

GWAS Analysis Utilizing PLINK & R

AUTHOR

Sarah Mughal

1. Load R Libraries

```
library(data.table)
library(ggplot2)
```

Warning: package 'ggplot2' was built under R version 4.5.2

```
library(dplyr)
```

Attaching package: 'dplyr'

The following objects are masked from 'package:data.table':

between, first, last

The following objects are masked from 'package:stats':

filter, lag

The following objects are masked from 'package:base':

intersect, setdiff, setequal, union

```
library(qqman)
```

For example usage please run: `vignette('qqman')`

Citation appreciated but not required:

Turner, (2018). qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. Journal of Open Source Software, 3(25), 731, <https://doi.org/10.21105/joss.00731>.

2. Setup file structure

```
try(dir.create("../output"))
```

Warning in dir.create("../output"): '../output' already exists

```
try(dir.create("../software"))
```

Warning in dir.create("../software"): '../software' already exists

3. Download PLINK 1.9

```
if (!file.exists("../software/plink")) {  
  message("[*] Downloading PLINK 1.9...")  
  curwd <- getwd()  
  dir.create("../software", showWarnings = FALSE)  
  setwd("../software")  
  
  download.file(  
    "https://s3.amazonaws.com/plink1-assets/plink_mac_20231211.zip",  
    "plink.zip"  
  )  
  unzip("plink.zip")  
  
  system("chmod u+x plink")  
  system("ls -lh")  
  
  setwd(curwd)  
}
```

4. Allele Frequencies

```
system("../software/plink --bfile ../data/homework --freq --out ../output/stats_freq")
```

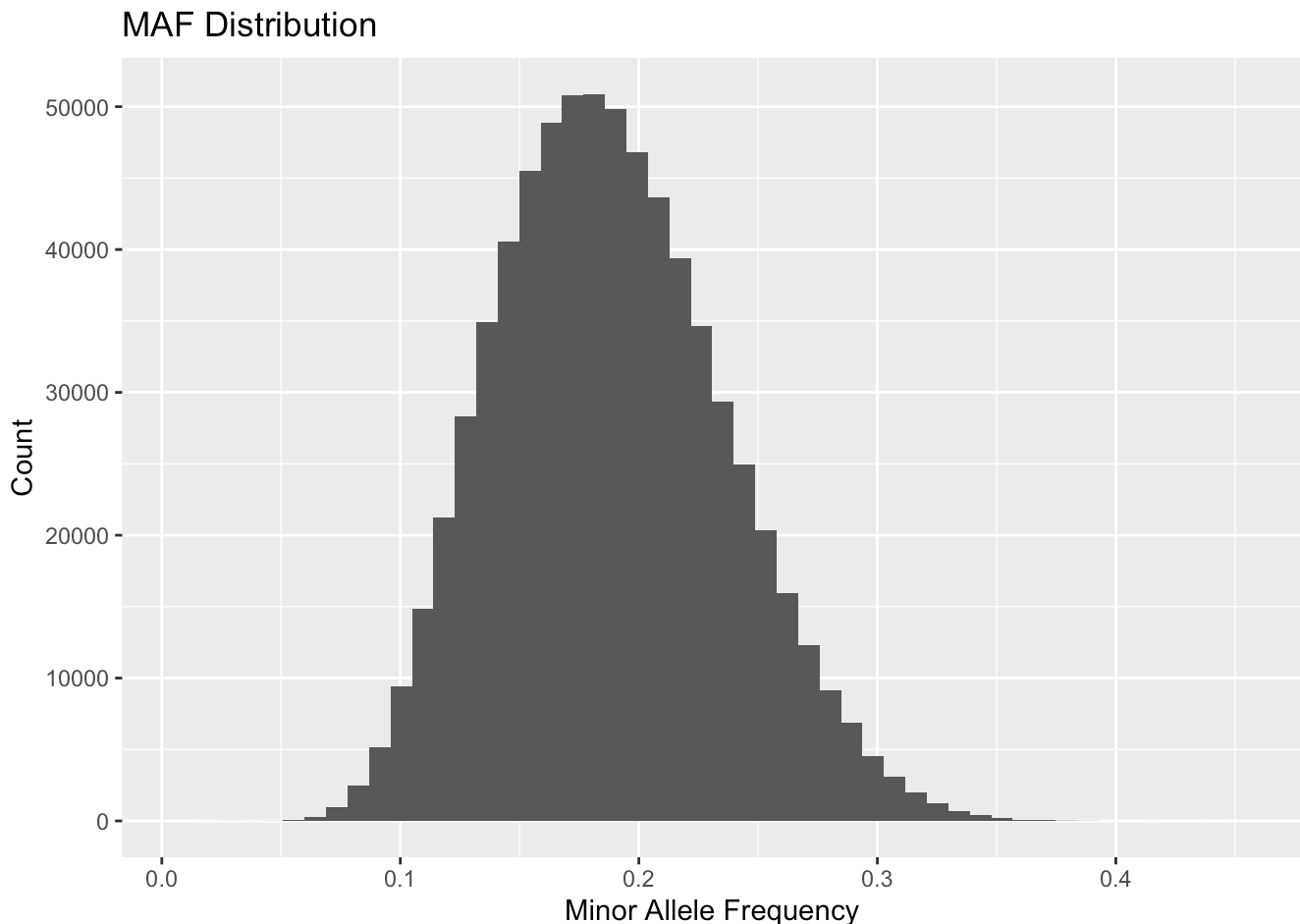
5. Load into R

```
freq <- fread("../output/stats_freq.frq")  
head(freq)
```

