# Rice Genome Wide Area Association Studies (GWAS): An Annotated Guide

Robert Gotwals, Computational Science Educator, NCSSM

Last compiled on November 07, 2023

## About this *Guide*

The purpose of this guide is to provide a line-by-line, code chunk-by-code chunk description of the R code needed to perform a genome-wide association study (GWAS) analysis of breeding data between two species of rice (*Oryza sativa*): *Curinga* x *O. rufipogon*. This *Guide* does not provide a description of GWAS analyses; that is found in the curricular materials. This *Guide* will annotate the required code for conducting the analyses.

This *Guide* assumes that you have downloaded and installed the R software (https://cran.rstudio.com/ (https://cran.rstudio.com/)) and the interface to R, RStudio (https://posit.co/download/rstudio-desktop/ (https://posit.co/download/rstudio-desktop/)). It also assumes that, in R, you have installed two “packages”; “statgenGWAS”d and “tidyverse. To install this, run the command”install.packages(“statgenGWAS,”tidyverse”) at the console command line. You only have to do tis once.

# About this activity

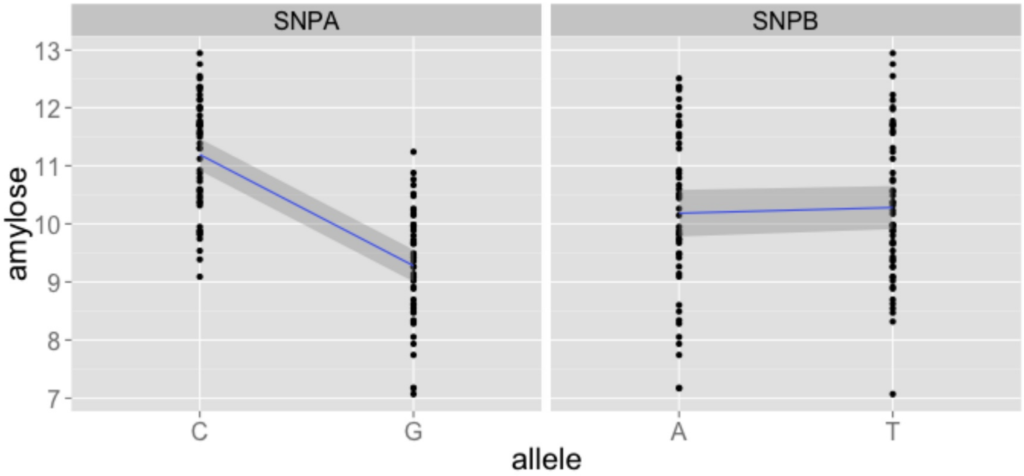
[NOTE! Success on this activity assumes that you have read the background reading of GWAS in the “Computing the Rice Genome” handout.]

In this activity, you are asked to perform a genome-wide association study – GWAS – on a large collection of rice genomes. The original dataset has 413 rice varieties from 82 countries, with approximately 44,000 genetic mutations – single nucleotide polymorphisms, or SNPS – on each of the different types of rice.

Let’s look at a simple example. Plants contain a compound called “amylose”, one of two compounds (the other being amylopectin) that make up wnat we call “starch”. Starch is a primary source of energy for humans (and animals), and is found in plants such as wheat, potatoes, bread, and, yes, rice. Starch is a carbohydrate, or “carb”. Some nutritionists argue that “carbs” are bad for you, but they are a critical source of energy from food.

There are thousands of genetic mutations, or SNPs, that characterize a given phenotype, or trait, such as amylose content. The graphic below show an example of two arbitrary SNPs. SNP-A shows that plants with an “A” allele (adenosine) tend to have higher amylose content than plants with a “G” (guanine) allele. Likewise, for SNP-B, higher amylose content is found in plants with an “A” (adenine) allele than plants with a “T” (thymine) allele.

**The fact that the slope between the two alleles is not flat tells you that there is an association, or relationship, between which allele is present and the amount of amylose content a given plant has.**



Map of two SNPs found in amylose
Like with the QTL activity, we want to focus in on a given trait – such as amylose content – and look for where that trait might be found among the 12 chromosomes of rice. We are going to look at conducting three different GWAS studies: 1. A GWAS study on the dataset without any corrections 2. A GWAS study using information from a Principal Component Analysis (PCA) study to reduce the number of SNPs shown. 3. A GWAS study using what is known as a Kinship Matrix.

The Wikipedia entry on GWAS is excellent: https://en.wikipedia.org/wiki/Genome-wide_association_study (https://en.wikipedia.org/wiki/Genome-wide_association_study)

# Load the libraries

```
rm(list=ls())
library(tidyverse)
library(statgenGWAS)
library(ggplot2)
```

# Load the data.

This dataset contains single nucleotide polymorphism data

```
url <- "http://chemistry.ncssm.edu/data/gwas/SNPlab.Rdata?raw=true"
download.file(url, destfile = "SNPlab.Rdata", mode="wb")
load("SNPlab.Rdata")
```

