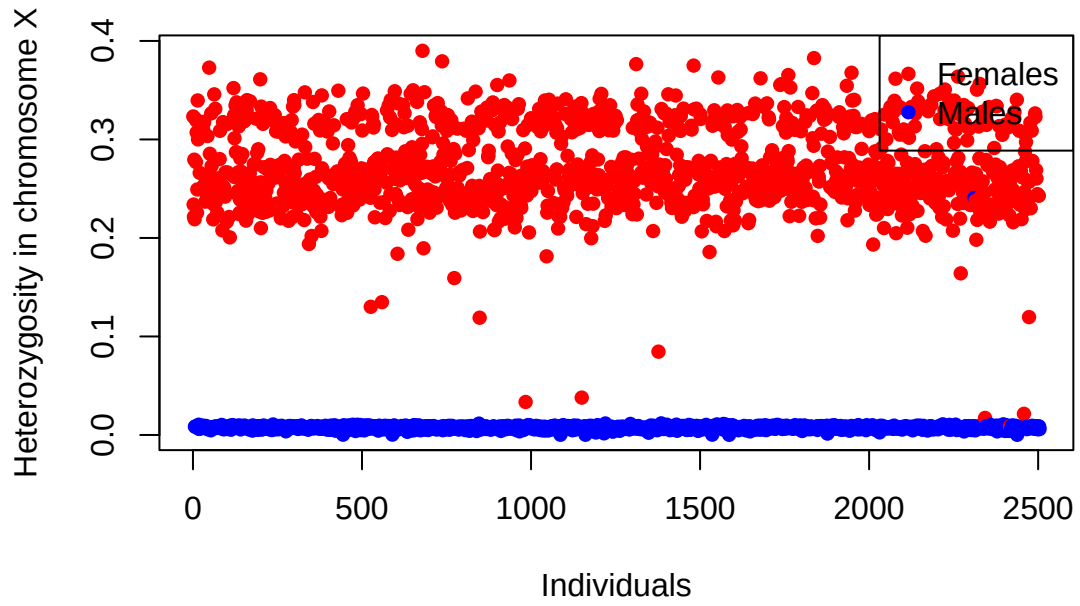
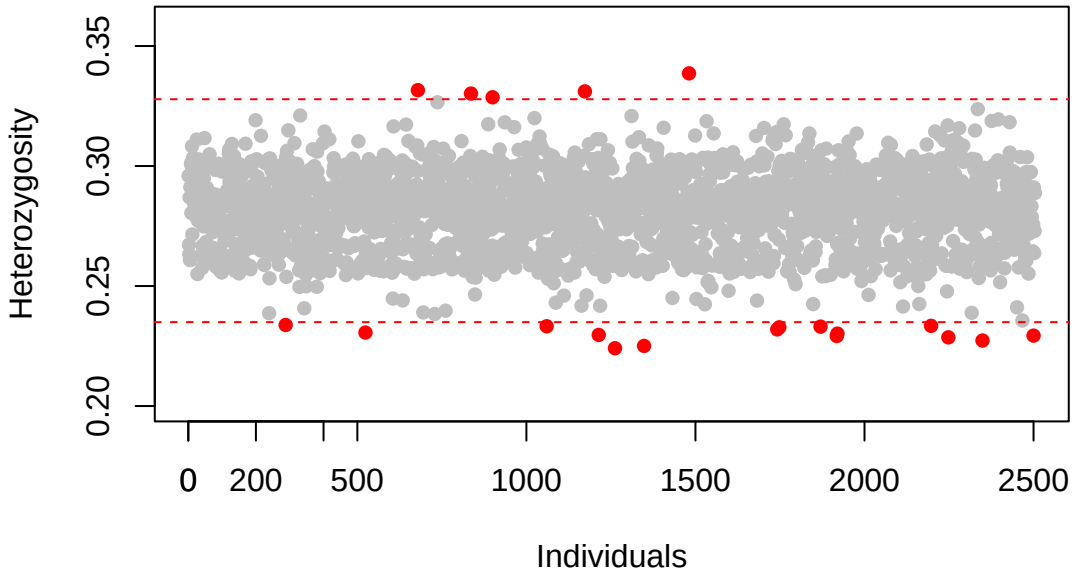


Figure 2 Heterozygosity of individuals

```
# Plot heterozygosity
plot(info.indv$Heterozygosity, col = het.col, pch = 16, xlab = "Individuals",
    ylab = "Heterozygosity", ylim = c(0.2, 0.36))
axis(side = 1, at = seq(0, 500, by = 200))
# Adding horizontal lines to set up thresholds
abline(h = upper.limit, lty = 2, col = "red")
abline(h = lower.limit, lty = 2, col = "red")
```