	CHR	SNP	A1	A2	MAF	NCHROBS
	<int>	<char>	<char>	<char>	<num>	<int>
1:	1	rs0	G	A	0.1247	20000
2:	1	rs1	G	A	0.1018	20000
3:	1	rs2	G	A	0.1118	20000
4:	1	rs3	G	A	0.1698	20000
5:	1	rs4	G	A	0.1390	20000
6:	1	rs5	G	A	0.2044	20000

TODO: Create a histogram of MAFs

```
ggplot(freq, aes(x = MAF)) +  
  geom_histogram(bins = 50) +  
  labs(title = "MAF Distribution", x = "Minor Allele Frequency", y = "Count")
```



QUESTION: What is the range of the frequency you see in these plots, and why?

ANSWER: The minor allele frequencies range approximately from 0.05 to 0.35, with most variants centered around ~0.18–0.20. Unlike real genetic data, which is typically skewed toward rare variants, this distribution appears bell-shaped. This pattern is likely due to the simulated nature of the dataset, where allele frequencies were generated from a specific distribution rather than reflecting natural population genetics processes.

6. Delete the dataset to save memory (optional)

```
rm(freq)
```

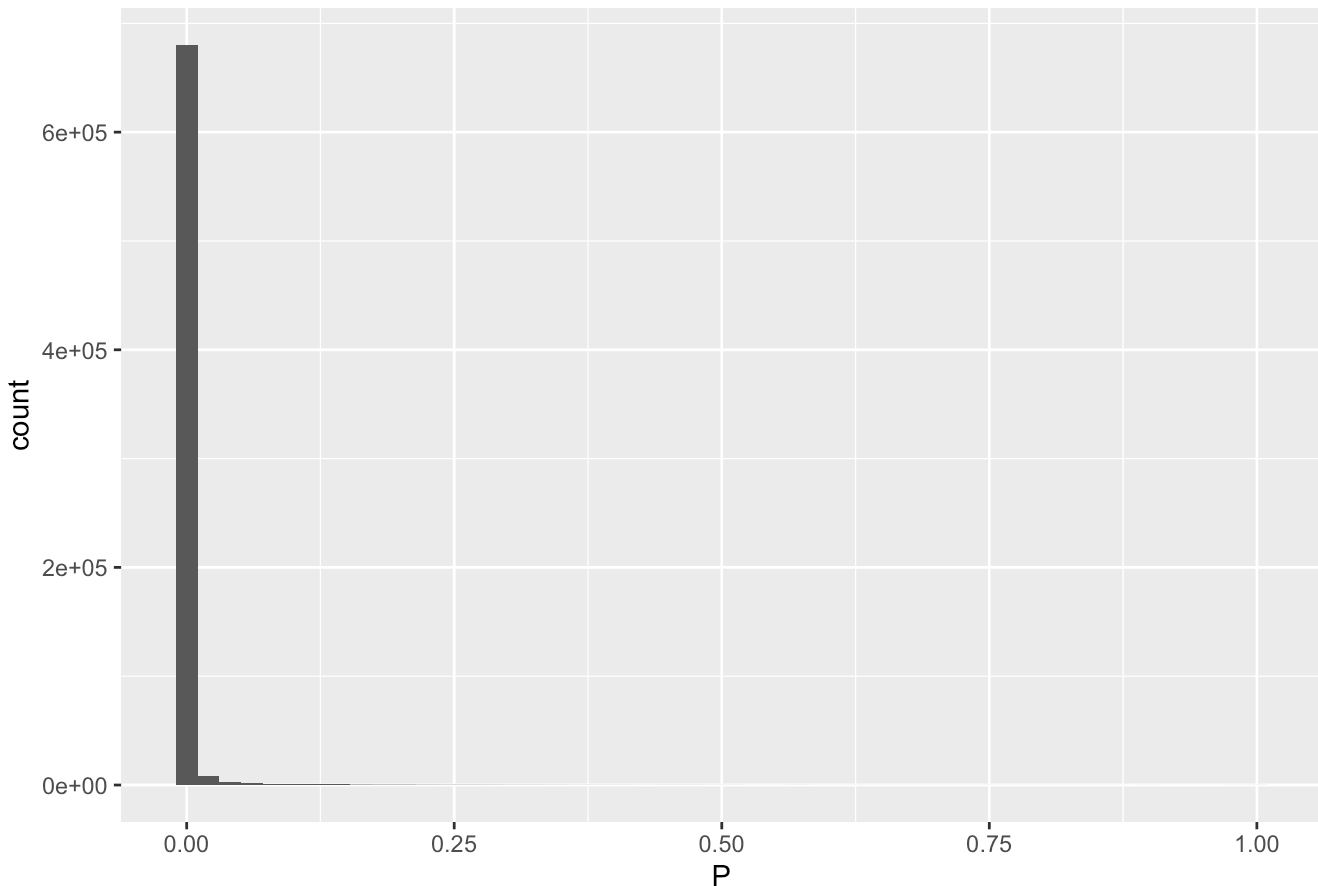
7. Hardy-Weinberg Equilibrium

```
system("../software/plink --bfile ../data/homework --hardy --out ../output/stats_hwe")  
hwe <- fread("../output/stats_hwe.hwe")
```

