

BCB 546 Homework 2: R Assignment

Weixia Deng

Data Processing

```
library(tidyverse)
# 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", "ZMLLR", "ZMMMR")
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
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")
}

```

Data Visualization

```

library(tidyverse)

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

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

for (i in 1:length(filels)){

```

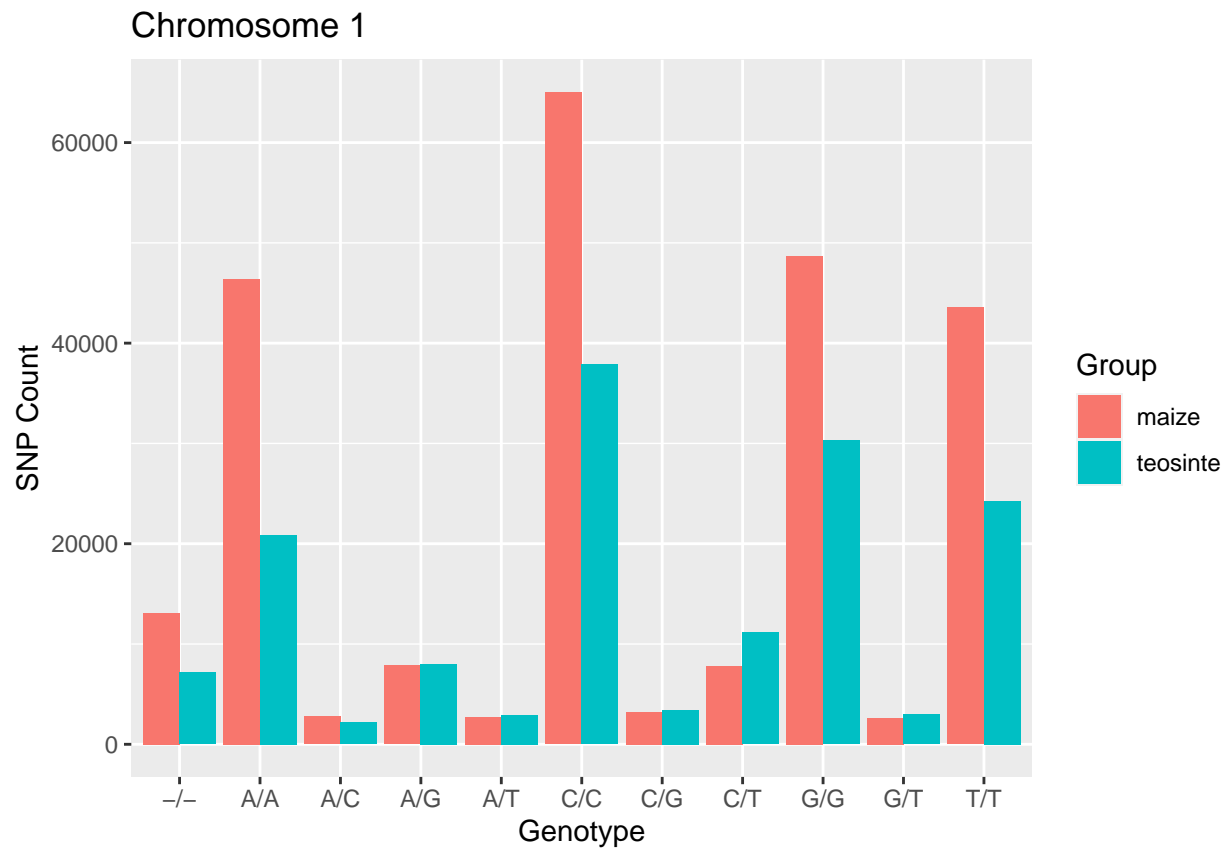
```

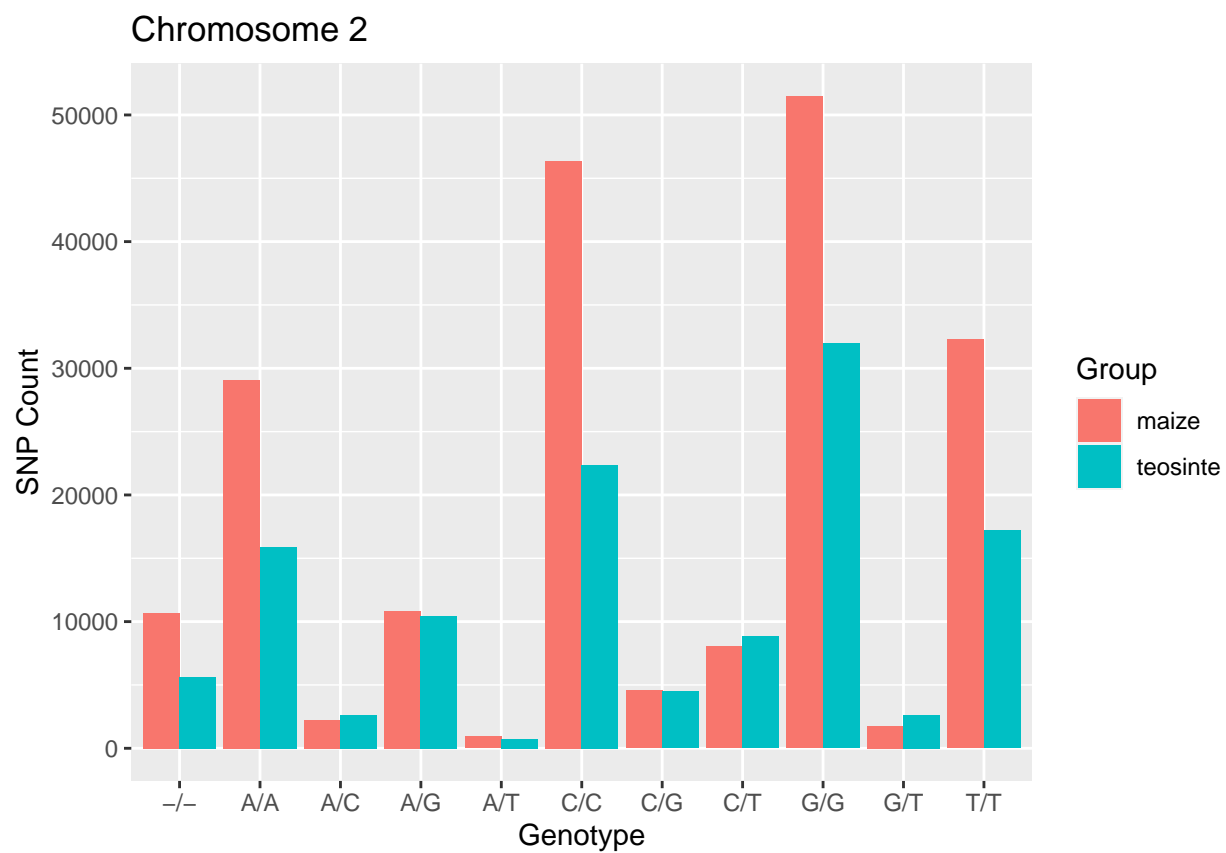
n <- paste("output/", filels[i], sep = "")
assign(dfls[i], read.table(n, sep = "\t", header = TRUE))
}

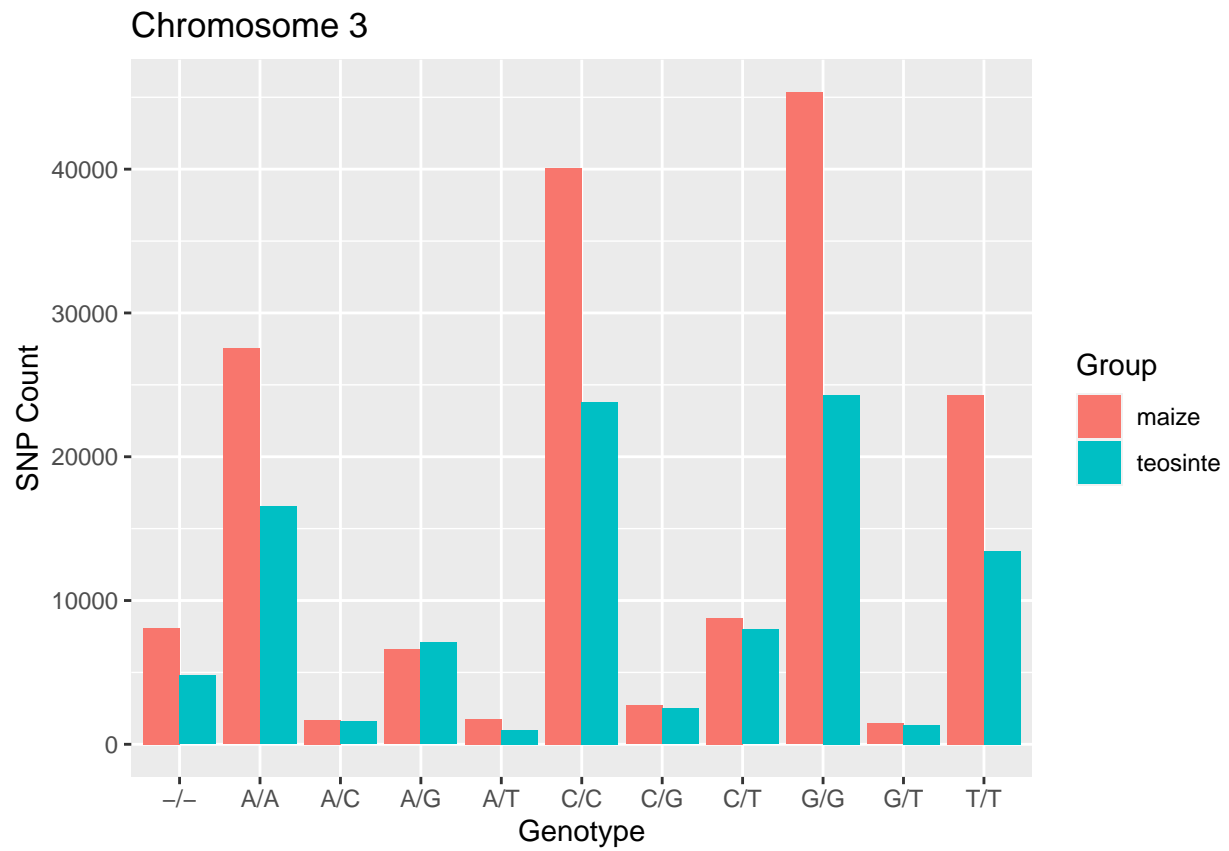
# Maize and Teosinte Visualization
## On each chromosome (1-10)
ls.maize <- list(md1, md2, md3, md4, md5, md6, md7, md8, md9, md10)
ls.teosinte <- list(td1, td2, td3, td4, td5, td6, td7, td8, td9, td10)

