Assignment2

Ashenafi Beyi

3/19/2021

STEP 1. Importa data sets

Import genotypes and snp\_positions data from repository

fang\_genotypes <- read.table("https://raw.githubusercontent.com/EEOB-BioData/BCB546-Spring2021/master/assignments/UNIX\_Assignment/fang\_et\_al\_genotypes.txt",  
 header=T, sep="\t", stringsAsFactors =F)  
dim(fang\_genotypes)

## [1] 2782 986

snp\_position <- read.table("https://raw.githubusercontent.com/EEOB-BioData/BCB546-Spring2021/master/assignments/UNIX\_Assignment/snp\_position.txt",  
 header=T, sep="\t", stringsAsFactors =F)  
dim(snp\_position)

## [1] 983 15

head(snp\_position)

## SNP\_ID cdv\_marker\_id Chromosome Position alt\_pos mult\_positions amplicon  
## 1 abph1.20 5976 2 27403404 abph1  
## 2 abph1.22 5978 2 27403892 abph1  
## 3 ae1.3 6605 5 167889790 ae1  
## 4 ae1.4 6606 5 167889682 ae1  
## 5 ae1.5 6607 5 167889821 ae1  
## 6 an1.4 5982 1 240498509 an1  
## cdv\_map\_feature.name gene candidate.random Genaissance\_daa\_id  
## 1 AB042260 abph1 candidate 8393  
## 2 AB042260 abph1 candidate 8394  
## 3 ae1 ae1 candidate 8395  
## 4 ae1 ae1 candidate 8396  
## 5 ae1 ae1 candidate 8397  
## 6 an1 an1 candidate 8398  
## Sequenom\_daa\_id count\_amplicons count\_cmf count\_gene  
## 1 10474 1 1 1  
## 2 10475 0 0 0  
## 3 10477 1 1 1  
## 4 10478 0 0 0  
## 5 10479 0 0 0  
## 6 10481 1 1 1

Summary > fang\_genotypes has 2782 observations/rows and 986 variables/columns > snp\_position has 983 observations or rows and 15 variables or columns

STEP 2. Data processing

install.packages(“dplyr”)

library("tidyverse")

## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3 v purrr 0.3.4  
## v tibble 3.0.6 v dplyr 1.0.4  
## v tidyr 1.1.2 v stringr 1.4.0  
## v readr 1.4.0 v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(dplyr)  
snp\_position.selected <- select(snp\_position, c("SNP\_ID", "Chromosome", "Position"))  
str(snp\_position.selected) # 983 rows and 3 columns

## 'data.frame': 983 obs. of 3 variables:  
## $ SNP\_ID : chr "abph1.20" "abph1.22" "ae1.3" "ae1.4" ...  
## $ Chromosome: chr "2" "2" "5" "5" ...  
## $ Position : chr "27403404" "27403892" "167889790" "167889682" ...

glimpse(snp\_position.selected) # similar function with str

## Rows: 983  
## Columns: 3  
## $ SNP\_ID <chr> "abph1.20", "abph1.22", "ae1.3", "ae1.4", "ae1.5", "an1.4",~  
## $ Chromosome <chr> "2", "2", "5", "5", "5", "1", "3", "3", "4", "4", "4", "4",~  
## $ Position <chr> "27403404", "27403892", "167889790", "167889682", "16788982~

Number of rows and columns

nrow(fang\_genotypes) # 2782

## [1] 2782

ncol(fang\_genotypes) # 986

## [1] 986

NROW(na.omit(fang\_genotypes)) # 2782

## [1] 2782

NCOL(na.omit(fang\_genotypes)) # 986

## [1] 986

Genotypes in the column Group, summary and table functions can be used alike.

levels(as.factor(fang\_genotypes$Group)) # 15 groups: "TRIPS" "ZDIPL" "ZLUXR" "ZMHUE" "ZMMIL" "ZMMLR"

## [1] "TRIPS" "ZDIPL" "ZLUXR" "ZMHUE" "ZMMIL" "ZMMLR" "ZMMMR" "ZMPBA" "ZMPIL"  
## [10] "ZMPJA" "ZMXCH" "ZMXCP" "ZMXIL" "ZMXNO" "ZMXNT" "ZPERR"

# "ZMMMR" "ZMPBA" "ZMPIL" "ZMPJA" "ZMXCH" "ZMXCP" "ZMXIL"   
 # "ZMXNO" "ZMXNT" "ZPERR"  
summary(as.factor(fang\_genotypes$Group))

