BCB 546 Homework 2: R Assignment

Weixia Deng

Part I

Data Inspection

Genotype Data

```
library(tidyverse)
file.info("fang_et_al_genotypes.txt")$size
# Read fang_et_al_genotypes file
fang <- read.table("fang_et_al_genotypes.txt", sep = "\t", header = TRUE)
dim(fang)
```

```
## [1] 11051939
## [1] 2782  986
```

The `fang_et_al_genotypes.txt` file size is 11051939 bytes and has 986 columns and 2782 observations.

```
fang$Group <- as.factor(fang$Group)
table(fang$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
```

The frequency count of each group is shown above.

```
str(fang[,1:15])
```

```
## 'data.frame':    2782 obs. of  15 variables:
##  $ Sample_ID: chr  "SL-15" "SL-16" "SL-11" "SL-12" ...
##  $ JG_OTU   : chr  "T-aust-1" "T-aust-2" "T-brav-1" "T-brav-2" ...
##  $ Group    : Factor w/ 16 levels "TRIPS","ZDIPL",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ abph1.20 : chr  "?/?" "?/?" "?/?" "?/?" ...
##  $ abph1.22 : chr  "?/?" "?/?" "?/?" "?/?" ...
##  $ ae1.3    : chr  "T/T" "T/T" "T/T" "T/T" ...
##  $ ae1.4    : chr  "G/G" "?/?" "G/G" "G/G" ...
##  $ ae1.5    : chr  "T/T" "T/T" "T/T" "T/T" ...
##  $ an1.4    : chr  "C/C" "C/C" "?/?" "C/C" ...
##  $ ba1.6    : chr  "?/?" "A/G" "G/G" "G/G" ...
##  $ ba1.9    : chr  "G/G" "G/G" "G/G" "G/G" ...
##  $ bt2.5    : chr  "?/?" "?/?" "C/C" "C/C" ...
##  $ bt2.7    : chr  "A/A" "A/A" "A/A" "A/A" ...
##  $ bt2.8    : chr  "?/?" "?/?" "?/?" "?/?" ...
##  $ Fea2.1   : chr  "C/C" "C/C" "?/?" "?/?" ...
```

Partial data structure of genotype data is shown above.

SNP Data

```
file.info("snp_position.txt")$size
# Read SNP file
snp <- read.table("snp_position.txt", sep = "\t", header = TRUE)
dim(snp)
```

```
## [1] 82763
## [1] 983  15
```

The `snp_position.txt` file size is 82763 and has 15 columns and 983 observations.

```
str(snp)
```

```
## 'data.frame':    983 obs. of  15 variables:
##  $ SNP_ID              : chr  "abph1.20" "abph1.22" "ae1.3" "ae1.4" ...
##  $ cdv_marker_id       : int  5976 5978 6605 6606 6607 5982 3463 3466 5983 5985 ...
##  $ Chromosome          : chr  "2" "2" "5" "5" ...
##  $ Position            : chr  "27403404" "27403892" "167889790" "167889682" ...
##  $ alt_pos             : chr  "" "" "" "" ...
##  $ mult_positions      : chr  "" "" "" "" ...
##  $ amplicon            : chr  "abph1" "abph1" "ae1" "ae1" ...
##  $ cdv_map_feature.name: chr  "AB042260" "AB042260" "ae1" "ae1" ...
##  $ gene                : chr  "abph1" "abph1" "ae1" "ae1" ...
##  $ candidate.random    : chr  "candidate" "candidate" "candidate" "candidate" ...
##  $ Genaissance_daa_id  : int  8393 8394 8395 8396 8397 8398 8399 8400 8401 8402 ...
##  $ Sequenom_daa_id     : int  10474 10475 10477 10478 10479 10481 10482 10483 10486 10487 ...
##  $ count_amplicons     : int  1 0 1 0 0 1 1 0 1 0 ...
##  $ count_cmf           : int  1 0 1 0 0 1 0 0 1 0 ...
##  $ count_gene          : int  1 0 1 0 0 1 1 0 1 0 ...
```

The data structure of SNP data is shown above.

Data Processing

```
# Read fang_et_al_genotypes file
fang <- read.table("fang_et_al_genotypes.txt", sep = "\t", header = TRUE)
# Group of maize and teosinte
maize <- c("ZMMIL", "ZMMLR", "ZMMMR")
```

```r
teosinte <- c("ZMPBA", "ZMPIL", "ZMPJA")
# Find index of maize and teosinte from third column of fang_et_al_genotypes
maize.idx <- fang$Group %in% maize
teosinte.idx <- fang$Group %in% teosinte
# Subset maize and teosinte from fang_et_al_genotypes data
# Only keep genotypes data for each group
# (drop Sample_ID, JG_OTU and Group columns)
fang.maize <- fang[maize.idx, -c(1:3)]
fang.teosinte <- fang[teosinte.idx, -c(1:3)]
# Add header of maize and teosinte (genotypes) to Subset of data
fang.maize <- rbind(colnames(fang.maize), fang.maize)
fang.teosinte <- rbind(colnames(fang.teosinte), fang.teosinte)
# Transpose two subsets of genotype data
# First column is the genotype
maize.t <- t(fang.maize) %>% as.data.frame()
teosinte.t <- t(fang.teosinte) %>% as.data.frame()

# Read snp_position file
snp <- read.table("snp_position.txt", sep = "\t", header = TRUE)
# Only keep SNP id (first column),
# chromosome location (third column),
# nucleotide location (fourth column)
snp.sub <- snp[,c(1,3,4)]
# Remove Position values are "", "multiple", "unknown"
position.idx <- snp.sub$Position %in% c("", "multiple", "unknown")
snp.sub <- snp.sub[!position.idx,]
# Set chromosome as factor, Position as numeric
snp.sub$Chromosome <- as.factor(snp.sub$Chromosome)
snp.sub$Position <- as.numeric(snp.sub$Position)

# Maize data
for (i in 1:10){
  # Subset SNP by Chromosome 1 to 10
  chromosome.idx <- snp.sub$Chromosome == i
  # Merge subset SNP data with maize genotype data by genotype
  df <- merge(x = snp.sub[chromosome.idx,], y = maize.t,
              by.x = "SNP_ID", by.y = "1")
  # Sort position by increasing order
  df.1 <- df[order(df$Position, decreasing = FALSE),]
  # Save df.1 to output folder
  n.1 <- paste("output/maize-increase-", i, ".txt", sep = "")
  write.table(df.1, file = n.1, sep = "\t")
  # Sort position by decreasing order
  df.2 <- df[order(df$Position, decreasing = TRUE),]
  # Replace missing data "?" to "-"
  df.2[,4:ncol(df.2)] <- lapply(df.2[,4:ncol(df.2)],
                                function(x) str_replace_all(x, "\\?", "-"))
  # Save df.2 to output folder
  n.2 <- paste("output/maize-decrease-", i, ".txt", sep = "")
  write.table(df.2, file = n.2, sep = "\t")
}

# Teosinte data
```