TODO: Create a histogram of the P-Values

```
ggplot(hwe, aes(x = P)) +  
  geom_histogram(bins = 50) +  
  labs(title = "HWE P-value Distribution")
```

HWE P-value Distribution



```
median(hwe$P, na.rm = TRUE)
```

```
[1] 5.017e-13
```

QUESTION: What is the median P-value?

ANSWER: The median Hardy-Weinberg P-value is approximately 5.017e-13

QUESTION: Any guesses why this might be happening?

ANSWER: The extremely small median P-value suggests widespread deviation from Hardy-Weinberg equilibrium across SNPs. In real data, this might indicate genotyping errors, population stratification, inbreeding, or selection. However, since this dataset is simulated, it is likely that the genotype data were generated in a way that does not strictly follow Hardy-Weinberg assumptions, leading to systematic deviation.

8. Linkage Disequilibrium Pruning

```
system("../software/plink --bfile ../data/homework --maf 0.05 --indep-pairwise 50 5 0.2 --
```

Because nearby SNPs on the same chromosome tend to be inherited together (linkage disequilibrium), we prune correlated variants prior to PCA to ensure that the principal components reflect genome-wide ancestry structure rather than local chromosomal linkage effects.

9. PCA on Pruned SNPs

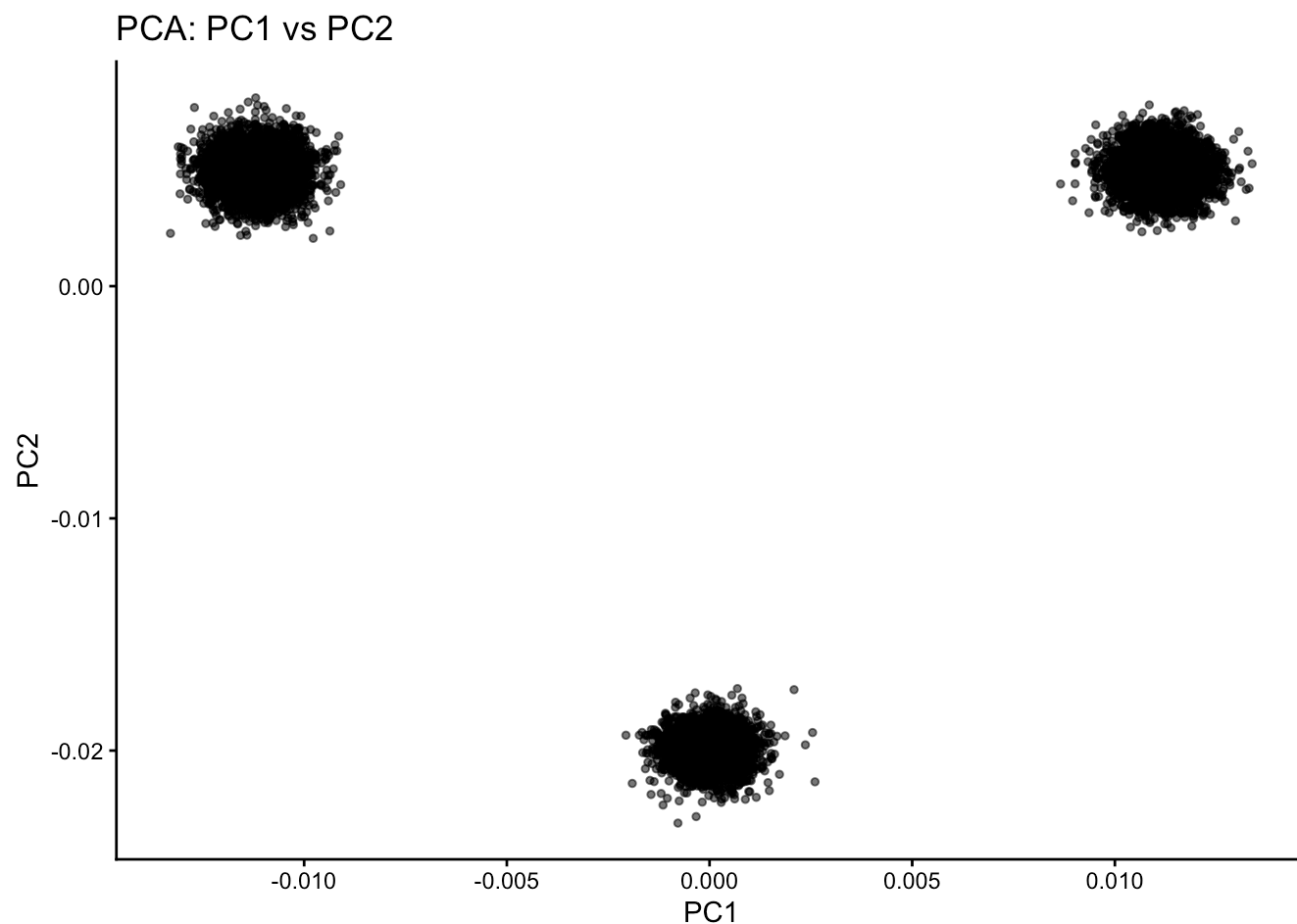
```
system("../software/plink --bfile ../data/homework --extract ../output/pruning.prune.in --
```

