QG_Midterm

Yuchen Sun

2023-03-29

```
library(dplyr)
library(MASS) # for matrix calculations
library(ggplot2)
wd <- "C:/Yuchen/WCM/Courses/2023Spring/CMPB5007/Midterm/"
```

1. Import the phenotype data from the file 'midterm phenotypes.txt' and

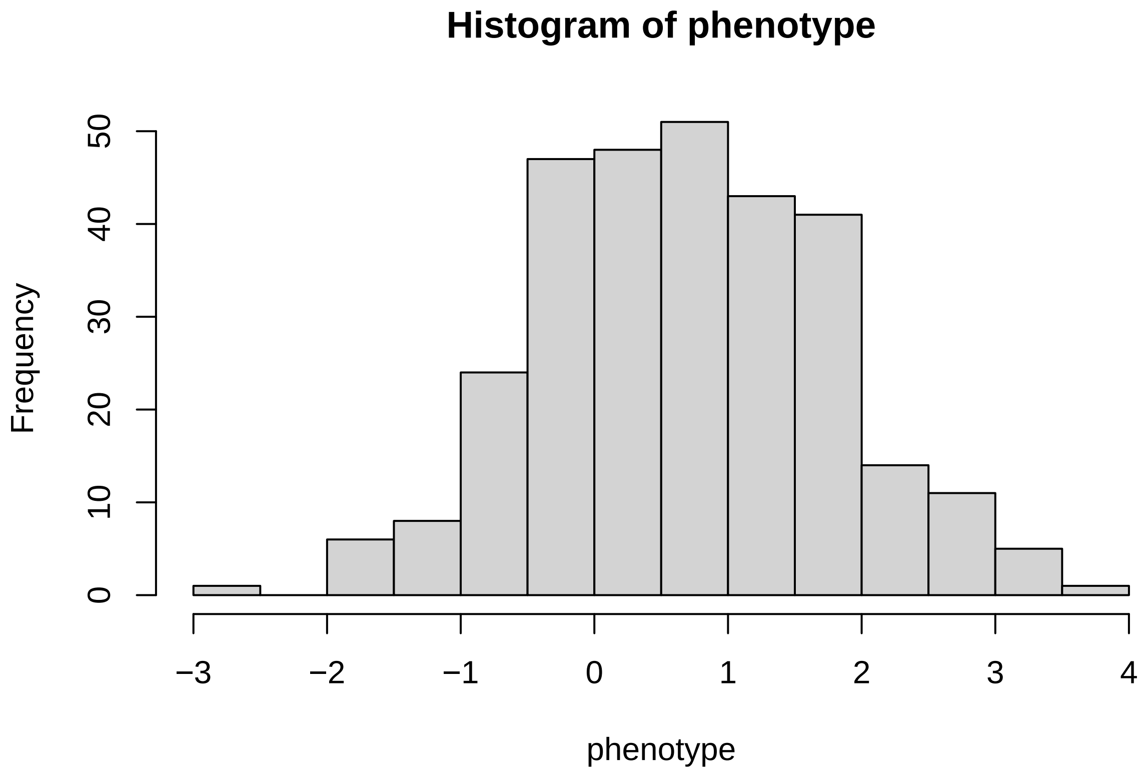
- ***(a). Calculate and report the total sample size n***

```
phenotype <- read.table(paste0(wd,"midterm_phenotypes.txt"))$V1
n <- length(phenotype)
cat("Total sample size n =", n)
```

```
## Total sample size n = 300
```