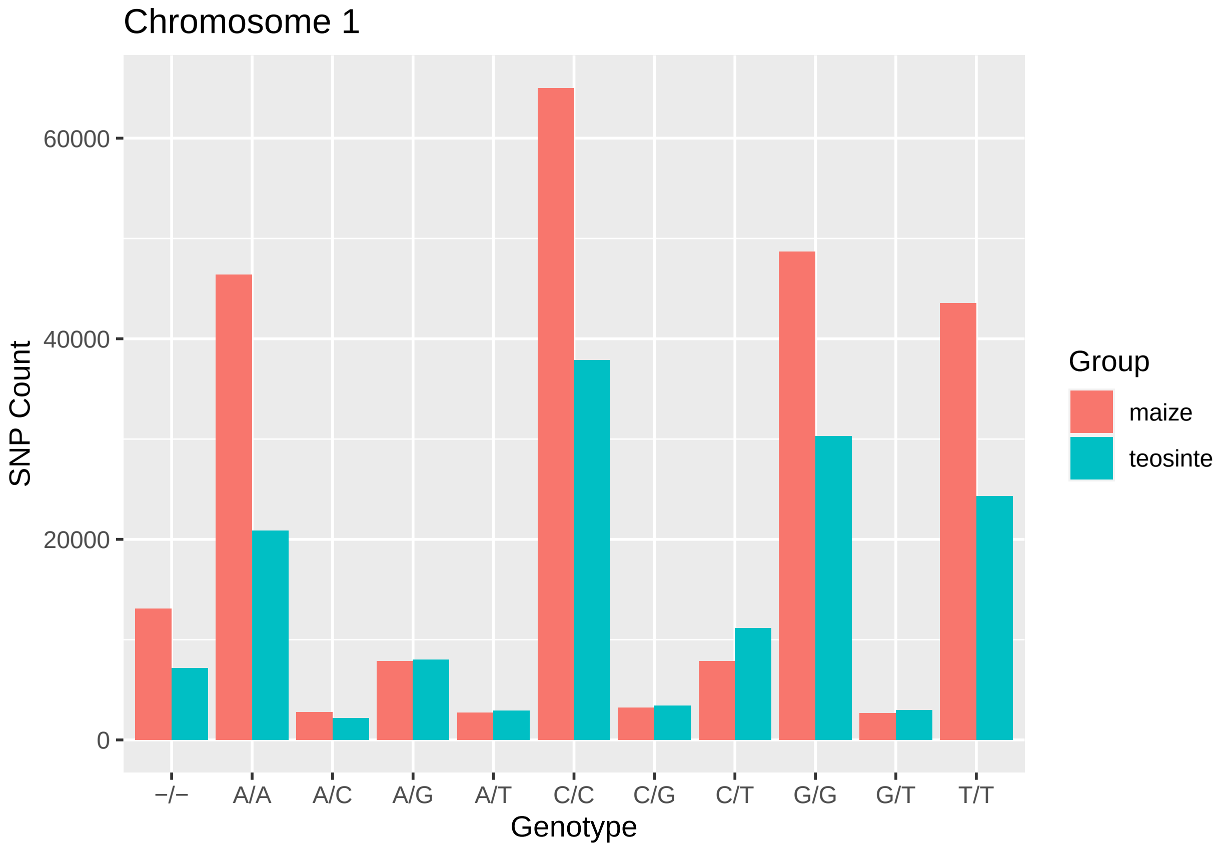
```r
for (i in 1:10){
  # Subset SNP by Chromosome 1 to 10
  chromosome.idx <- snp.sub$Chromosome == i
  # Merge subset SNP data with teosinte genotype data by genotype
  df <- merge(x = snp.sub[chromosome.idx,], y = teosinte.t,
              by.x = "SNP_ID", by.y = "1")
  # Sort position by increasing order
  df.1 <- df[order(df$Position, decreasing = FALSE),]
  # Save df.1 to output folder
  n.1 <- paste("output/teosinte-increase-", i, ".txt", sep = "")
  write.table(df.1, file = n.1, sep = "\t")
  # Sort position by decreasing order
  df.2 <- df[order(df$Position, decreasing = TRUE),]
  # Replace missing data "?" to "-"
  df.2[,4:ncol(df.2)] <- lapply(df.2[,4:ncol(df.2)],
                                function(x) str_replace_all(x, "\\?", "-"))
  # Save df.2 to output folder
  n.2 <- paste("output/teosinte-decrease-", i, ".txt", sep = "")
  write.table(df.2, file = n.2, sep = "\t")
}
```