10. Load PCA Eigenvectors

```
pcs <- read.table("../output/pca_results.eigenvec", header = FALSE)  
colnames(pcs) <- c("FID", "IID", paste0("PC", 1:10))
```

TODO: Plot at least PC1 vs. PC2

```
library(ggplot2)  
  
ggplot(pcs, aes(PC1, PC2)) +  
  geom_point(alpha = 0.6, size = 1) +  
  labs(title = "PCA: PC1 vs PC2") +  
  theme_classic()
```



```
# How much variance does each PC explain  
eigenvals <- scan("../output/pca_results.eigenval")  
eigenvals / sum(eigenvals)
```

```
[1] 0.44235172 0.27627356 0.03527511 0.03521911 0.03518510 0.03516881  
[7] 0.03516474 0.03513704 0.03511973 0.03510507
```

The PCA plot shows three clearly separated clusters along PC1 and PC2, indicating strong population structure within the simulated dataset. This suggests the presence of multiple genetically distinct subgroups. Such structure can confound association analyses if not properly adjusted for, which justifies including principal components as covariates in the GWAS.

11. Load & Prepare Phenotype

```
phe <- fread("../data/homework.fam", header = FALSE)  
head(phe)
```

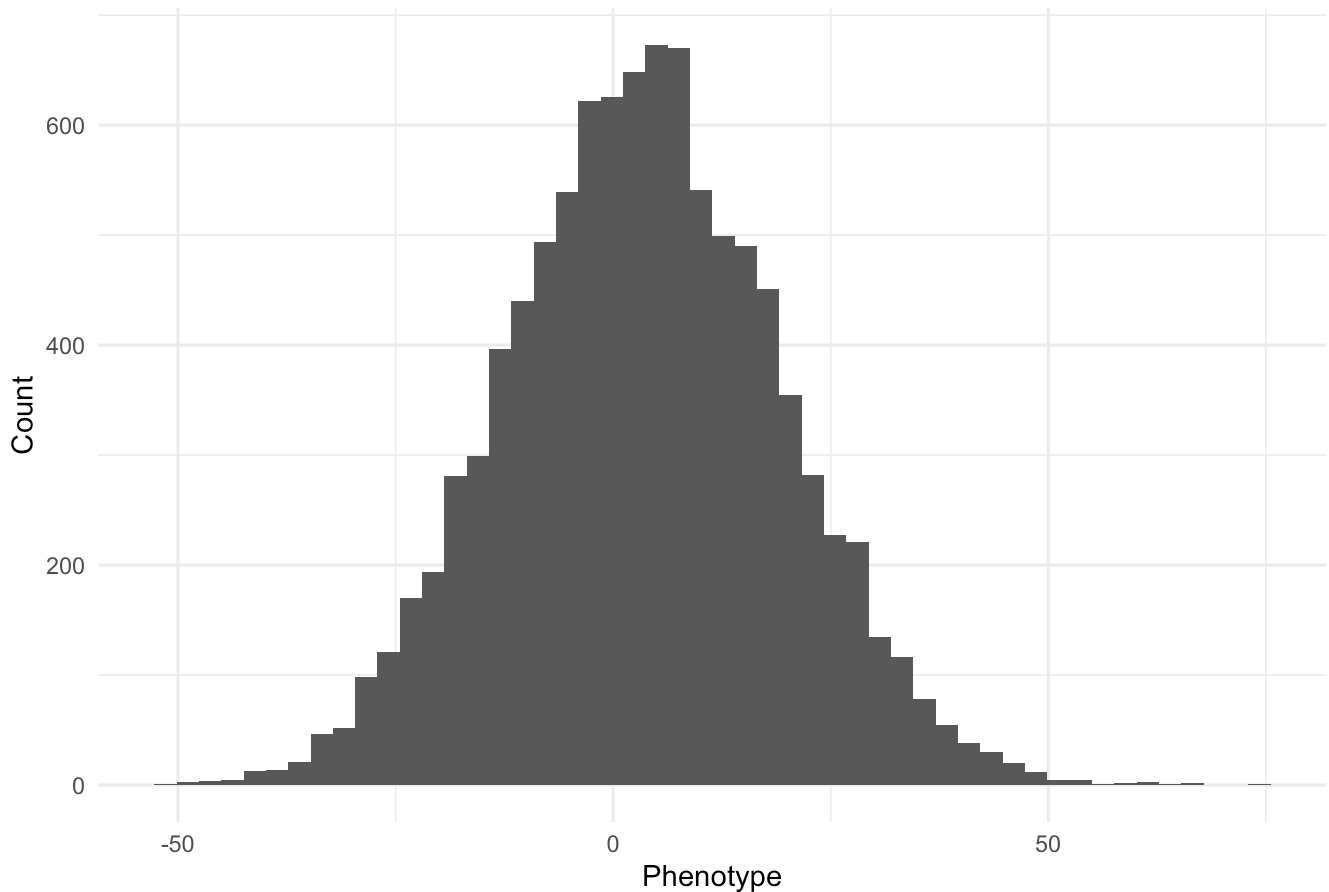
	V1	V2	V3	V4	V5	V6
	<int>	<int>	<int>	<int>	<int>	<num>
1:	1	1	0	0	1	5.9097139
2:	2	2	0	0	1	-12.8233179
3:	3	3	0	0	1	8.6122768
4:	4	4	0	0	1	2.5378888
5:	5	5	0	0	1	18.6831344
6:	6	6	0	0	1	-0.8993228

```
# Keep FID, IID, phenotype column (column 6)  
phe <- phe[, .(FID = V1, IID = V2, Pheno = V6)]
```

TODO: Histogram of Phenotype

```
ggplot(phe, aes(x = Pheno)) +  
  geom_histogram(bins = 50) +  
  labs(title = "Phenotype Distribution",  
        x = "Phenotype",  
        y = "Count") +  
  theme_minimal()
```

Phenotype Distribution



The phenotype appears approximately normally distributed, supporting the use of linear regression for the GWAS analysis.

TODO: Summarize the Phenotype

```
summary(phe$Pheno)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-51.450	-7.340	3.393	3.403	14.111	74.256

The phenotype is continuous, with a mean of approximately 3.4 and a range from -51.5 to 74.3. The distribution appears roughly symmetric, supporting the use of linear regression in the GWAS analysis.

12. Write Phenotype File for PLINK

```
fwrite(phe, file="../output/homework.phe", quote=FALSE, sep="\t")
```

13. GWAS A: Naive (No Covariates)

```
system("../software/plink --bfile ../data/homework --maf 0.005 --linear --pheno ../output
```

14. Load Naive Results + Sanity Check

