Assignment 4

Student Name: Amanda Everitt
Student ID: 998974934

SETUP:

```
knitr::opts_chunk$set(cache=TRUE, autodep=TRUE)
data.geno <- read.csv("Rice_44K_genotypes.csv.gz", row.names=1, na.strings=c("NA","00"))
data.geno.2500 <- data.geno[,sample(2500)]
dim(data.geno.2500)
```

```
## [1]  413 2500
```

```
geno.numeric <- data.matrix(data.geno.2500) #convert the data matrix to numbers
#head(geno.numeric[,1:20])
genDist <- as.matrix(dist(geno.numeric)) #calculate the Euclidian distance between each rice variety
geno.mds <- as.data.frame(cmdscale(genDist)) #perform the multi-dimensional scaling
head(geno.mds) #now we have 2 dimensions
```

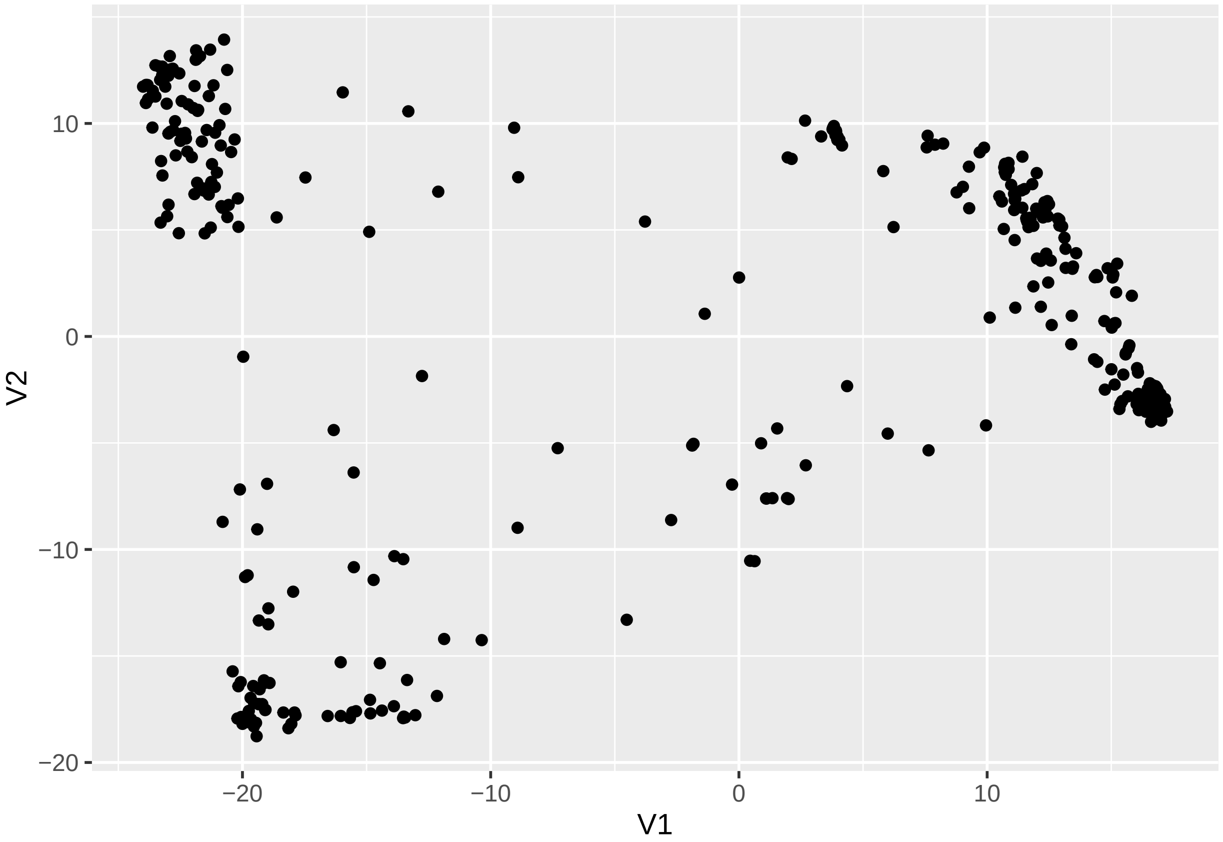
```
##                 V1         V2
## NSFTV1  16.970245  -3.606206
## NSFTV3 -23.802451  11.143276
## NSFTV4 -15.513342 -10.833049
## NSFTV5   1.351032  -7.592767
## NSFTV6 -13.520755 -10.455302
## NSFTV7  11.863364   2.350348
```

```
#install.packages("ggplot2")
library(ggplot2)
```

EXERCISE 1: Is there any evidence for populations structure (different sub populations)? If so, how many sub populations do you think the MDS plot reveals? What do you make of the individuals that are between the major groups?

> There is evidence of sub populations by the groupings within the scatter plot. There appears to be three sub populations. The individuals that are between the major groups are most likely individuals that share some, but not all, of the major SNPS that are causing the sub populations to group as they did.

```
ggplot(geno.mds, aes(V1,V2)) + geom_point()
```



EXERCISE 2:

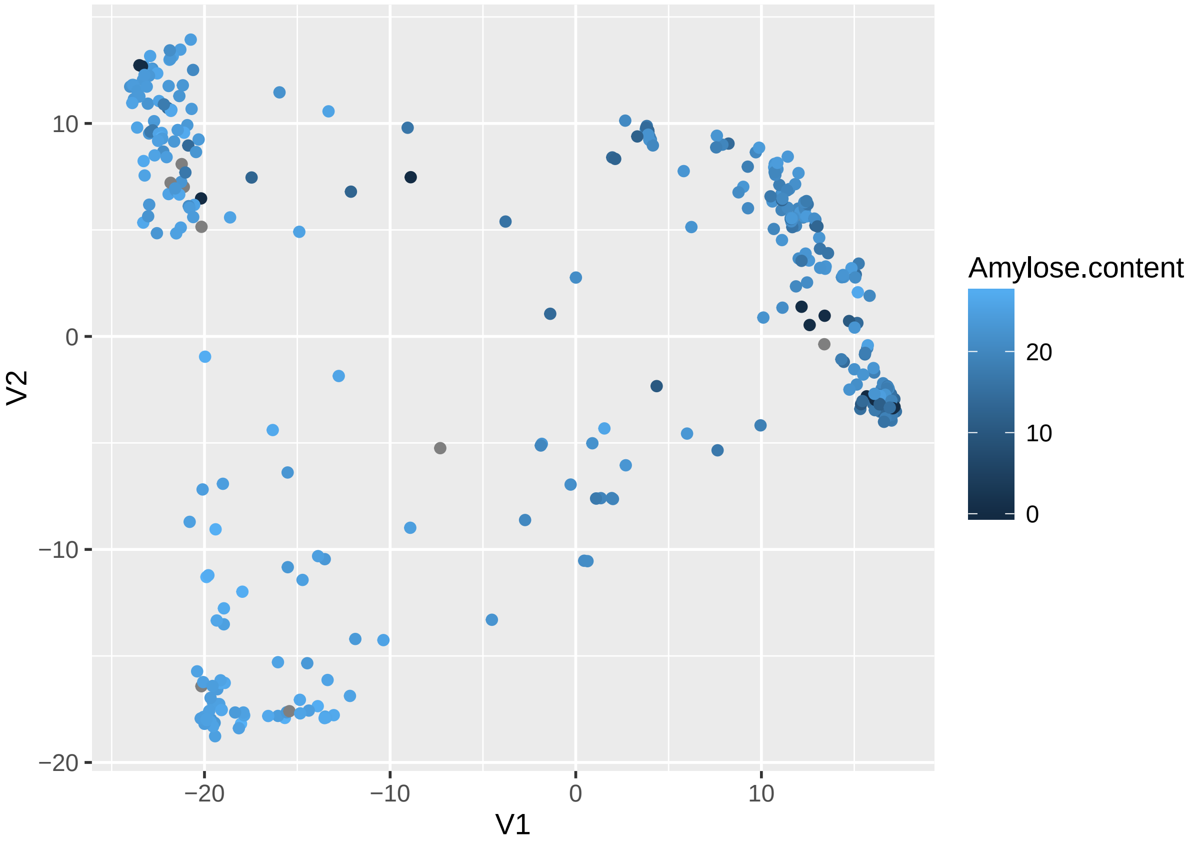
- Use the `read.csv()` `head()` and `summary()` functions that you learned earlier to import and look at this file. Import the file into an object called "data.pheno".
- Use merge() to merge the MDS scaled genotype data with the phenotype data. Here the column that we are merging on is the "row.name" column. So you can use `by="row.names"` or `by=1` in your call to merge. Use summary and head to look at the new object and make sure that it is as you expect.

```
data.pheno <- read.csv("RiceDiversity.44K.MSU6.Phenotypes.csv", row.names=1, na.strings=c("NA","00"))
data.geno.pheno <- merge(geno.mds, data.pheno ,by= "row.names")
#summary(data.geno.pheno)
```

