Genome-Wide Association Study (GWAS) Analysis

GWAS Assignment

1. Overview (10 marks)
Your report should begin with a brief overall summary of the cohort (how many cases / controls /males / females /founders / non-founders etc.). How many SNPs are included in the study?

```
plink --bfile gwas --out gwas_summary –noweb
```

This command generated two output files

gwas_summary.log: A log file containing execution details.

306,102 variants loaded from .bim file.

4,000 people (2,000 males, 2,000 females) loaded from .fam.

4,000 phenotype values loaded from .fam.

The log confirms that the dataset consists of 306,102 variants, 4,000 individuals, and 4,000 phenotype values. All individuals are founders, with 0 non-founders.

gwas_summary.frq: This file contains allele frequency data for each SNP. Example:

CHR SNP A1 A2 MAF NCHROBS

1 rs3934834 T C 0.09966 7636

```
Get-Content gwas.fam | ForEach-Object { ($_ -split "\s+")[5] } |
Group-Object | Sort-Object Name
```

This command was used to calculate the distribution of cases (phenotype = 2) and controls (phenotype = 1),

Count Name

----- ----

2000 1 (Controls)

2000 2 (Cases)

Analysis: A total of 306,102 SNPs and 4,000 individuals (2,000 cases and 2,000 controls) provide ample power for detecting associations with common genetic variants.The genotyping rate of **98.3%** (as noted in the log file) indicates minimal missing data, which is

crucial for reliable statistical analysis. Equal representation of males (2,000) and females (2,000) ensures that the study can account for potential sex-based effects. All 4,000 participants are founders (0 non-founders), suggesting no familial relationships, reducing potential confounding effects. A perfect 1:1 ratio of cases and controls strengthens the robustness of statistical comparisons. The dataset's size, quality, and balance make it highly suitable for a genome-wide association study (GWAS). The high genotyping rate and absence of familial relationships enhance the reliability of the analysis.

Q2 QC tests (15 marks)
i) How many SNPs is individual A2038 missing data for? (hint: use the grep linux command)
ii) For how many individuals is the SNP rs2493272 data missing?
iii) Create two plots which summarise the overall missingness for the data, one for SNPS and one
for individuals. Would you consider the missingness rates in general to be high or low? What might
this indicate about the data?

i)

```
PS C:\Users\boghe\Downloads\GWAS> plink --bfile gwas --missing --
noweb --out missing_data
```

Output: This PLINK command is used to analyze missing genotype data in the GWAS dataset. It generates the following output files:

missing_data.imiss
Summarizes missing data for each individual, including the fraction of missing genotypes.

missing_data.lmiss
Summarizes missing data for each SNP (variant), including the fraction of missing genotypes.

missing_data.log
A log file documenting the execution details of the command

```
PS C:\Users\boghe\Downloads\GWAS> @(Get-Content
missing_data.imiss | Select-Object -First 1) + (Select-String -
Pattern "A2038" -Path missing_data.imiss | ForEach-Object {
$_.Line })
```

This command:
Selects the header row from the missing_data.imiss file, which contains column namesa nd
Searches for the identifier A2038 in the file and retrieves the corresponding row of data.

FID IID MISS_PHENO N_MISS N_GENO F_MISS

37 A2038 N 5211 306102 0.01702

Individual A2038 is missing data for 5211 SNPs.

ii) `PS C:\Users\boghe\Downloads\GWAS> Get-Content missing_data.lmiss | Select-Object -First 1`

`PS C:\Users\boghe\Downloads\GWAS> Select-String -Pattern "rs2493272" -Path missing_data.lmiss`

CHR SNP N_MISS N_GENO F_MISS

missing_data.lmiss:202: 1 rs2493272 111 4000 0.02775

N_MISS = 111: The number of individuals for whom rs2493272 has missing data is 111.