Check the dimensions

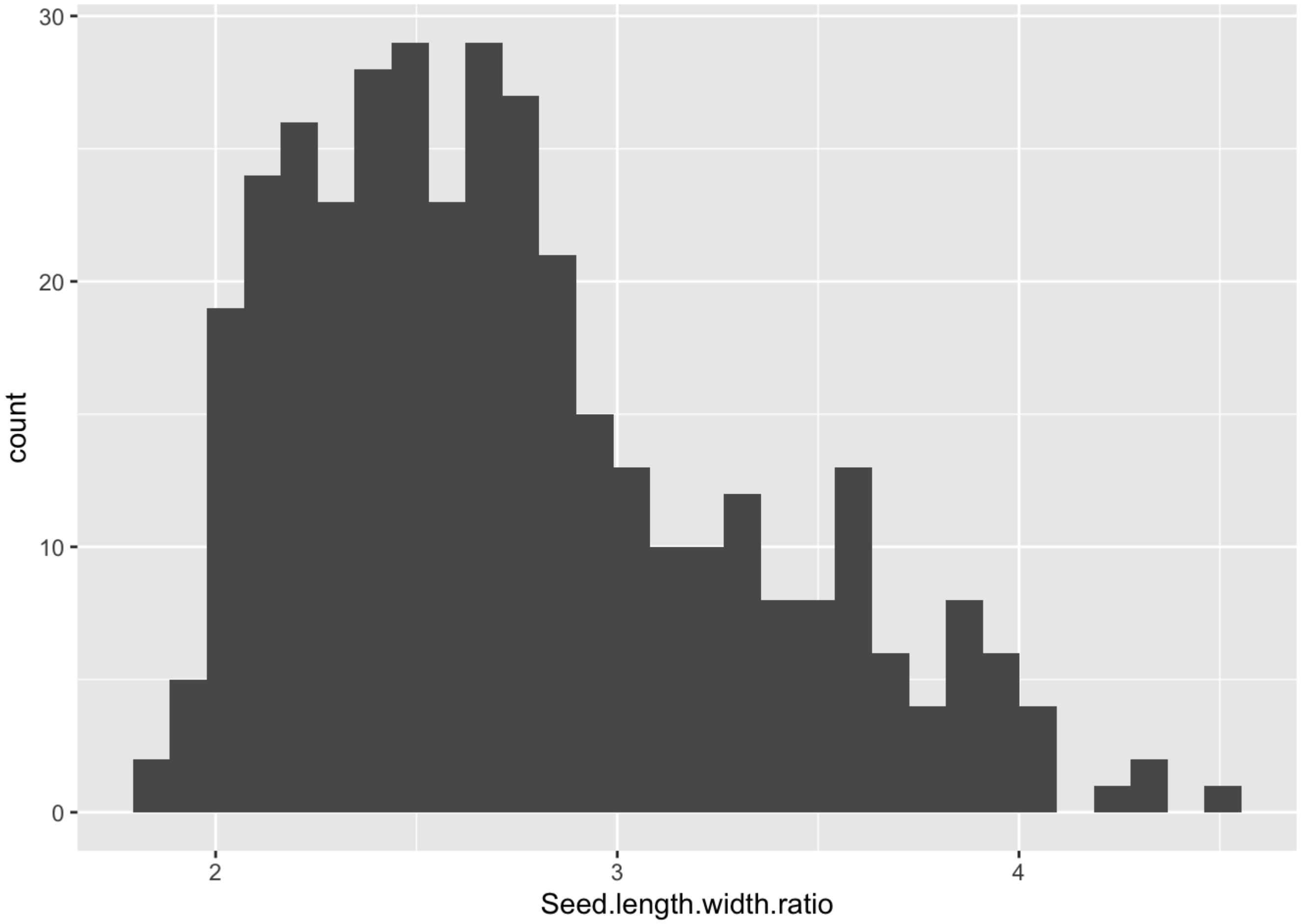
```
## [1] "There are  413  rows and  56  columns of data"
```

The names of the phenotypes are as follows:

```
##  [1] "Country_of_Origin"             "Region"                         
##  [3] "Alu.Tol"                       "Flowering.time.at.Arkansas"     
##  [5] "Flowering.time.at.Faridpur"    "Flowering.time.at.Aberdeen"     
##  [7] "FT.ratio.of.Arkansas.Aberdeen" "FT.ratio.of.Faridpur.Aberdeen"  
##  [9] "Culm.habit"                    "Leaf.pubescence"                
## [11] "Flag.leaf.length"              "Flag.leaf.width"                
## [13] "Awn.presence"                  "Panicle.number.per.plant"       
## [15] "Plant.height"                  "Panicle.length"                 
## [17] "Primary.panicle.branch.number" "Seed.number.per.panicle"        
## [19] "Florets.per.panicle"           "Panicle.fertility"              
## [21] "Seed.length"                   "Seed.width"                     
## [23] "Seed.volume"                   "Seed.surface.area"              
## [25] "Brown.rice.seed.length"        "Brown.rice.seed.width"          
## [27] "Brown.rice.surface.area"       "Brown.rice.volume"              
## [29] "Seed.length.width.ratio"       "Brown.rice.length.width.ratio"  
## [31] "Seed.color"                    "Pericarp.color"                 
## [33] "Straighthead.suseptability"    "Blast.resistance"               
## [35] "Amylose.content"               "Alkali.spreading.value"         
## [37] "Protein.content"
```

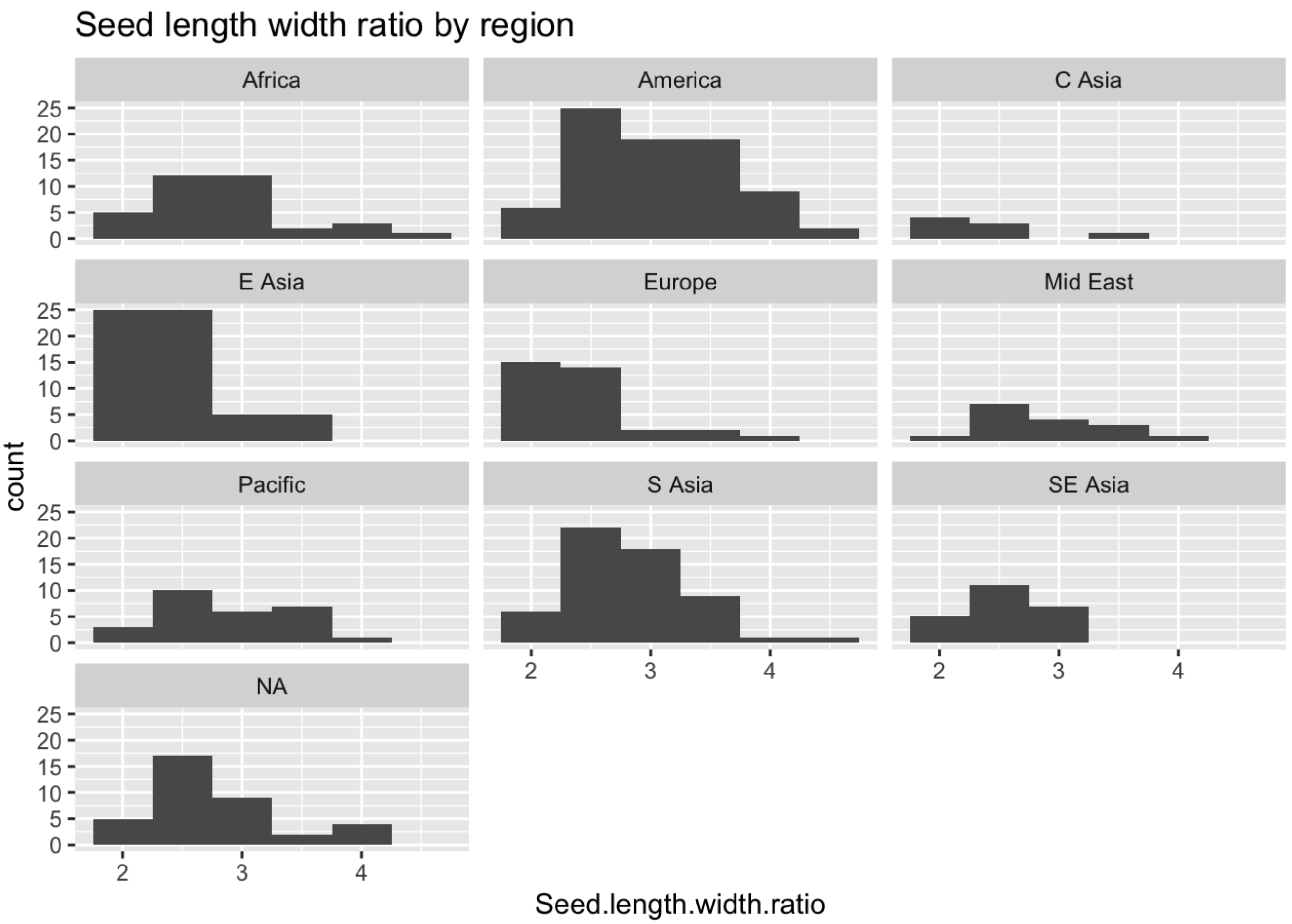
One of the defining characteristics of rice is the grain or seed shape (short, medium, or long grain). Let’s examine the distribution of seed length to width ratios. You will recall we can make a histogram as follows:

```
pheno.geno.pca.pop %>%
  ggplot(aes(x=`Seed.length.width.ratio`)) +
  geom_histogram()
```