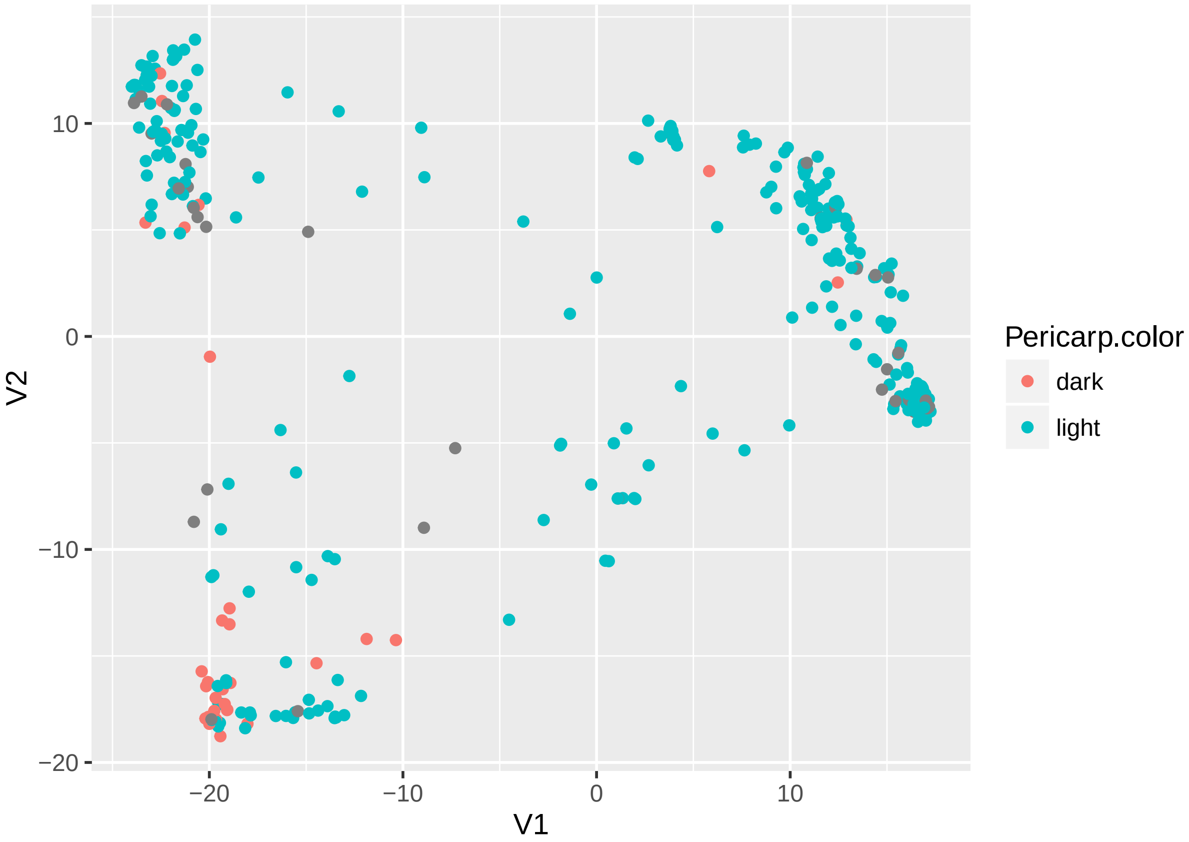
EXERCISE 3: Prepare three different plots to explore if subgroups vary by 1) Amylose content; 2) Pericarp color; 3) Region. Do any of these seem to be associated with the different population groups? Briefly discuss.

> It seems the subgroups cluster by region, but not by Amylose content of Pericarp color. In the scatter plot colored by region, the bottom left region largely seems to be made up of S. Asia. Also, the right middle group appears to contain most plants from America.

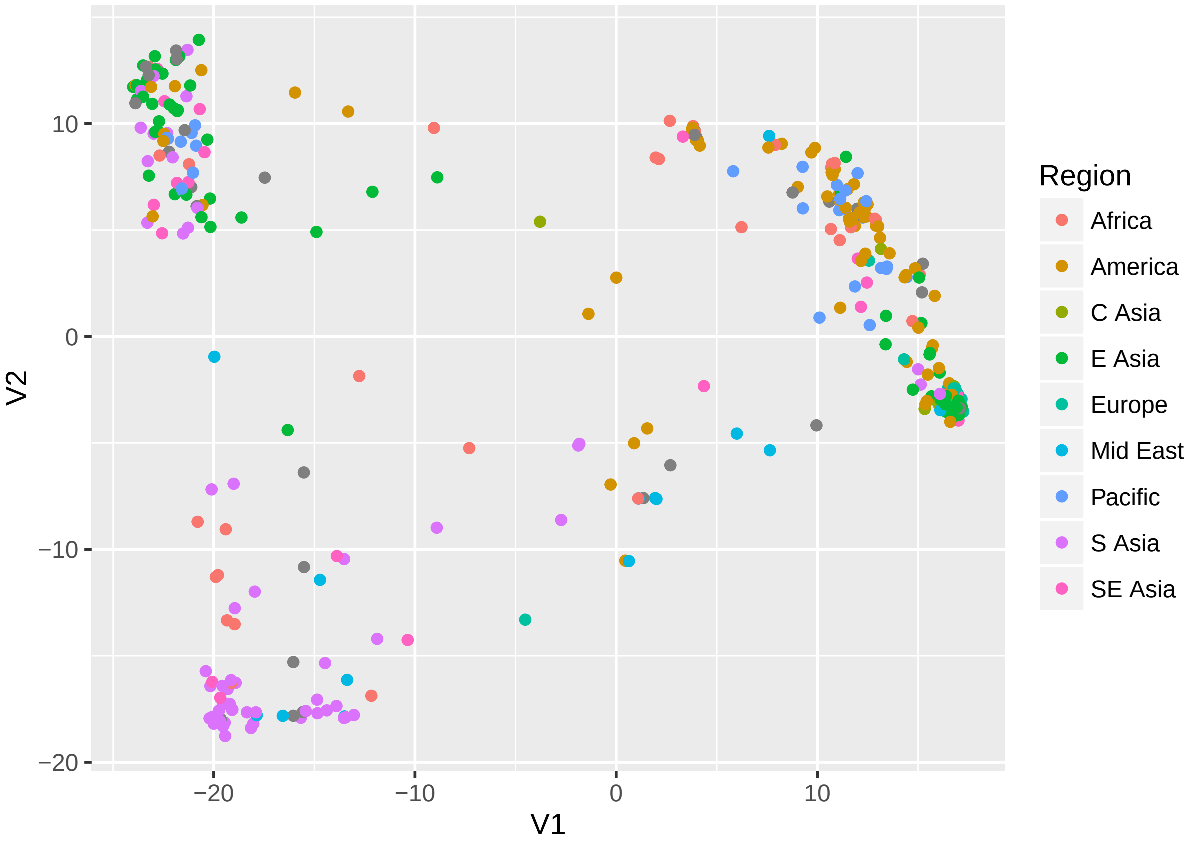
```
ggplot(data.geno.pheno, aes(V1,V2, color = Amylose.content)) + geom_point()
```



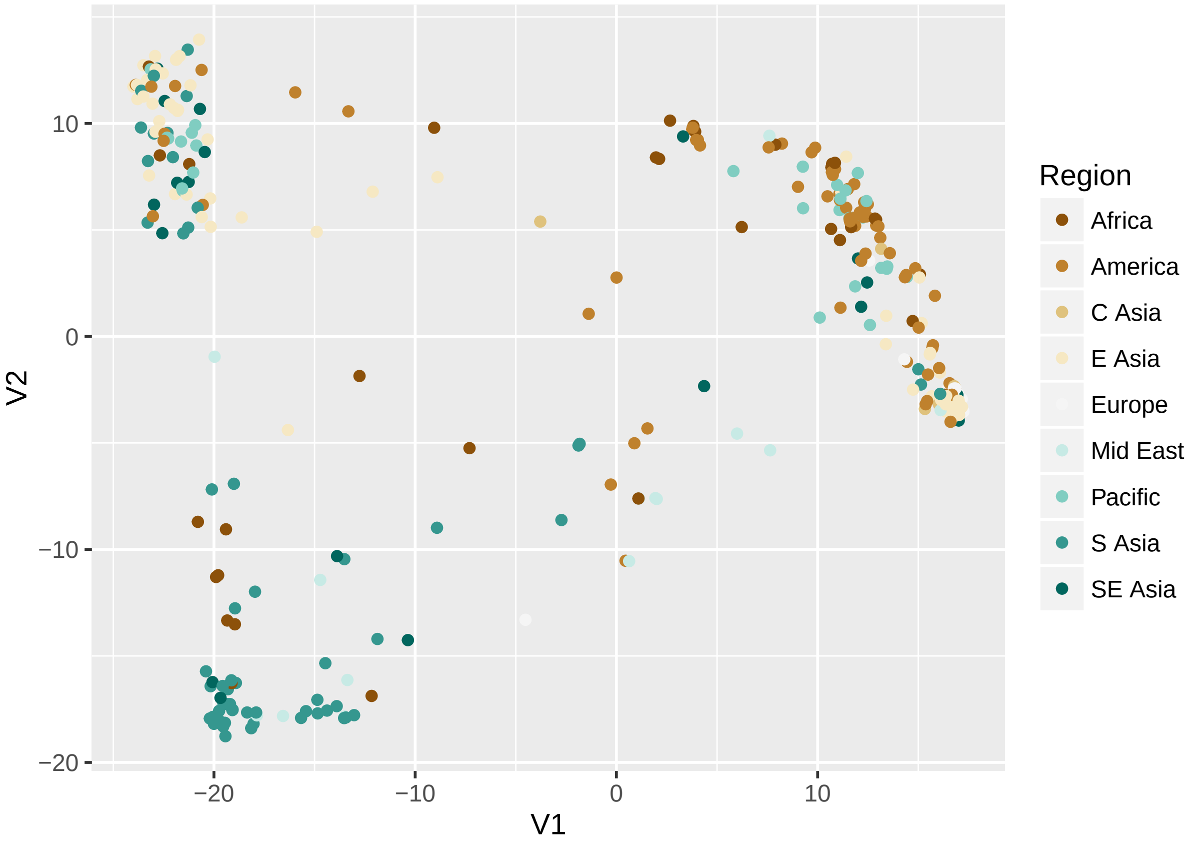
```
ggplot(data.geno.pheno, aes(V1,V2, color = Pericarp.color)) + geom_point()
```



```
p <- ggplot(data.geno.pheno, aes(V1,V2, color = Region)) + geom_point()
p
```