iii)

```
library(ggplot2)
```

Load lmiss and imiss files into R

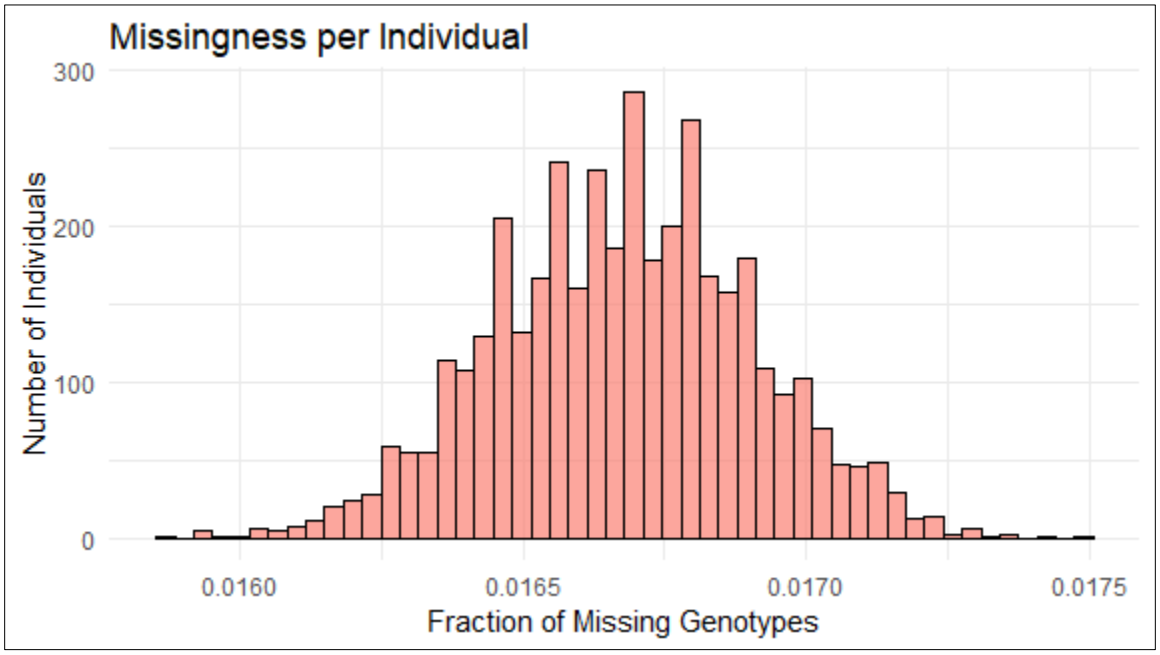
```
snps_missing <- read.table("missing_data.lmiss", header = FALSE,

col.names = c("CHR", "SNP", "N_MISS", "N_TOTAL", "F_MISS"))

snps_missing$F_MISS <- as.numeric(snps_missing$F_MISS)

individuals_missing <- read.table("missing_data.imiss", header =
FALSE,

col.names = c("FID", "IID", "MISS_PHENO", "N_MISS", "N_GENO",
"F_MISS"))

individuals_missing$F_MISS <-
as.numeric(individuals_missing$F_MISS)
```

