QGG Midterm 2019

Solution

April 23, 2019

Question 1

part a

```
phenotypes <- read.csv("~/Downloads/midterm2019_pheno+pop.csv",
                        stringsAsFactors = F, header = F)
```

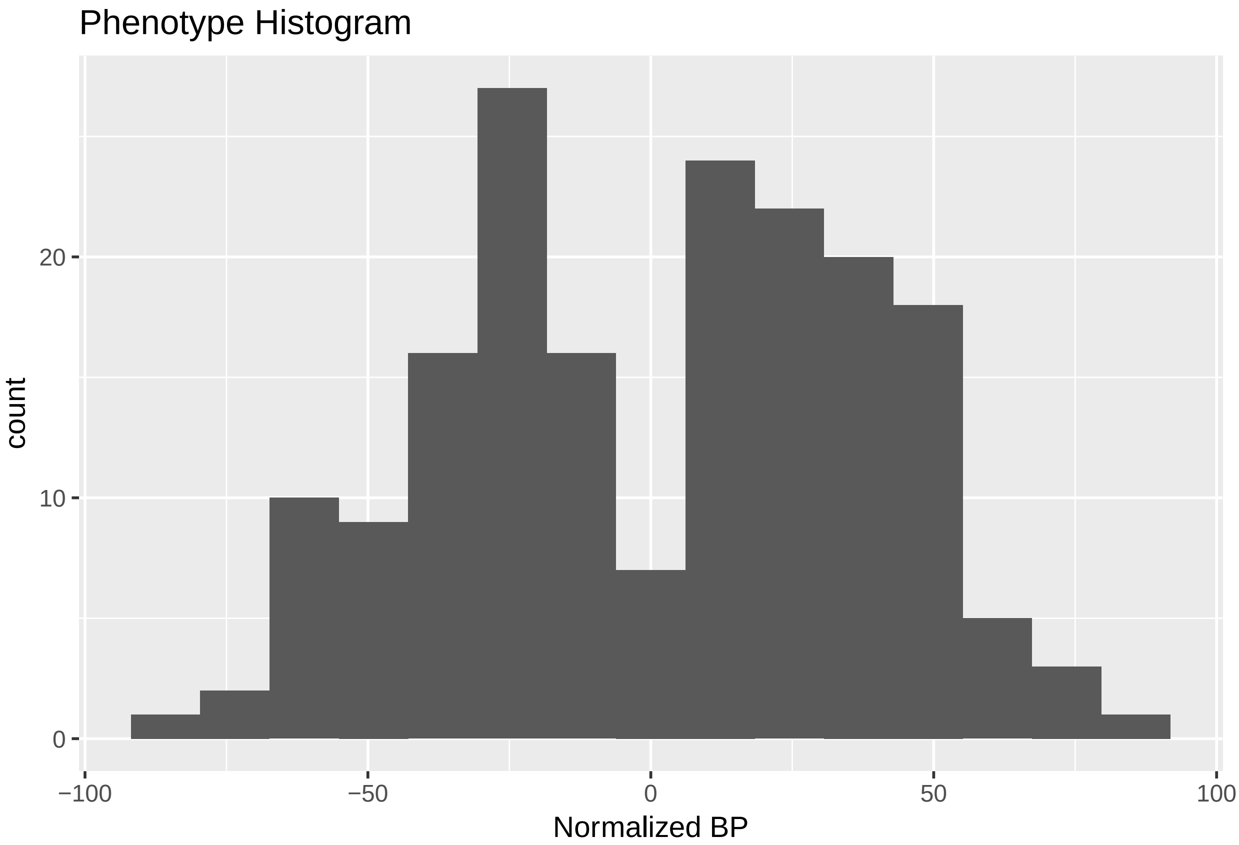
part b

```
cat("Part B - The number of samples is: ",nrow(phenotypes))
```

```
## Part B - The number of samples is:  181
```

part c

```
library(ggplot2)
ggplot(phenotypes,aes(V2))+geom_histogram(bins=15)+
  labs(x="Normalized BP",title="Phenotype Histogram")
```



part d Among the acceptable answers: The histogram of the phenotypes looks like it has two peaks instead of one, and therefore looks bimodal or like two normal distributions that overlap, where this could be attributable to the individuals in this sample being taken from two different populations, which we could surmise have very different means for the phenotype. A linear regression could still be appropriate for these data even if the overall phenotype histogram looks bimodal, since the overall probability model for a genetic regression is a normal distribution for distinct groups (determined by the independent X variables) each with a different overall mean, such that inclusion of a covariate appropriately coded X for population in the regression model could produce an appropriate model with mean shifted populations that are normally distributed (or conversely, if the impact of population were removed by modeling the population structure as a covariate, the phenotype would be expected to look normal).