- ***(b). Plot a histogram of the phenotypes***

```
hist(phenotype)
```



2. Import the genotype data from the file 'midterm genotypes.txt'

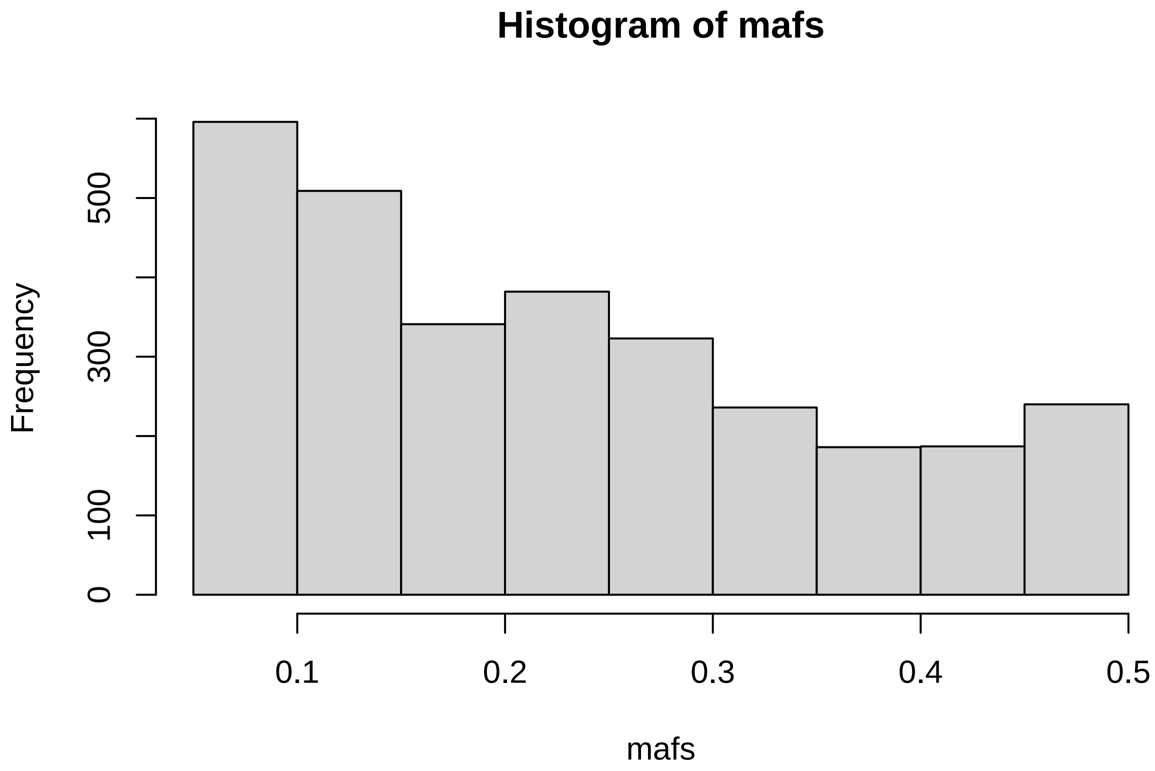
- *(a). Calculate and report the number of SNPs N*

```
genotype <- read.csv(paste0(wd,"midterm_genotypes.txt"),header=T)
N <- dim(genotype)[2]
cat("Number of SNPs N =", N)
```

```
## Number of SNPs N = 3000
```

- *(b). Calculate the MAF for every SNP and plot a histogram of the MAFs*

```
calc_maf <- function(snps) {
  count <- table(snps)
  return (min(count)/sum(count))
}
mafs <- apply(genotype,2,calc_maf)
hist(mafs)
```