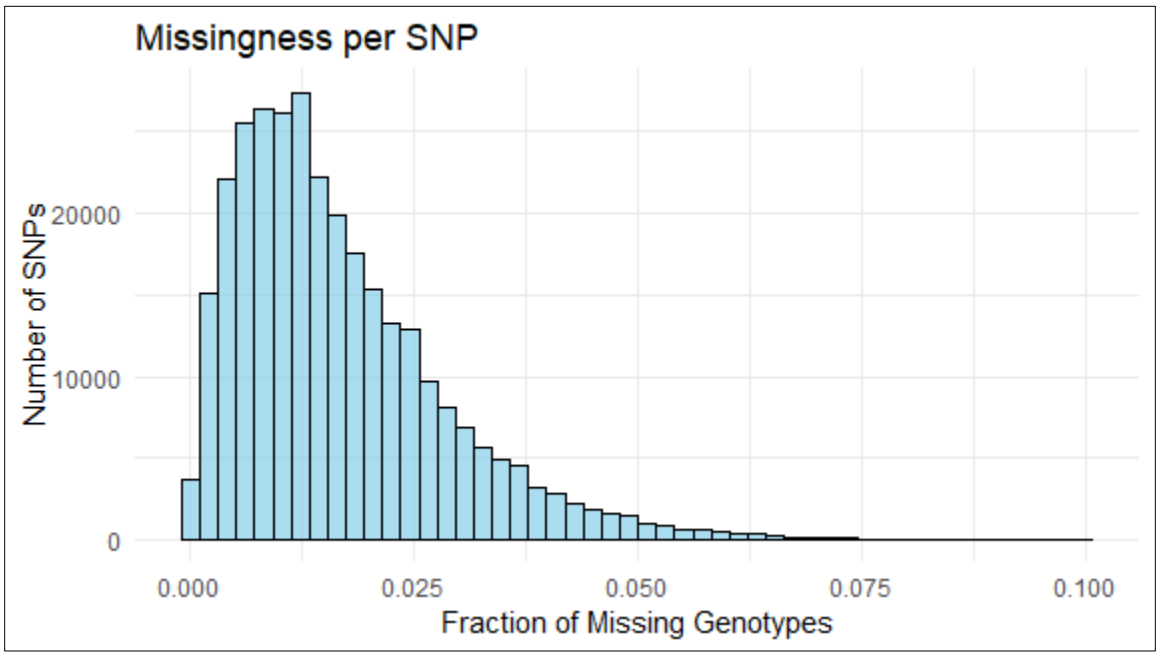
Plot missingness per SNP:

```
ggplot(snps_missing, aes(x = F_MISS)) +

geom_histogram(bins = 50, fill = "skyblue", color = "black",
alpha = 0.7) +

labs(title = "Missingness per SNP", x = "Fraction of Missing
Genotypes", y = "Number of SNPs") +

theme_minimal()
```

Plot missingness per Individual:

```
ggplot(individuals_missing, aes(x = F_MISS)) +

geom_histogram(bins = 50, fill = "salmon", color = "black", alpha
= 0.7) +

labs(title = "Missingness per Individual", x = "Fraction of
Missing Genotypes", y = "Number of Individuals") +
theme_minimal()
```





Analysis: The histogram of SNP missingness shows a peak around 0.01, indicating that the majority of SNPs have approximately 1% of their genotypes missing. However, the distribution has a slight rightward skew, suggesting that a small number of SNPs have higher levels of missing data. To improve data quality, quality control steps can be applied to filter out these SNPs. The histogram of individual missingness follows a roughly normal distribution, with most individuals showing missingness between 0.0160 and 0.0175. This represents a relatively low level of missing data, as missingness below 5% is generally considered acceptable. Overall, the low level of missingness in this dataset indicates high data quality and suggests that the dataset is reliable for analysis.

Q3 Allele frequencies (10 marks)
i) Which is the minor allele for SNP rs4970357 and what is its frequency?
ii) Create a plot which shows the overall distribution of MAF

i)

```
plink --bfile gwas --freq --snp rs4970357 --noweb --out
snp_frequency
```

This PLINK command calculates the allele frequencies for a specific SNP, rs4970357, in the GWAS dataset. It outputs the results in a file named snp_frequency.frq

CHR SNP A1 A2 MAF NCHROBS

1 rs4970357 C A 0.05028 7856

The minor allele frequency (MAF) is 0.05028, meaning the minor allele (A) occurs in approximately 5.03% of the population.