## TRIPS ZDIPL ZLUXR ZMHUE ZMMIL ZMMLR ZMMMR ZMPBA ZMPIL ZMPJA ZMXCH ZMXCP ZMXIL   
## 22 15 17 10 290 1256 27 900 41 34 75 69 6   
## ZMXNO ZMXNT ZPERR   
## 7 4 9

table(as.factor(fang\_genotypes$Group))

##   
## TRIPS ZDIPL ZLUXR ZMHUE ZMMIL ZMMLR ZMMMR ZMPBA ZMPIL ZMPJA ZMXCH ZMXCP ZMXIL   
## 22 15 17 10 290 1256 27 900 41 34 75 69 6   
## ZMXNO ZMXNT ZPERR   
## 7 4 9

Number of chromosomes in snp\_position data

summary(as.factor(snp\_position$Chromosome))

## 1 10 2 3 4 5 6 7   
## 155 53 127 107 91 122 76 97   
## 8 9 multiple unknown   
## 62 60 6 27

Subset data 3 maize genotypes > transpose > merge with snp data

maize\_genotypes <- filter(fang\_genotypes, Group == 'ZMMIL' | Group == 'ZMMLR' | Group == 'ZMMMR')  
dim(maize\_genotypes) # 1573 rows/observations 986 columns/variables

## [1] 1573 986

levels(as.factor(maize\_genotypes$Group)) # "ZMMIL" "ZMMLR" "ZMMMR"

## [1] "ZMMIL" "ZMMLR" "ZMMMR"

summary(as.factor(maize\_genotypes$Group)) # ZMMIL = 290, ZMMLR = 1256, ZMMMR = 27

## ZMMIL ZMMLR ZMMMR   
## 290 1256 27

#Transpose

resource for reading: <https://tibble.tidyverse.org/reference/rownames.html>

library(tidyverse)  
maize\_genotypes <- column\_to\_rownames(maize\_genotypes, var = "Sample\_ID")

maize\_genotypes.tr <- t(maize\_genotypes)%>%as.data.frame()%>%rownames\_to\_column(., var = "SNP\_ID")  
maize\_genotypes.tr <- maize\_genotypes.tr[3:nrow(maize\_genotypes.tr),]  
  
maize\_snp <- merge(snp\_position.selected, maize\_genotypes.tr, by="SNP\_ID")  
maize\_snp <- select(maize\_snp, SNP\_ID, Chromosome, Position, everything())  
dim(maize\_snp) # 983 528

## [1] 983 1576

table(as.factor(maize\_snp$Chromosome))

##   
## 1 10 2 3 4 5 6 7   
## 155 53 127 107 91 122 76 97   
## 8 9 multiple unknown   
## 62 60 6 27

Check for missed chromosome

sum(maize\_snp$Chromosome == "") # 0

## [1] 0

Subset data 3 teosinte genotypes > transpose > merge with snp data

teosinte\_genotypes <- filter(fang\_genotypes, Group == 'ZMPBA' | Group == 'ZMPIL' | Group == 'ZMPJA')  
dim(teosinte\_genotypes) # 975 rows/observations and 986 columns/variables

## [1] 975 986

levels(as.factor(teosinte\_genotypes$Group)) # "ZMPBA" "ZMPIL" "ZMPJA"

## [1] "ZMPBA" "ZMPIL" "ZMPJA"

summary(as.factor(teosinte\_genotypes$Group)) # ZMPBA = 900, ZMPIL = 41, ZMPJA = 34

## ZMPBA ZMPIL ZMPJA   
## 900 41 34

Transpose

teosinte\_genotypes <- column\_to\_rownames(teosinte\_genotypes, var = "Sample\_ID")

teosinte\_genotypes.tr <- t(teosinte\_genotypes)%>%as.data.frame()%>%rownames\_to\_column(., var = "SNP\_ID")  
teosinte\_genotypes.tr <- teosinte\_genotypes.tr[3:nrow(teosinte\_genotypes.tr),]  
  
teosinte\_snp <- merge(snp\_position.selected, teosinte\_genotypes.tr, by="SNP\_ID")  
  
teosinte\_snp <- select(teosinte\_snp, SNP\_ID, Chromosome, Position, everything())  
  
dim(teosinte\_snp) # 983 528

## [1] 983 978

The merged dataset has 983 rows/observations and 528 columns/variables.

check missed values in chromosome

Check for missed values

sum(teosinte\_snp$Chromosome == "")

## [1] 0

STEP 3. Subset joined data set by ‘Chromosome’ Order them by ’position’and replace missed value by “?” or “-”

Subset data by chromosome, order in INCREASING by “position” and replace missed values in columns 4 to 978 by “?”