part e Any one of the following answers (and others are possible): (1) The phenotype does not fall into discrete classes, (2) The phenotype does not follow a discrete (e.g., Bernoulli error) rather a normal error, (3) The phenotype relationship with the independent variables does not appear to follow a logistic curve relationship.}

Question 2

part a

```
genotypes <- read.csv("~/Downloads/midterm2019_genotypes_v2.csv",
                      stringsAsFactors = F, header = F)
```

part 2b

```
cat("Part C - the number of genotypes is: ", ncol(genotypes))
```

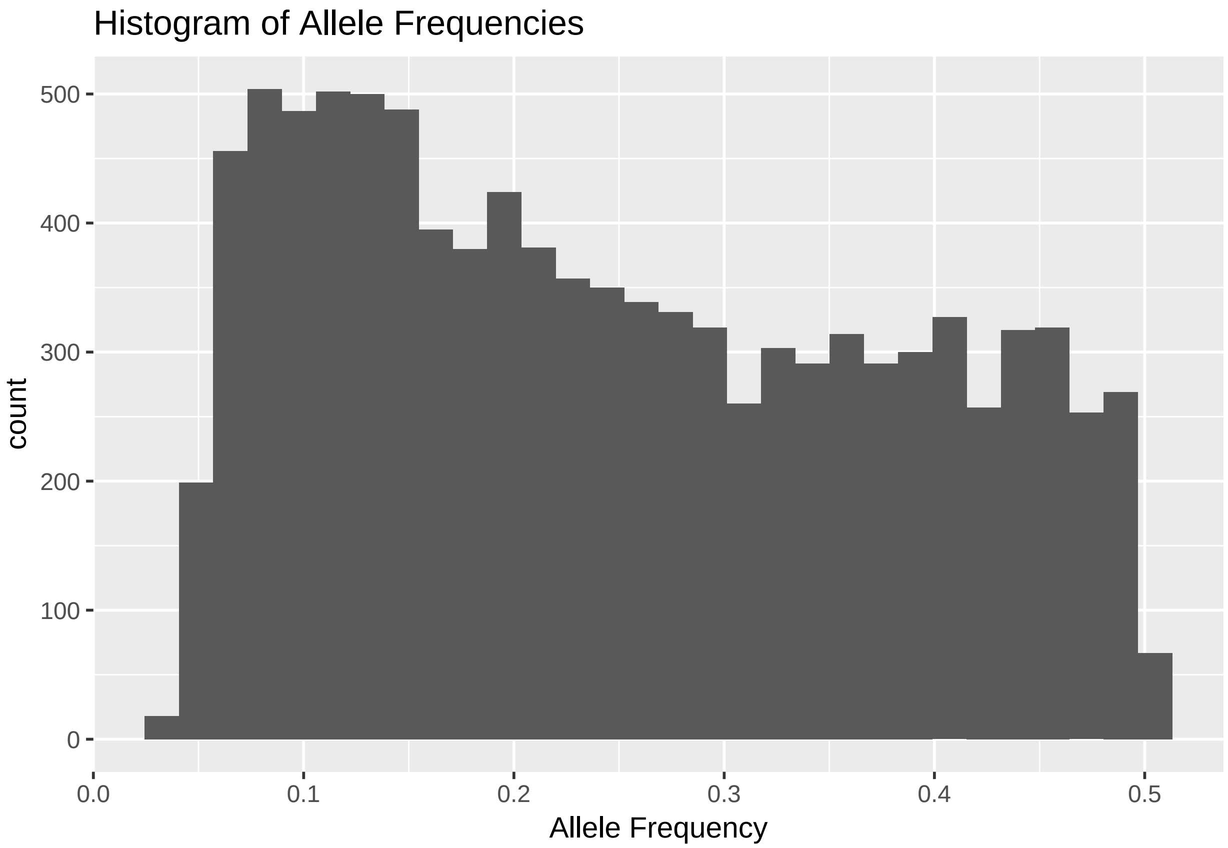
```
## Part C - the number of genotypes is:  9998
```