ii)

```
library(ggplot2)
```

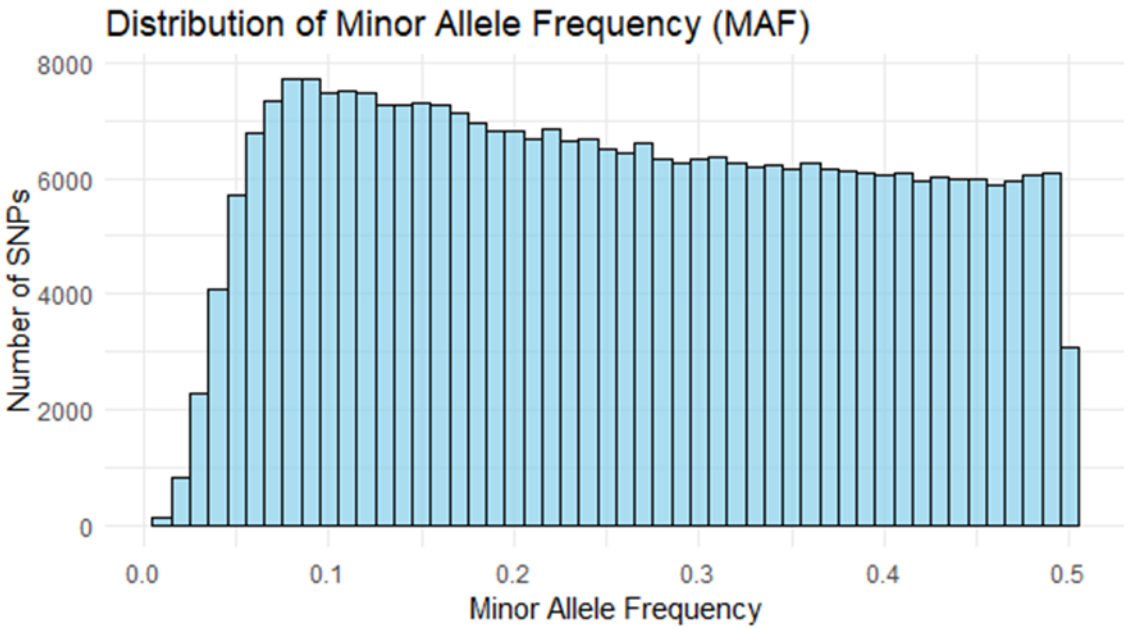
Load MAF Data

```
maf_data <- read.table("maf_data.frq", header = TRUE)

head(maf_data)
```

Creates Histogram plot

```
ggplot(maf_data, aes(x = MAF)) +

geom_histogram(bins = 50, fill = "skyblue", color = "black",
alpha = 0.7) +

labs(title = "Distribution of Minor Allele Frequency (MAF)",
```

```
x = "Minor Allele Frequency",

y = "Number of SNPs") +

theme_minimal()
```



Q4 Other QC steps (10 marks)
i) Choose two additional QC steps to carry out on the data explaining your choice and visualising
any results as appropriate.

```
PS C:\Users\boghe\Downloads\GWAS> plink --bfile gwas --geno 0.05
--make-bed --out SNP.miss –noweb
```

This PLINK command applies a quality control filter to remove SNPs (single nucleotide polymorphisms) with a high rate of missing genotypes.

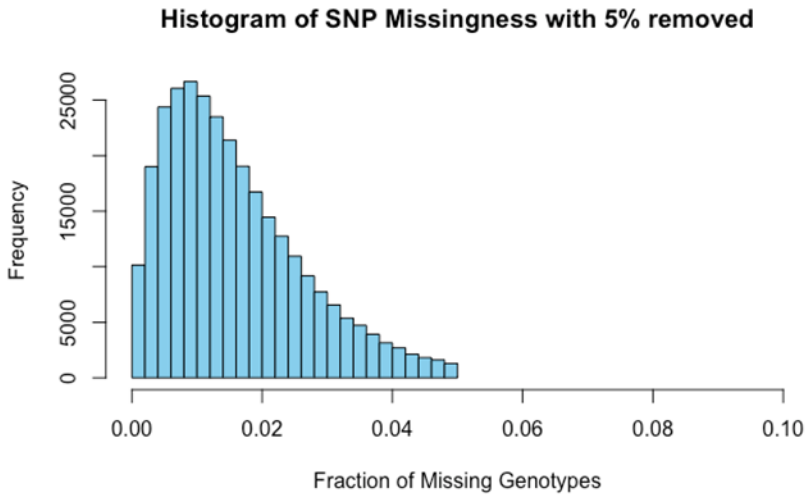
Filters out SNPs with a missing genotype rate greater than 5%. SNPs where more than 5% of the data is missing will be excluded.

Create a new binary dataset after applying the filter, including. bed, .bim, and .fam files.

Total genotyping rate is 0.983323.

5552 variants removed due to missing genotype data (--geno).

300550 variants and 4000 people pass filters and QC.