Maize genotypes

maize\_chromosome.incr1 <- subset(maize\_snp, Chromosome==1)%>%arrange(Position)%>%  
 mutate\_at(4:978, ~replace(., is.na(.), "?"))  
maize\_chromosome.incr2 <- subset(maize\_snp, Chromosome==2)%>%arrange(Position)%>%  
 mutate\_at(4:978, ~replace(., is.na(.), "?"))  
maize\_chromosome.incr3 <- subset(maize\_snp, Chromosome==3)%>%arrange(Position)%>%  
 mutate\_at(4:978, ~replace(., is.na(.), "?"))  
maize\_chromosome.incr4 <- subset(maize\_snp, Chromosome==4)%>%arrange(Position)%>%  
 mutate\_at(4:978, ~replace(., is.na(.), "?"))  
maize\_chromosome.incr5 <- subset(maize\_snp, Chromosome==5)%>%arrange(Position)%>%  
 mutate\_at(4:978, ~replace(., is.na(.), "?"))  
maize\_chromosome.incr6 <- subset(maize\_snp, Chromosome==6)%>%arrange(Position)%>%  
 mutate\_at(4:978, ~replace(., is.na(.), "?"))  
maize\_chromosome.incr7 <- subset(maize\_snp, Chromosome==7)%>%arrange(Position)%>%  
 mutate\_at(4:978, ~replace(., is.na(.), "?"))  
maize\_chromosome.incr8 <- subset(maize\_snp, Chromosome==8)%>%arrange(Position)%>%  
 mutate\_at(4:978, ~replace(., is.na(.), "?"))  
maize\_chromosome.incr9 <- subset(maize\_snp, Chromosome==9)%>%arrange(Position)%>%  
 mutate\_at(4:978, ~replace(., is.na(.), "?"))  
maize\_chromosome.incr10 <- subset(maize\_snp, Chromosome==10)%>%arrange(Position)%>%  
 mutate\_at(4:978, ~replace(., is.na(.), "?"))

Teosinte genotypes

teosinte\_chromosome.incr1 <- subset(teosinte\_snp, Chromosome==1)%>%arrange(Position) %>%   
 mutate\_at(4:978, ~replace(., is.na(.), "?"))  
teosinte\_chromosome.incr2 <- subset(teosinte\_snp, Chromosome==2)%>%arrange(Position) %>%   
 mutate\_at(4:978, ~replace(., is.na(.), "?"))  
teosinte\_chromosome.incr3 <- subset(teosinte\_snp, Chromosome==3)%>%arrange(Position) %>%   
 mutate\_at(4:978, ~replace(., is.na(.), "?"))  
teosinte\_chromosome.incr4 <- subset(teosinte\_snp, Chromosome==4)%>%arrange(Position) %>%   
 mutate\_at(4:978, ~replace(., is.na(.), "?"))  
teosinte\_chromosome.incr5 <- subset(teosinte\_snp, Chromosome==5)%>%arrange(Position) %>%   
 mutate\_at(4:978, ~replace(., is.na(.), "?"))  
teosinte\_chromosome.incr6 <- subset(teosinte\_snp, Chromosome==6)%>%arrange(Position) %>%   
 mutate\_at(4:978, ~replace(., is.na(.), "?"))  
teosinte\_chromosome.incr7 <- subset(teosinte\_snp, Chromosome==7)%>%arrange(Position) %>%   
 mutate\_at(4:978, ~replace(., is.na(.), "?"))  
teosinte\_chromosome.incr8 <- subset(teosinte\_snp, Chromosome==8)%>%arrange(Position) %>%   
 mutate\_at(4:978, ~replace(., is.na(.), "?"))  
teosinte\_chromosome.incr9 <- subset(teosinte\_snp, Chromosome==9)%>%arrange(Position) %>%   
 mutate\_at(4:978, ~replace(., is.na(.), "?"))  
teosinte\_chromosome.incr10 <- subset(teosinte\_snp, Chromosome==10)%>%arrange(Position) %>%   
 mutate\_at(4:978, ~replace(., is.na(.), "?"))