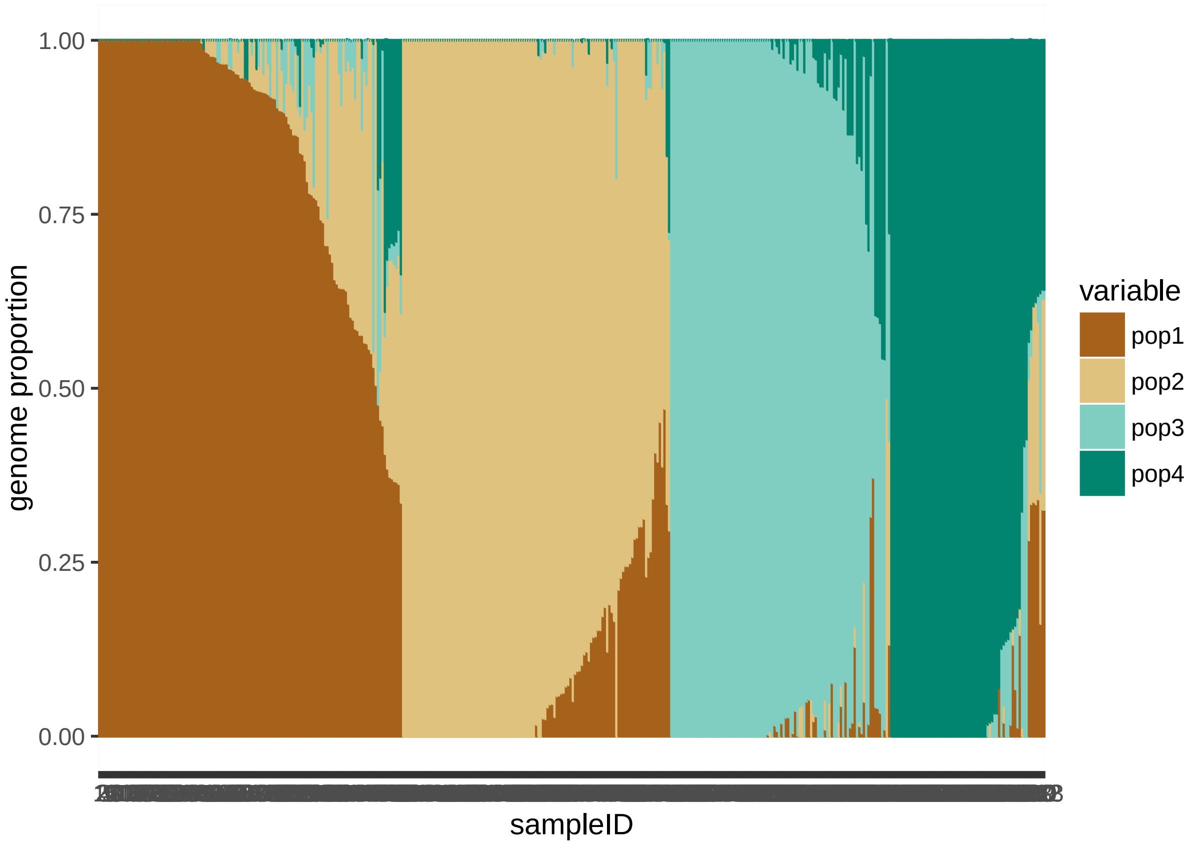


```
p + scale_color_brewer(type="div")
```



Setup for Exercise 4:

```
data.geno.2500.c <- apply(data.geno.2500,2,as.character) #Convert to character matrix
data.geno.2500.ps <- matrix("",nrow=nrow(data.geno.2500.c)*2,ncol=ncol(data.geno.2500.c))
#Create a new Matrix to hold reformatted data
for (i in 1:nrow(data.geno.2500.c)) {
  data.geno.2500.ps[(i-1)*2+1,] <- substr(data.geno.2500.c[i,],1,1)
  data.geno.2500.ps[(i-1)*2+2,] <- substr(data.geno.2500.c[i,],2,2)
}   #for each row of genotypes, create 2 rows, one with the first allele and one with the second allele
library(PSMix)
load("ps4.2500.RData")
ps4.df <- as.data.frame(cbind(round(ps4$AmPr,3),ps4$AmId))
colnames(ps4.df) <- c(paste("pop",1:(ncol(ps4.df)-1),sep=""),"popID")
maxGenome <- apply(ps4$AmPr,1,max)
ps4.df <- ps4.df[order(ps4.df$popID,-maxGenome),]
ps4.df$sampleID <- factor(1:413)
library(reshape2)
ps4.df.melt <- melt(ps4.df,id.vars=c("popID","sampleID"))
ggplot(aes(x=sampleID, y=value, color=variable, fill=variable), data=ps4.df.melt) + geom_bar(stat="iden
```

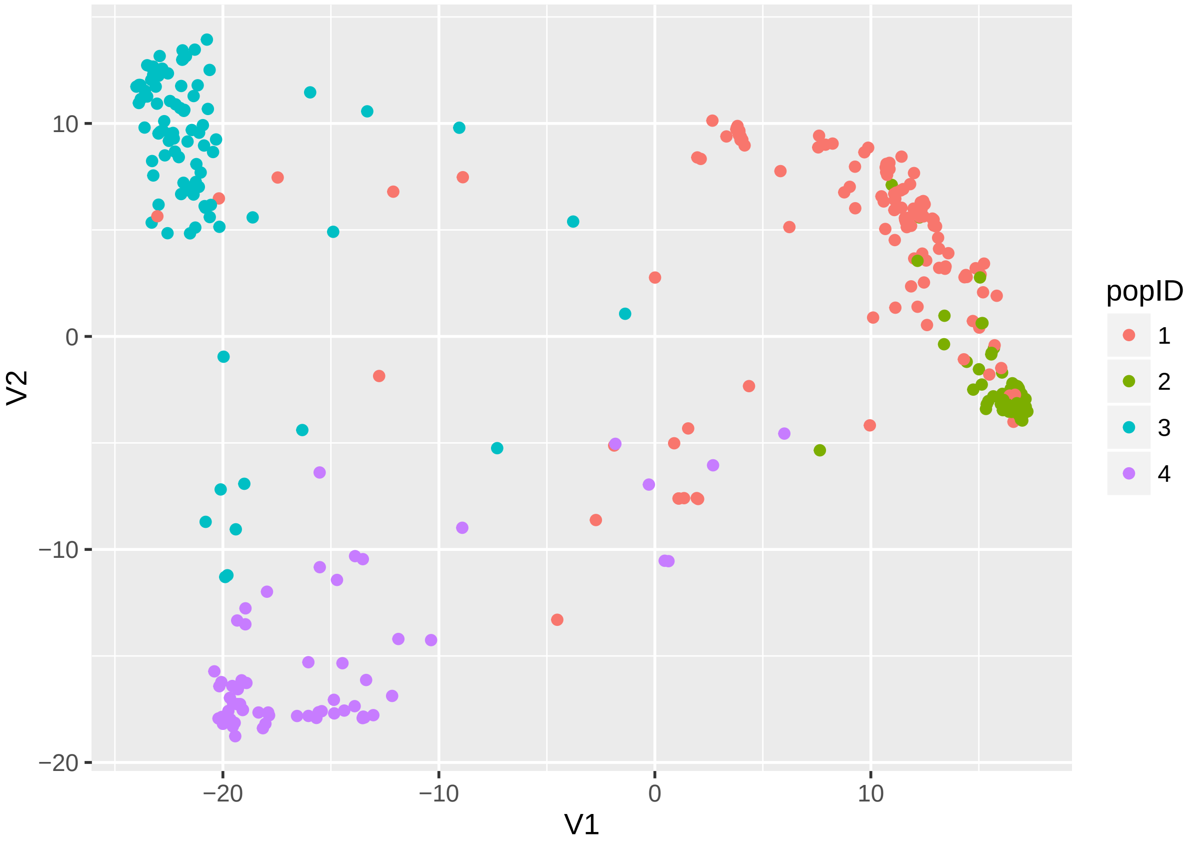


There does appear to be evidence of admixture because some of the individuals are clearly comprised of multiple populations (to a varying degree).