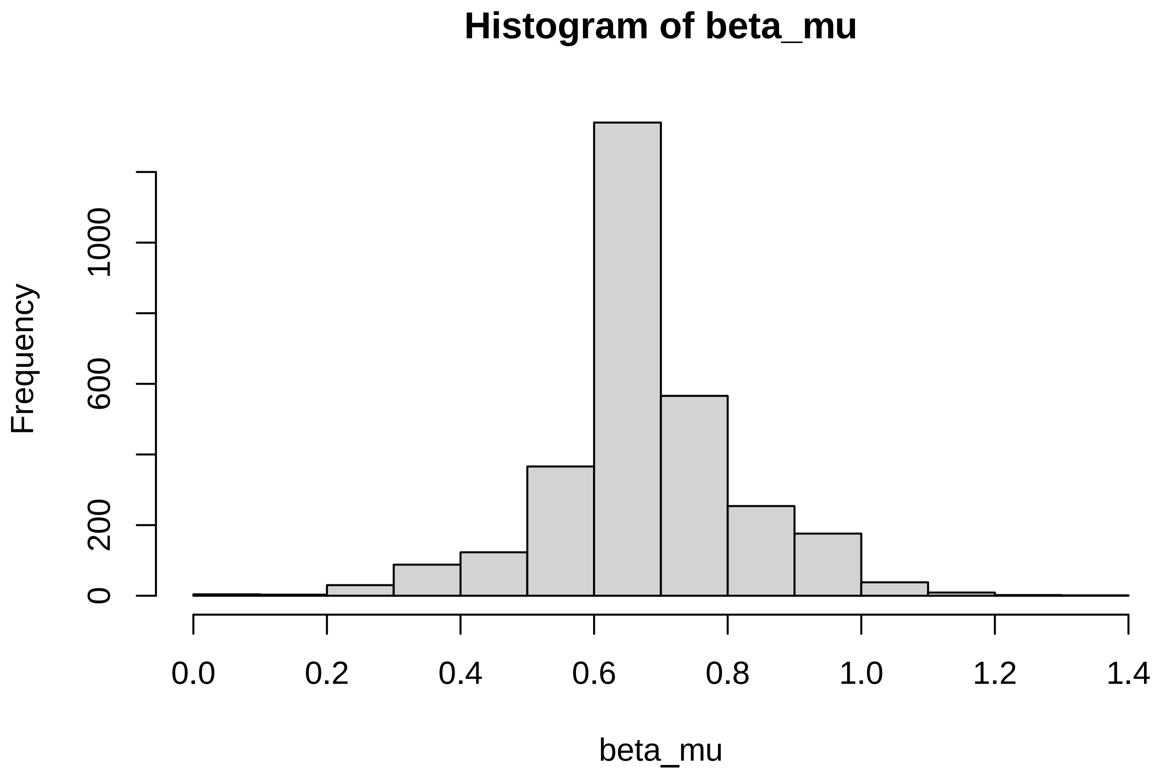
3. Write code to calculate $MLE(\hat{\beta}) = [\hat{\beta}_\mu, \hat{\beta}_a, \hat{\beta}_d]$ ***for each SNP and***

```
xa_convert <- function(snps) {
  count <- table(snps)
  minor <- names(count[count==min(count)])[1]
  xa <- snps==minor
  return (xa[seq(1,2*n,2)]+xa[seq(2,2*n,2)])
}

xd_convert <- function(snps) {
  count <- table(snps)
  minor <- names(count[count==min(count)])[1]
  xd <- snps==minor
  return (xd[seq(1,2*n,2)]!=xd[seq(2,2*n,2)])
}

xa <- matrix(NA,nrow=n,ncol=N)
xd <- matrix(NA,nrow=n,ncol=N)
for (i in 1:N) {
  xa[,i] <- xa_convert(genotype[,i])
  xd[,i] <- xd_convert(genotype[,i])
}
xa <- xa-1
xd <- xd*2-1
```

```
calc_mle <- function(y, xa, xd) {
  x <- cbind(1,xa,xd)
  mle <- ginv(t(x)%*%x)%*%t(x)%*%y
  return (mle)
}

beta <- matrix(NA,nrow=3,ncol=N)
for (i in 1:N) {
  beta[,i] <- calc_mle(phenotype,xa[,i],xd[,i])
}
```

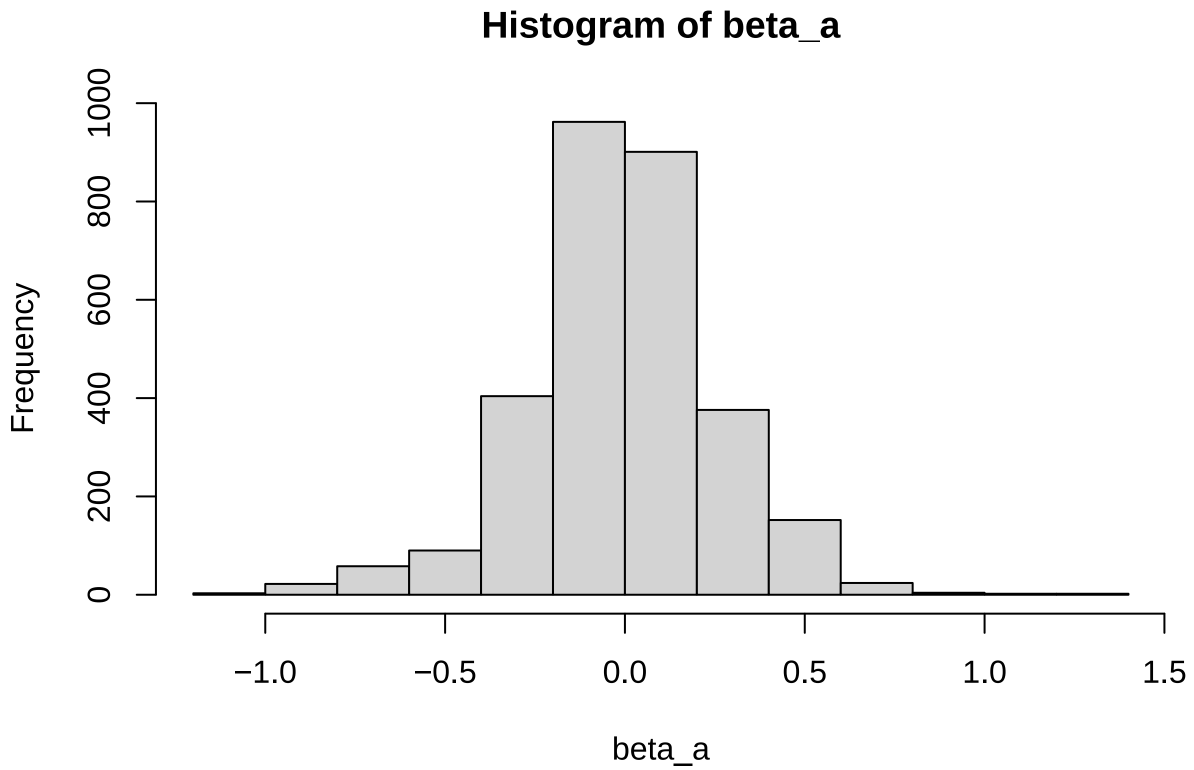
- ***(a). Plot a histogram of all the*** $\hat{\beta}_\mu$

```
beta_mu <- beta[1,]
hist(beta_mu)
```