Subset data by chromosome, order in DECREASING by “position” and replace missed values in columns 4 to 978 by “-”

Maize genotypes

maize\_chromosome.dec1 <- subset(maize\_snp, Chromosome==1)%>%arrange(desc(Position))%>%  
 mutate\_at(4:978, ~replace(., is.na(.), "-"))  
maize\_chromosome.dec2 <- subset(maize\_snp, Chromosome==2)%>%arrange(desc(Position))%>%  
 mutate\_at(4:978, ~replace(., is.na(.), "-"))  
maize\_chromosome.dec3 <- subset(maize\_snp, Chromosome==3)%>%arrange(desc(Position))%>%  
 mutate\_at(4:978, ~replace(., is.na(.), "-"))  
maize\_chromosome.dec4 <- subset(maize\_snp, Chromosome==4)%>%arrange(desc(Position))%>%  
 mutate\_at(4:978, ~replace(., is.na(.), "-"))  
maize\_chromosome.dec5 <- subset(maize\_snp, Chromosome==5)%>%arrange(desc(Position))%>%  
 mutate\_at(4:978, ~replace(., is.na(.), "-"))  
maize\_chromosome.dec6 <- subset(maize\_snp, Chromosome==6)%>%arrange(desc(Position))%>%  
 mutate\_at(4:978, ~replace(., is.na(.), "-"))  
maize\_chromosome.dec7 <- subset(maize\_snp, Chromosome==7)%>%arrange(desc(Position))%>%  
 mutate\_at(4:978, ~replace(., is.na(.), "-"))  
maize\_chromosome.dec8 <- subset(maize\_snp, Chromosome==8)%>%arrange(desc(Position))%>%  
 mutate\_at(4:978, ~replace(., is.na(.), "-"))  
maize\_chromosome.dec9 <- subset(maize\_snp, Chromosome==9)%>%arrange(desc(Position))%>%  
 mutate\_at(4:978, ~replace(., is.na(.), "-"))  
maize\_chromosome.dec10 <- subset(maize\_snp, Chromosome==10)%>%arrange(desc(Position))%>%  
 mutate\_at(4:978, ~replace(., is.na(.), "-"))

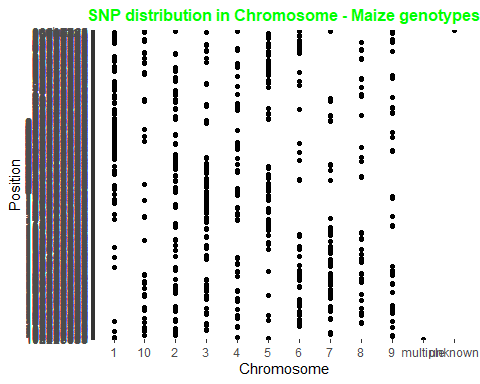
Teosinte genotypes

teosinte\_chromosome.dec1 <- subset(teosinte\_snp, Chromosome==1)%>%arrange(desc(Position))%>%  
 mutate\_at(4:978, ~replace(., is.na(.), "-"))  
teosinte\_chromosome.dec2 <- subset(teosinte\_snp, Chromosome==2)%>%arrange(desc(Position))%>%  
 mutate\_at(4:978, ~replace(., is.na(.), "-"))  
teosinte\_chromosome.dec3 <- subset(teosinte\_snp, Chromosome==3)%>%arrange(desc(Position))%>%  
 mutate\_at(4:978, ~replace(., is.na(.), "-"))  
teosinte\_chromosome.dec4 <- subset(teosinte\_snp, Chromosome==4)%>%arrange(desc(Position))%>%  
 mutate\_at(4:978, ~replace(., is.na(.), "-"))  
teosinte\_chromosome.dec5 <- subset(teosinte\_snp, Chromosome==5)%>%arrange(desc(Position))%>%  
 mutate\_at(4:978, ~replace(., is.na(.), "-"))  
teosinte\_chromosome.dec6 <- subset(teosinte\_snp, Chromosome==6)%>%arrange(desc(Position))%>%  
 mutate\_at(4:978, ~replace(., is.na(.), "-"))  
teosinte\_chromosome.dec7 <- subset(teosinte\_snp, Chromosome==7)%>%arrange(desc(Position))%>%  
 mutate\_at(4:978, ~replace(., is.na(.), "-"))  
teosinte\_chromosome.dec8 <- subset(teosinte\_snp, Chromosome==8)%>%arrange(desc(Position))%>%  
 mutate\_at(4:978, ~replace(., is.na(.), "-"))  
teosinte\_chromosome.dec9 <- subset(teosinte\_snp, Chromosome==9)%>%arrange(desc(Position))%>%  
 mutate\_at(4:978, ~replace(., is.na(.), "-"))  
teosinte\_chromosome.dec10 <- subset(teosinte\_snp, Chromosome==10)%>%arrange(desc(Position))%>%  
 mutate\_at(4:978, ~replace(., is.na(.), "-"))