Appendix 6 Identity-by-descent analysis to compute related individuals

```
# Conversion from PLINK format to GDS format
snpgdsBED2GDS("./ega_synthetic_snps_600k/ega_synthetic_snps_600k.bed",
    "./ega_synthetic_snps_600k/ega_synthetic_snps_600k.fam",
    "./ega_synthetic_snps_600k/ega_synthetic_snps_600k.bim",
    out = "obGDS_")
```

```
## Start file conversion from PLINK BED to SNP GDS ...
##     BED file: './ega_synthetic_snps_600k/ega_synthetic_snps_600k.bed'
##         SNP-major mode (Sample X SNP), 340.5M
##     FAM file: './ega_synthetic_snps_600k/ega_synthetic_snps_600k.fam'
##     BIM file: './ega_synthetic_snps_600k/ega_synthetic_snps_600k.bim'
## Fri Feb  4 10:14:47 2022     (store sample id, snp id, position, and chromosome)
##     start writing: 2504 samples, 569834 SNPs ...
## [..................................................]  0%, ETC: ---        [==========================
## Fri Feb  4 10:15:02 2022     Done.
## Optimize the access efficiency ...
## Clean up the fragments of GDS file:
##     open the file 'obGDS_' (343.8M)
##     # of fragments: 39
##     save to 'obGDS_.tmp'
##     rename 'obGDS_.tmp' (343.8M, reduced: 252B)
##     # of fragments: 18
```

```
# Open GDS file
genofile <- snpgdsOpen("obGDS_")

# Let's remove recursively SNPs in Linkage Desequilibrium
# (because they are correlated and this is not useful for
# the analysis)
set.seed(12345678)
snps.qc <- colnames(geno.qc.snps)
snp.prune <- snpgdsLDpruning(genofile, ld.threshold = 0.2, snp.id = snps.qc)
```

```
## SNP pruning based on LD:
## Excluding 74,662 SNPs (non-autosomes or non-selection)
## Excluding 0 SNP (monomorphic: TRUE, MAF: NaN, missing rate: NaN)
##     # of samples: 2,504
##     # of SNPs: 495,172
##     using 1 thread
##     sliding window: 500,000 basepairs, Inf SNPs
##     |LD| threshold: 0.2
##     method: composite
## Chromosome 1: 6.56%, 5,737/87,453
## Chromosome 2: 13.28%, 5,511/41,512
## Chromosome 3: 13.24%, 4,784/36,135
## Chromosome 4: 12.37%, 4,638/37,485
## Chromosome 5: 13.11%, 4,171/31,814
## Chromosome 6: 12.44%, 4,224/33,958
## Chromosome 7: 13.40%, 3,998/29,834
## Chromosome 8: 12.94%, 3,567/27,575
## Chromosome 9: 14.98%, 3,275/21,868
## Chromosome 10: 14.05%, 3,631/25,852
## Chromosome 11: 13.32%, 3,288/24,679
```

```
## Chromosome 12: 14.41%, 3,424/23,767
## Chromosome 13: 13.91%, 2,579/18,535
## Chromosome 14: 14.87%, 2,436/16,381
## Chromosome 15: 14.92%, 2,206/14,783
## Chromosome 16: 16.29%, 2,547/15,640
## Chromosome 17: 16.50%, 2,324/14,086
## Chromosome 18: 16.06%, 2,294/14,285
## Chromosome 19: 16.50%, 2,013/12,202
## Chromosome 20: 17.15%, 1,871/10,908
## Chromosome 21: 15.55%, 1,148/7,384
## Chromosome 22: 16.35%, 1,180/7,216
## 70,846 markers are selected in total.
```

```
snps.ibd <- unlist(snp.prune, use.names = FALSE)

# Now compute IDB values
ibd <- snpgdsIBDMoM(genofile, kinship = TRUE, snp.id = snps.ibd,
    num.thread = 1)
```

```
## IBD analysis (PLINK method of moment) on genotypes:
## Excluding 498,988 SNPs (non-autosomes or non-selection)
## Excluding 0 SNP (monomorphic: TRUE, MAF: NaN, missing rate: NaN)
##     # of samples: 2,504
##     # of SNPs: 70,846
##     using 1 thread
## PLINK IBD:    the sum of all selected genotypes (0,1,2) = 62285170
## Fri Feb  4 10:16:04 2022    (internal increment: 4480)
## [..................................................]  0%, ETC: ---        [==========================
## Fri Feb  4 10:16:24 2022    Done.
```

```
ibd.kin <- snpgdsIBDSelection(ibd)

# RELATED INDIVIDUALS kinship score is higher than 0.1, the
# individuals are related (so this is the threshold)
ibd.kin.thres <- subset(ibd.kin, kinship > 0.1)
head(ibd.kin.thres)
```

```
##            ID1     ID2        k0         k1   kinship
## 44905  HG00116 HG00120 0.6582519 0.24658166 0.1092286
## 194658 HG00238 HG00240 0.7055773 0.18816258 0.1001707
## 505992 HG00475 HG00542 0.6317763 0.19418622 0.1355653
## 565256 HG00581 HG00607 0.7304523 0.08575837 0.1133343
## 569781 HG00584 HG00595 0.7452908 0.07012813 0.1098226
## 725845 HG00851 HG00881 0.7112033 0.10709317 0.1176250
```

```
# Achieve IDs of related individuals. In my case there are
# two
ids.rel <- SNPassoc:::related(ibd.kin.thres)
length(ids.rel)
```

```
## [1] 64
```

Appendix 7 Filtering individuals after QC

```
# Now let's apply all the filters computed before
use <- info.indv$Call.rate > 0.95 & abs(info.indv$Heterozygosity) >
    lower.limit & abs(info.indv$Heterozygosity) < upper.limit &
    !sex.discrep & !rownames(info.indv) %in% ids.rel
mask.indiv <- use & !is.na(use)
geno.qc <- geno.qc.snps[mask.indiv, ]
geno.qc
```

```
## A SnpMatrix with  2401 rows and  511336 columns
## Row names:  HG02583 ... HG02941
## Col names:  rs806731 ... rs200631149
```

```
pheno.qc <- pheno[mask.indiv, ]
identical(rownames(pheno.qc), rownames(geno.qc))
```

```
## [1] TRUE
```

```
dim(pheno)
```

```
## [1] 2504   78
```

```
dim(pheno.qc)
```

```
## [1] 2401   78
```

```
dim(geno)
```

```
## [1]   2504 569834
```

```
dim(geno.qc)
```

```
## [1]   2401 511336
```

```
# Removed SNPs
sum(!mask.snps)
```

```
## [1] 58498
```

```
# Removed individuals
sum(!mask.indiv)
```