part 2c

```
#start by using the table function, which tells how many of each genotype there are
#we then add two times the maximum homozygote frequency with the hetrozygous count
#to return maf we subtract from 1
#we must specify the homozygote because the hertrozygous could be the most common
#we use the maximum because if both homozygotes are not present the minimum value is also the maximum
#which would largely thow off our minimum count, although this problem does not occur with maximum

maf_calc <- function(x){
  tab_x <- table(x)
  af <- 1-((2*max(tab_x[names(tab_x) %in% c(0,2)])+sum(x==1))/(2*nrow(genotypes)))
  return(af)
}

af <- apply(genotypes, 2, maf_calc)
ggplot(data.frame(af),aes(af))+geom_histogram(bins=30)+
  labs(x="Allele Frequency", title="Histogram of Allele Frequencies")
```



part 2d Power is defined as the probability of correctly rejecting the null hypothesis in a statistical test when the null hypothesis is false. Mathematically:

$$power = \int_{c_\alpha}^{-\infty} Pr(T(x)|\theta)dT(x)$$

part e Full credit given for a simple answer such as: We expect the power to be related to the MAF, with lower power for a lower MAF and higher power for a higher MAF. A more complex answer may note that having a lower MAF means that your sample size for one of the genotypes is vastly lower than the others. This randomly drawn group may then have characteristics that are unrepresentative of the full distribution - similar to how any small sample size has a greater variance around the true mean of the distribution. The association however, will assume this small sample to well-represent the true distribution, thereby drawing a possibly false conclusion.

Question 3

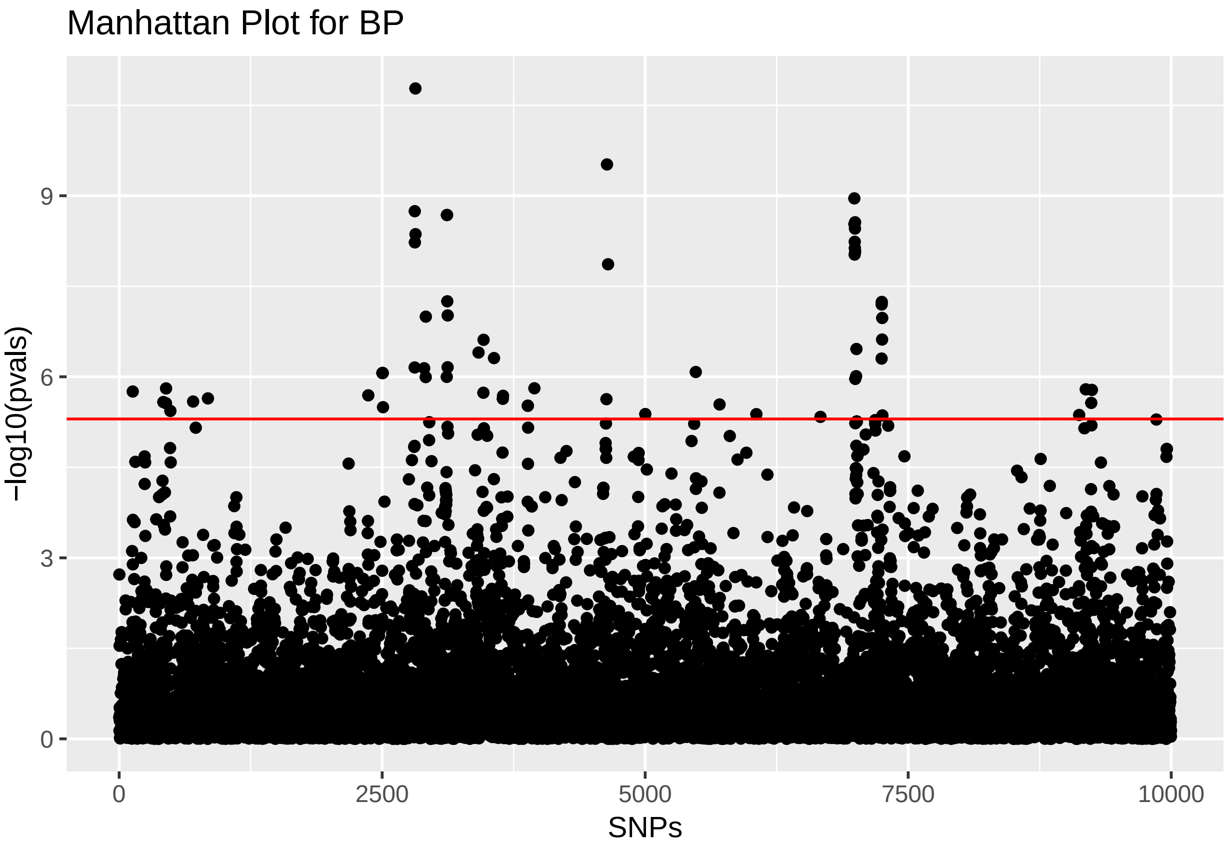
part 3a

```r
library(MASS)

pval_calculator <- function(xa_input, pheno_input){
    xa_input <- xa_input - 1
    xd_input <- 1 - 2*abs(xa_input)
    n_samples <- length(xa_input)
    X_mx <- cbind(1,xa_input,xd_input)
    MLE_beta <- ginv(t(X_mx) %*% X_mx) %*% t(X_mx) %*% pheno_input
    y_hat <- X_mx %*% MLE_beta
    SSM <- sum((y_hat - mean(pheno_input))^2)
    SSE <- sum((pheno_input - y_hat)^2)
    df_M <- 2
    df_E <- n_samples - 3
    MSM <- SSM / df_M
    MSE <- SSE / df_E
    Fstatistic <- MSM / MSE
    pval <- pf(Fstatistic, df_M, df_E,lower.tail = FALSE)
    return(pval)
}
pvals <- apply(genotypes,2,pval_calculator,phenotypes[,2])
```