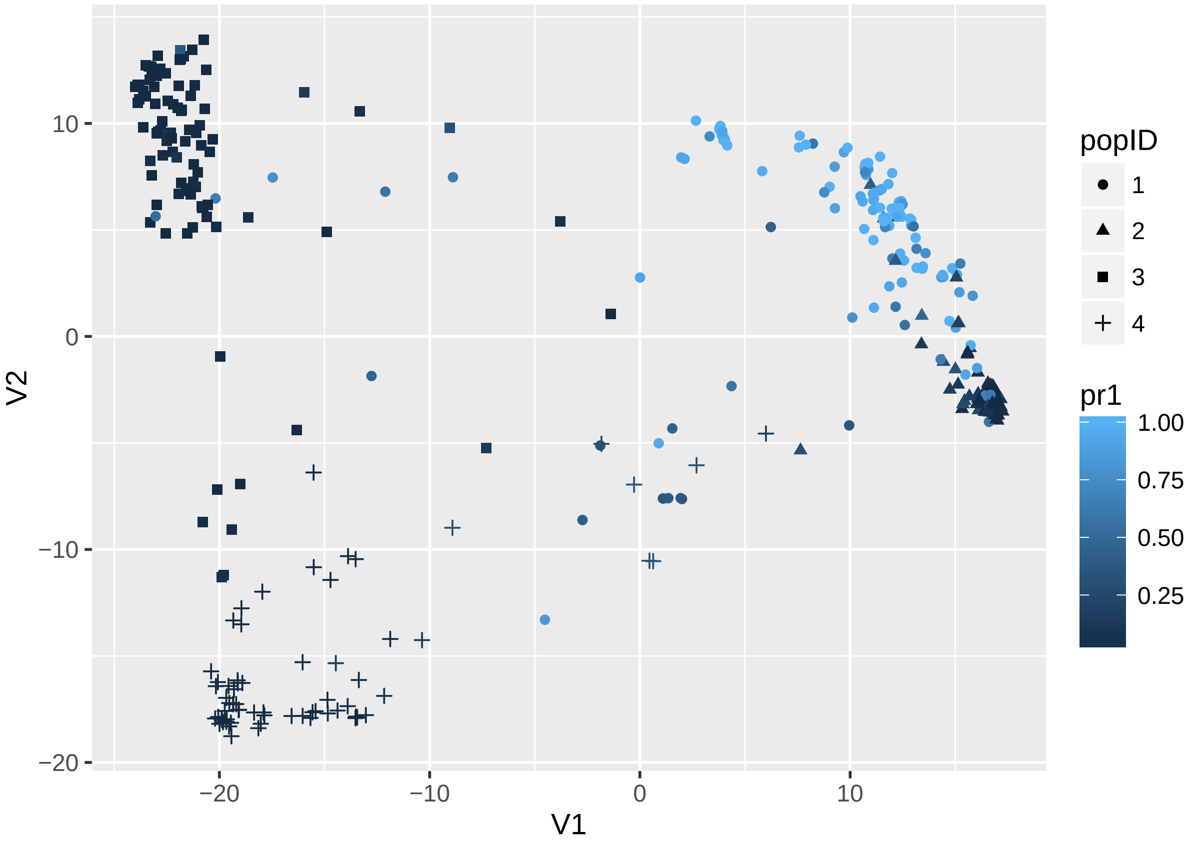
EXERCISE 4: Re plot the MDS data, but include the population assignment in an informative way. How do the populations assignments relate to the MDS plot?

The MDS plot clearly made subgroups based on the population assignment as seen in the first figure below. The second figure also shows the gradual effect of containing containing a higher probability of a certain population (here, population 1).

```
geno.mds$popID <- factor(ps4$AmId)
#head(geno.mds$popID)
colnames(ps4$AmPr) <- paste("pr",1:4,sep="") #Give proportions useful names
geno.mds <- cbind(geno.mds,ps4$AmPr) #add the admixture proportions for future use
#head(geno.mds)
ggplot(geno.mds, aes(V1,V2, color = popID)) + geom_point()
```



```
ggplot(geno.mds, aes(V1,V2, color = pr1, shape= popID)) + geom_point()
```

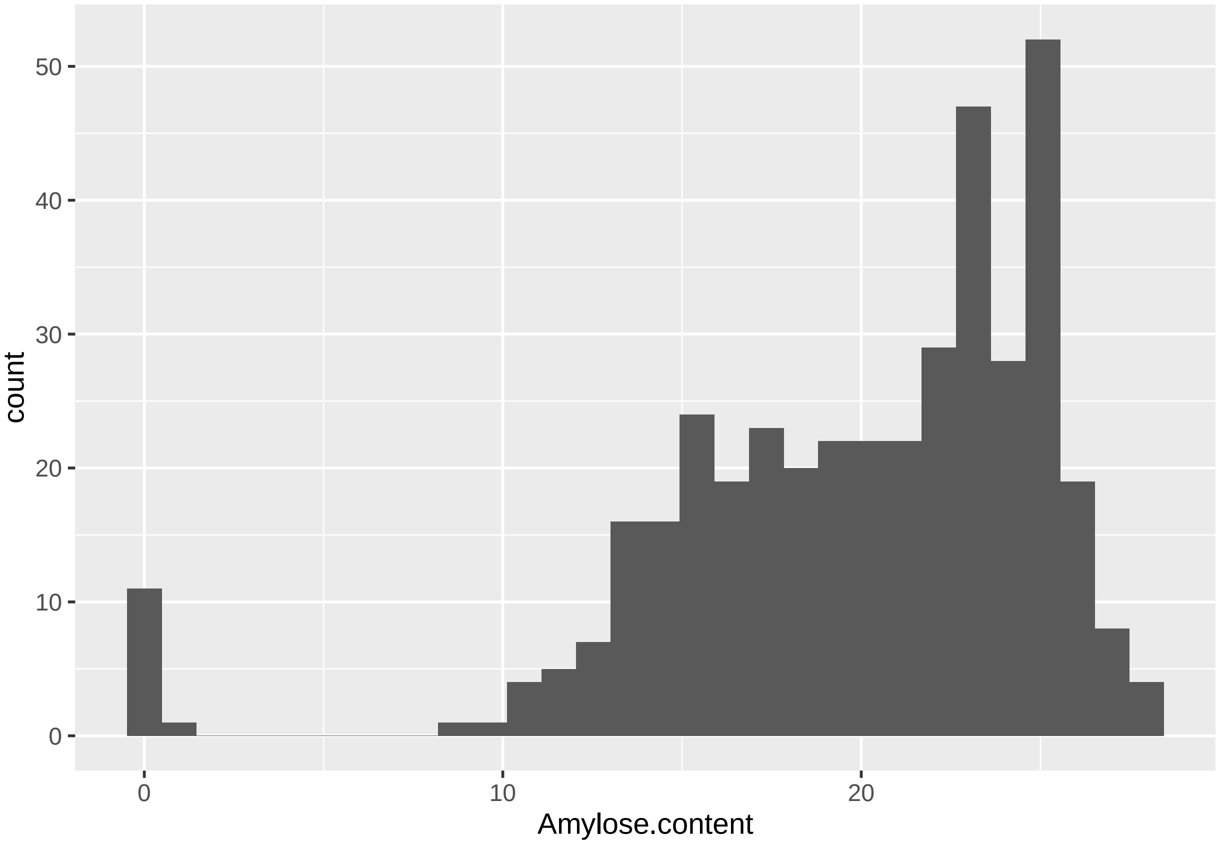


Setup for Exercise 5:

```
#load("data_from_SNP_lab.RData")
data.pheno.mds <- merge(geno.mds,data.pheno,by="row.names",all=T) #even if you already have this object
qplot(x=Amylose.content,data=data.pheno.mds,geom="histogram")
```