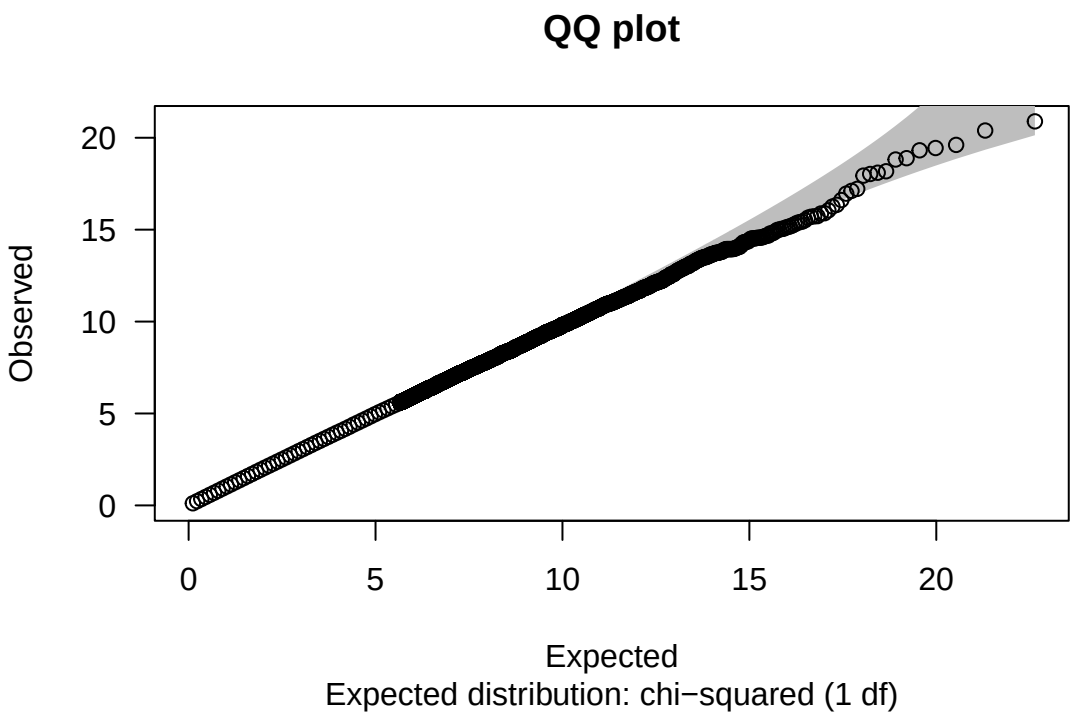
```
## [1] 103
```

Appendix 8 Single association analysis of alcohol frequencies (non-adjusted) and chi squared test

```
# Compute single SNP test
res <- single.snp.tests(HighLow, data = pheno.qc, snp.data = geno.qc)

# Compute chi squared test
chi2 <- chi.squared(res, df = 1)
```

Figure 3 QQplot of single snp test analysis results for high and low frequencies of alcohol intake without covariates

```
# Compute QQplot
qq.chisq(chi2)
```




```
##            N      omitted       lambda
## 5.113360e+05 0.000000e+00 1.012312e+00
```

Appendix 9 PCA analysis

```
pca <- snpgdsPCA(genofile, sample.id = rownames(geno.qc), snp.id = snps.ibd,
    num.thread = 1)
```

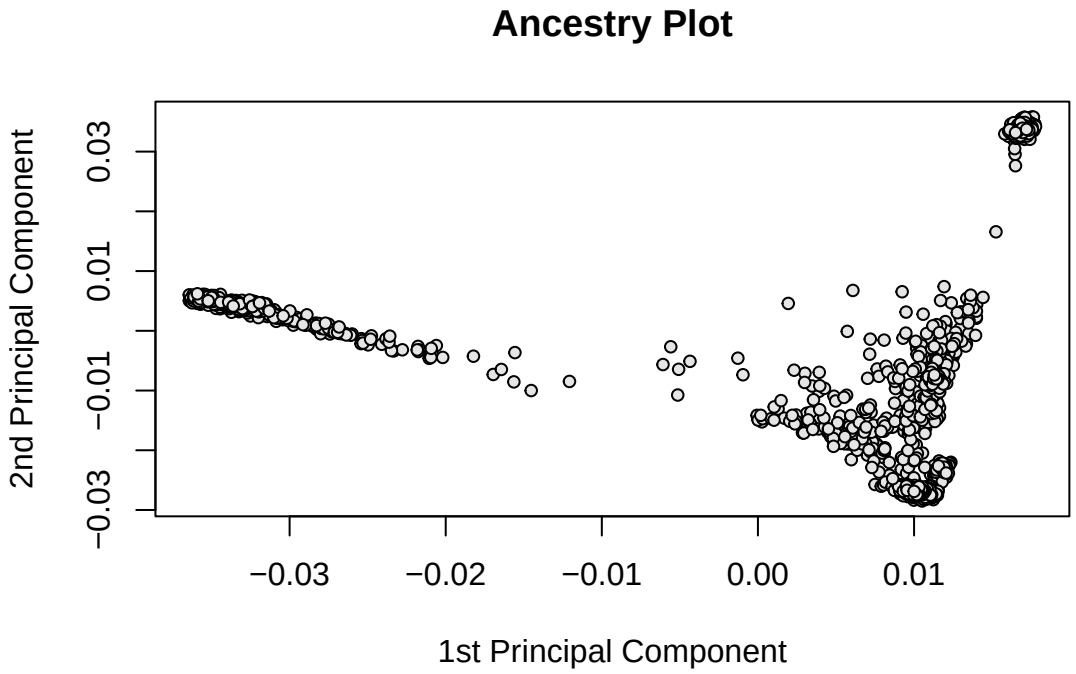
```
## Principal Component Analysis (PCA) on genotypes:
## Excluding 498,988 SNPs (non-autosomes or non-selection)
## Excluding 0 SNP (monomorphic: TRUE, MAF: NaN, missing rate: NaN)
##     # of samples: 2,401
##     # of SNPs: 70,846
##     using 1 thread
##     # of principal components: 32
## PCA:    the sum of all selected genotypes (0,1,2) = 60151479
## CPU capabilities: Double-Precision SSE2
```

```
## Fri Feb  4 10:17:03 2022    (internal increment: 148)
## [..............................................]  0%, ETC: ---        [==========================
## Fri Feb  4 10:18:46 2022    Begin (eigenvalues and eigenvectors)
## Fri Feb  4 10:18:52 2022    Done.
```

```
# Add just the first five principal components
pheno.qc <- data.frame(pheno.qc, pca$eigenvect[, 1:5])

# After performing QC and PCA close the GDS file
closefn.gds(genofile)
```