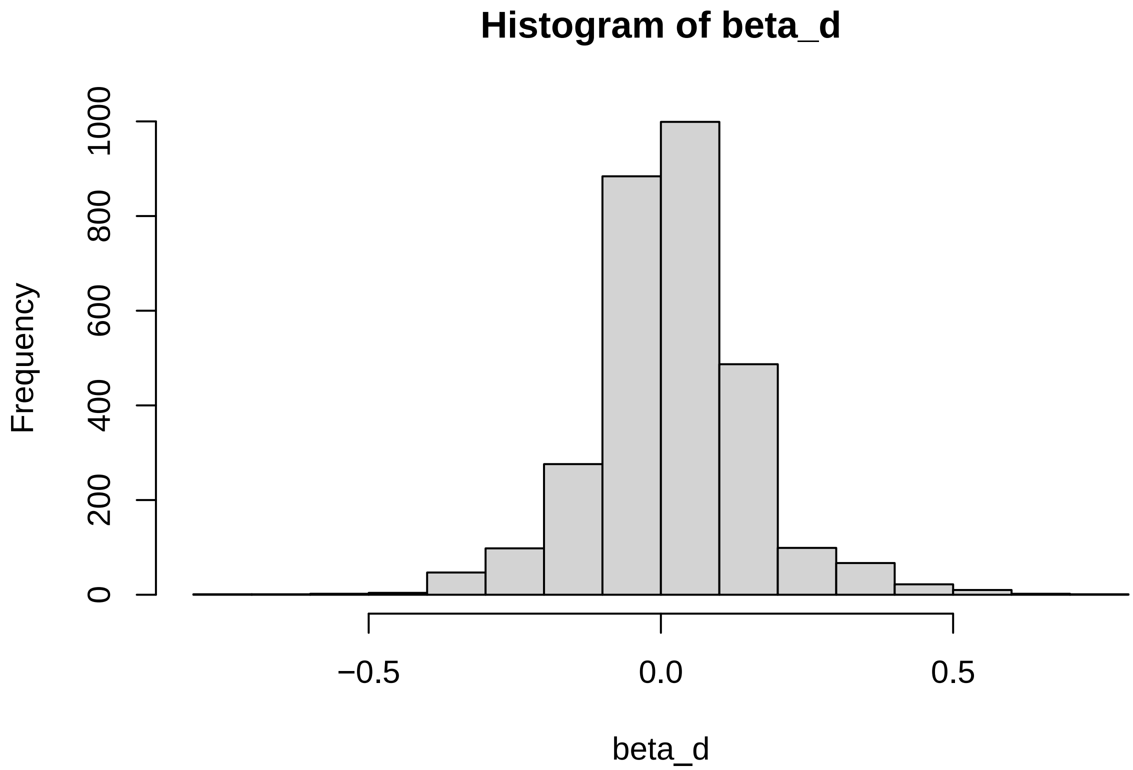
- ***(b). Plot a histogram of all the*** $\hat{\beta}_a$

```
beta_a <- beta[2,]
hist(beta_a)
```