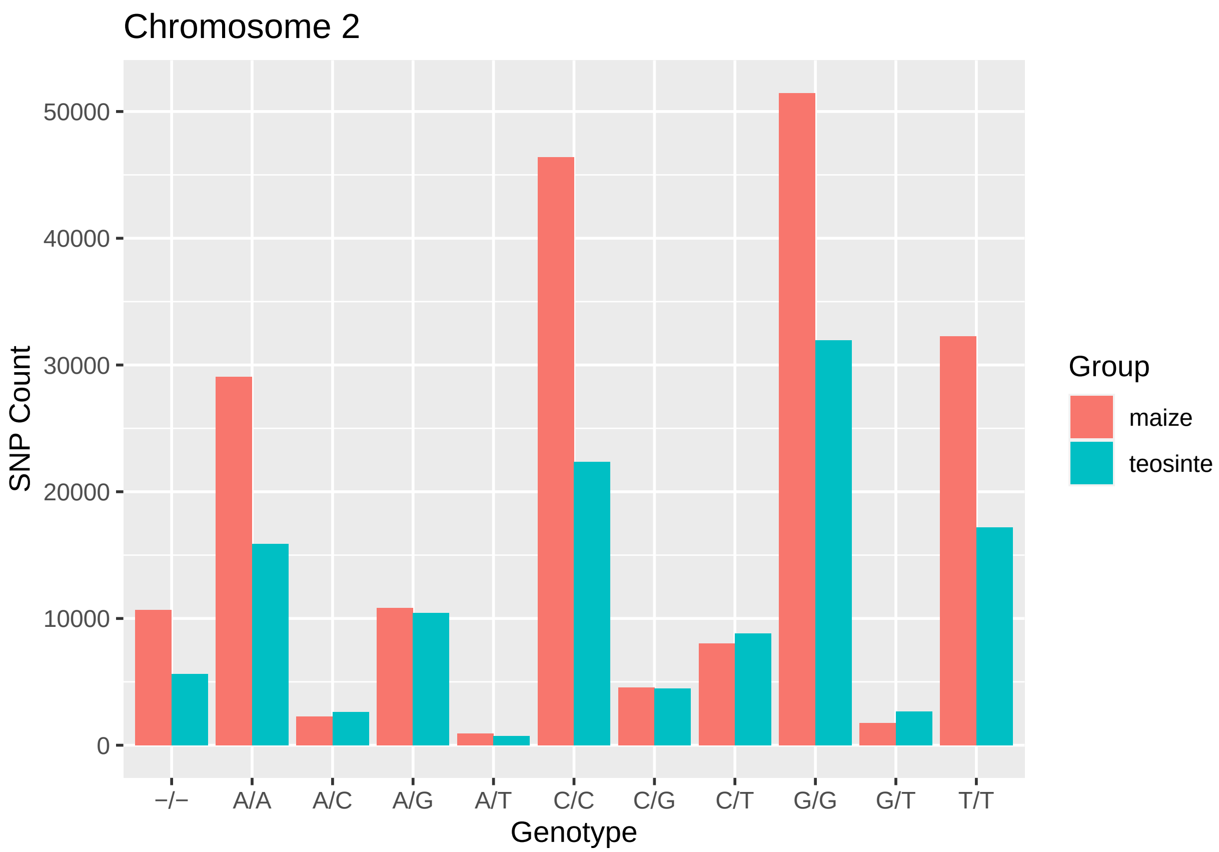
Part II

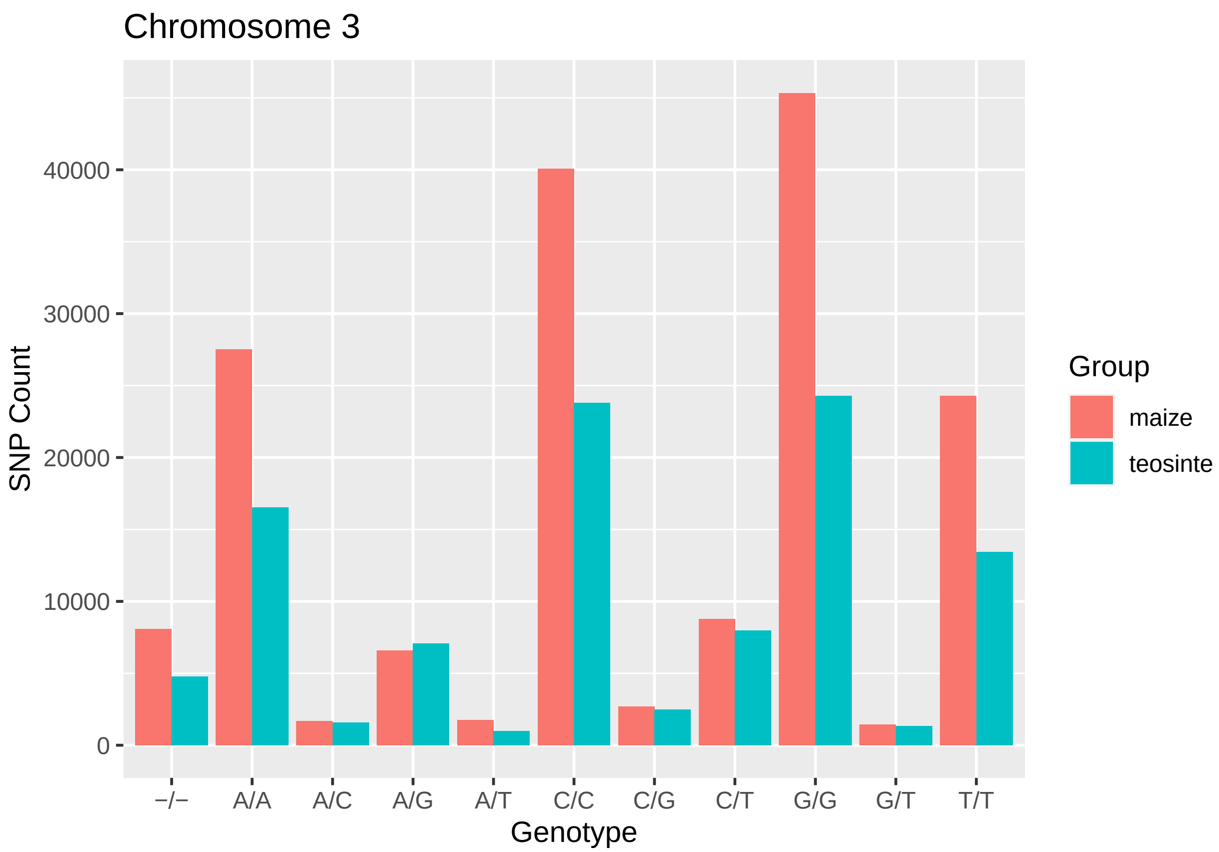
Data Visualization

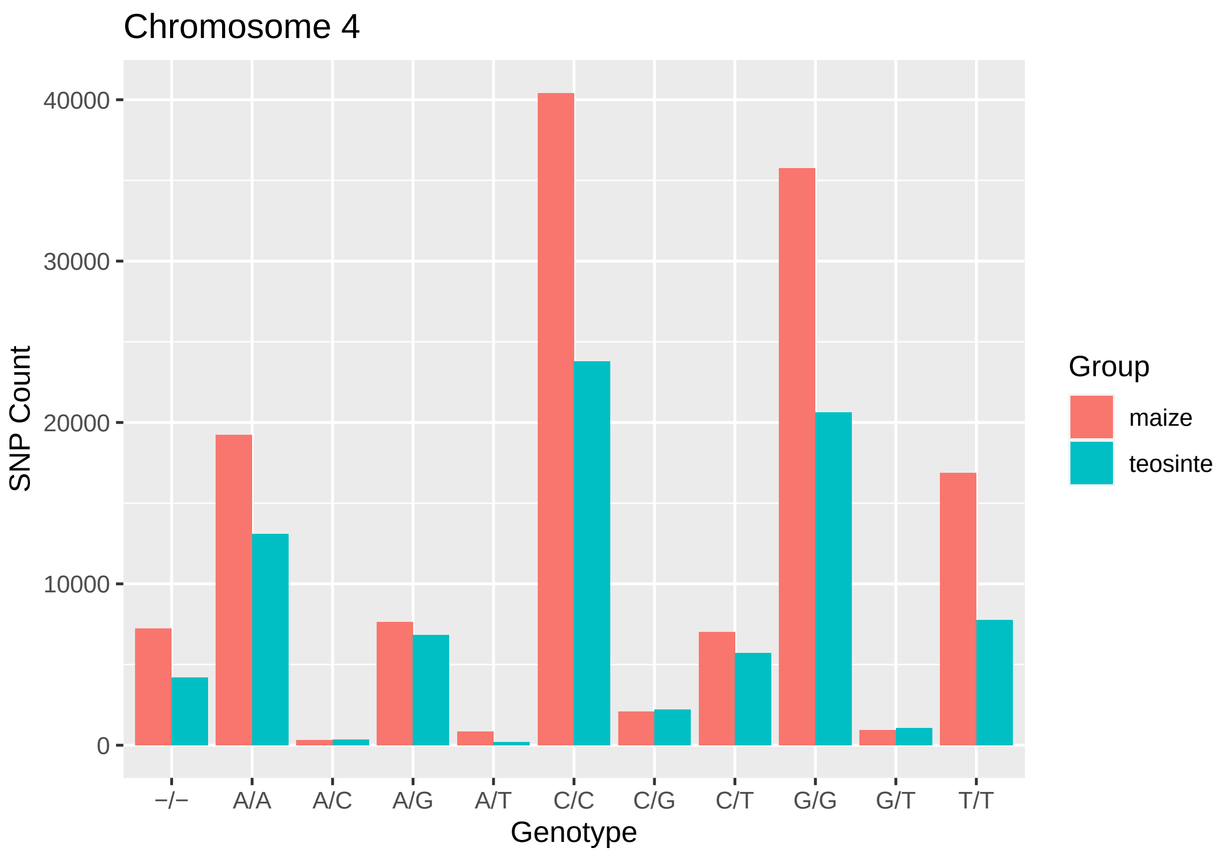
SNPs per chromosome (on and across chromosomes)