```
gwas_res <- fread("../output/gwas_naive.assoc.linear", header = TRUE)
table(gwas_res$CHR)
```

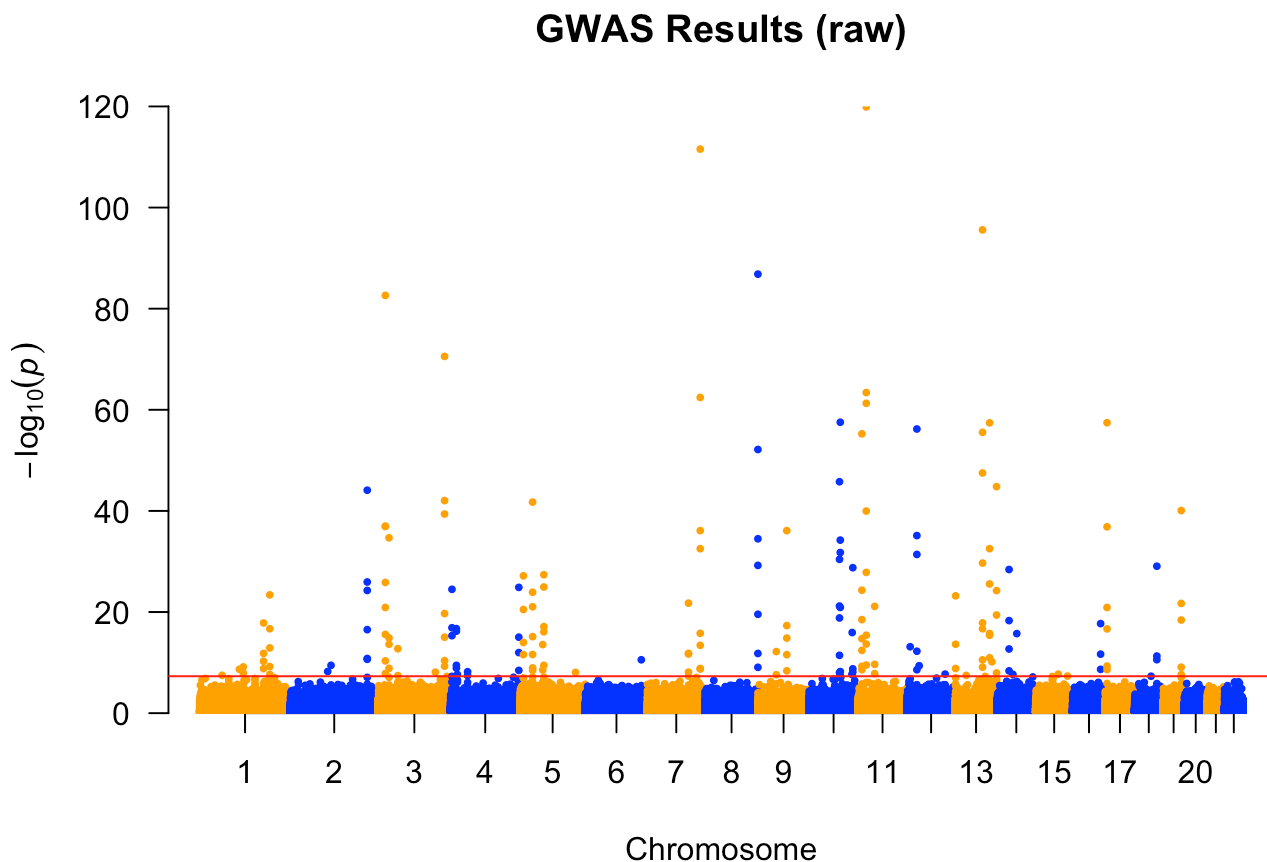
```

 1      2      3      4      5      6      7      8      9     10     11     12     13