- ***(c). Plot a histogram of all the*** $\hat{\beta}_d$

```
beta_d <- beta[3,]
hist(beta_d)
```



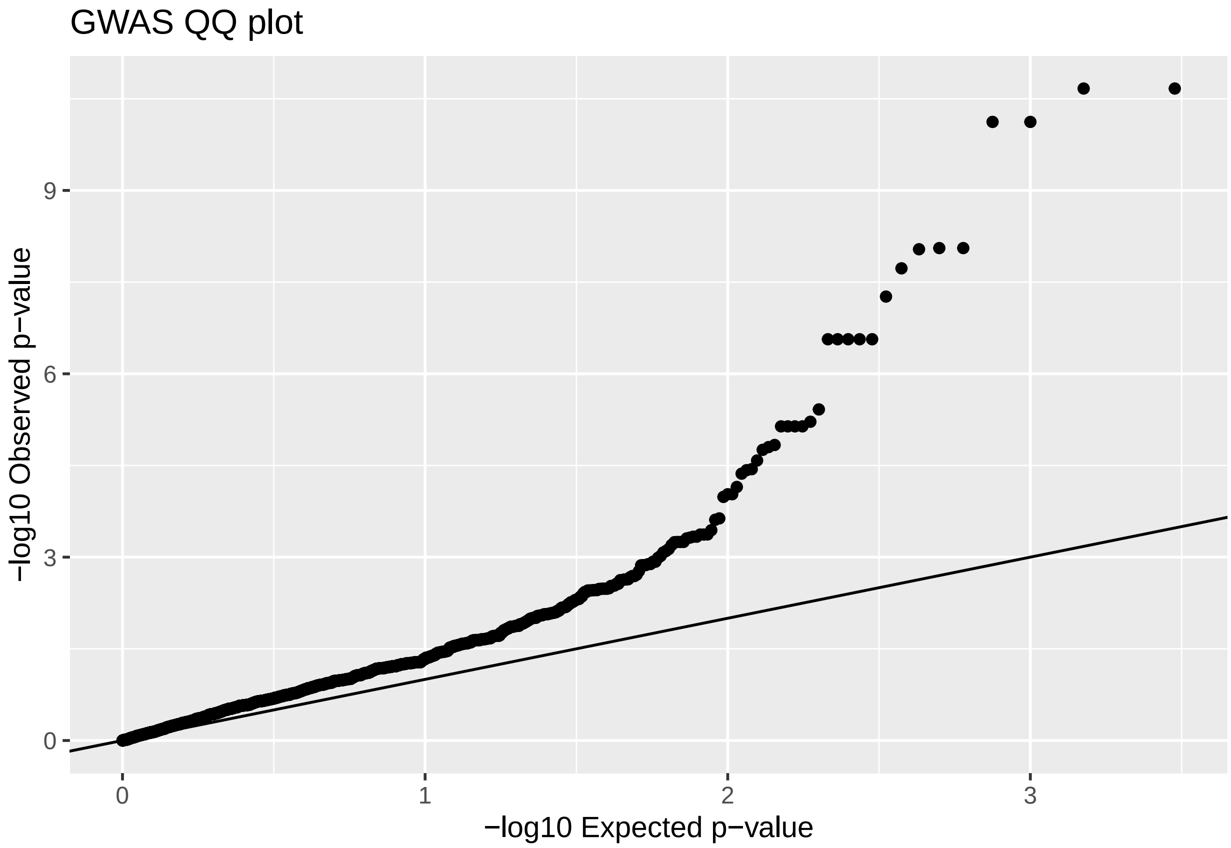
4. For each SNP, calculate p-values for the null hypothesis $H_0 : \beta_a = 0 \cap \beta_d = 0$ versus the alternative hypothesis $H_A : \beta_a \neq 0 \cup \beta_d \neq 0$ when applying the genetic linear regression model.

```
calc_pval <- function(y,xa,xd,beta) {
  x <- cbind(1,xa,xd)
  yhat <- x%*%beta
  df1 <- 2
  df2 <- n-3
  msm <- sum((yhat-mean(y))^2)/df1
  mse <- sum((y-yhat)^2)/df2
  f <- msm/mse
  pval <- pf(f,df1,df2,lower.tail=F)
  return (pval)
}

pvals <- rep(0,N)
for (i in 1:N) {
  pvals[i] <- calc_pval(phenotype,xa[,i],xd[,i],beta[,i])
}
```