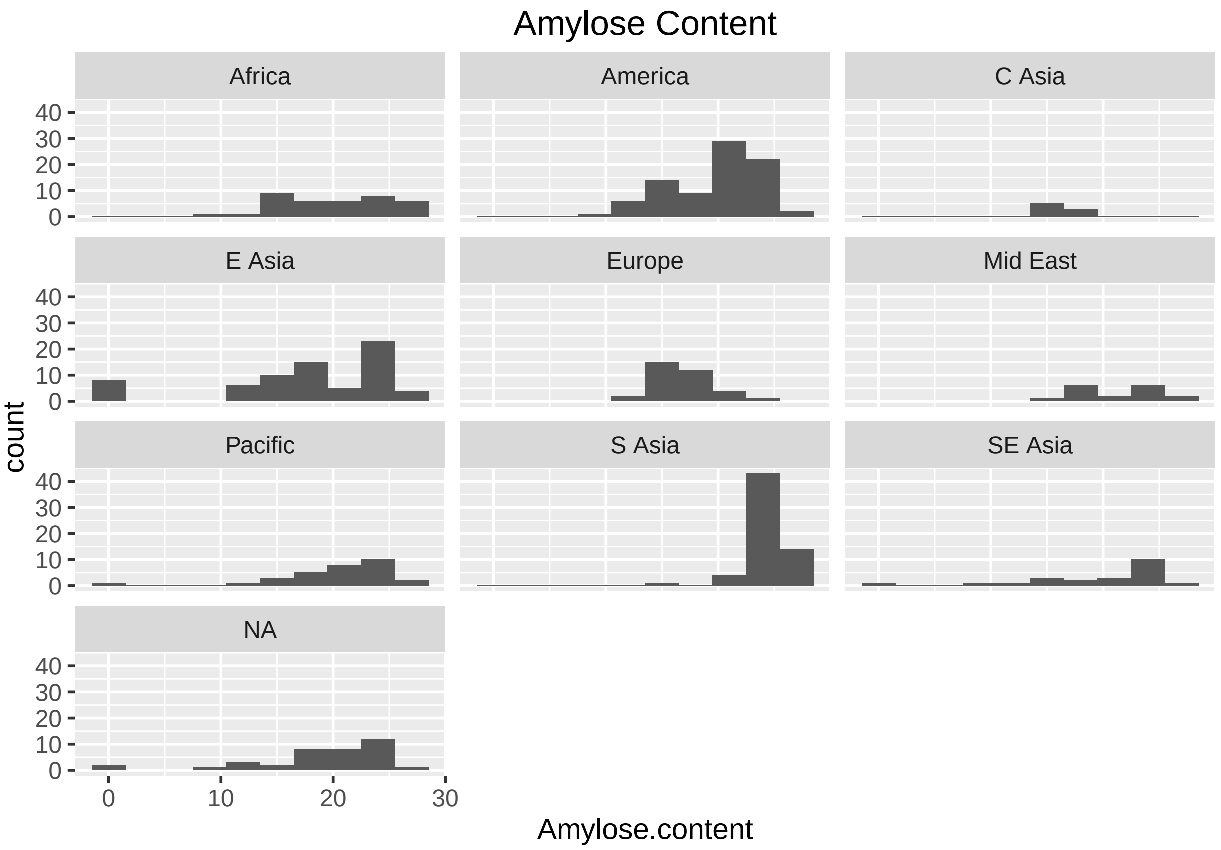
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 12 rows containing non-finite values (stat_bin).
```



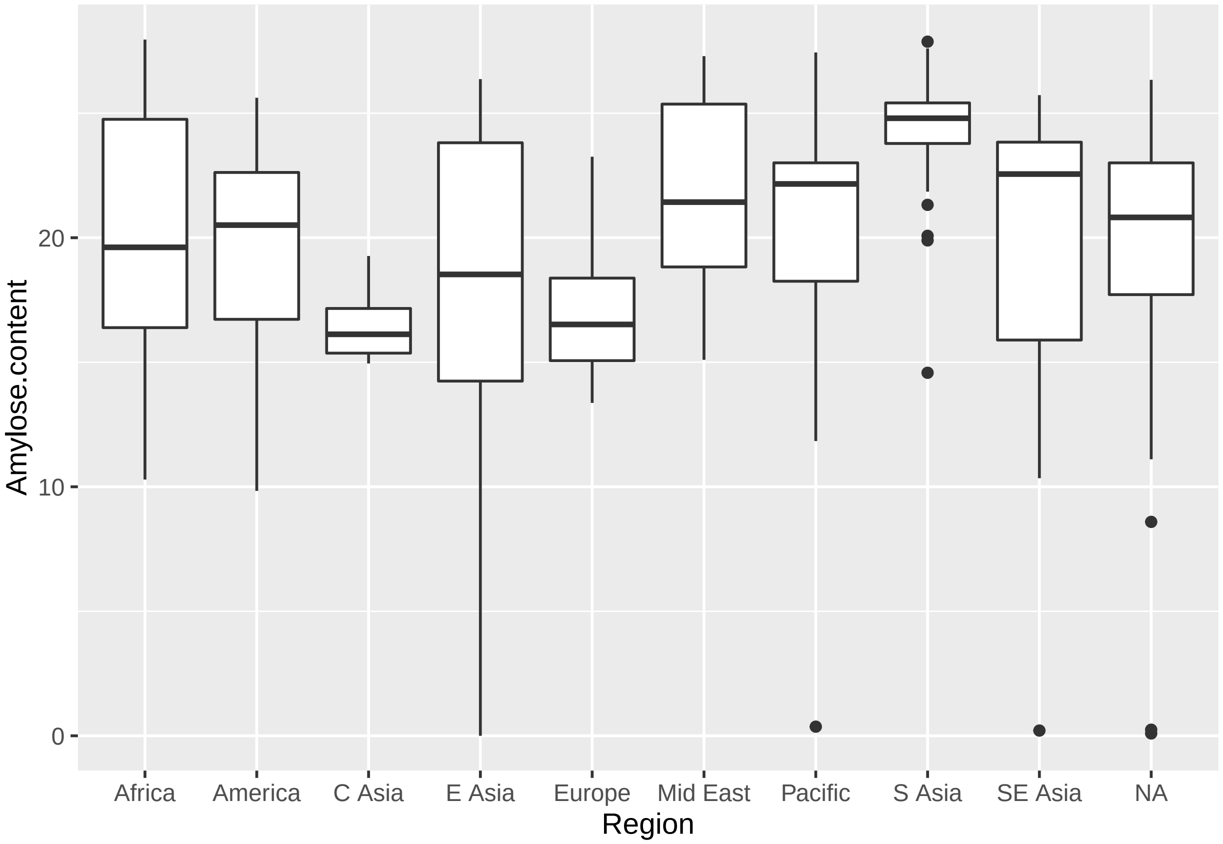
```
pl <- ggplot(data=data.pheno.mds,aes(x=Amylose.content)) #create the basic plot object
pl <- pl + geom_histogram(binwidth=3) + facet_wrap(facets= ~ Region, ncol=3) + ggtitle("Amylose Content
pl #display the plot
```

```
## Warning: Removed 12 rows containing non-finite values (stat_bin).
```



```
qplot(x=Region,y=Amylose.content,geom="boxplot",data=data.pheno.mds)
```

```
## Warning: Removed 12 rows containing non-finite values (stat_boxplot).
```



Exercise 5:

- Plot your chosen trait data
- as a **single histogram** for all of the data
- as **separate histograms** for each of the 4 population assignments made by PSMix
- as a **boxplot** separated by population.
- Based on these histograms do you think that your trait varies by population?
- **BONUS** Try using the "violin" geom. What is this showing?

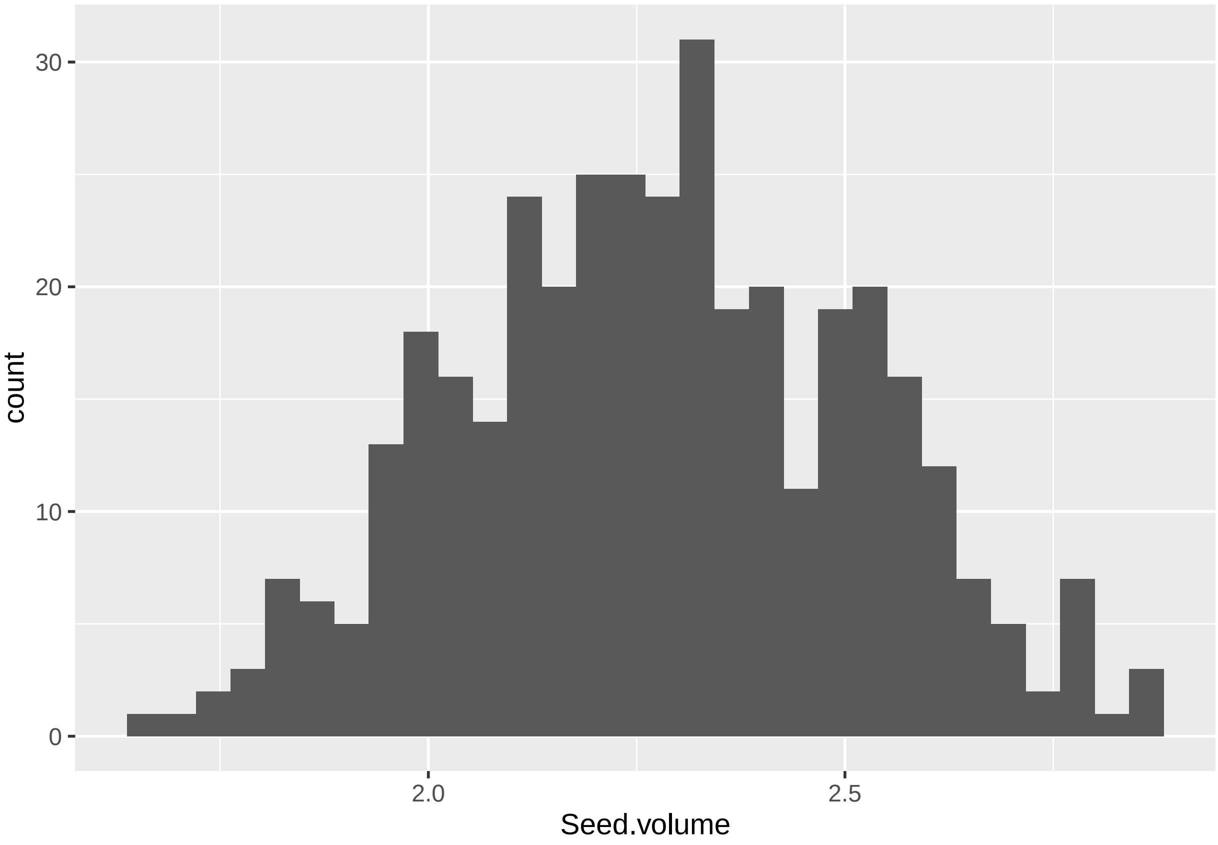
Yes, I do think the seed volume varies by population. This is especially clear in the box plot by population that clearly shows the median difference between population 2 and 4.

The violin plot is showing the probability of density of seed volume for each respective population. For example, it appeaers that population 4 has the highest probability of having a higher than average seed volume whereas population 3 favors having a lower than average seed volume.