part 3b

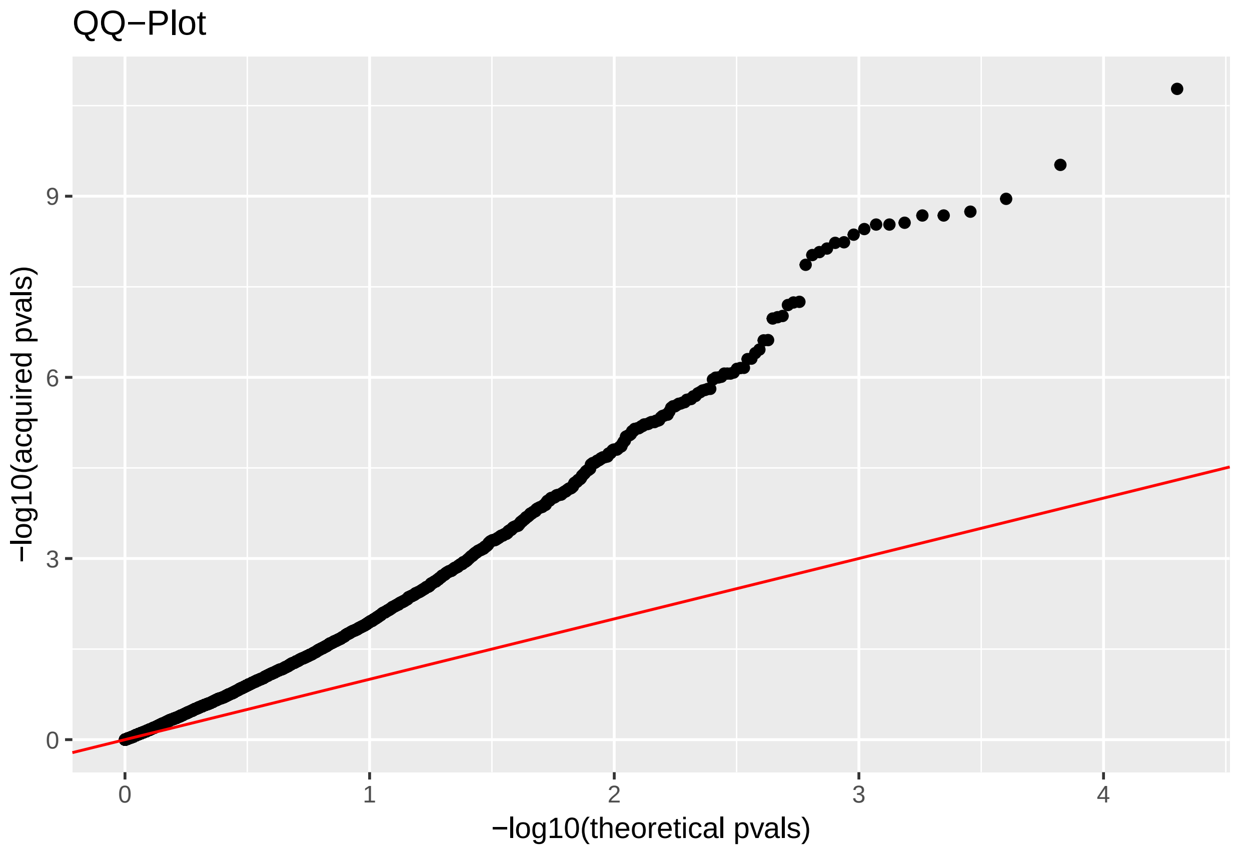
```
ggplot(data.frame(pvals),aes(1:length(pvals),-log10(pvals)))+
  geom_point()+labs(x="SNPs",title="Manhattan Plot for BP")+
  geom_hline(yintercept = -log10(0.05/ncol(genotypes)),color="red")
```