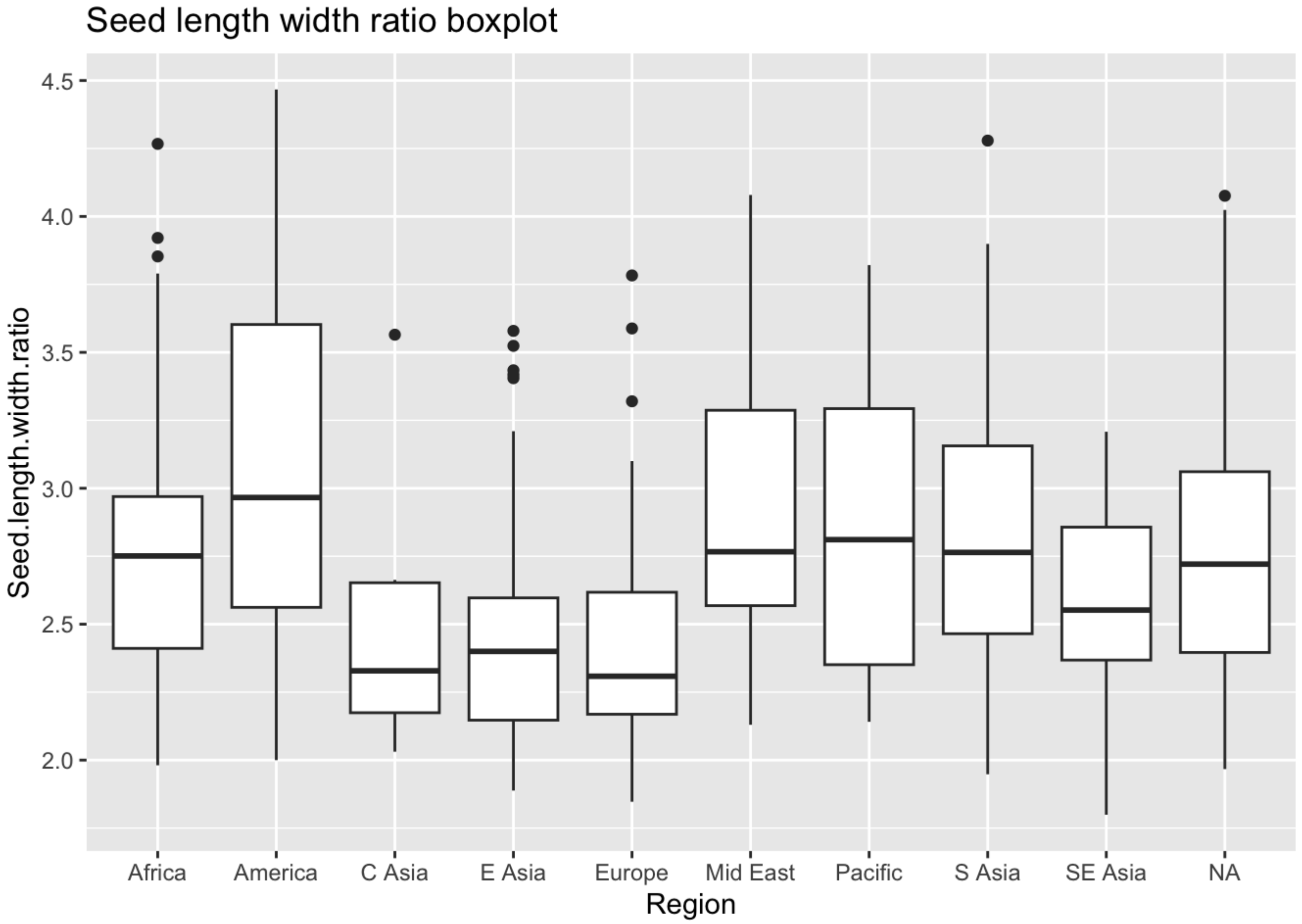


It might be interesting to ask if the distributions look similar for different regions.
We can easily produce separate histograms for each level of a factor (as you saw in the ggplot tutorial). As a refresher, first we create a plot object, called pl, using the ggplot() function. Then we add additional information to the plot. the mapping=aes() argument tells R about the "aesthetics" of the plot, in other words which variable should be mapped to which aspect of the plot.

```
pl <- ggplot(data=pheno.geno.pca.pop, aes(x=Seed.length.width.ratio))
pl <- pl + geom_histogram(binwidth=.5)
pl <- pl + facet_wrap(facets= ~ Region, ncol=3)
pl <- pl + ggtitle("Seed length width ratio by region")
pl #display the plot
```