```
qplot(x=Seed.volume,data=data.pheno.mds,geom="histogram")
```

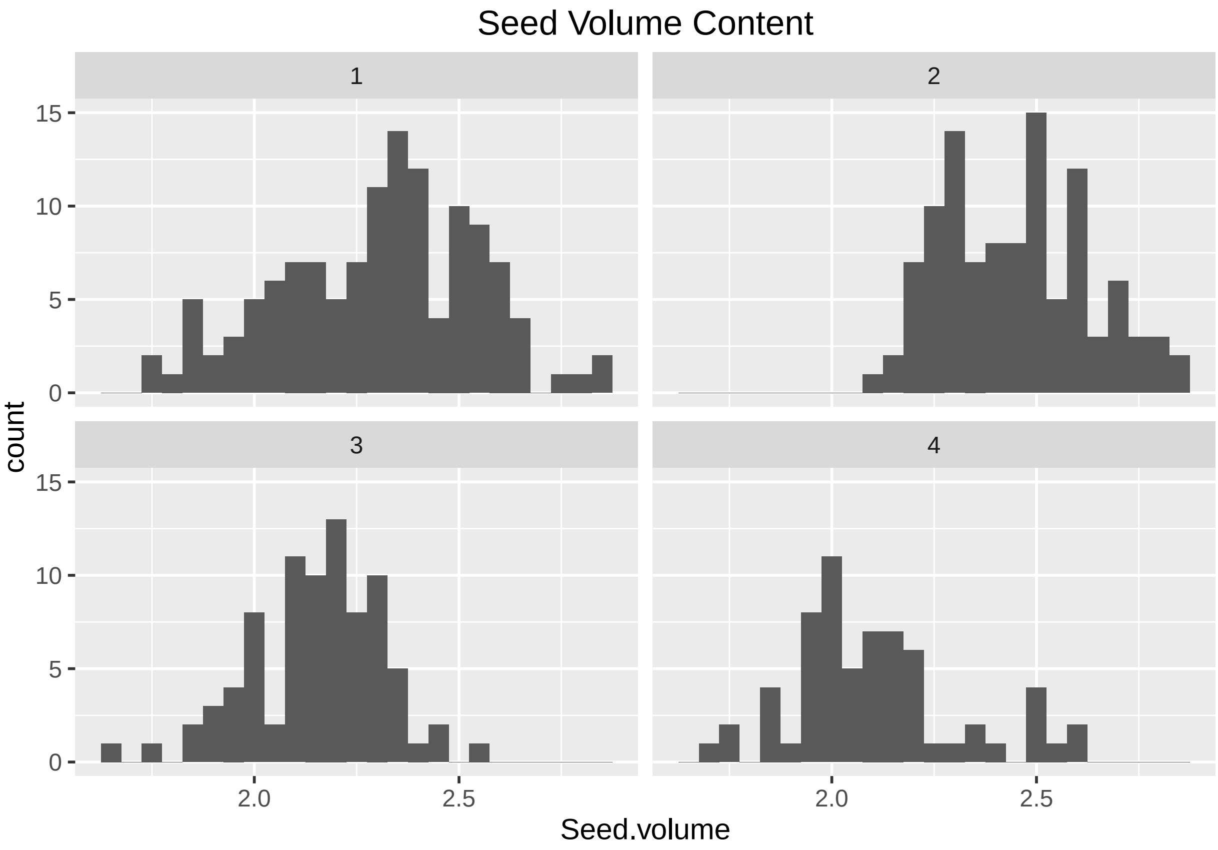
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 36 rows containing non-finite values (stat_bin).
```



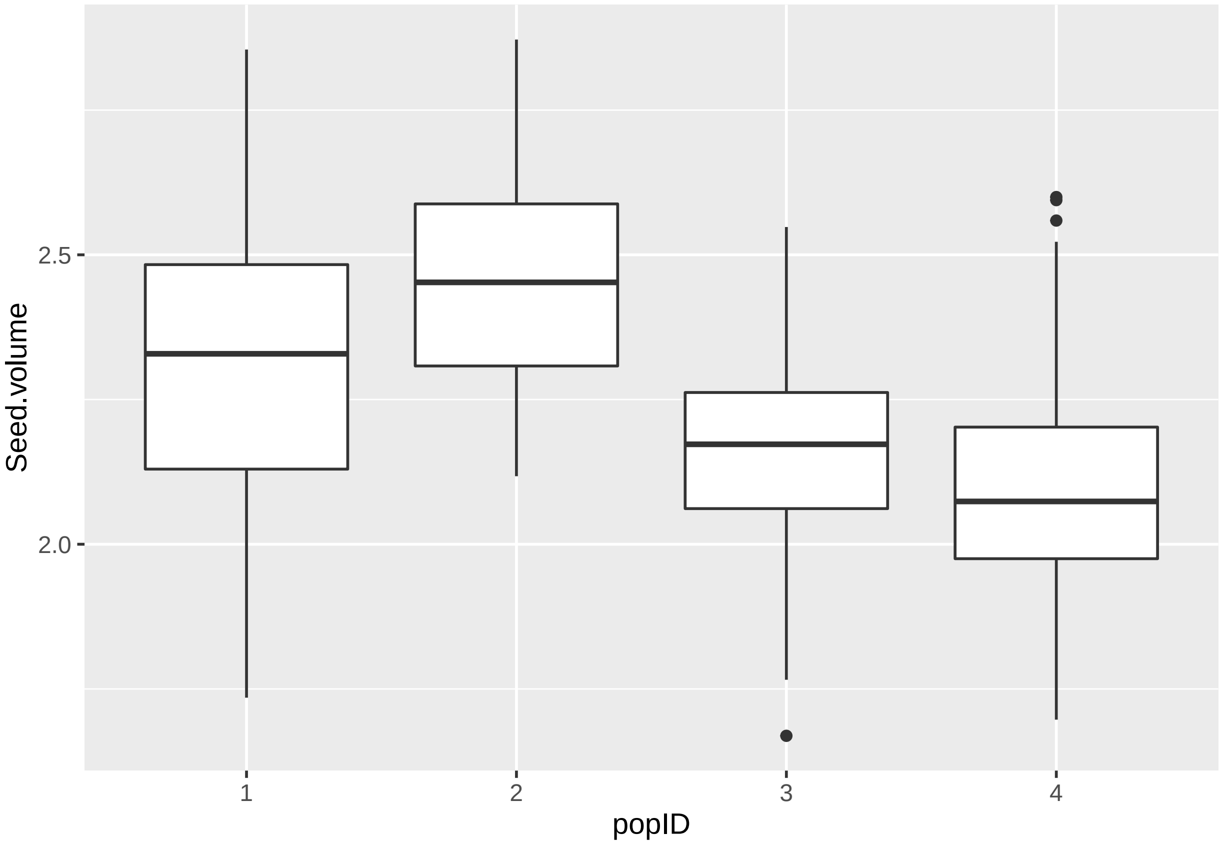
```
plot <- ggplot(data=data.pheno.mds,aes(x=Seed.volume))
plot <- plot + geom_histogram(binwidth=0.05) + facet_wrap(facets= ~ popID) + ggtitle("Seed Volume Conte
plot
```

```
## Warning: Removed 36 rows containing non-finite values (stat_bin).
```



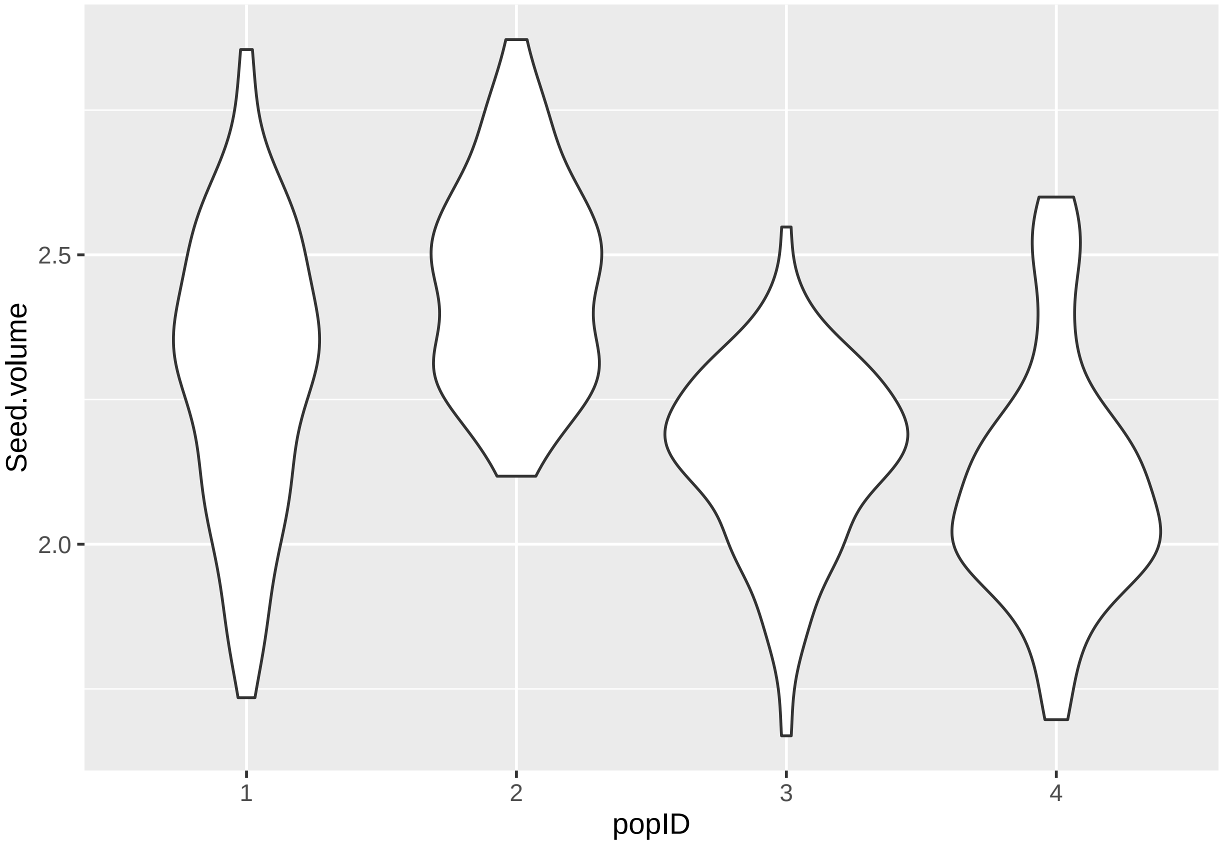
```
qplot(x=popID,y=Seed.volume,geom="boxplot",data=data.pheno.mds)
```

```
## Warning: Removed 36 rows containing non-finite values (stat_boxplot).
```



```
qplot(popID, Seed.volume, data = data.pheno.mds, geom = "violin")
```

```
## Warning: Removed 36 rows containing non-finite values (stat_ydensity).
```



Set-up for Excercise 6:

```
mean(data.pheno.mds$Amylose.content,na.rm=T)
```

```
## [1] 19.88496
```

```
tapply(X=data.pheno.mds$Amylose.content,INDEX=data.pheno.mds$Region,FUN=min,na.rm=T)
```

```
##     Africa    America     C Asia     E Asia     Europe   Mid East 
## 10.2900000  9.8333333 14.9500000  0.0000000 13.3666667 15.0966667 
##    Pacific     S Asia    SE Asia 
##  0.3666667 14.5766667  0.2100000
```

```
aov1 <- aov(Amylose.content ~ Region,data=data.pheno.mds) #1-way ANOVA for Amylose.content by Region
summary(aov1)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)    
## Region        8   2036  254.49   10.63 2.05e-13 ***
## Residuals   355   8495   23.93                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 49 observations deleted due to missingness
```

Exercise 6: Obtain the mean of your trait for each of the 4 PSMix populations. Perform an ANOVA for your trait to test if it varies significantly by population. Show your code, the ANOVA output, and provide an interpretation. Discuss: Do your results present a problem for GWAS?