Question 4

part 4a

```
qqDf <- data.frame(exp = sort(pvals),theo = sort(qunif(ppoints(length(pvals)))))
ggplot(qqDf,aes(-log10(theo), -log10(exp)))+geom_point()+
  geom_abline(slope = 1, color = "red")+
  labs(title="QQ-Plot",x="-log10(theoretical pvals)",y="-log10(acquired pvals)")
```



Part 4b Among the acceptable answers: The QQ plot leaves the 45 degree line very early, indicating that the statistical model applied individually to each genotype produced too many low p-values than expected if the null hypothesis were correct for the majority of genotypes and that we did not achieve an appropriate fit, since we expect the null hypothesis to be correct for the majority of genotypes if we achieved appropriate model fit for our GWAS data. We could expect a QQ plot of this type if we had an unaccounted for covariate such as unaccounted for populations structure, i.e., two populations with distinct phenotype means and different frequencies of alleles for many of the genotypes, such that a statistical model fit without a population covariate would effectively result in a statistical test for whether the genotypes differed in frequency between the two populations and not whether a genotype was in LD with a causal genotype. Given that we know there are two populations represented in this sample, this seems like a reasonable explanation for the shape of the QQ.

Question 5

Part 5a

```
popFreq <- table(substr(phenotypes[,1],1,1))
cat("There are ",popFreq[1]," people in the first pop and ",
    popFreq[2]," people in the second")
```

```
## There are  89  people in the first pop and  92  people in the second
```

Part 5b Among the acceptable answers: A PCA can intuitively be considered a rotation of axes for highly multivariate data, such that if the data were projected on to the first new axis / PC this would produce the greatest possible variance of all possible new axes, the next PC would be orthogonal to the first and be the next greatest variance, etc. Since individuals in distinct populations will tend to `separate' orcluster’ the samples into two groups in such a multivariate plot, the first PC will tend to point in the direction of these clusters (i.e., clusters at different ends of the PC) since this will produce high variance along the PC, which allows us to visually separate the two clusters once plotted on the first couple of PCs.