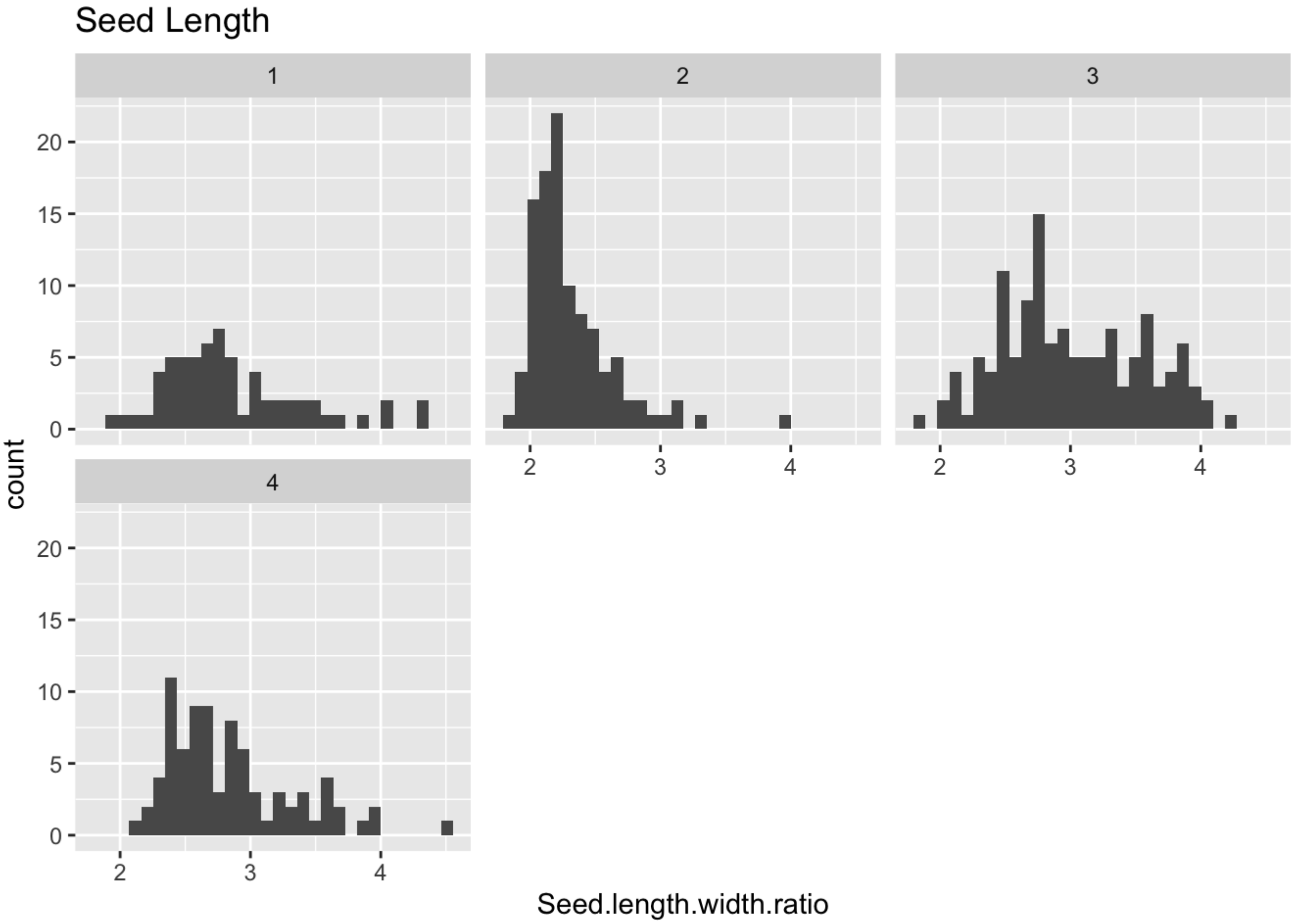


```
pheno.geno.pca.pop %>%
  ggplot(aes(x=Region,y=`Seed.length.width.ratio`)) +
  geom_boxplot() + ggtitle("Seed length width ratio boxplot")
```



For the rest of the lab, choose a trait to work on. I recommend one of `Flowering.time.at.Aberdeen`, `Leaf.pubescence`, `Seed.length.width.ratio`, `Seed.width`, `Brown.rice.Seed.length.width.ratio`, `Brown.rice.seed.width`, `Brown.rice.length.width.ratio`.

```
pl <- ggplot(data=pheno.geno.pca.pop,aes(x=Seed.length.width.ratio))
pl <- pl + geom_histogram()
pl <- pl + facet_wrap(facets= ~ assignedPop, ncol=3)
pl <- pl + ggtitle("Seed Length")
pl #display the plot
```



For the seed length width trait, there appear to be differences in the means between regions.

# GWAS

A Genome Wide Association Study (GWAS) looks for significant associations between allele state at SNPs and phenotypic traits. It tests each SNP in turn.

## Run the GWAS

We will compare GWAS run with a few different population structure corrections and methods

1. No correction
2. Population PCAs as covariate
3. Kinship matrix as correction

### PCAs: Principal Component Analysis

One of the techniques used to deal with large amounts of data, particularly data with many variables (this dataset has 56 different variables) is with a machine learning technique known as *Principal Component Analysis*, or PCA for short. PCA looks at the data in terms of *variance*, or how spread out the data is, and reduces the dataset from n-dimensions (in our case, 56-dimensions) to two or three dimensions, which are much easier to plot or otherwise analyze. The new dimensions are known as “PC1”, “PC2”, etc. Typically, the majority of the data is captured in the PC1 and PC2 variables.

In this activity, we’re going to run PCA on the data, and use it as a “cofactor”, or covariate, on the the original dataset. This helps us to only include the parts of the data that are particularly significant. For our convariates, we’ll use the first four PCs – PC1, PC2, PC3, and PC4.

## Kinship Matrix

A Kinship matrix is one way to adjust for population structure. Kinship refers to the genetic relatedness between individuals. For example, if you have siblings your kinship with them is 50%. Similarly your kinship with each of your parents is 50%.

Since we do not know the pedigree of these rice strains, we can instead determine kinship from the observed genotypic data.