Figure 4 PCA analysis plot of first and second principal components

```
with(pca, plot(eigenvect[, 1], eigenvect[, 2], xlab = "1st Principal Component",
    ylab = "2nd Principal Component", main = "Ancestry Plot",
    pch = 21, bg = "gray90", cex = 0.8))
```



Appendix 10 Snp rhs test with smoke covariate and chi squared test

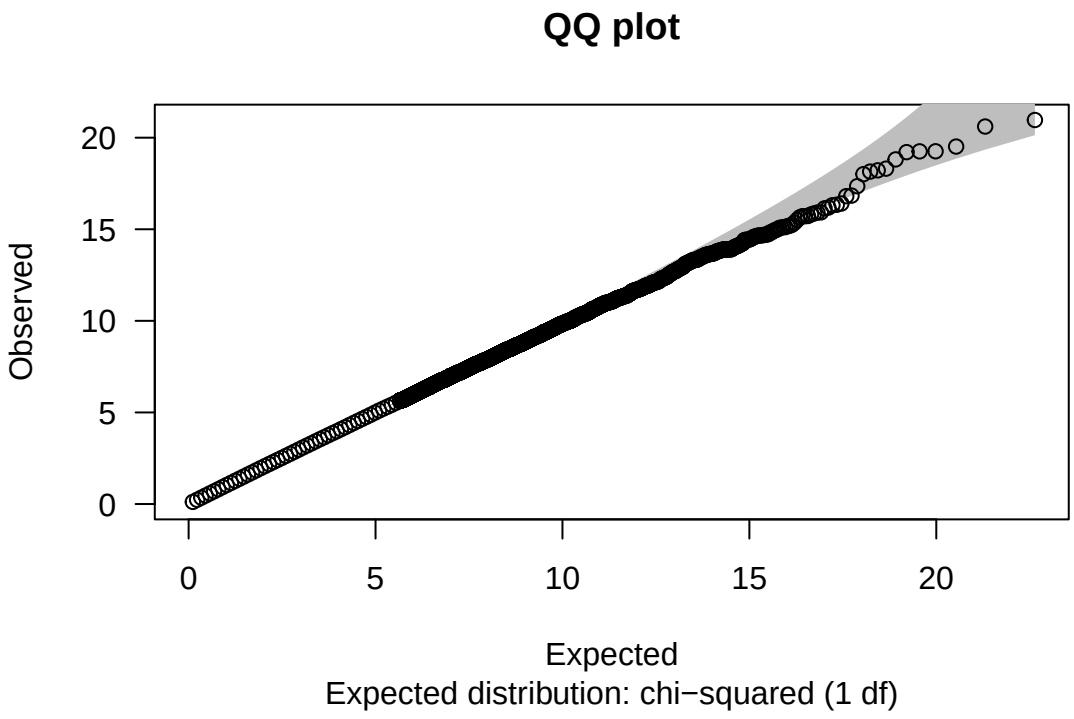
```
# The model is slighly adjusted. The X2 component shows
# lambda values closer to 1.
res.adj <- snp.rhs.tests(HighLow ~ pheno.qc$Smoke, data = pheno.qc,
    snp.data = geno.qc)

# Compute chi squared test
chi2 <- chi.squared(res.adj)

# No significant SNPs are found if using standard
# Bonferroni correction.  Just when using a threshold of
# 5e-5, significant SNPs are found.
bonf.sig <- 5e-05
```

```
ps <- p.value(res.adj)
res.adj[ps < bonf.sig & !is.na(ps), ]
```

```
##            Chi.squared Df      p.value
## rs487616       16.80741  1 4.137155e-05
## rs12662272     19.21078  1 1.170507e-05
## rs56403353     18.30906  1 1.878121e-05
## rs917180       18.14844  1 2.043368e-05
## rs35991633     18.21668  1 1.971447e-05
## rs1835844      20.96492  1 4.677712e-06
## rs1588481      19.51462  1 9.983258e-06
## rs308605       20.60854  1 5.634410e-06
## rs9539443      17.35034  1 3.108431e-05
## rs72727737     19.25306  1 1.144869e-05
## rs1549605      16.83701  1 4.073112e-05
## rs62113363     19.25772  1 1.142074e-05
## rs6075815      18.81278  1 1.441980e-05
## rs7052248      18.00789  1 2.199916e-05
```

Figure 5 QQplot of snp rhs test analysis results for high and low frequencies of alcohol intake with ever smoke covariate

```
qq.chisq(chi2)
```