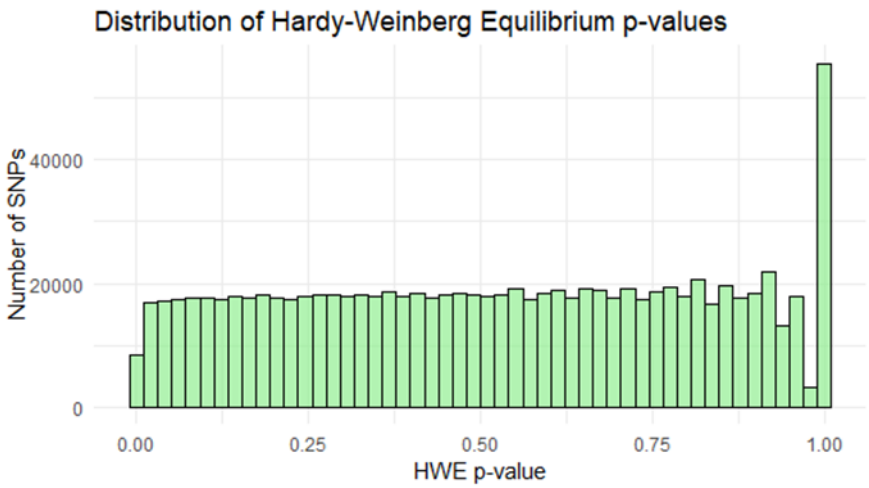
QC2

The Hardy-Weinberg Equilibrium (HWE) test assesses whether allele frequencies in a population conform to the expected distribution under random mating conditions. Significant deviations from HWE can suggest issues such as genotyping errors, population stratification, or the influence of selection pressures. In case-control studies, it is standard practice to evaluate HWE specifically within the control group.

```
PS C:\Users\boghe\Downloads\GWAS> plink --bfile SNP.miss --hardy
--noweb --out QC.data
```

This command applies a filter to remove SNPs that significantly deviate from Hardy-Weinberg Equilibrium (HWE), which may indicate genotyping errors or population structure issues.

Total genotyping rate is 0.984114.
307 variants removed due to Hardy-Weinberg equilibrium test (--hwe).
300,243 variants and 4,000 individuals pass filters and QC.



The consistent distribution of p-values between 0 and 0.9 indicates that the majority of SNPs align well with Hardy-Weinberg Equilibrium (HWE) and show no notable deviations. The noticeable increase at a p-value of 1 reflects a large proportion of SNPs that fully meet HWE expectations. However, SNPs with p-values approaching zero require closer examination to decide if they should be excluded due to potential issues with data quality or biological influences such as selection pressures or population structure.

Q5

```
plink --bfile QC.data --snp rs9651273 --model --out model.results
–noweb
```

This command performs association testing on the specific SNP rs9651273 under different genetic models. model.results.model: Contains the p-values and association statistics for SNP rs9651273 under the various genetic models.

```
Get-Content "genetic_model_results.model" | Select-Object -First
1
```

```
Select-String "rs9651273" "genetic_model_results.model"
```

This command retrieves the header row from the genetic_model_results.model file. The header provides column names describing the association results for each genetic model tested.

"CHR	*SNP*	*A1*	*A2*	*TEST*	*AFF*	*UNAFF*	*CHISQ*	*DF*	*P*
1	*rs9651273*	*A*	*G*	*GENO*	*322/1013/650*	*305/939/714*	*6.085*	*2*	*0.04773*

1	*rs9651273*	*A*	*G*	*TREND*	*1657/2313*	*1549/2367*	*3.995*	*1*	*0.04563*
1	*rs9651273*	*A*	*G*	*ALLELIC*	*1657/2313*	*1549/2367*	*3.892*	*1*	*0.04853*
1	*rs9651273*	*A*	*G*	*DOM*	*1335/650*	*1244/714*	*6.029*	*1*	*0.01407*
1	*rs9651273*	*A*	*G*	*REC*	*322/1663*	*305/1653*	*0.3062*	*1*	*0.58"*

The SNP exhibits the lowest p-value under the dominant (DOM) model (0.01407), which is below the 0.05 threshold. This indicates a statistically significant association between the SNP and the phenotype

6i

```
PS C:\Users\boghe\Downloads\GWAS> plink --bfile QC.data --assoc --adjust --out assoc.adjusted –noweb
```

This command performs association testing on the filtered dataset (QC.data), adjusts the p-values for multiple testing, and saves the results to assoc.adjusted.