60806 58860 48644 46455 44023 41591 38672 35510 34294 32835 32835 32348 27970
 14     15     16     17     18     19     20     21     22
26025 24808 21890 19701 18971 14350 15323 11674 12415

```

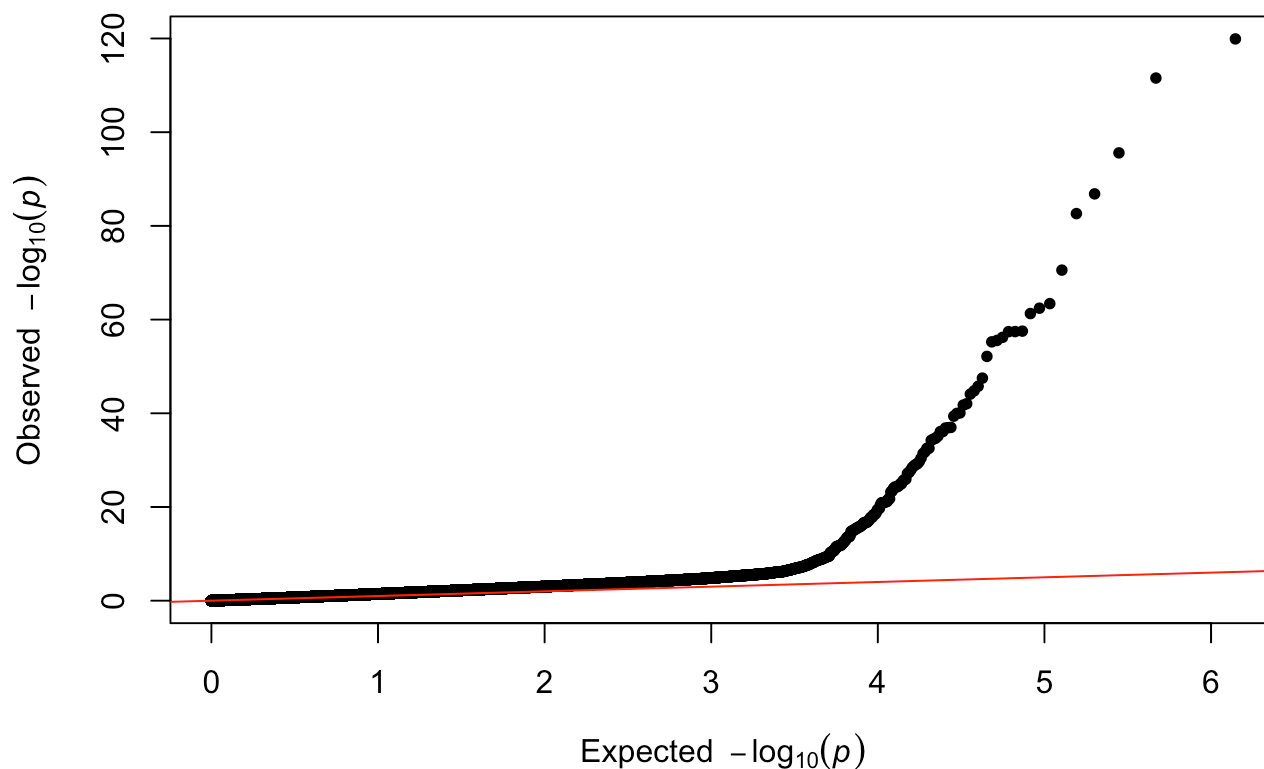
15. Manhattan + QQ Plot for Naive