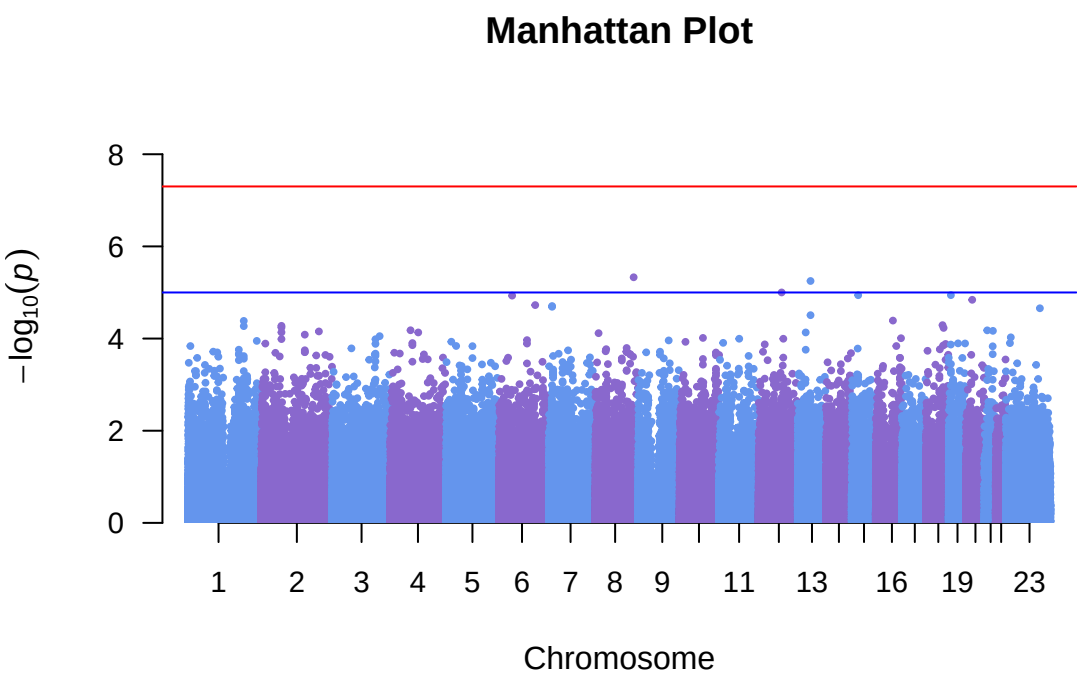

```
##           N     omitted      lambda
## 5.11336e+05 0.00000e+00 1.02225e+00
```

Appendix 11 Preparing Manhattan

```
pvals <- data.frame(SNP = annotation$snp.name, CHR = annotation$chromosome,
    BP = annotation$position, P = p.value(res.adj))
# missing data is not allowed
pvals <- subset(pvals, !is.na(CHR) & !is.na(P))
```

Figure 6 Manhattan plot for SNPs associated with high frequency of alcohol intake. The red line represents standard Bonferroni correction. The blue line represents the adjusted threshold.

```
# As spotted, red line corresponds to standard Bonferroni
# correction (7.30103) Nevertheless, SNPs are only
# significant if threshold is settled to 5 (blue line)
plt <- manhattan(pvals, main = "Manhattan Plot", ylim = c(0,
    9), cex = 0.6, cex.axis = 0.9, col = c("cornflowerblue",
    "mediumpurple3"))
```