The mean of the seed volume is 2.29 and the mean of the average seed volumes by population is 2.16. With a p value of 2e-16, seed volume does vary significntly by population. This result does present a problem for GWAS because we need to correct for the natural population associations from nonrandom mating before we can decipher any signifant SNPs.

```
mean(data.pheno.mds$Seed.volume,na.rm=T)
```

```
## [1] 2.278356
```

```
tapply(X=data.pheno.mds$Seed.volume,INDEX=data.pheno.mds$popID,FUN=min,na.rm=T)
```

```
##        1        2        3        4
## 1.734861 2.117516 1.669027 1.696892
```

```
aov2 <- aov(Seed.volume ~ popID,data=data.pheno.mds)
summary(aov2)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## popID         3  6.492  2.1639   52.97 <2e-16 ***
## Residuals   373 15.236  0.0408
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 36 observations deleted due to missingness
```

Set up for Exercise 7:

```
snpinfo <- read.csv("snpInfo.csv",row.names=1)
head(snpinfo) #note one column for chromosome and one for position (in base pairs)
```

```
##        snp chr   pos
## 1 X1_13147   1 13147
## 2 X1_73192   1 73192
## 3 X1_74969   1 74969
## 4 X1_75852   1 75852
## 5 X1_75953   1 75953
## 6 X1_91016   1 91016
```

```
rownames(data.pheno.mds) <- data.pheno.mds$Row.names
data.geno.pheno <- merge(data.pheno.mds,data.geno,by="row.names")
```

```
## Warning in merge.data.frame(data.pheno.mds, data.geno, by = "row.names"):
## column name 'Row.names' is duplicated in the result
```

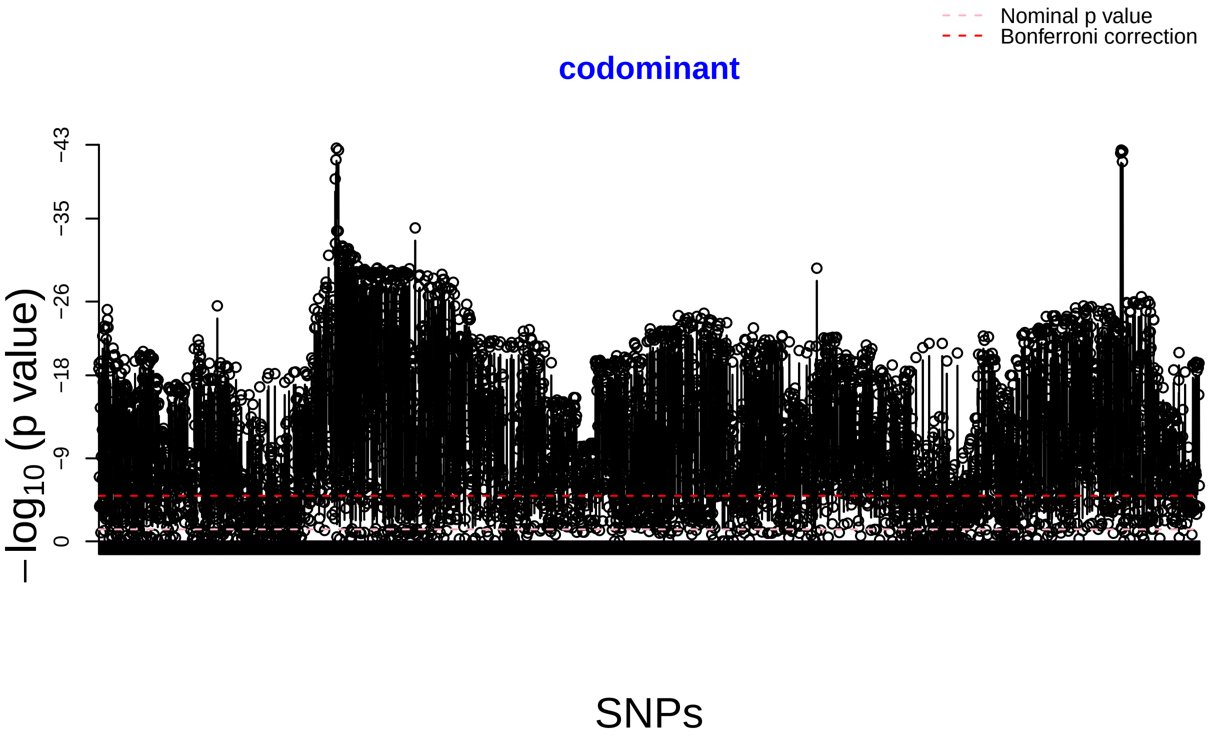
```
library(SNPassoc)
```

```
## Loading required package: haplo.stats
```

```
## Loading required package: survival
## Loading required package: mvtnorm

## Loading required package: parallel
```

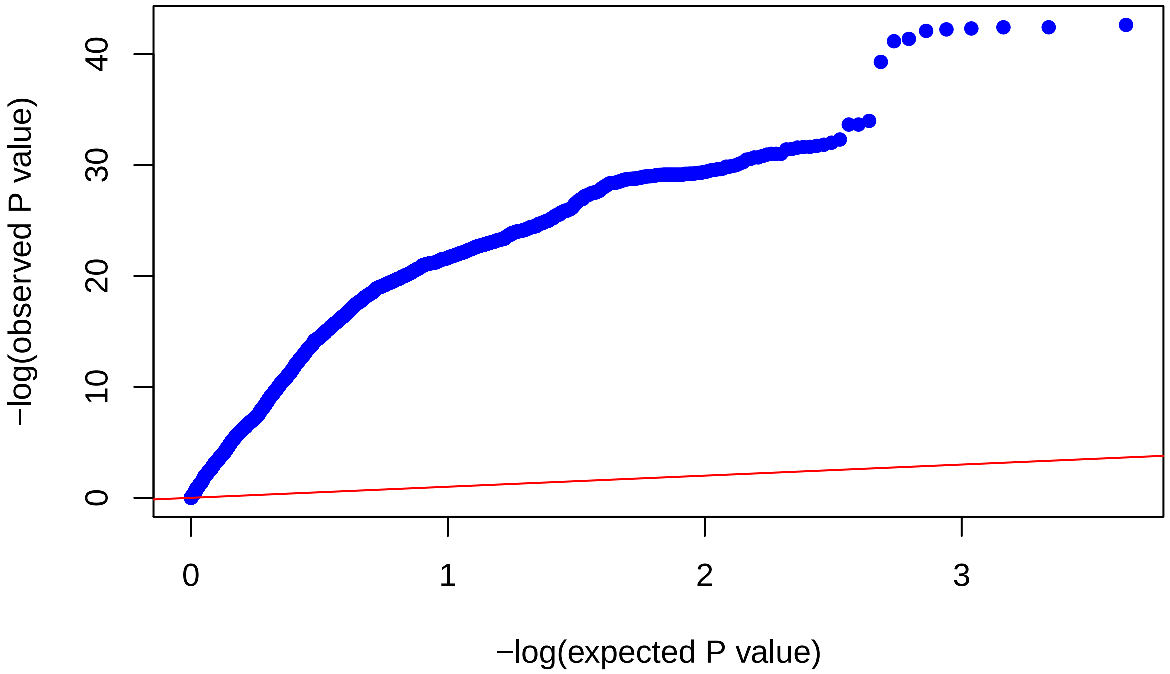
```
data.geno.pheno3 <- data.geno.pheno[,c(1:47,grep("X3_",colnames(data.geno.pheno)))]
snpinfo3 <- snpinfo[snpinfo$chr==3,]
snps3 <- setupSNP(data.geno.pheno3,48:ncol(data.geno.pheno3),sort=T,info=snpinfo3,sep="")
#analysis without population structure correction
#this takes ~ 5 minutes to run.
wg3 <- WGassociation(Seed.volume,data=snps3,model="co",genotypingRate=50)
plot(wg3,print.label.SNPs=FALSE)
```