5. For the p-values you calculated in question [4]

- ***(a). Produce a QQ plot for these p-values (label your plot and your axes using informative names!)***

```r
obs_pval <- sort(pvals)
exp_pval <- qunif(seq(0,1,length.out=length(obs_pval)+2),min=0,max=1)
exp_pval <- exp_pval[exp_pval!=0 & exp_pval!=1]

dfqq <- data.frame(observed=-log10(obs_pval),expected=-log10(exp_pval))
ggplot(dfqq,aes(x=expected,y=observed)) + geom_point() +
  geom_abline(intercept=0,slope=1) +
  labs(x="-log10 Expected p-value",
       y="-log10 Observed p-value",
       title="GWAS QQ plot")
```



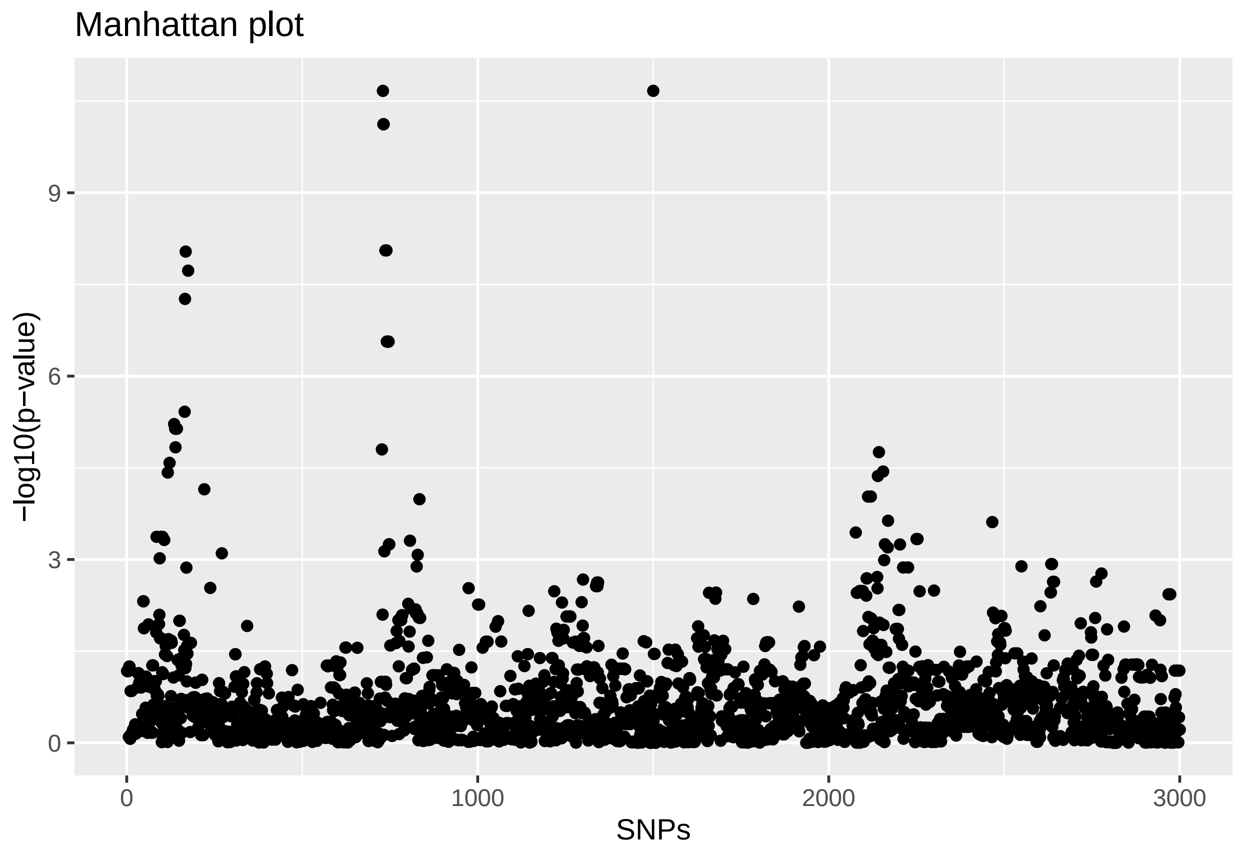
- ***(b). USING NO MORE THAN TWO SENTENCES answer the following question: based on this QQ plot, do you think you have achieved an appropriate model fit with your analysis and why do you think this is the case?***

  Yes, I believe the appropriate model fit is achieved with this analysis, because the points on the left of the plot generally follows the null trend, and only a few points on the right side of the plot has high deviations from the trend (these points have significant p-values). The points on the right of the plot may indicate causal variants / linkage disequilibrium.