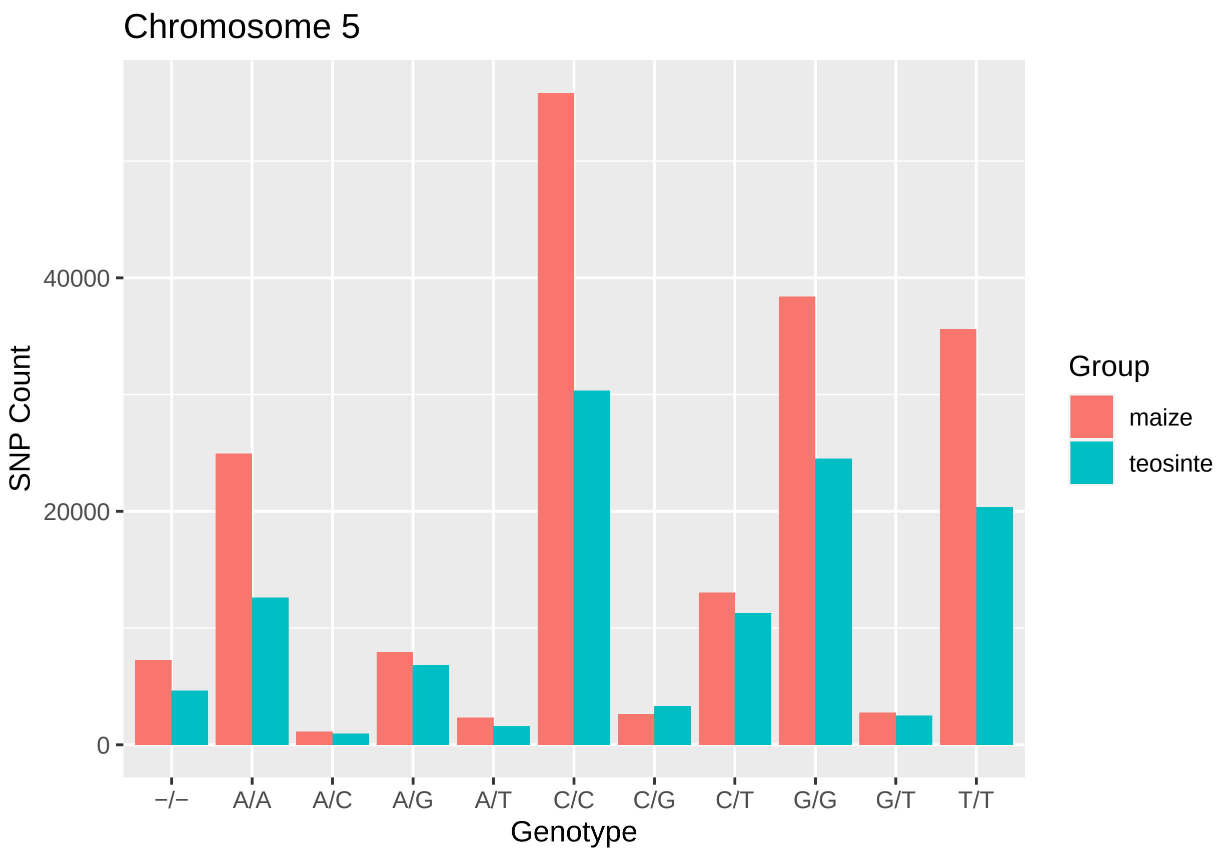
```r
library(tidyverse)

# Open 40 output files
filels <- list.files("output/")

dfls <- gsub(".txt", "", filels) %>% gsub("maize", "m",.) %>%
  gsub("teosinte", "t",.) %>% gsub("increase", "i",.) %>%
  gsub("decrease", "d",.) %>% gsub("-", "",.)

for (i in 1:length(filels)){
  n <- paste("output/", filels[i], sep = "")
  assign(dfls[i], read.table(n, sep = "\t", header = TRUE))
}

# Maize and Teosinte Visualization
## On each chromosome (1-10)
ls.maize <- list(md1, md2, md3, md4, md5, md6, md7, md8, md9, md10)
ls.teosinte <- list(td1, td2, td3, td4, td5, td6, td7, td8, td9, td10)

for (i in 1:10){
  # Loop through chromosome 1-10
  md <- ls.maize[[i]]
  td <- ls.teosinte[[i]]
  # Find the frequency count of genotype by groups maize and teosinte
  m.freq <- md[,4:ncol(md)] %>% unlist() %>% table() %>% as.data.frame()
  m.freq$Group <- "maize"
  t.freq <- td[,4:ncol(td)] %>% unlist() %>% table() %>% as.data.frame()
  t.freq$Group <- "teosinte"
  # Plot SNP of each chromosome by two groups
  freq <- rbind(m.freq, t.freq)
  freq$Group <- as.factor(freq$Group)
  n <- paste("Chromosome", i, sep = " ")
  print(ggplot(freq, aes(x = ., y = Freq, fill = Group)) +
    geom_bar(stat = "identity", position = position_dodge()) +
    labs(title = n, x = "Genotype", y = "SNP Count"))
}
```