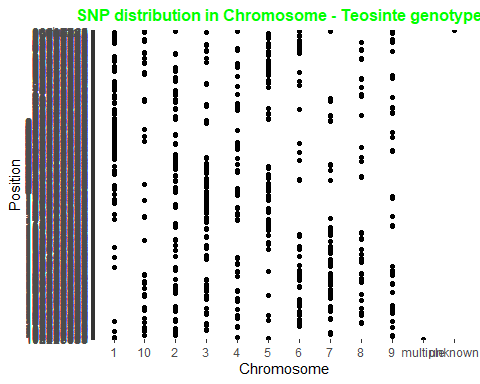
Step 4. Data vizualization

1. Bar charts for SNPs per chromosome in maize and teosinte genotypes

library(ggplot2)  
maize\_snp\_bar <- ggplot(data = maize\_snp) + geom\_point(mapping=aes(x=Chromosome, y=Position)) +  
 ggtitle("SNP distribution in Chromosome - Maize genotypes")  
maize\_snp\_bar + theme(plot.title = element\_text(color = "green", size = 12, face = "bold", hjust = 0.5))



teosinte\_snp\_bar <- ggplot(data = teosinte\_snp) + geom\_point(mapping=aes(x=Chromosome, y=Position)) +  
 ggtitle("SNP distribution in Chromosome - Teosinte genotypes")  
teosinte\_snp\_bar + theme(plot.title = element\_text(color = "green", size = 12, face = "bold", hjust = 0.5))



1. Bar charts showing homozygous and heterozygous distribution in selected genes Four catageroies: homozygous, hetrozygous, missed, and NA (e.g. A/C, G/T)

Gene1 = ZDP\_0752a

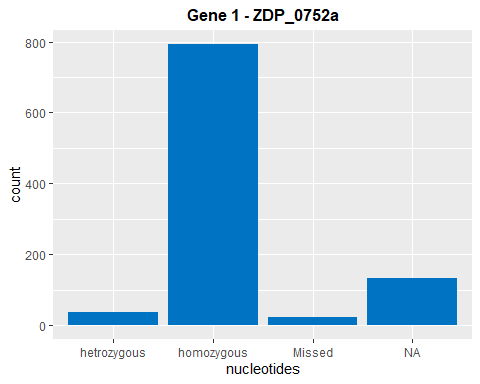
library(tidyr)  
library(tidyverse)

Gene1 = ZDP\_0752a

ZDP\_0752a <- transmute(maize\_snp, nucleotides =case\_when(   
 ZDP\_0752a %in% c('A/A', 'C/C', 'G/G', 'T/T')~"homozygous",  
 ZDP\_0752a %in% c('A/T', 'C/G', 'G/C', 'T/A') ~"hetrozygous",  
 ZDP\_0752a %in% '?/?' ~ "Missed"))  
table(ZDP\_0752a)

## ZDP\_0752a  
## hetrozygous homozygous Missed   
## 36 793 22

# hetrozygous homozygous Missed   
# 36 793 22   
Gene1 <- ggplot(ZDP\_0752a, aes(nucleotides)) + geom\_bar(fill="#0073C2FF") +  
 ggtitle ("Gene 1 - ZDP\_0752a")  
Gene1 + theme(plot.title = element\_text(color = "black", size = 12, face = "bold", hjust = 0.5))

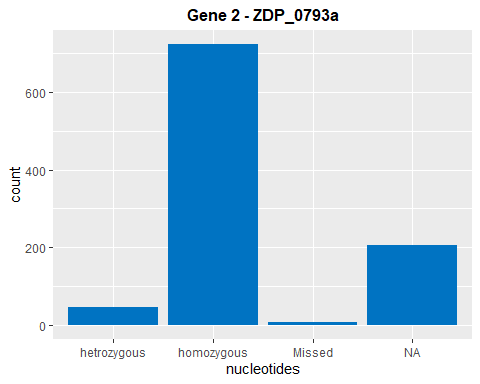


Gene2 = ZDP\_0793a

ZDP\_0793a <- transmute(maize\_snp, nucleotides =case\_when(   
 ZDP\_0793a %in% c('A/A', 'C/C', 'G/G', 'T/T')~"homozygous",  
 ZDP\_0793a %in% c('A/T', 'C/G', 'G/C', 'T/A') ~"hetrozygous",  
 ZDP\_0752a %in% '?/?' ~ "Missed"))  
table(ZDP\_0793a)