6. ***For the p-values you calculated in question [4]***

- ***(a). Produce a Manhattan plot***

```
dfmht <- data.frame(index=1:length(pvals),pval=pvals)
ggplot(dfmht,aes(x=index,y=-log10(pvals))) + geom_point() +
  labs(x="SNPs",y="-log10(p-value)",title="Manhattan plot")
```



- *(b). Report HOW MANY SNPs (not which, just how many!) you find to be significant when controlling the study-wide type 1 error of 0.05 using a Bonferroni correction*

```
n_test <- length(pvals)
ab <- 0.05/n_test
n_sig <- sum(pvals<ab)
cat("Number of significant SNPs with Bonferroni n_sig =", n_sig)
```

```
## Number of significant SNPs with Bonferroni n_sig = 22
```

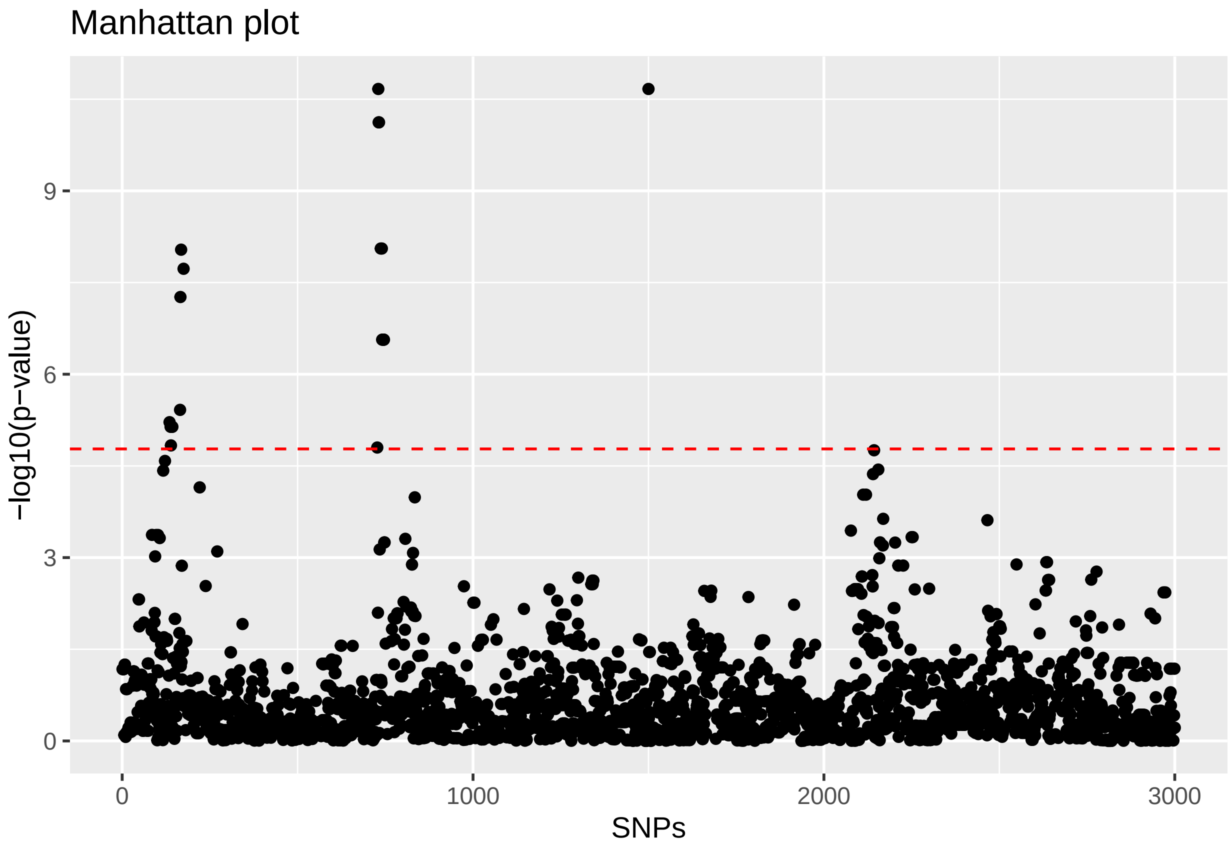
7. USING NO MORE THAN TWO SENTENCES answer the following question: based on your answer the question [6], how many distinct 'peaks' do you think you have identified and why?