## No correction

```
nullmat <- matrix(0, ncol=413,nrow=413, dimnames = dimnames(data.kinship))
gwas.noCorrection<- runSingleTraitGwas(gData = gData.rice.recode,
                                       traits = "Seed.length.width.ratio",
                                       kin = nullmat)
```

We can get a quick summary with:

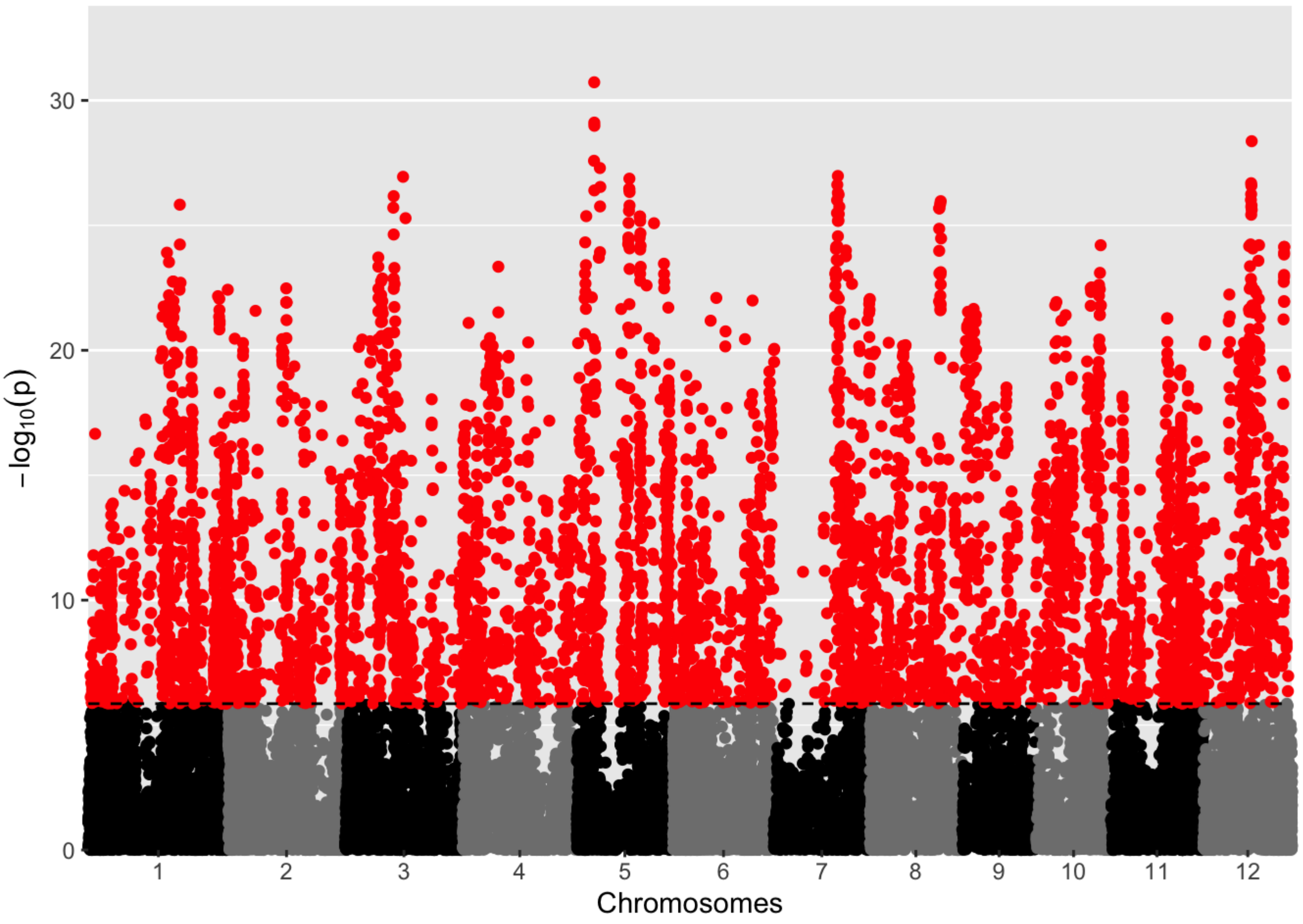
```
summary(gwas.noCorrection)
```

```
## data.pheno.small:
##  Traits analysed: Seed.length.width.ratio
##
##  Data are available for 35877 SNPs.
##   82 of them were not analyzed because their minor allele frequency is below 0.01
##
##  GLSMethod: single
##
##  Trait: Seed.length.width.ratio
##
##      Mixed model with only polygenic effects, and no marker effects:
##      Genetic variance: 1.926348e-05
##      Residual variance: 0.3070004
##
##      LOD-threshold: 5.854852
##      Number of significant SNPs: 6057
##      Smallest p-value among the significant SNPs: 1.87946e-31
##      Largest p-value among the significant SNPs: 1.392822e-06 (LOD-score: 5.856104)
##
##      No genomic control correction was applied
##      Genomic control inflation-factor: 11.302
```

What needs to be observed here are the number of single nucleotide polymorphisms (SNPs) that are identified. Without any corrective measures, there are over 6,000 significant SNPs identified. This is out of a total of 35,877 SNPs found in the dataset.

A Manhattan plot shows this large number of signficant SNPs.

```
plot(gwas.noCorrection, plotType = "manhattan")
```



In the Manhattan plot, each dot is a SNP. SNPs in red are above the significance threshold. If you see a lot of red, then that means that many, many SNPs were significant.

The Manhattan plot uses alternating gray and black shading to allow you to see the individual chromosomes more easily.