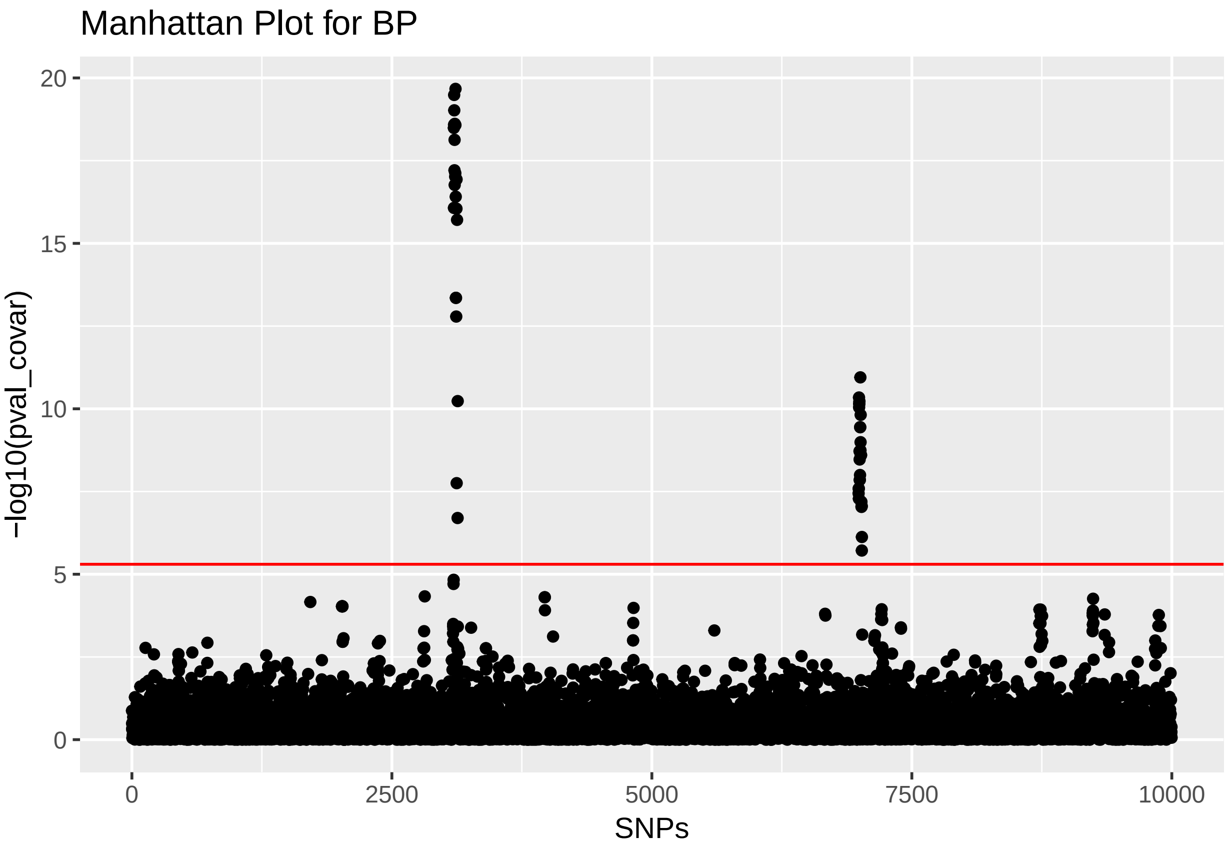
Question 6

Part 6a

```
pval_calculator_covar <- function(xa_input, pheno_input, z_input){
    xa_input <- xa_input - 1
    xd_input <- 1 - 2*abs(xa_input)
    n_samples <- length(xa_input)
    Z_mx <- cbind(1,z_input)
    XZ_mx <- cbind(1,xa_input,xd_input,z_input)
    MLE_beta_theta0 <- ginv(t(Z_mx)  %*% Z_mx)  %*% t(Z_mx)  %*% pheno_input
    MLE_beta_theta1 <- ginv(t(XZ_mx) %*% XZ_mx) %*% t(XZ_mx) %*% pheno_input
    y_hat_theta0 <- Z_mx  %*% MLE_beta_theta0
    y_hat_theta1 <- XZ_mx %*% MLE_beta_theta1
    SSE_theta0 <- sum((pheno_input - y_hat_theta0)^2)
    SSE_theta1 <- sum((pheno_input - y_hat_theta1)^2)
    df_M <- 2
    df_E <- n_samples - 3
    Fstatistic <- ((SSE_theta0-SSE_theta1)/df_M) / (SSE_theta1/df_E)
    pval <- pf(Fstatistic, df_M, df_E,lower.tail = FALSE)
    return(pval)
}

pop_num <- rep(1,nrow(genotypes))
pop_num[substr(phenotypes[,1],1,1)=="N"] <- -1
pval_covar <- apply(genotypes,2, pval_calculator_covar,phenotypes[,2],pop_num)
```