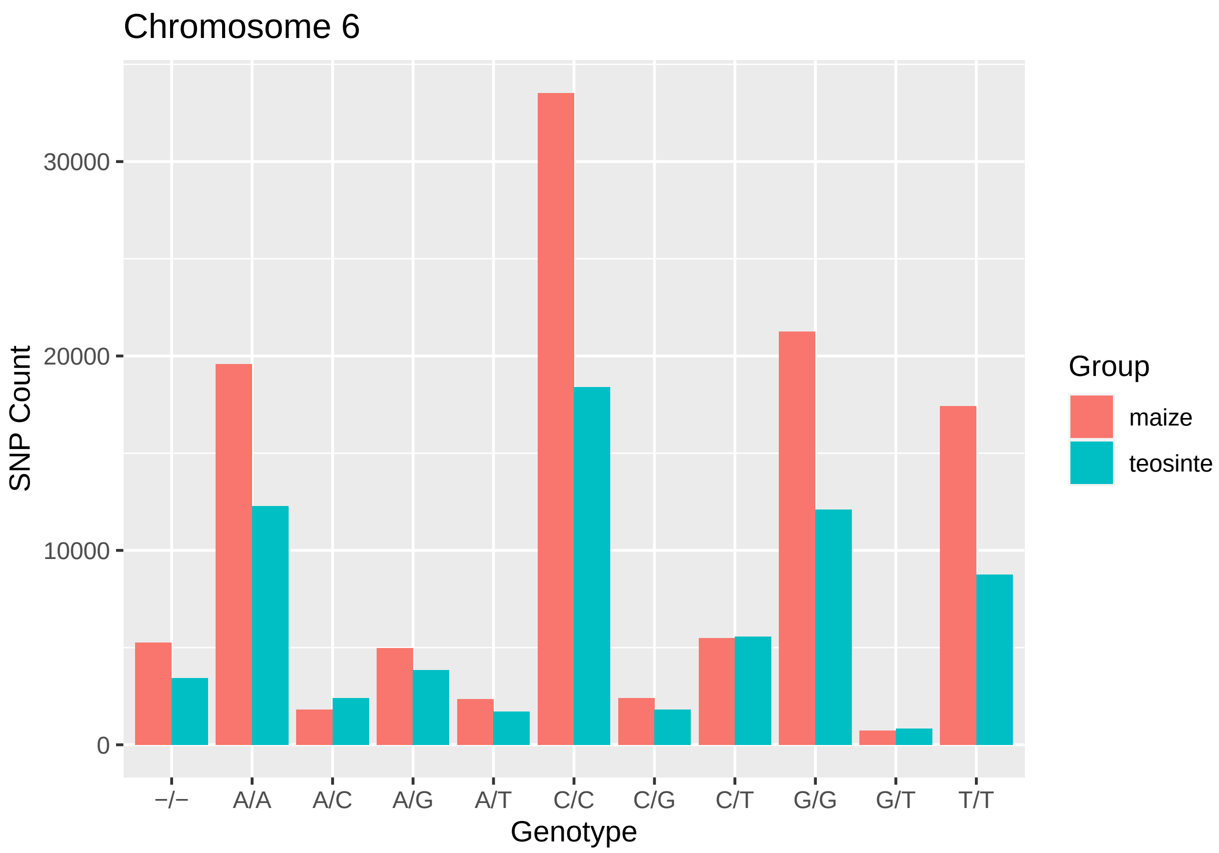


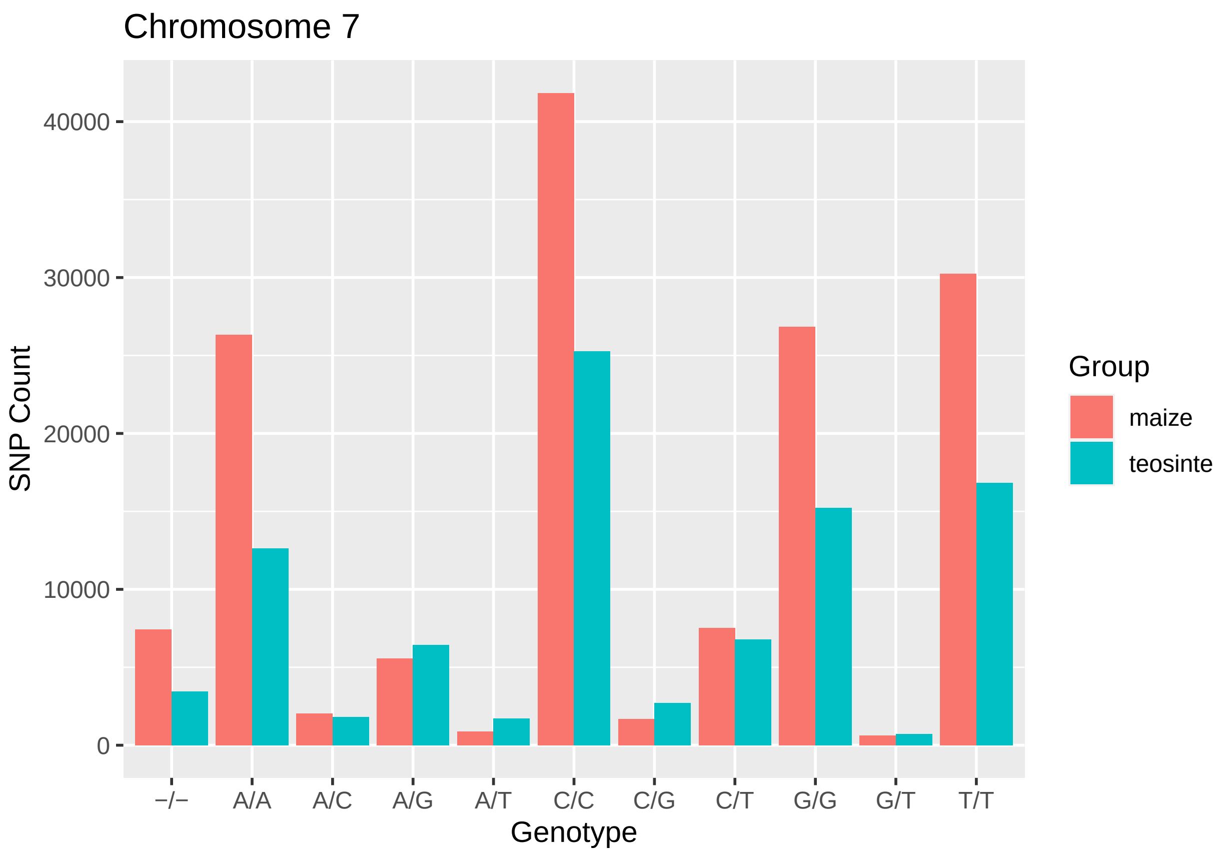

Chromosome 2
SNP Count
50000
40000
30000
20000
10000
0
−/−
A/A
A/C
A/G
A/T
C/C
C/G
C/T
G/G
G/T
T/T
Genotype
Group
maize
teosinte

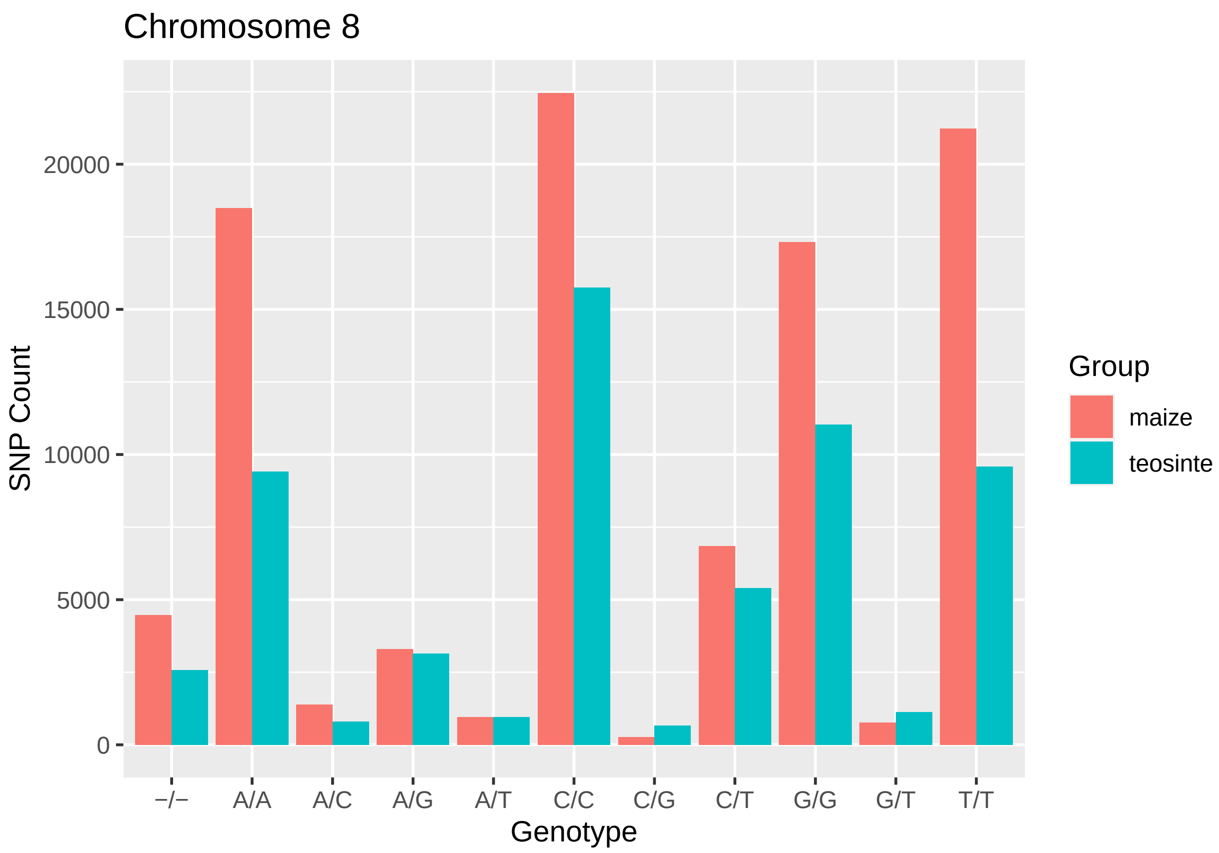

Chromosome 3
SNP Count
40000
30000
20000
10000
0
−/−
A/A
A/C
A/G
A/T
C/C
C/G
C/T
G/G
G/T
T/T
Genotype
Group
maize
teosinte