Appendix 12 Genetic score of top SNPs with Bonferroni correction of 5e-5

```
# Select top SNPs with Bonferroni correction of 5e-5
sig.pval <- subset(pvals, ps < bonf.sig & !is.na(ps))
topSNPs <- as.character(sig.pval$SNP)

# Subset TOP SNPs
geno.topSNPs <- geno.qc[, topSNPs]

# export top SNPs on SNPmatrix format
write.SnpMatrix(geno.topSNPs, file = "topSNPs.txt")
```

```
## [1] 2401   14
```

```r
# import top SNPs matrix al.top <-
# read.delim('topSNPs.txt', sep='')
al.top <- read.delim("topSNPs.txt", sep = "")
al.top <- cbind(al.top, pheno.qc)

# prepare data for SNPassoc (SNPs are coded as 0,1,2)
ii <- grep("^rs", names(al.top))
# This is used for genetic score
al.top.s <- setupSNP(al.top, colSNPs = ii, name.genotypes = c(0,
    1, 2))

# run association (just additive)
ans <- WGassociation(HighLow, al.top.s, model = "log-additive")
# run association (for all)
SNP.for.or <- WGassociation(HighLow, al.top.s)
# association(HighLow ~ rs1835844, al.top.s)

# Select just significant variants
sel <- labels(al.top.s)[additive(ans) < 5e-05]
alcohol.sel <- al.top.s[, sel]


# Data frame with SNPs and frequency groups
alcohol.sel <- data.frame(lapply(alcohol.sel, additive))
dd.end <- data.frame(casecontrol = al.top.s$HighLow, alcohol.sel)
# Filter first possible NA values in cascontrol column
dd.end.complete <- dd.end[complete.cases(dd.end), ]

# Computing multivariate GLM with selected SNPs for
# predicting high frequency of alcohol intake
mod <- stepAIC(glm(casecontrol ~ ., dd.end.complete, family = "binomial"),
    method = "forward", trace = 0)
summary(mod)
```

```
##
## Call:
## glm(formula = casecontrol ~ rs487616 + rs12662272 + rs56403353 +
##     rs35991633 + rs1835844 + rs1588481 + rs308605 + rs72727737 +
##     rs1549605 + rs62113363 + rs6075815 + rs7052248, family = "binomial",
##     data = dd.end.complete)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9215  -0.9643  -0.7136   1.1966   2.0716
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.63478    0.45487   1.396 0.162854
## rs487616     0.32572    0.08726   3.733 0.000190 ***
## rs12662272   0.17097    0.08089   2.114 0.034551 *
## rs56403353  -0.36974    0.14602  -2.532 0.011337 *
## rs35991633  -0.18895    0.06891  -2.742 0.006105 **
## rs1835844    0.23619    0.07159   3.299 0.000970 ***
```

```
## rs1588481   -0.18817    0.07374  -2.552 0.010715 *
## rs308605    -0.36765    0.07002  -5.251 1.52e-07 ***
## rs72727737  -0.28999    0.08377  -3.462 0.000537 ***
## rs1549605   -0.25962    0.07901  -3.286 0.001017 **
## rs62113363   0.31576    0.07348   4.297 1.73e-05 ***
## rs6075815   -0.30600    0.08273  -3.699 0.000217 ***
## rs7052248    0.26761    0.05722   4.677 2.92e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2672.1  on 1997  degrees of freedom
## Residual deviance: 2487.6  on 1985  degrees of freedom
## AIC: 2513.6
##
## Number of Fisher Scoring iterations: 4
```

```
# Retrieve significant SNPs for the analysis
snps.score <- names(coef(mod))[-1]

# Position of selected SNPs on the table with data
pos <- which(names(dd.end.complete) %in% snps.score)

# Compute risk scores
score <- riskScore(mod, data = dd.end.complete, cGenPreds = pos,
    Type = "unweighted")
table(score)
```

```
## score
##   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19
##   1   8  20  63 104 209 256 325 312 247 195 120  73  32  26   5   2
```

```
# GLM between scores and alcohol frequency
mod.lin <- glm(casecontrol ~ score, dd.end.complete, family = "binomial")
mod.lin
```

```
##
## Call:  glm(formula = casecontrol ~ score, family = "binomial", data = dd.end.complete)
##
## Coefficients:
## (Intercept)        score
##     -3.2027       0.2553
##
## Degrees of Freedom: 1997 Total (i.e. Null);  1996 Residual
## Null Deviance:       2672
## Residual Deviance: 2496  AIC: 2500
```

```
# This SNPs increase 29% the alcohol intake respect the
# rest
exp(coef(mod.lin)[2])
```

```
##    score
## 1.290809
```

```
# ROC plot preparation
predrisk <- predRisk(mod.lin, dd.end.complete)
```

Figure 7 Histogram of risk scores distribution for 12 top SNPs

```
# Distribution of the score across individuals
hist(score, col = "gray90")
```

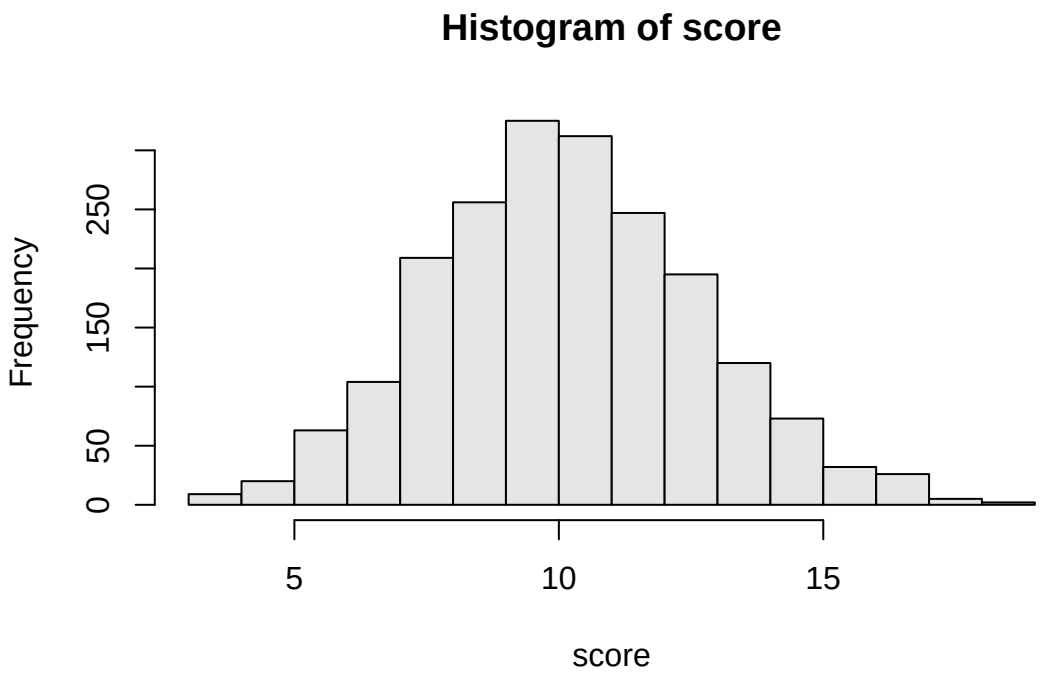
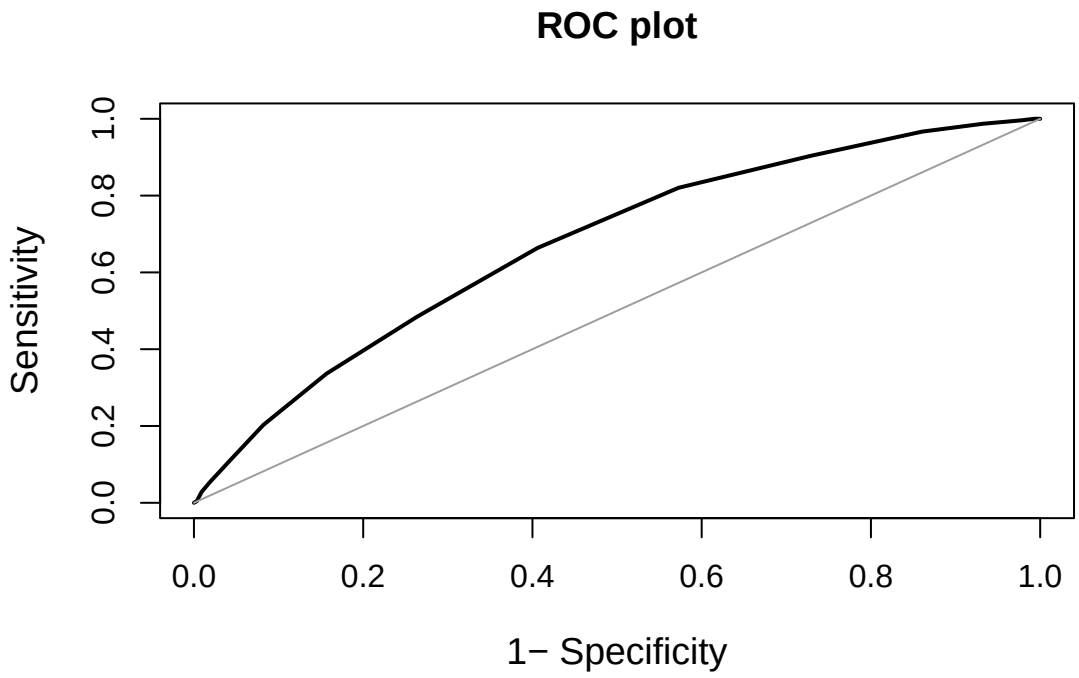