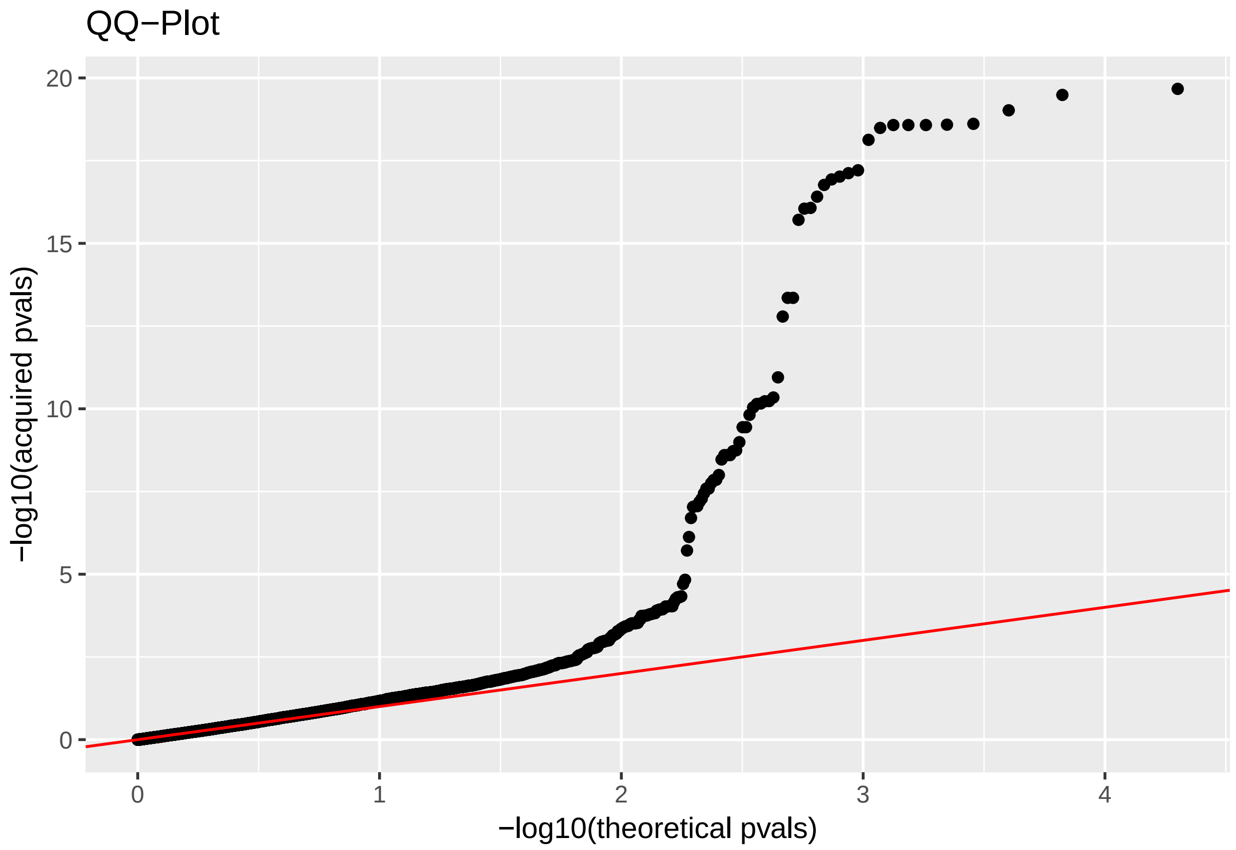
```
ggplot(data.frame(pval_covar),aes(1:length(pval_covar),-log10(pval_covar)))+
  geom_point()+labs(x="SNPs",title="Manhattan Plot for BP")+
  geom_hline(yintercept = -log10(0.05/ncol(genotypes)),color="red")
```



Question 7

Part 7a

```
qqDf <- data.frame(exp = sort(pval_covar),theo = sort(qunif(ppoints(length(pval_covar)))))
ggplot(qqDf,aes(-log10(theo), -log10(exp)))+geom_point()+geom_abline(slope = 1, color = "red")+
  labs(title="QQ-Plot",x="-log10(theoretical pvals)",y="-log10(acquired pvals)")
```



Part 7b Among the acceptable answers: The QQ plot hugs the 45 degree line for most of the (smaller) log 10 p-values and then has a 'tail' of higher log 10 p-values that leave the 45 degree line at the end, which is exactly the distribution of p-values we should expect if for most of genotypes the null hypothesis is correct and the null is not correct for one or just a few genotypes that are in LD with causal genotypes. This is exactly what we would expect to see if we achieved appropriate statistical model fit for our GWAS.

Question 8

Part 8a

```
cat("The Bonferonnni cut off is: ", 0.05/length(pvals))
```

```
## The Bonferonnni cut off is:  5.001e-06
```

Part 8b Many answers possible as long as they are justified, among the acceptable answers: I observed two separate peaks using the criteria that a set of p-values in order along the chromosome that were all above / do not fall below the Bonferroni cutoff indicate a peak, i.e., the two peaks are separated by genotypes with p-values that fall below the cutoff.

Part 8c Many answers possible as long as they are justified, including (possibly zero, one total, one each, several per peak) e.g., example of an acceptable answer: I believe each peak indicates one causal polymorphism because in humans, sets of genotypes are generally in close enough LD to tag a single causal polymorphism with a large enough effect to detect in a GWAS and such causal polymorphisms appear to be scattered throughout the genome / not in LD with each other.

Part 8d Among the acceptable answers: Since we expect to reject the null hypothesis for a statistical test of association for any non-causal genotype in LD with a causal genotype, and since many genotypes are expected to be in LD (i.e., in the same physical genomic location) with each other, plus even if the causal genotype is measured (i.e., not always the case) it may not produce the most significant p-value, we cannot determine which of the genotypes for which we have rejected the null is the causal genotype with certainty.