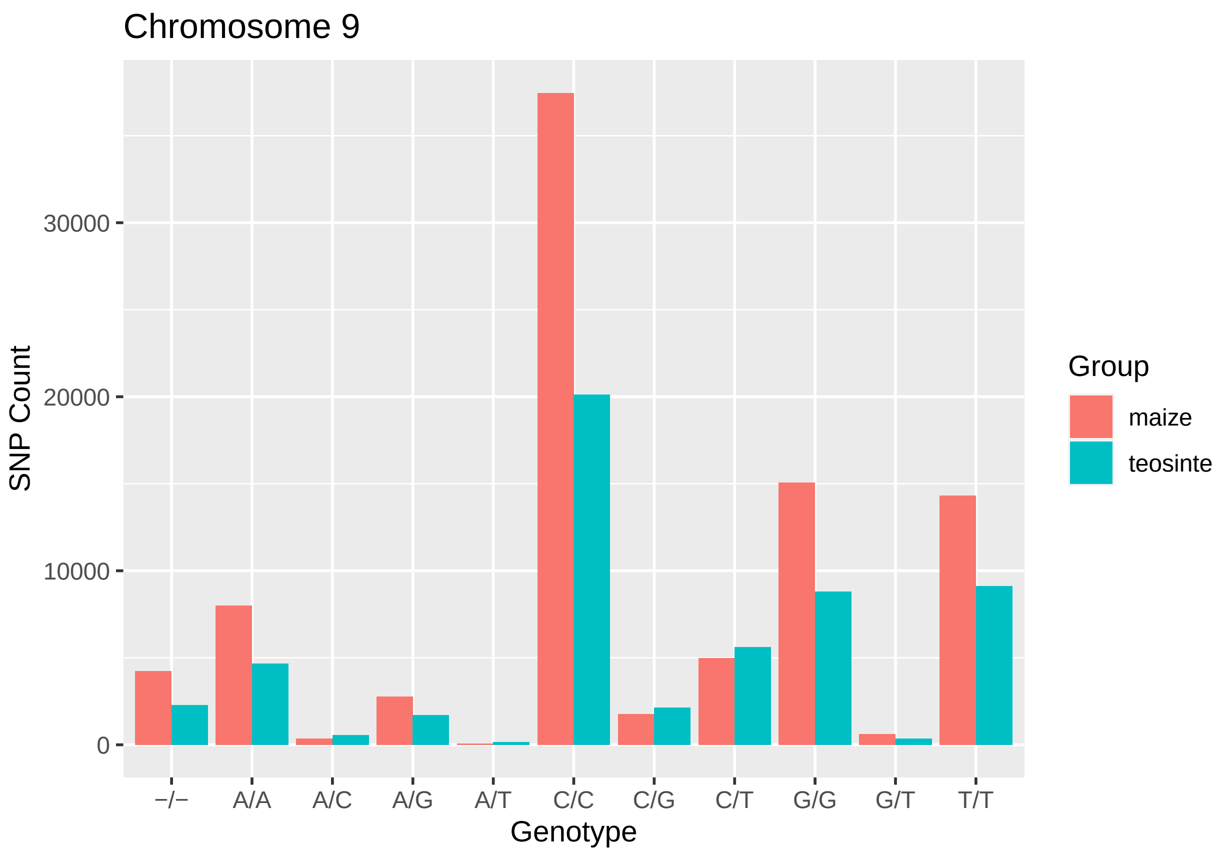

Chromosome 4
SNP Count
40000
30000
20000
10000
0
−/−
A/A
A/C
A/G
A/T
C/C
C/G
C/T
G/G
G/T
T/T
Genotype
Group
maize
teosinte

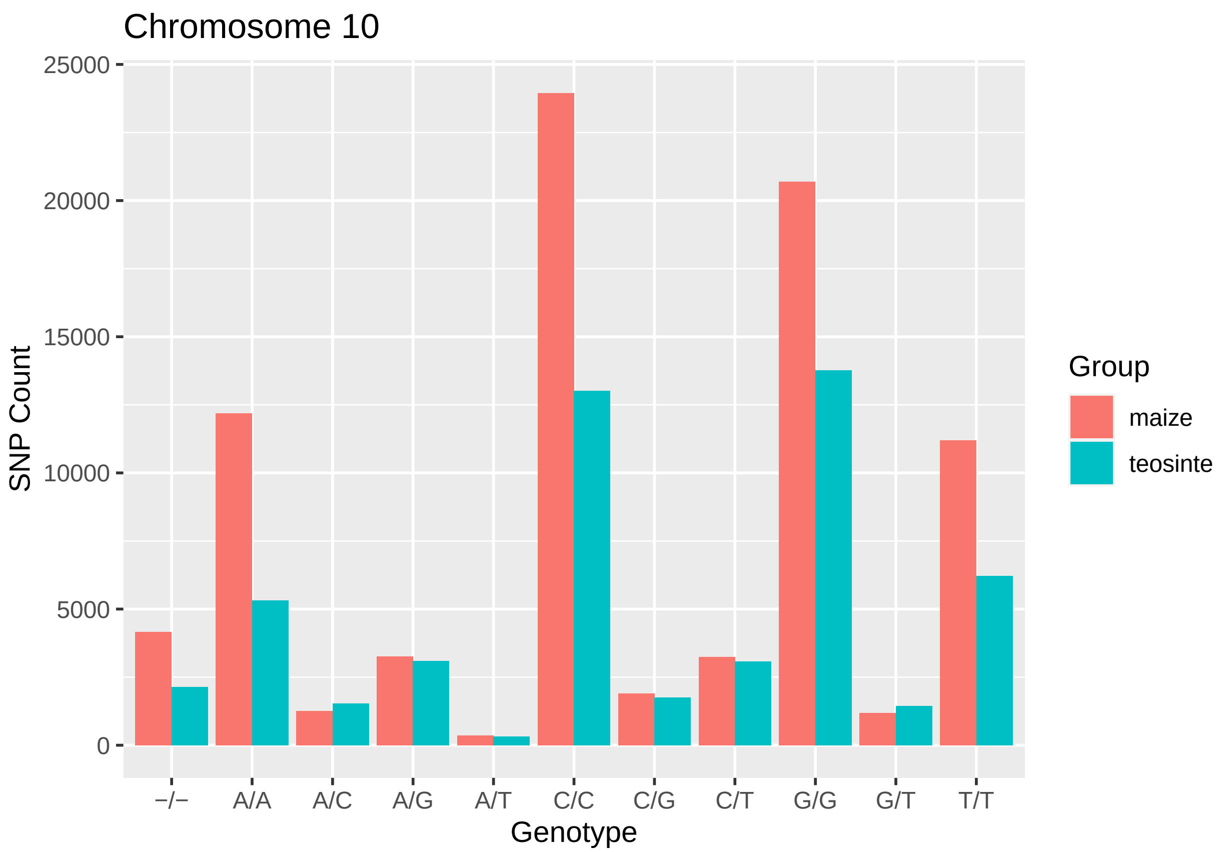

Chromosome 5
SNP Count
40000
20000
0
−/−
A/A
A/C
A/G
A/T
C/C
C/G
C/T
G/G
G/T
T/T
Genotype
Group
maize
teosinte


Chromosome 6
SNP Count
30000
20000
10000
0
−/−
A/A
A/C
A/G
A/T
C/C
C/G
C/T
G/G
G/T
T/T
Genotype
Group
maize
teosinte

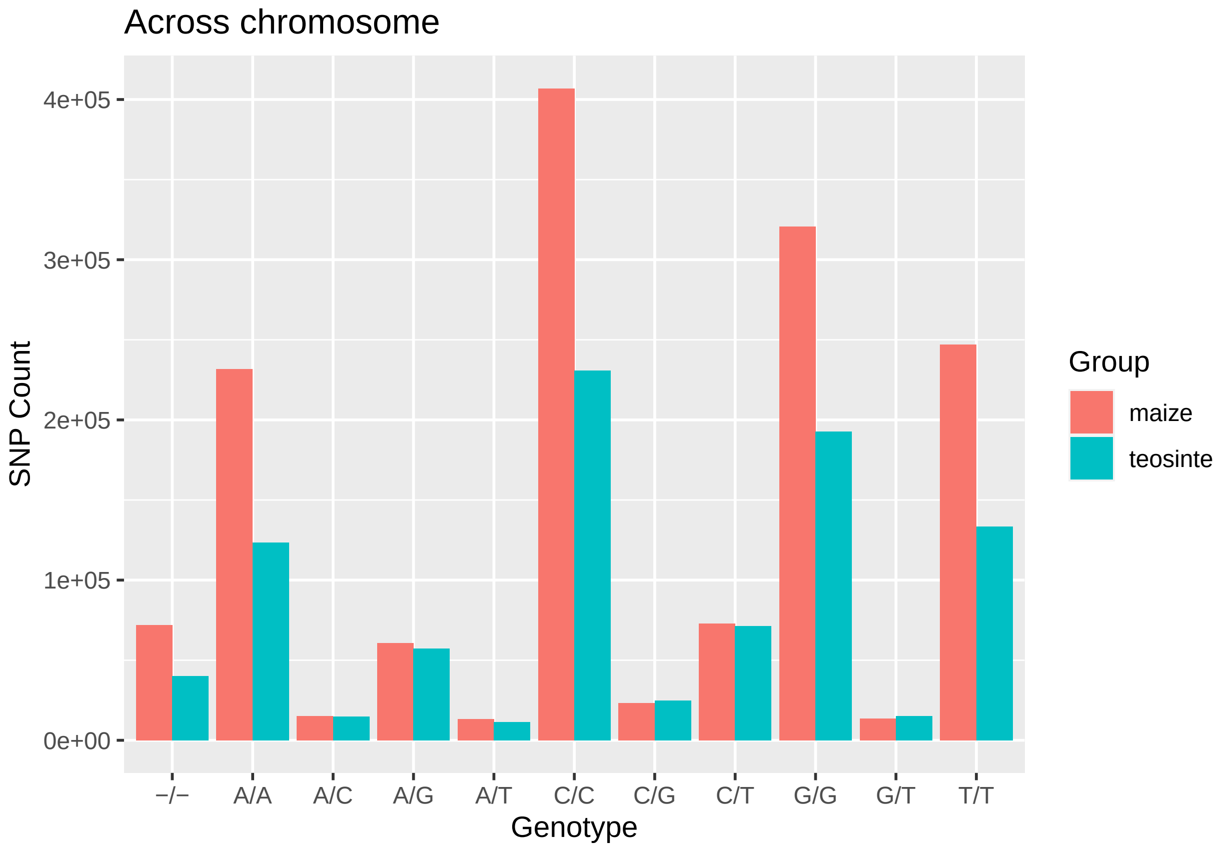

Chromosome 7
SNP Count
40000
30000
20000
10000
0
−/−
A/A
A/C
A/G
A/T
C/C
C/G
C/T
G/G
G/T
T/T
Genotype
Group
maize
teosinte