I believe only 3 or 4 peaks were detected, because judging from the the index of potential causal variants and the Manhattan plot, it appears that only three clusters are identified, which are around SNP 135~143, 165~175 (it could be that these two are from the same peak), 727~746, and 1500. Variants near each other are considered as the same peak because they are likely associated with linkage disequilibrium.

```
which(pvals<ab)
```

```
##  [1]  135  138  139  140  141  143  165  166  168  175  727  730  731  732  737
## [16]  740  741  743  744  745  746 1500
```

```
ggplot(dfmht,aes(x=index,y=-log10(pvals))) + geom_point() +
  labs(x="SNPs",y="-log10(p-value)",title="Manhattan plot") +
  geom_hline(yintercept=-log10(ab),color='red',lty=2)
```



8. Imagine you are explaining the outcome of your analysis to your biological collaborator who does not have a deep understanding of a GWAS. Answer the following:

- ***(a). What is a causal polymorphism?***

  A causal polymorphism is a change in genotype, or at a location in the DNA, that causes a change in the phenotype under certain conditions.

- ***(b). USING NO MORE THAN TWO SENTENCES describe why do you observe 'peaks' in your Manhattan plot?***

  A peak observed in Manhattan plot means that the genotype at the location where the peak occurs is highly associated with the change in phenotype. The peak is resulted by a significant p-value, which is a smaller p-value that results in a larger `-log10(p-value)` value.

- ***(c). USING NO MORE THAN TWO SENTENCES describe why the peaks in your Manhattan plot may indicate the genomic position of a causal polymorphism but not (necessarily) the actual causal polymorphism?***

The peaks in Manhattan plot indicates that the location detected is significantly close to the genomic position of the actual causal polymorphism, but since it's close, it might be that the phenotypes are not actually associated with that particular position, but with the actual causal polymorphism, which tends to be inherited together due to the close distance between them. Therefore, peaks can only indicate close genomic positions, but not actual causal polymorphism.

- ***(d). Provide one reason why a peak may NOT indicate the position of a causal polymorphism.***

  One reason is linkage disequilibrium, which results in the effect that non-causal variants that is close to the causal variant (genetic linkage) tend to also have the association with the change in phenotype.

9.

- ***(a). Provide a rigorous definition of the 'power' of a hypothesis test***

Power is the probability of correctly rejecting the null hypothesis when the alternative hypothesis is true (null hypothesis if false).

- ***(b). List three factors that could impact the power of a hypothesis test in a GWAS***
  - *Sample size*: a smaller sample size can show misleading properties and deviates from the general population trend. Therefore, with a larger sample size, the associations captured are more likely to be significant, which would increase the power of the hypothesis test.
  - *Genome (testing) size*: for GWAS there are millions or even billions of tests, depending on the genome size (or size of regions captured in the test). Since GWAS is corrected with multiple-testing, which generally lowers the power of hypothesis tests, and that the multiple-testing (e.g., Bonferroni correction) is dependent on the testing size, therefore genome (testing) size will impact the power of the hypothesis test.
  - *Linkage disequilibrium*: linkage disequilibrium associated the change in phenotype with many genetic locations that are close to the actual causal variant. This resulted in higher number of rejected null hypotheses, but could in turn compensate for the non-rejection if there exist a less significant causal variant near another significant causal variant, promoting the power of the test. Therefore, the existence of linkage disequilibrium has an impact on the power of the hypothesis test.

10.

- ***(a). Provide a rigorous definition of a random variable***

  A random variable is a function that maps the sample space to corresponding reals. It describes the outcomes of experiments with numbers and bridges between the sample space and the experimental outcomes.

- ***(b). Provide a rigorous definition of a statistic***

  A statistic is a function on a sample (takes a sample as input) and outputs a corresponding values.

- ***(c). Provide a rigorous definition of a p-value***

  P-value is the probability of obtaining a value of statistic that's more extreme than a threshold conditioned on the null hypothesis (a parameter would take a specific value) being true.