## ZDP\_0793a  
## hetrozygous homozygous Missed   
## 46 723 7

# hetrozygous homozygous Missed   
# 46 723 7   
  
Gene2 <- ggplot(ZDP\_0793a, aes(nucleotides)) + geom\_bar(fill="#0073C2FF") +   
 ggtitle("Gene 2 - ZDP\_0793a")  
Gene2 + theme(plot.title = element\_text(color = "black", size = 12, face = "bold", hjust = 0.5))



1. bar charts of chromosome counts and SNP distributions in the entire fang\_genotypes file

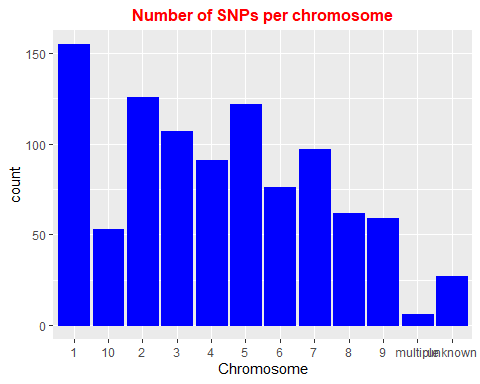
fang\_genotypes1 <- column\_to\_rownames(fang\_genotypes, var = "Sample\_ID")  
fang\_genotypes.tr <- t(fang\_genotypes1)%>%as.data.frame()%>%rownames\_to\_column(., var = "SNP\_ID")  
fang\_genotypes.tr <- fang\_genotypes.tr[3:nrow(maize\_genotypes.tr),]  
  
joined\_snp\_genotypes <- merge(snp\_position, fang\_genotypes.tr, by= "SNP\_ID")

Bar chart of Chromosome frequencies

chromosome\_count <- ggplot(joined\_snp\_genotypes, aes(x= Chromosome) ) +   
 geom\_histogram(stat= "Count", color = "blue", fill = "blue") + ggtitle ("Number of SNPs per chromosome")

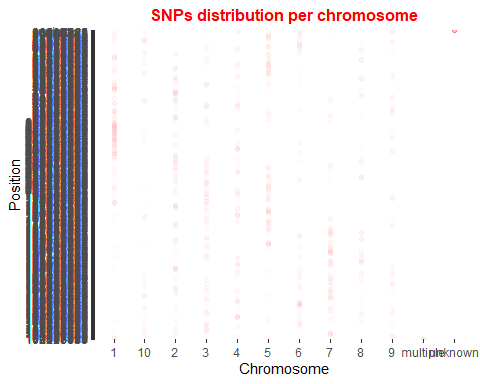
## Warning: Ignoring unknown parameters: binwidth, bins, pad

chromosome\_count + theme(plot.title = element\_text(color = "red", size = 12, face = "bold", hjust = 0.5))



Bar chart of SNP distribution across chromosomes

chromosome\_snp <- ggplot(joined\_snp\_genotypes, aes(x= Chromosome, y= Position))+   
 geom\_point(stat=, color = "red", alpha= 0.01)+ ggtitle ("SNPs distribution per chromosome")  
  
chromosome\_snp + theme(plot.title = element\_text(color = "red", size = 12, face = "bold", hjust = 0.5))



Summary of the work flows and what has been done

1. Two data sets were imported from the course repository
2. Data processing was conducted: i. Data size and structure were explored. ii. Three target variables selected in SNP\_position data set. iii. Target maize and teosinte genotypes were filtered, and transposed and merged with SNP\_position data set.
3. The merged SNP position and genotypes were subset by chromosome, order by “Position”, and missed values replaced by “?” or “-” for maize and teosinte genotypes separately.
4. Data vizualization
   1. Chromosome counts and SNP distribution in maize and teosinte genotypes files
   2. Two genes were selected for data vizualization exercise
      1. The selected gene was catagorized into “homozygous”, “hetrozygous” and “missed”
      2. Bar chart showing nucleotide distribution was plotted.
   3. The entire genotype data set was merged with snp position data. Bar charts depcting chromosome ferquencies and snp distribution across the chromsomes were plotted.