# PCA as population correction

Let's see if things improve if we correct for population structure. One methods is to include the PCs as covariates in the analysis. You can this with:

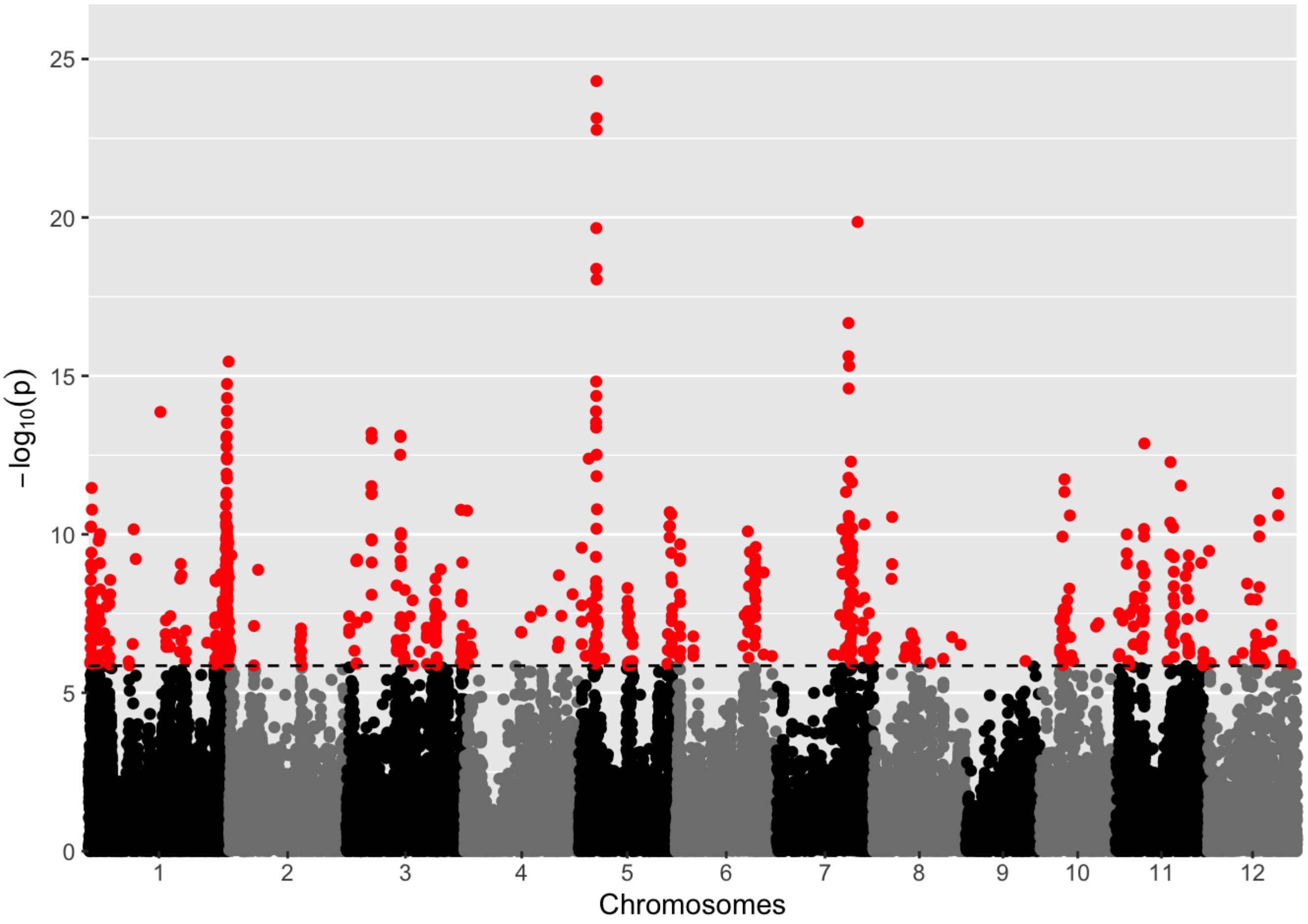
```
# run the GWAS
gwas.PCA <- runSingleTraitGwas(gData = gData.rice.recode,
                               traits = "Seed.length.width.ratio",
                               kin = nullmat,
                               covar = c("PC1", "PC2", "PC3", "PC4")
)
```

Again, you can summarize and plot with:

```
summary(gwas.PCA)
```

```
## data.pheno.small:
##  Traits analysed: Seed.length.width.ratio
##
##  Data are available for 35877 SNPs.
##   82 of them were not analyzed because their minor allele frequency is below 0.01
##
##  GLSMethod: single
##
##  Trait: Seed.length.width.ratio
##
##      Mixed model with only polygenic effects, and no marker effects:
##      Genetic variance: 3.454236e-05
##      Residual variance: 0.2025171
##
##      LOD-threshold: 5.854852
##      Number of significant SNPs: 778
##      Smallest p-value among the significant SNPs: 4.961236e-25
##      Largest p-value among the significant SNPs: 1.394354e-06 (LOD-score: 5.855627)
##
##      No genomic control correction was applied
##      Genomic control inflation-factor: 3.619
```

```
plot(gwas.PCA, plotType = "manhattan")
```



With the PCA-correction, we can see that the number of significant SNPs has been reduced to almost 800 (778), a huge improvement over the 6,000 or so significant SNPs we got with the no-correction model.