Figure 8 ROC plot of risk scores from 12 top SNPs

```
plotROC(data = dd.end.complete, cOutcome = 1, predrisk = predrisk)
```



```
## AUC [95% CI] for the model 1 :  0.673 [ 0.65  -  0.697 ]
```

Appendix 13 Locuszoom preparation file

```
# Locus zoom preparation file for three SNPs
locuszoom <- sig.pval[c(1, 2, 6), c(1, 4)]
colnames(locuszoom) <- c("MarkerName", "P.value")
write.table(locuszoom, file = "locuszoom.txt", sep = "\t", dec = ".",
    row.names = FALSE, quote = FALSE)
```

Appendix 14 Gathering data for 12 top SNPs

```
# Select best SNPs after SNP risk score
topSNPs <- snps.score

# Compute OR for three different models
orAdditive <- odds(SNP.for.or[topSNPs, ], model = "log-additive")
orDominant <- odds(SNP.for.or[topSNPs, ], model = "dominant")
orRecessive <- odds(SNP.for.or[topSNPs, ], model = "recessive")

# Gather information for the top SNPs table
top.SNPs.position <- annotation[topSNPs, 4]
top.SNPs.chr <- annotation[topSNPs, 1]
top.SNPs.MAF <- info.snps[topSNPs, 5]
top.SNPs.minorAl <- annotation[topSNPs, 6]
top.SNPs.gene <- c("-", "SUPT3H", "-", "MGC4859", "CCDC26", "LINC02424 and LOC105369859",
    "-", "PLA2G4E", "CNOT1", "-", "-", "SEPTIN6")

# Create the table with data of top SNPs
top.table <- data.frame(top.SNPs.position, top.SNPs.chr, top.SNPs.gene,
    top.SNPs.MAF, top.SNPs.minorAl, orAdditive$OR, orDominant$OR,
    orRecessive$OR)
rownames(top.table) <- topSNPs
colnames(top.table) <- c("Genetic position", "Chromosome", "Gene",
    "MAF", "Minor-allele", "OR-Additive", "OR-Dominant", "OR-Recessive")
```

Table 1 Top SNPs table for Additive, Dominant and Recessive models.

```
input <- kable(top.table, align = c(rep("c", times = 12)), format = "latex",
    booktabs = TRUE)
kable_styling(input, latex_options = "scale_down")
```

	Genetic position	Chromosome	Gene	MAF	Minor-allele	OR-Additive	OR-Dominant	OR-Recessive
rs487616	189614321	1	-	0.1891077	A	1.41	1.87	1.47
rs12662272	45123889	6	SUPT3H	0.2694009	C	1.39	1.81	1.43
rs56403353	124073745	6	-	0.0510450	C	0.57	0.51	0.52
rs35991633	10673337	7	MGC4859	0.4897549	G	0.76	0.71	0.68
rs1835844	130500331	8	CCDC26	0.3353438	C	1.36	1.90	1.37
rs1588481	78742314	12	LINC02424 and LOC105369859	0.2778224	G	0.74	0.57	0.72
rs308605	62757700	13	-	0.4090543	A	0.74	0.64	0.70
rs72727737	42344471	15	PLA2G4E	0.1983105	C	0.71	0.37	0.73
rs1549605	58649139	16	CNOT1	0.2376456	T	0.74	0.67	0.69
rs62113363	7644207	19	-	0.3190036	T	1.36	1.71	1.42
rs6075815	21519011	20	-	0.2054684	G	0.71	0.62	0.67
rs7052248	118766049	23	SEPTIN6	0.4399116	A	1.26	1.45	1.43