Load association into R:

```
assoc.results <- read.table("assoc.data.assoc", header=TRUE)
```

Apply Bonferroni and FDR corrections:

```
assoc.results$Bonf <- p.adjust(assoc.results$P, method = "bonferroni")

assoc.results$FDR <- p.adjust(assoc.results$P, method = "fdr"

sum(assoc.results$Bonf < 0.05 & assoc.results$FDR < 0.05)
```

[1] 9

This command produced the output "9," indicating that nine SNPs were considered significant under both the Bonferroni and FDR corrections. The Bonferroni method minimizes the risk of false positives, while the FDR method manages the rate of false discoveries. These significant SNPs are highly likely to have a true association with the phenotype.

ii

Genomic control (GC) is a method used to assess and account for population structure in genome-wide association studies (GWAS). Population structure can lead to inflated test statistics, which may result in false-positive associations. This inflation is quantified using the lambda factor, which measures the extent to which test statistics deviate from their expected distribution under the null hypothesis. A lambda value significantly greater than 1 indicates the presence of population stratification affecting the analysis.

Genomic inflation factor (based on median chi-squared) is 1.00949

The lambda value of 1.00949 is nearly equal to 1, indicating minimal or no population stratification. This suggests that the case and control groups are predominantly uniform regarding ancestry, and any population structure that exists is unlikely to influence the results significantly.

Q7

```
PS C:\Users\boghe\Downloads\GWAS> plink --bfile gwas --logistic --adjust --noweb --out logistic_regression_without_covariates
```

This command runs logistic regression to identify SNP-phenotype associations without considering covariates and applies multiple testing correction.

```
PS C:\Users\boghe\Downloads\GWAS> plink --bfile gwas --logistic --covar gwas.covar --sex --adjust --out logistic_results_with_covariates –noweb
```

This command performs logistic regression with additional covariates (e.g., sex and variables from gwas.covar), adjusting for confounders, and applies multiple testing correction.

```
PS C:\Users\boghe\Downloads\GWAS> Select-String "Genomic inflation factor" logistic_regression_with_covariates.log
```

logistic_regression_with_covariates.log:47:Genomic inflation factor (based on median chi-squared) is 1.01129

```
PS C:\Users\boghe\Downloads\GWAS> Select-String "Genomic inflation factor" logistic_regression_without_covariates.log
```

logistic_regression_without_covariates.log:44:Genomic inflation factor (based on median chi-squared) is 1.01133

With Covariates: λ = 1.01129. When age and sex were included as covariates in the logistic regression model, the inflation factor slightly decreased compared to the analysis without covariates. Without Covariates: λ = 1.01133 When the model did not include covariates, the inflation factor was marginally higher. The slight decrease in λ with covariates indicates that including age and sex accounts for a small amount of variation in the data that could otherwise contribute to inflation. Both values (~1.011) are very close to 1, which suggests minimal genomic inflation. This implies that There is little to no evidence of population stratification or systematic bias in the data. The change in λ is negligible, meaning the covariates have a limited impact on addressing inflation in this dataset.