# Kinship Matrix

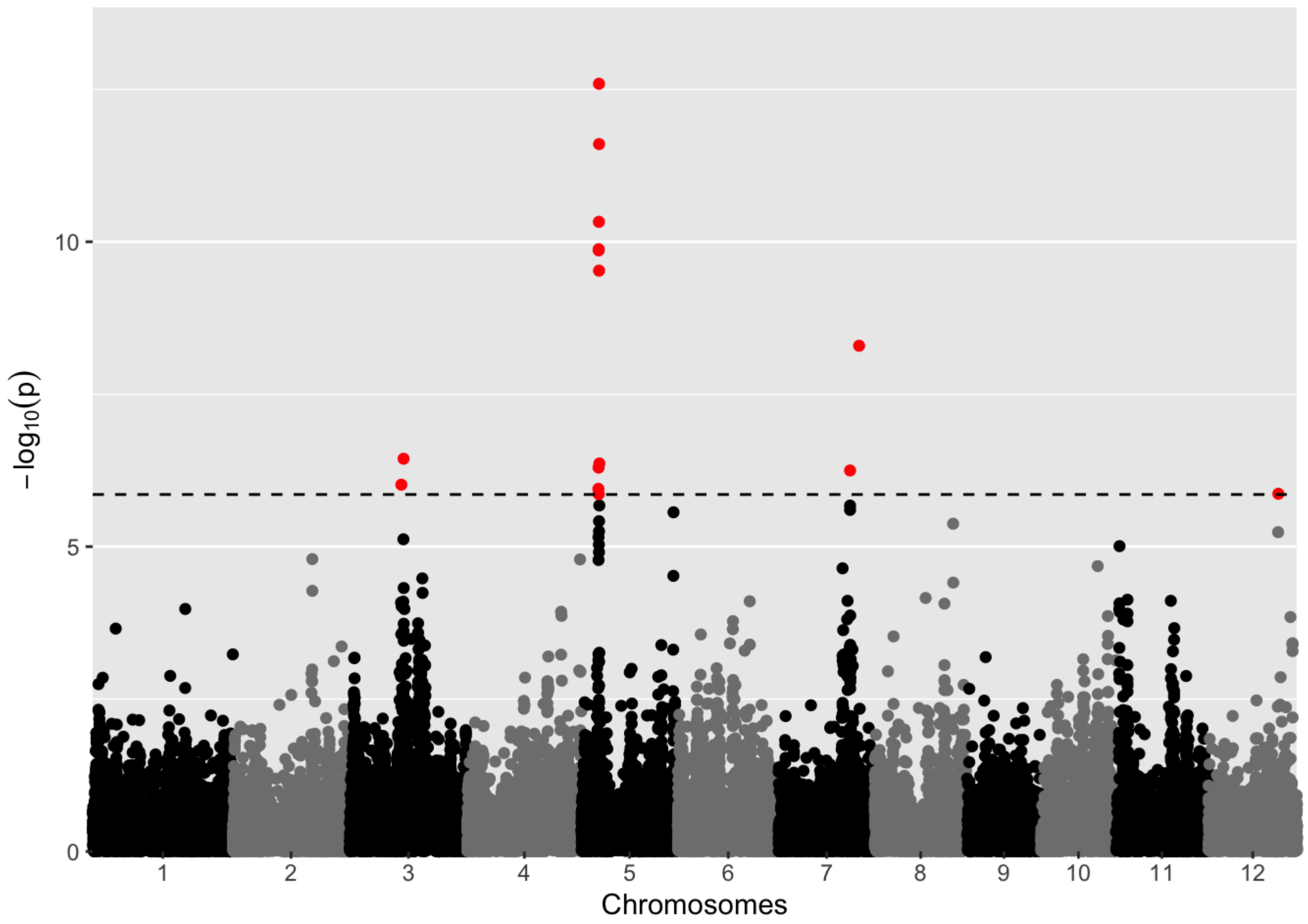
An alternative strategy for population structure correction is to use a kinship matrix, which provides the genetic relatedness of each strain to all of the others.

```
# run the GWAS
gwas.K <- runSingleTraitGwas(gData = gData.rice.recode,
                              traits = "Seed.length.width.ratio",
                             kin = data.kinship)
```

```
summary(gwas.K)
```

```
## data.pheno.small:
##  Traits analysed: Seed.length.width.ratio
##
##  Data are available for 35877 SNPs.
##   82 of them were not analyzed because their minor allele frequency is below 0.01
##
##  GLSMethod: single
##
##  Trait: Seed.length.width.ratio
##
##      Mixed model with only polygenic effects, and no marker effects:
##      Genetic variance: 0.1805641
##      Residual variance: 0.009703427
##
##      LOD-threshold: 5.854852
##      Number of significant SNPs: 15
##      Smallest p-value among the significant SNPs: 2.557938e-13
##      Largest p-value among the significant SNPs: 1.376259e-06 (LOD-score: 5.8613)
##
##      No genomic control correction was applied
##      Genomic control inflation-factor: 0.846
```

```
plot(gwas.K, plotType = "manhattan")
```



With the Kinship matrix model, the summary shows that we have 15 significant SNPs, a number much easier to navigate.

** Compare the Q-Q and Manhattan plots of the no correction, PCA correction, and kinship matrix runs. Based on the no-correction results, do you think a correction was needed? Did the corrections make a difference? If so, which one worked better? How did this effect the number of “significant” SNPs in the Manhattan plot? (In the Manhattan plot the horizontal line represents the significance threshold. If you don’t see any dots above the line, nothing was).

# get the significant SNPs

Let’s retrieve the significant SNPs for our trait. Choose which ever model you think was best, and then extract the SNPs with:

```
sigSnps <- gwas.K$signSnp[[1]]
sigSnps
```

| trait | snp | c... | pos | allFreq | pValue | effect |
|---|---|---|---|---|---|---|
| <chr> | <chr> | <int> | <int> | <dbl> | <dbl> | <dbl> |
| Seed.length.width.ratio | 3_16074273 | 3 | 16074273 | 0.36074271 | 9.648273e-07 | -0.1584059 |
| Seed.length.width.ratio | 3_16759297 | 3 | 16759297 | 0.22546419 | 3.615754e-07 | -0.2136852 |
| Seed.length.width.ratio | 5_5239714 | 5 | 5239714 | 0.25729443 | 1.128044e-06 | 0.1655768 |

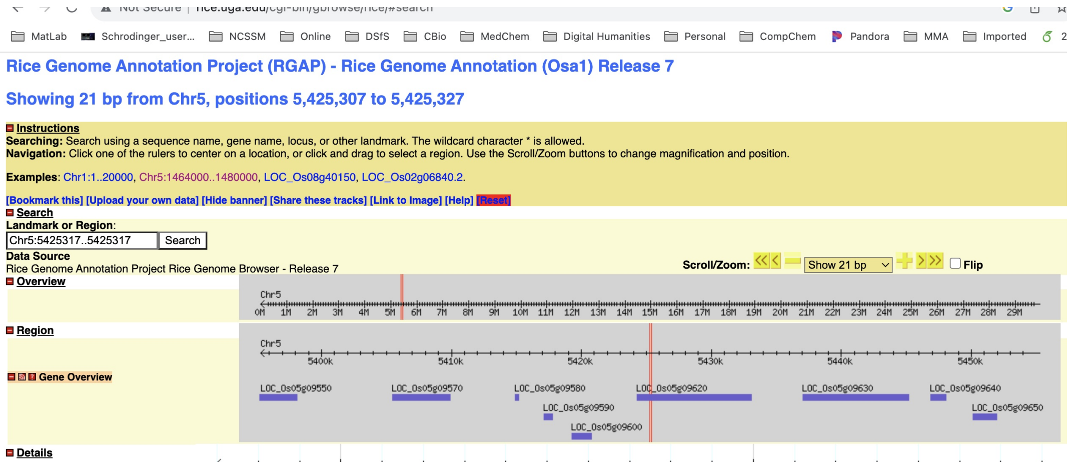
| trait<br><chr> | snp<br><chr> | c...<br><int> | pos<br><int> | allFreq<br><dbl> | pValue<br><dbl> | effect<br><dbl> |
|---|---|---|---|---|---|---|
| Seed.length.width.ratio | 5_5294319 | 5 | 5294319 | 0.40450928 | 5.061484e-07 | 0.1602426 |
| Seed.length.width.ratio | 5_5338183 | 5 | 5338183 | 0.27586207 | 1.376259e-06 | 0.1761666 |
| Seed.length.width.ratio | 5_5341363 | 5 | 5341363 | 0.39124668 | 1.324712e-10 | 0.2058287 |
| Seed.length.width.ratio | 5_5343859 | 5 | 5343859 | 0.40318302 | 1.383723e-10 | 0.2037028 |
| Seed.length.width.ratio | 5_5396733 | 5 | 5396733 | 0.45092838 | 4.708428e-11 | -0.1871518 |
| Seed.length.width.ratio | 5_5425317 | 5 | 5425317 | 0.43766578 | 2.557938e-13 | -0.1904083 |
| Seed.length.width.ratio | 5_5431746 | 5 | 5431746 | 0.46419098 | 2.972041e-10 | -0.1724215 |