Figure 9 Locuszoom plot of rs487616

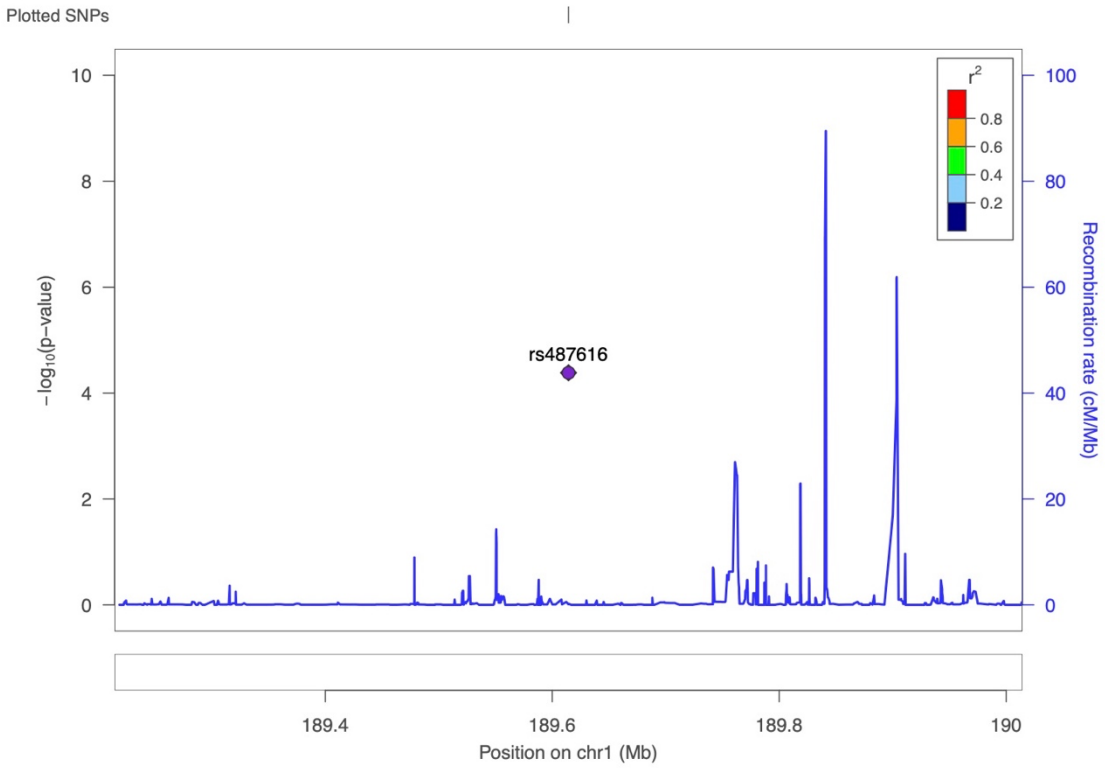


Figure 10 Locuszoom plot of rs12662272

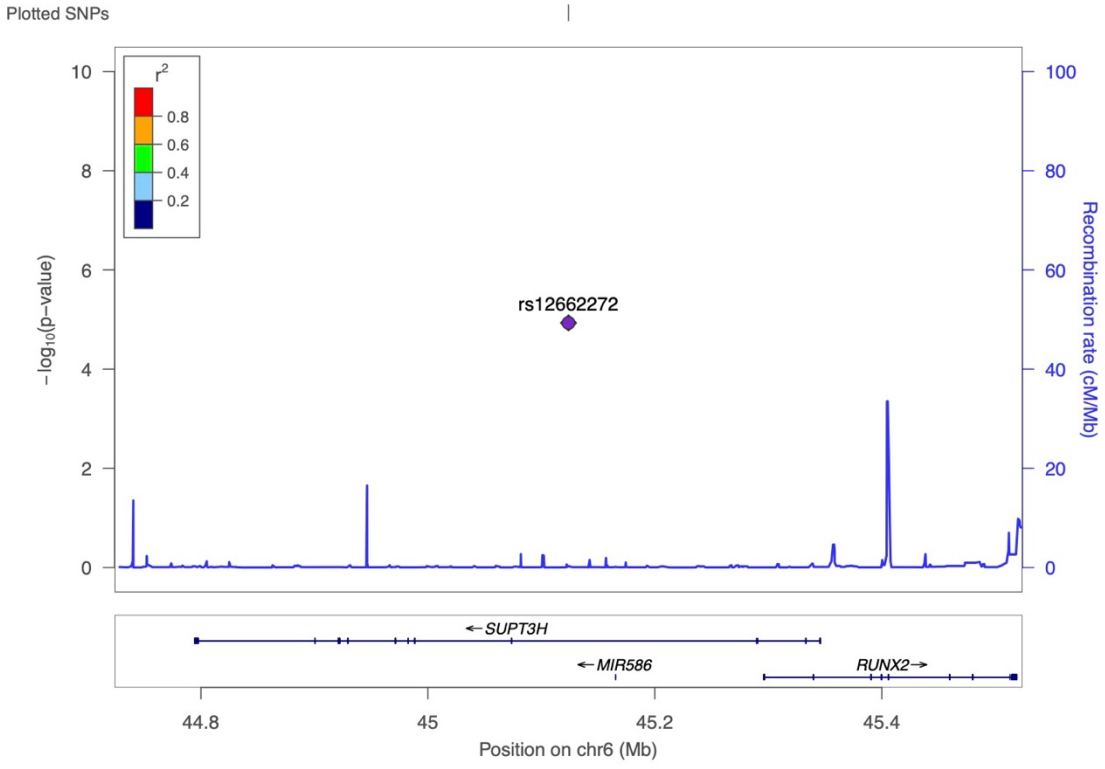


Figure 11 Locuszoom plot of rs1835844