Chromosome 8
SNP Count
0
5000
10000
15000
20000
−/−
A/A
A/C
A/G
A/T
C/C
C/G
C/T
G/G
G/T
T/T
Genotype
Group
maize
teosinte


Chromosome 9
SNP Count
30000
20000
10000
0
−/−
A/A
A/C
A/G
A/T
C/C
C/G
C/T
G/G
G/T
T/T
Genotype
Group
maize
teosinte



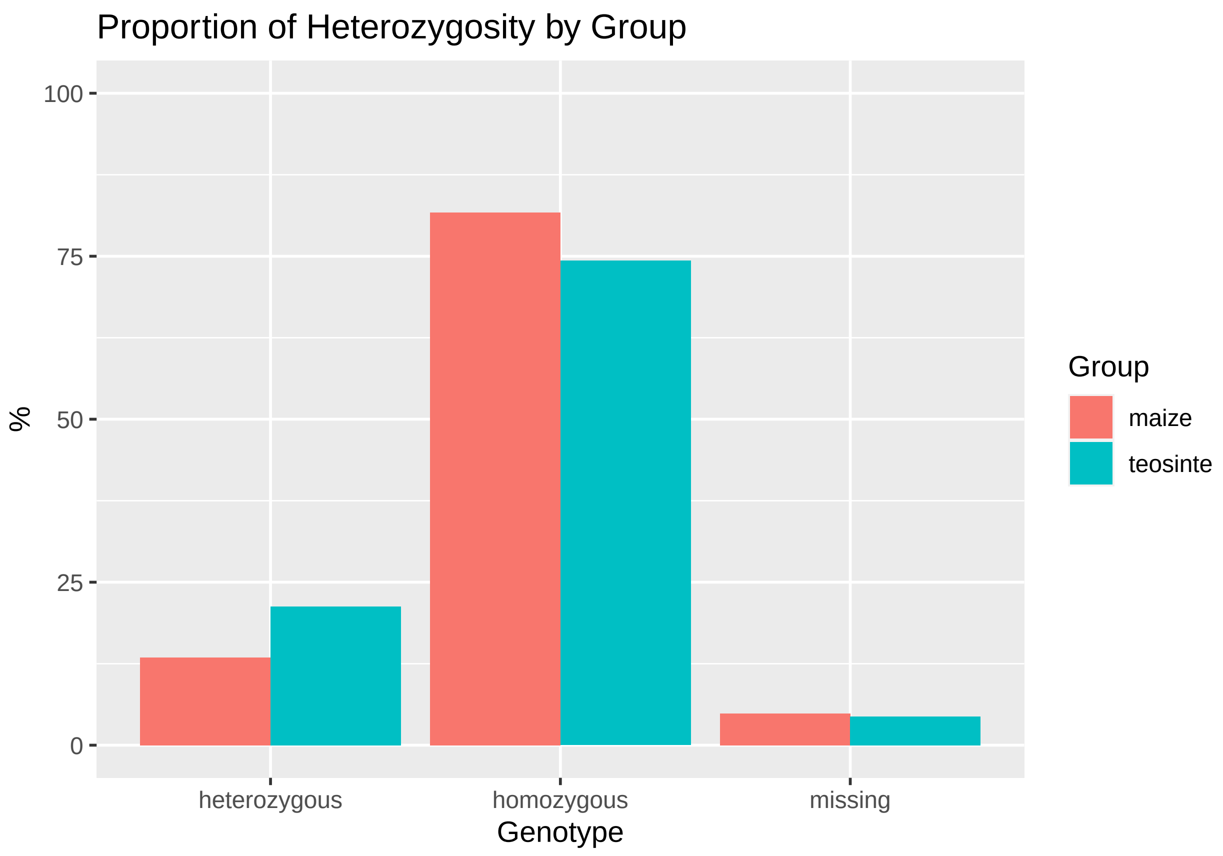
```
## Across chromosome
maize <- rbind(md1, md2, md3, md4, md5, md6, md7, md8, md9, md10)
teosinte <- rbind(td1, td2, td3, td4, td5, td6, td7, td8, td9, td10)
# Find the frequency count of genotype by groups maize and teosinte
m.freq <- maize[,4:ncol(maize)] %>% unlist() %>% table() %>% as.data.frame()
m.freq$Group <- "maize"
t.freq <- teosinte[,4:ncol(teosinte)] %>% unlist() %>%
  table() %>% as.data.frame()
t.freq$Group <- "teosinte"
# Plot SNP of across chromosome by two groups
freq <- rbind(m.freq, t.freq)
freq$Group <- as.factor(freq$Group)
ggplot(freq, aes(x = ., y = Freq, fill = Group)) +
  geom_bar(stat = "identity", position = position_dodge()) +
  labs(title = "Across chromosome", x = "Genotype", y = "SNP Count")
```



Missing data and amount of heterozygosity

```
# Missing data and amount of heterozygosity
proportion <- list(Group = c(), Genotype = c(), Prop = c())
homozygous <- c("A/A", "C/C", "G/G", "T/T")
missing <- c("-/-")
# Group maize
genotype <- m.freq$.
total <- sum(m.freq$Freq)
# proportion of missing data
miss.idx <- genotype %in% missing
miss.p <- sum(m.freq$Freq[miss.idx]) / total
# proportion of homozygous
homo.idx <- genotype %in% homozygous
homo.p <- sum(m.freq$Freq[homo.idx]) / total
# proportion of heterozygous
hetero.idx <- !(miss.idx | homo.idx)
hetero.p <- sum(m.freq$Freq[hetero.idx]) / total

proportion$Group[1:3] <- rep("maize", 3)
proportion$Genotype[1:3] <- c("missing", "homozygous", "heterozygous")
proportion$Prop[1:3] <- c(miss.p, homo.p, hetero.p)

# Group Teosinte
```