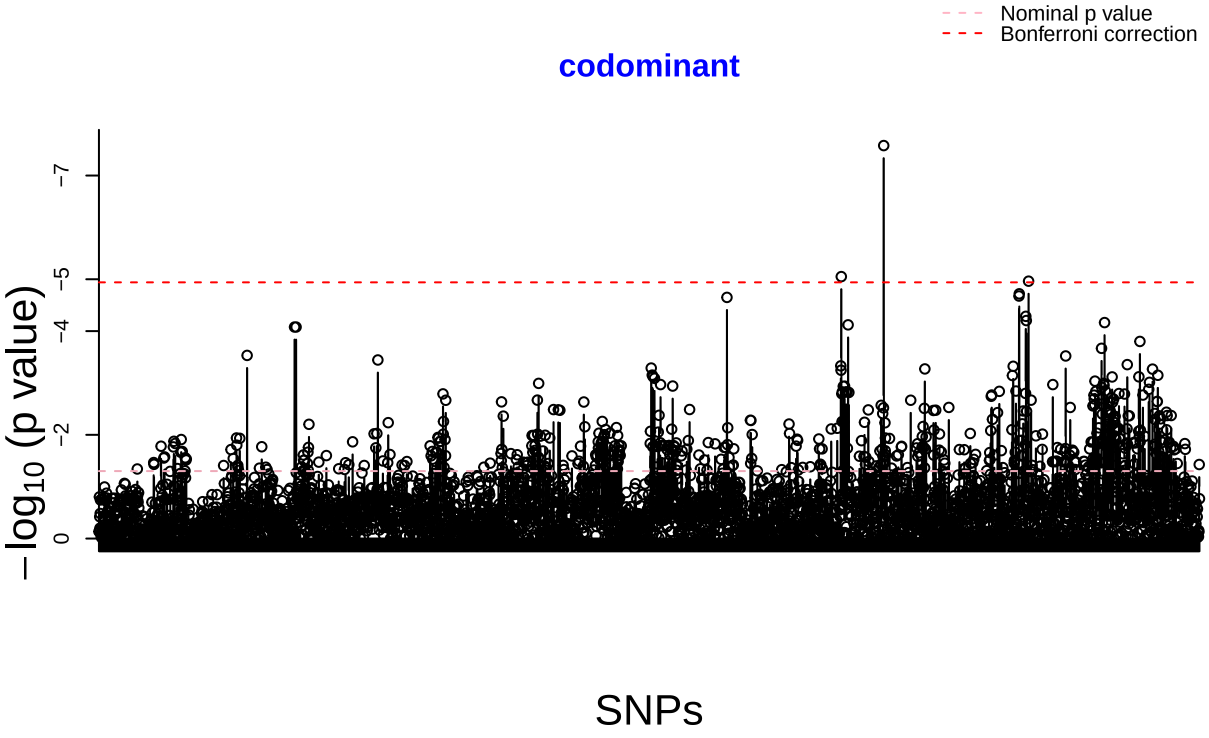
```
#the p-values for the co-dominant model are extracted by using the codominant() function
#determine the number of significant SNPs (p < 0.00001):
sum(codominant(wg3) < 1e-5)
```

```
## [1] 3042
```

```
qqpval(codominant(wg3))
```



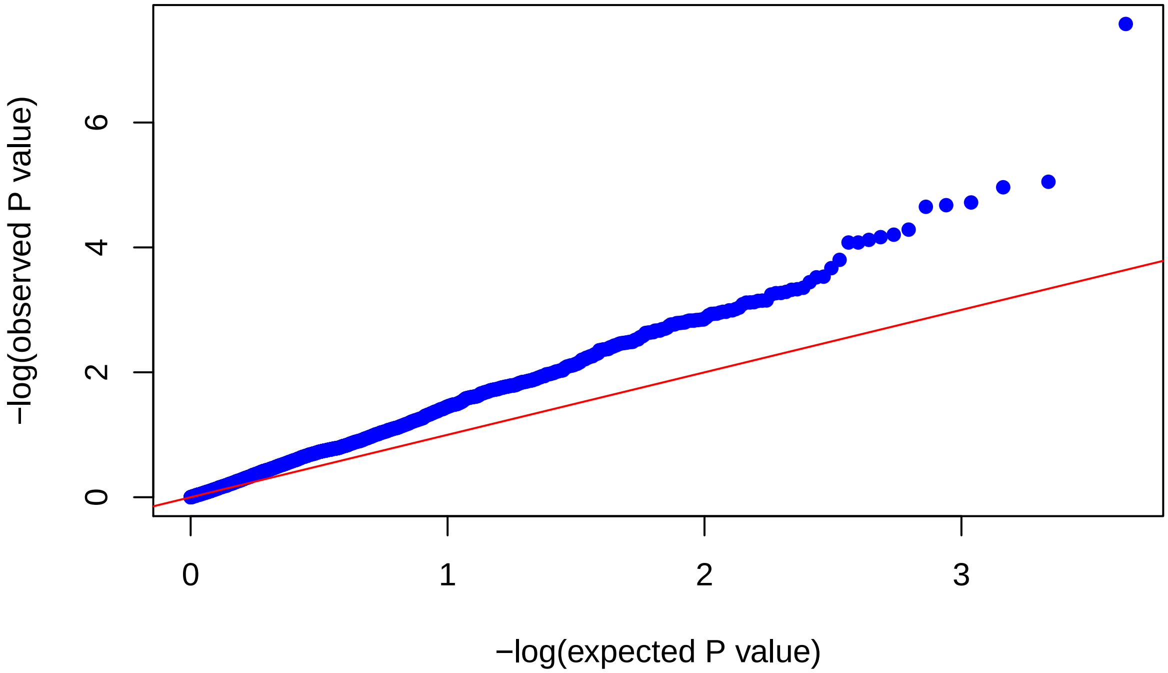
```
wg3.corrected <- WGassociation(Alu.Tol ~ pr1 + pr2 + pr3 + pr4,data=snps3,model="co",genotypingRate=50)
#analysis with population structure correction.
plot(wg3.corrected,print.label.SNPs=FALSE)
```



```
sum(codominant(wg3.corrected) < 1e-5)
```

```
## [1] 2
```

```
qqpval(codominant(wg3.corrected))
```

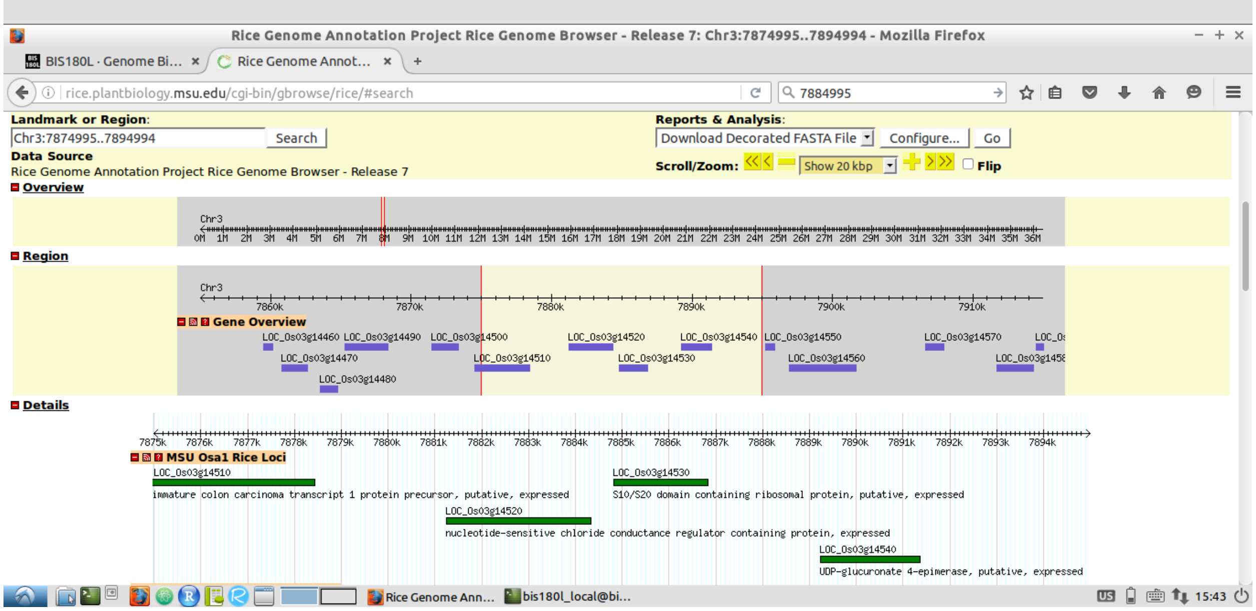


Exercise 7: Describe and discuss the differences in the analysis with and without population structure correction. Which do you think is the better one to follow-up on, and why?

> By incorporating the population structure correction, the analysis is preformed on a more uniform data set. This is because it will remove false positives caused by population structure. This is why the image which incorporates the population structure is much clearer and why the amount of significant SNPs is heavily reduced. It would be better to follow-up on the analysis with population structure correction because you are more likely to find SNPs which are truly associated and not due to false positives.

Exercise 8: Report the SNP you chose and the three closest genes. These are candidate genes for determining the phenotype of your trait of interest in the rice population. Briefly discuss these genes as possible candidates for the GWAS peak. (Include a screenshot of the genome browser)

> SNP chosen: Chr3:7884995 because it had the highest p-value of 7.324748e-18 in the corrected data set.
>
> The closest three genes are: LOC_Os03g14510, LOC_Os03g14530, and LOC_Os03g14520. Most likely, the SNP will either effect LOC_Os03g14520 by falling within the gene, or LOC_Os03g14530 by effecting the promoter upstream.



Rice Genome Annotation Project Rice Genome Browser - Release 7: Chr3:7874995..7894994 - Mozilla Firefox

BIS180L · Genome Bi… ×

Rice Genome Annot… ×

rice.plantbiology.msu.edu/cgi-bin/gbrowse/rice/#search

7884995

Landmark or Region: Chr3:7874995..7894994 Search

Data Source

Rice Genome Annotation Project Rice Genome Browser - Release 7

Reports & Analysis: Download Decorated FASTA File Configure... Go

Scroll/Zoom: Show 20 kbp ☐ **Flip**

Overview

Region

Gene Overview

Details

MSU Osa1 Rice Loci

LOC_Os03g14510

immature colon carcinoma transcript 1 protein precursor, putative, expressed

LOC_Os03g14530

S10/S20 domain containing ribosomal protein, putative, expressed

LOC_Os03g14520

nucleotide-sensitive chloride conductance regulator containing protein, expressed

LOC_Os03g14540

UDP-glucuronate 4-epimerase, putative, expressed