1-10 of 15 rows | 1-8 of 12 columns Previous **1** 2 Next

In this table, some of the column names should be obvious. Here are some of the others:

- allFreq: the minor allele frequency at that SNP
- effect: the change in the trait value associated with the SNP
- effectSE: the standard error of the estimated effect
- RLR2: likelihood-ratio-based $R^2$ as defined in Sun et al. (2010)
- LOD: -log10(P) (which isn't a true LOD score…)
- propSnpVar: of the total variance in the trait, what proportion is "explained" by this SNP

Looking at the list (and you will need to use the right arrow key on the table header to see the LOD score), I find that the 9th entry has a LOD score of 12.592110. This is a SNP on Chromosome 5, at location 5425317 (basepairs). The Rice Genome Browser (http://rice.uga.edu/ (http://rice.uga.edu/)) produces this graphic with a search of Chr5:5425317..5425317


Not Secure | rice.uga.edu/cgi-bin/gbrowse/rice/#search

Rice Genome Annotation Project (RGAP) - Rice Genome Annotation (Osa1) Release 7

Showing 21 bp from Chr5, positions 5,425,307 to 5,425,327

Instructions

**Searching:** Search using a sequence name, gene name, locus, or other landmark. The wildcard character * is allowed.
**Navigation:** Click one of the rulers to center on a location, or click and drag to select a region. Use the Scroll/Zoom buttons to change magnification and position.

**Examples**: Chr1:1..20000, Chr5:1464000..1480000, LOC_Os08g40150, LOC_Os02g06840.2.

[Bookmark this] [Upload your own data] [Hide banner] [Share these tracks] [Link to Image] [Help] [Reset]

Search

**Landmark or Region**: Chr5:5425317..5425317

**Data Source**
Rice Genome Annotation Project Rice Genome Browser - Release 7

**Scroll/Zoom:** Show 21 bp ☐ Flip

Rice genome database

**Analyses:**

Look for genes close to the most significant SNP using the rice genome browser (http://rice.uga.edu/cgi-bin/gbrowse/rice/). Pick a significant SNP from your analysis and enter its chromosome and position in the search box. The browser wants you to enter a start and stop position, so for example, you should enter "Chr3:30449857..30449857" and then choose "show 20kb" from the pulldown menu on the right hand side. 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 in your answer**

*Hint: You can include an image in your knitted Rmd file with*_ on its own line, where `MyImage.jpg` is the path to your image._

The three closest genes are:

- LOC_Os05g09550. This is a “Der-1 like family domain containing protein”. The SNP is actually in this gene. This gene encodes for a protein with a transmembrane domain. DER1 proteins, at least in yeast, are involved in degradation of proteins in the ER. possibly this could affect secretion of a flowering time signal.
- LOC_Os05g09540. This gene is unannotated and has no homology, so hard to say…
- LOC_Os05g09530. This is a protease, so it could be involved in signaling related to flowering.

# Your task

Below is a list of all of the phenotypes for rice from this dataset. This *Guide* uses “Seed.length.width.ratio” as the example. *Your task* is to choose another phenotype, replace it everywhere that is appropriate, and *knit* the code. Make your changes in the file “RiceGWASDataAnalysis.Rmd” markdown file. Write about GWAS and your results. As described above, look for genes close to the most significant SNP using the rice genome browser. Pick a significant SNP from your analysis and enter its chromosome and position in the search box. The browser wants you to enter a start and stop position, so for example, you should enter “Chr3:30449857..30449857” and then choose “show 20kb” from the pulldown menu on the right hand side. 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 in your answer

Hint: You can include an image in your knitted Rmd file with “MyImage.jpg” on its own line, where “MyImage.jpg” is the path to your image.

```
##  [1] "Country_of_Origin"               "Region"
##  [3] "Alu.Tol"                         "Flowering.time.at.Arkansas"
##  [5] "Flowering.time.at.Faridpur"      "Flowering.time.at.Aberdeen"
##  [7] "FT.ratio.of.Arkansas.Aberdeen"   "FT.ratio.of.Faridpur.Aberdeen"
##  [9] "Culm.habit"                      "Leaf.pubescence"
## [11] "Flag.leaf.length"                "Flag.leaf.width"
## [13] "Awn.presence"                    "Panicle.number.per.plant"
## [15] "Plant.height"                    "Panicle.length"
## [17] "Primary.panicle.branch.number"   "Seed.number.per.panicle"
## [19] "Florets.per.panicle"             "Panicle.fertility"
## [21] "Seed.length"                     "Seed.width"
## [23] "Seed.volume"                     "Seed.surface.area"
## [25] "Brown.rice.seed.length"          "Brown.rice.seed.width"
## [27] "Brown.rice.surface.area"         "Brown.rice.volume"
## [29] "Seed.length.width.ratio"         "Brown.rice.length.width.ratio"
## [31] "Seed.color"                      "Pericarp.color"
## [33] "Straighthead.suseptability"      "Blast.resistance"
## [35] "Amylose.content"                 "Alkali.spreading.value"
## [37] "Protein.content"
```

For this project, use the file *RiceGWASDataAnalysis.Rmd*. You should not have to modify code other than changing the phenotype as appropriate, but you are expected to do some writing and other analyses as appropriate.

*PROGRAMMING NOTE! It is STRONGLY recommended that you put all files in a folder on your DESKTOP called "RiceGWAS". This will help to ensure that all of the files, especially those downloaded from the Internet, are in the correct location.*