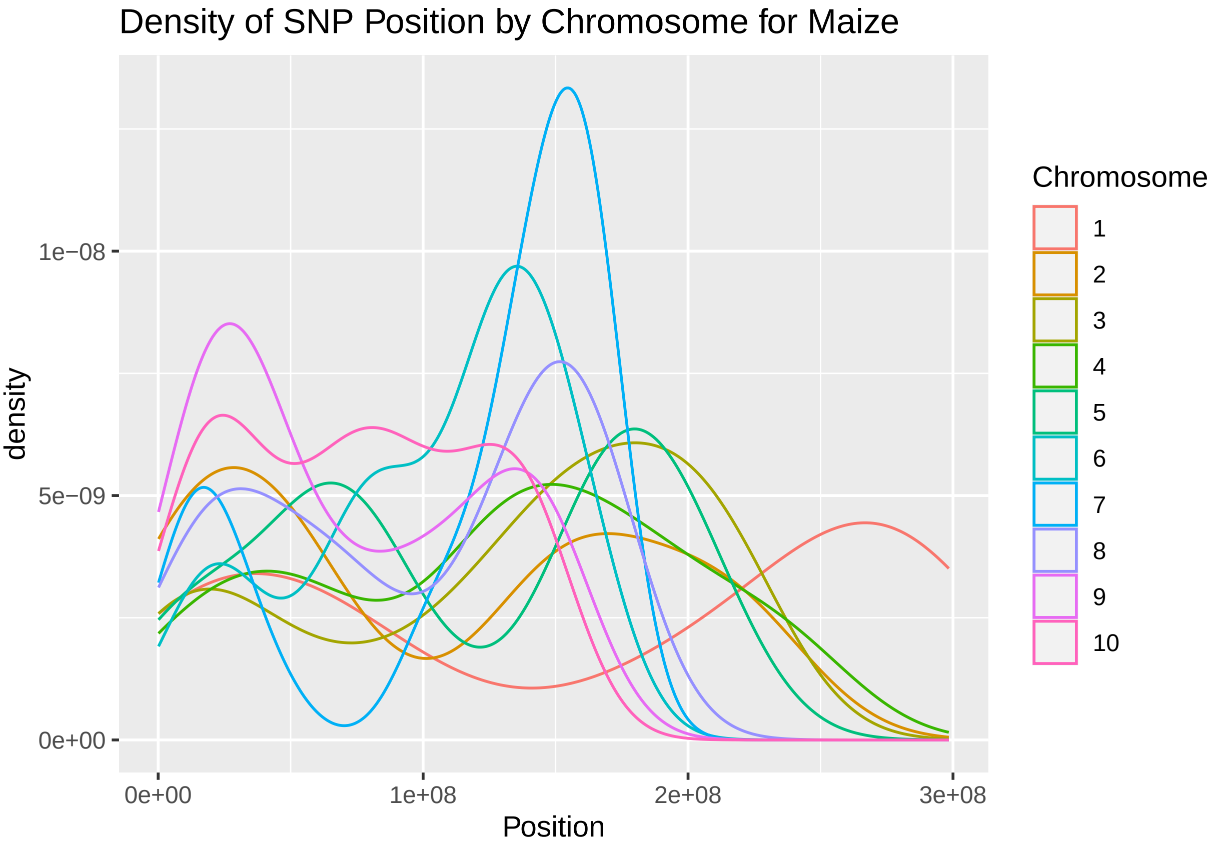
```
genotype <- t.freq$.
total <- sum(t.freq$Freq)
# proportion of missing data
miss.idx <- genotype %in% missing
miss.p <- sum(t.freq$Freq[miss.idx]) / total
# proportion of homozygous
homo.idx <- genotype %in% homozygous
homo.p <- sum(t.freq$Freq[homo.idx]) / total
# proportion of heterozygous
hetero.idx <- !(miss.idx | homo.idx)
hetero.p <- sum(t.freq$Freq[hetero.idx]) / total

proportion$Group[4:6] <- rep("teosinte", 3)
proportion$Genotype[4:6] <- c("missing", "homozygous", "heterozygous")
proportion$Prop[4:6] <- c(miss.p, homo.p, hetero.p)

proportion <- as.data.frame(proportion)
proportion$Group <- as.factor(proportion$Group)
proportion$Genotype <- as.factor(proportion$Genotype)
proportion$Prop <- proportion$Prop * 100

ggplot(proportion, aes(x = Genotype, y = Prop, fill = Group)) +
  geom_bar(stat = "identity", position = position_dodge()) + ylim(0, 100) +
  labs(title = "Proportion of Heterozygosity by Group",
       x = "Genotype", y = "%")
```



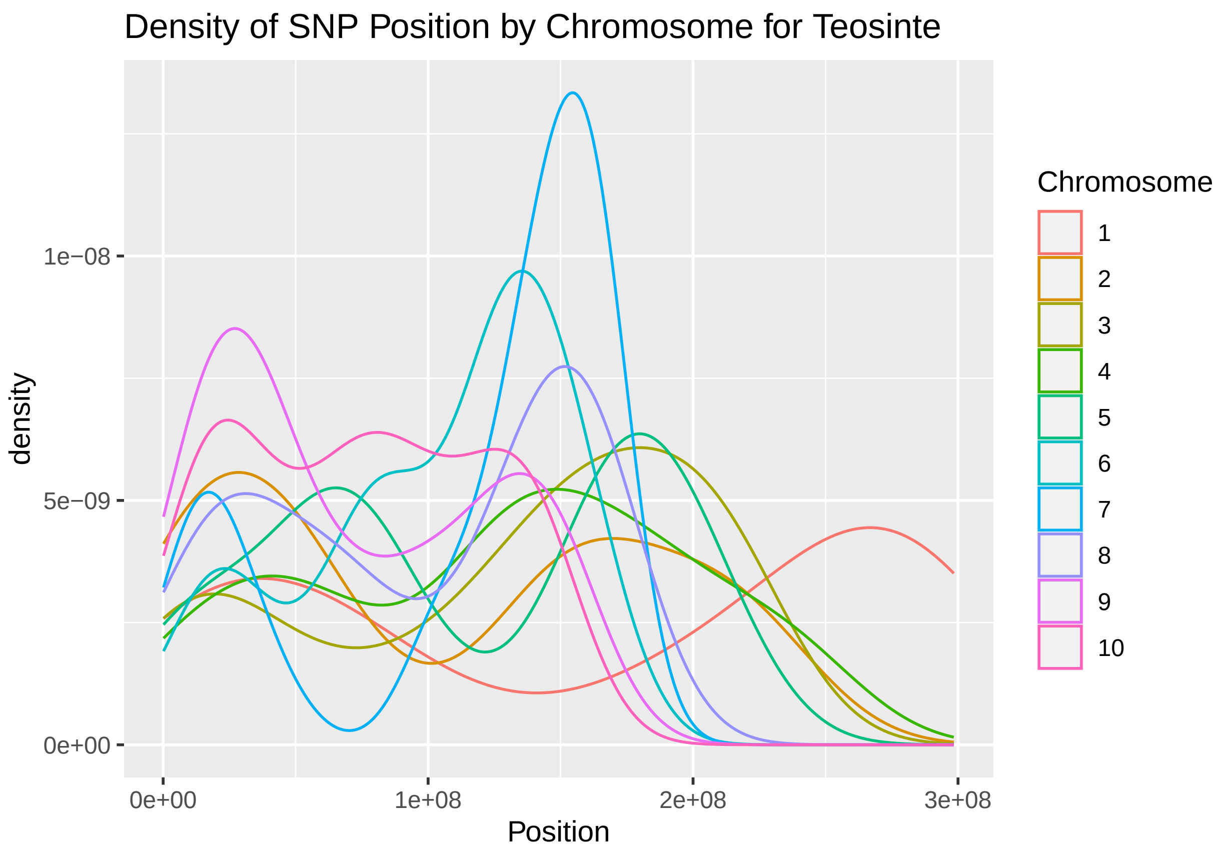
Summary of above plots: Over all, there more SNP positions in maize than teosinte individuals. And there are more homozygous than heterozygous in both maize and teosinte.

Your own visualization

```
maize$Chromosome <- as.factor(maize$Chromosome)
ggplot(maize, aes(x = Position, color = Chromosome)) + geom_density() +
  labs(title = "Density of SNP Position by Chromosome for Maize")
```



```
teosinte$Chromosome <- as.factor(teosinte$Chromosome)
ggplot(teosinte, aes(x = Position, color = Chromosome)) + geom_density() +
  labs(title = "Density of SNP Position by Chromosome for Teosinte")
```


Density of SNP Position by Chromosome for Teosinte
density
1e−08
5e−09
0e+00
0e+00
1e+08
2e+08
3e+08
Position
Chromosome
1
2
3
4
5
6
7
8
9
10