Question 9

Part 9a

```r
peak1 <- pval_covar[3096:3133]
peak2 <- pval_covar[6988:7020]
causal1 <- peak1[peak1==min(peak1)]
causal2 <- peak2[peak2==min(peak2)]
cat("The two most significant polymorphisms in each peak are",names(causal1),"\n and",
    names(causal2), "with pvalules",causal1, "and",causal2,", respectively")
```

```
## The two most significant polymorphisms in each peak are V3112
##  and V7005 with pvalules 2.145867e-20 and 1.121738e-11 , respectively
```

Part 9b Many possible answers where any one from the following (non-exhaustive) list will be accepted:

1. By chance, the phenotypes are more strongly associated with a tag SNP than the causal SNP.
2. An error in measuring some of the phenotypes, resulted in the phenotypes being more strongly associated with a tag SNP than the causal SNP.
3. The causal SNP has a smaller MAF than the tag SNP (and the right structure of LD exists to produce a slightly smaller p-value for the tag)
4. There was a genotyping error that changed the value of a few of the causal SNP genotypes, leading to a less significant p-value.
5. An unaccounted for or partially accounted for covariate is more correlated with the tag than the causal SNP.
6. A tag SNP is tagging two or more causal genotypes

Part 9c

```r
getCor <- function( index){
  cat("For index",index ,"the two correlations are",
      cor(genotypes[,index],genotypes[,index-1]), "\n and",
      cor(genotypes[,index],genotypes[,index+1]),"\n")
}
getCor(3112)
```

```
## For index 3112 the two correlations are 0.9723772
##  and -0.2754027
```

```r
getCor(7005)
```

```
## For index 7005 the two correlations are 0.9717368
##  and 0.8837464
```

```r
library(genetics)
```

```
## Warning: package 'genetics' was built under R version 3.5.2

## Loading required package: combinat

##
## Attaching package: 'combinat'
## The following object is masked from 'package:utils':
##
##     combn

## Loading required package: gdata

## gdata: read.xls support for 'XLS' (Excel 97-2004) files ENABLED.

##

## gdata: read.xls support for 'XLSX' (Excel 2007+) files ENABLED.

##
## Attaching package: 'gdata'

## The following object is masked from 'package:stats':
##
##     nobs

## The following object is masked from 'package:utils':
##
##     object.size

## The following object is masked from 'package:base':
##
##     startsWith

## Loading required package: gtools

## Loading required package: mvtnorm

##

## NOTE: THIS PACKAGE IS NOW OBSOLETE.

##

##   The R-Genetics project has developed an set of enhanced genetics

##   packages to replace 'genetics'. Please visit the project homepage

##   at http://rgenetics.org for informtion.

##

##
## Attaching package: 'genetics'

## The following objects are masked from 'package:base':
##
##     %in%, as.factor, order
```

```r
get_r <- function(index){
  mainVariant <- gsub(x=gsub(x=gsub(x=genotypes[,index],0,"C/C"),1,"C/A"),2,"A/A")
  plusOne <- gsub(x=gsub(x=gsub(x=genotypes[,index+1],0,"C/C"),1,"C/A"),2,"A/A")
  minusOne <- gsub(x=gsub(x=gsub(x=genotypes[,index-1],0,"C/C"),1,"C/A"),2,"A/A")
  plus_r <- LD(genotype(mainVariant),genotype(plusOne))$r
  minus_r <- LD(genotype(mainVariant),genotype(minusOne))$r
  cat("For index",index ,"the two correlations are", plus_r, "and", minus_r,"\n")
}

get_r(3112)
```

```
## For index 3112 the two correlations are 0.329932 and 0.971726
```

```r
get_r(7055)
```

```
## For index 7055 the two correlations are -0.4301944 and 0.9996883
```

Part 9d Among the acceptable answers: Since the result of LD is that genotypes in LD will tend to have the same state in an individual, no matter how we code the dummy X variables for these genotypes, we expect either large values of the X's to occur more frequently with each other and small values of the X's to occur more frequently with each other than large with small, producing a large positive value for the correlation, or for large values of X to occur with small values of X but not large with large or small with small, producing a large negative correlation, i.e., for neither outcome would we expect a correlation near zero.

Question 10

Part 10a Among the acceptable answers: A causal polymorphism for a given phenotype is a polymorphic site in the genome where directly swapping one allele for another produces a change in value of the phenotype under some condition, or symbolically:

$$A1- > A2 => \Delta\bar{Y}|Z$$

Part 10b Many possible answers, an example of an acceptable answer: A CRISPR experiment performed one of an identical twin to swap one allele of the putative causal polymorphism for the other, where the twins were then raised under exactly the same conditions and BP was measured at the same timepoint.

Part 10c p-value - the probability of obtaining the observed value of the test statistic or more extreme conditional on the null hypothesis being true.

Part 10d Many possible answers where any three from the following (non-exhaustive) list will be accepted:

1. Type I error
2. Genotyping error
3. Phenotyping error
4. Coding or other mistake in the analytics
5. Disequilibrium without linkage disequilibrium
6. Unaccounted for covariate