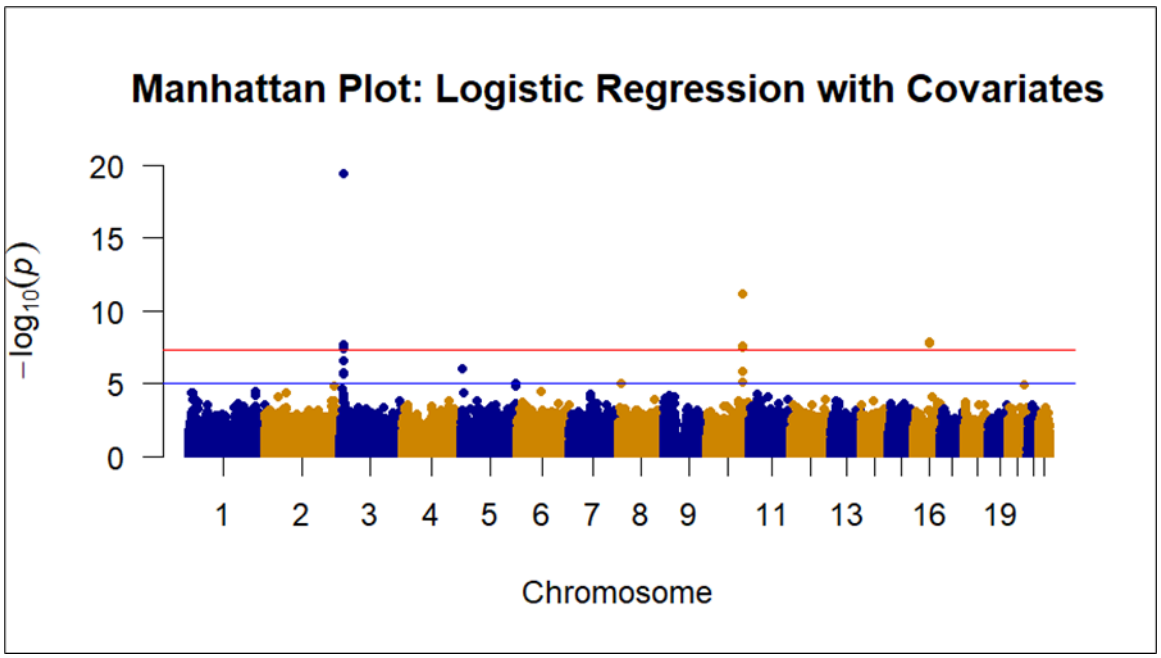
Q8

```
library(qqman)
```

Association test results were loaded in and plotted

```
logistic <-
read.table("logistic_regression_with_covariates.assoc.logistic" ,
header = TRUE)
```

```
**Output:** manhattan(logistic, main = "Manhattan Plot", ylim =
c(0,20), cex = 0.3, cex.axis = 0.9, col = c("blue","grey"),
chrlabs = c(1:20, "P", "Q"))
```



Observations:

SNPs above the red line (e.g., on chromosomes 2, 3, and 11) show strong evidence of association with the phenotype. These are the most likely candidates for further investigation. SNPs between the blue and red lines suggest potential associations, but they do not reach the genome-wide significance threshold. These may still warrant follow-up, especially if they are in regions of biological relevance. The SNP distribution across chromosomes appears uniform, indicating no major issues with data quality or analysis. There is a particularly strong signal on chromosome 3 with a very high -log10(p) value (>15), which may represent a highly significant genetic variant.

Significant SNPs identified:

rs7615580 rs6768587 rs2028760 rs6802898 rs7901695 rs7903146 rs7904519 rs8050136 rs3751812

Several single nucleotide polymorphisms (SNPs) identified are strongly associated with metabolic disorders, particularly type 2 diabetes mellitus (T2DM) and obesity. Among these, rs7903146 in the TCF7L2 gene is one of the most significant markers for T2DM risk, with the T allele impairing insulin secretion and glucose tolerance due to β-cell dysfunction [1]. Other SNPs in the TCF7L2 gene, such as rs7901695, rs7904519, rs6768587, rs2028760, and rs6802898, have been linked to T2DM in genome-wide association studies, demonstrating consistent effects on glucose regulation and insulin activity [2,3]. SNPs in the FTO gene, including rs8050136, rs3751812, and rs7615580, are associated with obesity, with risk alleles contributing to increased BMI, adiposity, and altered energy homeostasis, factors that indirectly heighten T2DM risk [4,5]. These findings underscore the pivotal roles of TCF7L2 and FTO variants in influencing susceptibility to T2DM and obesity, offering valuable insights for risk stratification and potential therapeutic targeting.

Bibliography

Lyssenko, V., et al. (2007). "The rs7903146 Variant in the TCF7L2 Gene Increases Risk of Type 2 Diabetes by Impairing Insulin Secretion." Diabetes, 56(12), 3105-3110.

Grant, S.F., et al. (2006). "Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes." Nature Genetics, 38(3), 320-323.

Zeggini, E., et al. (2007). "Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes." Science, 316(5829), 1336-1341.

Claussnitzer, M., et al. (2015). "FTO Obesity Variant Circuitry and Adipocyte Browning in Humans." The New England Journal of Medicine, 373(10), 895-907.

Loos, R.J., et al. (2008). "Common variants near MC4R are associated with fat mass, weight, and risk of obesity." Nature Genetics, 40(6), 768-775.