```
manhattan(gwas_res,
  chr="CHR", bp="BP", p="P", snp="SNP",
  main="GWAS Results (raw)",
  col=c("orange", "blue"),
  suggestiveline = FALSE,
  genomewideline = -log10(5e-8),
  cex = 0.6)
```



```
qq(gwas_res$P, main="Q-Q: Naive")
```

Q-Q: Naive



16. Write Covariates for PLINK

```
covar <- pcs[, c("FID", "IID", paste0("PC", 1:5))]
fwrite(covar, "../output/covariates.txt", row.names = FALSE, quote = FALSE, sep = "\t")
```

17. GWAS B: GWAS Adjusted for Ancestry (Covariates)

```
system("../software/plink --bfile ../data/homework --linear --pheno ../output/homework.ph
```

18. Load Adjusted Results + Sanity Check

```
gwas_res <- fread("../output/gwas_corrected.assoc.linear")
table(gwas_res$CHR)
```

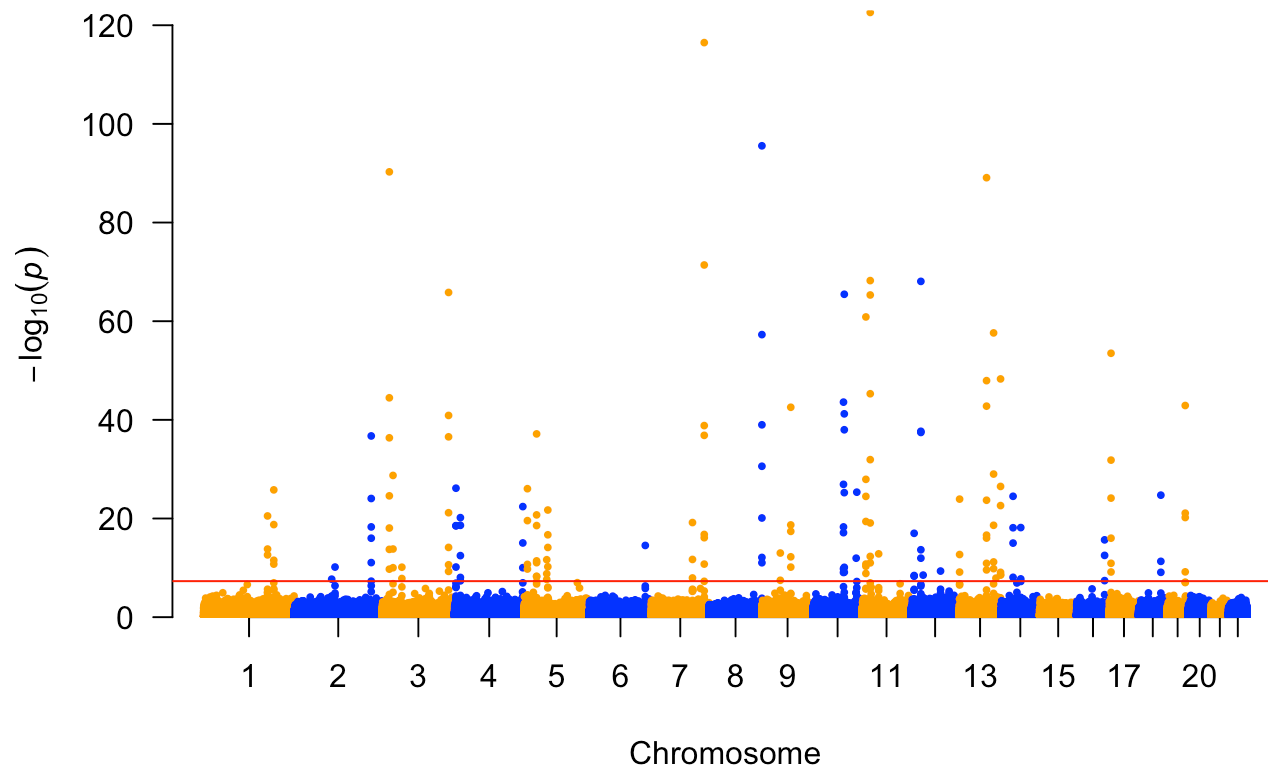
```

 1    2    3    4    5    6    7    8    9   10   11   12   13
60806 58860 48644 46455 44023 41591 38672 35510 34294 32835 32835 32348 27970
 14   15   16   17   18   19   20   21   22
26025 24808 21890 19701 18971 14350 15323 11674 12415
```

TODO: Manhattan + QQ Plot for Adjusted

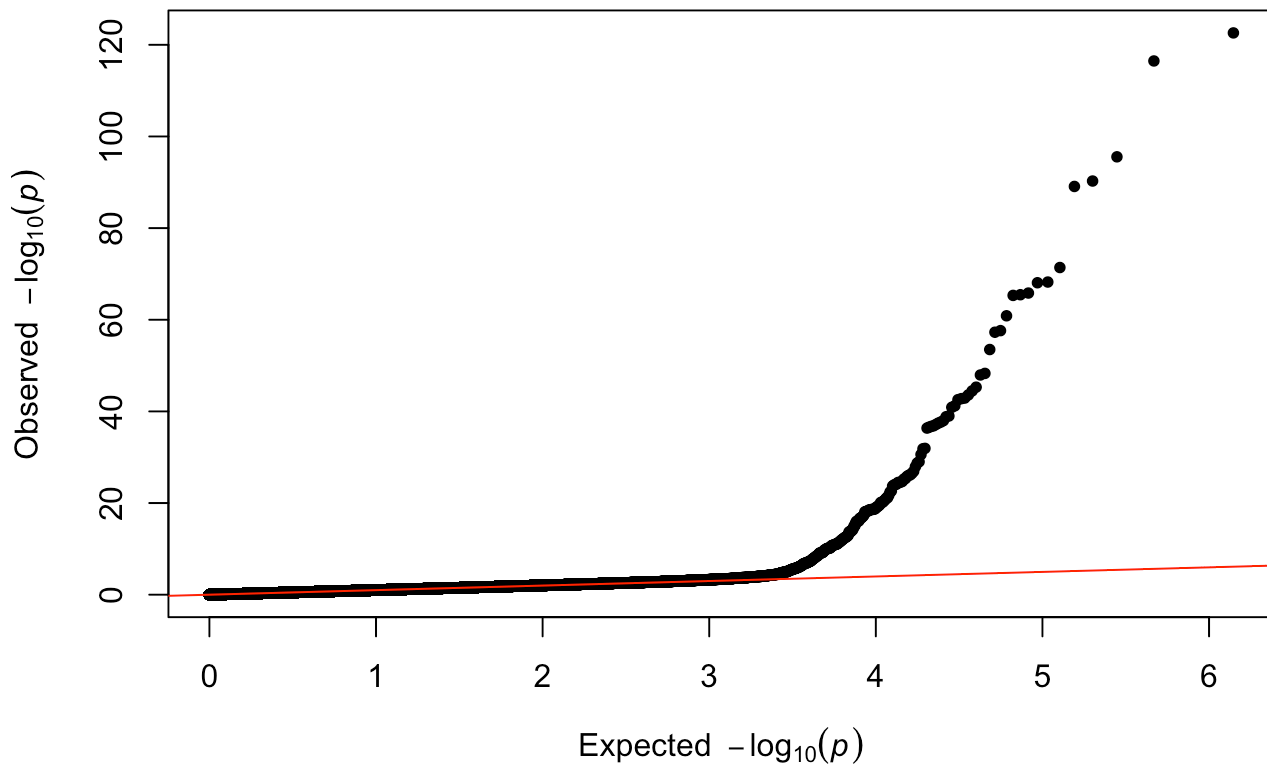
```
manhattan(gwas_res,
  chr="CHR", bp="BP", p="P", snp="SNP",
  main="GWAS Results (PC-adjusted)",
  col=c("orange", "blue"),
  suggestiveline = FALSE,
  genomewideline = -log10(5e-8),
  cex = 0.6)
```

GWAS Results (PC-adjusted)



```
qq(gwas_res$P, main="Q-Q: Corrected")
```

Q-Q: Corrected



The naive GWAS showed strong inflation, meaning many SNPs appeared significant even when they likely were not. The Q-Q plot clearly deviated from what we would expect under the null hypothesis, suggesting widespread false positives caused by population structure. After adjusting for ancestry using principal components, the Q-Q plot moved much closer to the expected line, indicating that much of the inflation was due to confounding rather than true genetic effects. Although some strong signals remained, the overall number of false positives was reduced, demonstrating the importance of correcting for population structure in GWAS.