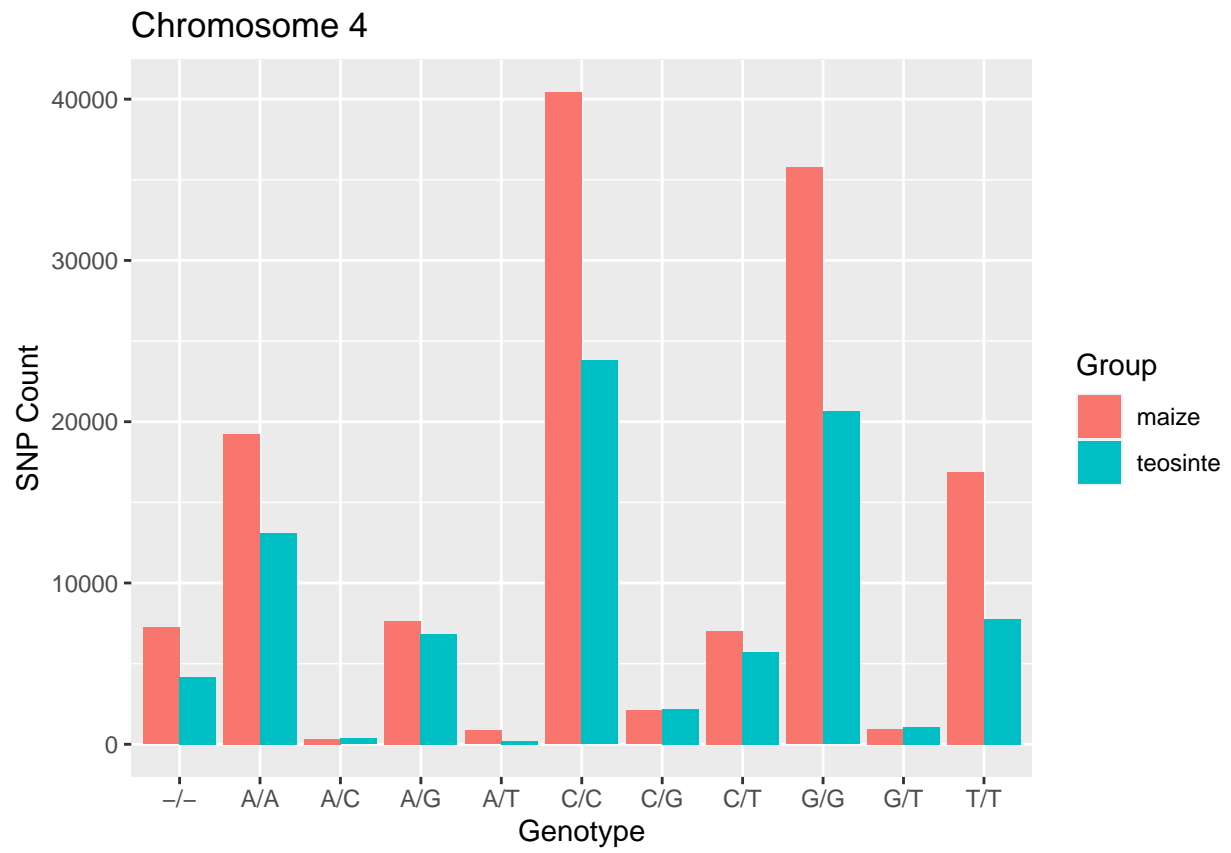
for (i in 1:10){
  # Loop through chromosome 1-10
  md <- ls.maize[[i]]
  td <- ls.teosinte[[i]]
  # Find the frequency count of genotype by groups maize and teosinte
  m.freq <- md[,4:ncol(md)] %>% unlist() %>% table() %>% as.data.frame()
  m.freq$Group <- "maize"
  t.freq <- td[,4:ncol(td)] %>% unlist() %>% table() %>% as.data.frame()
  t.freq$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"))
}

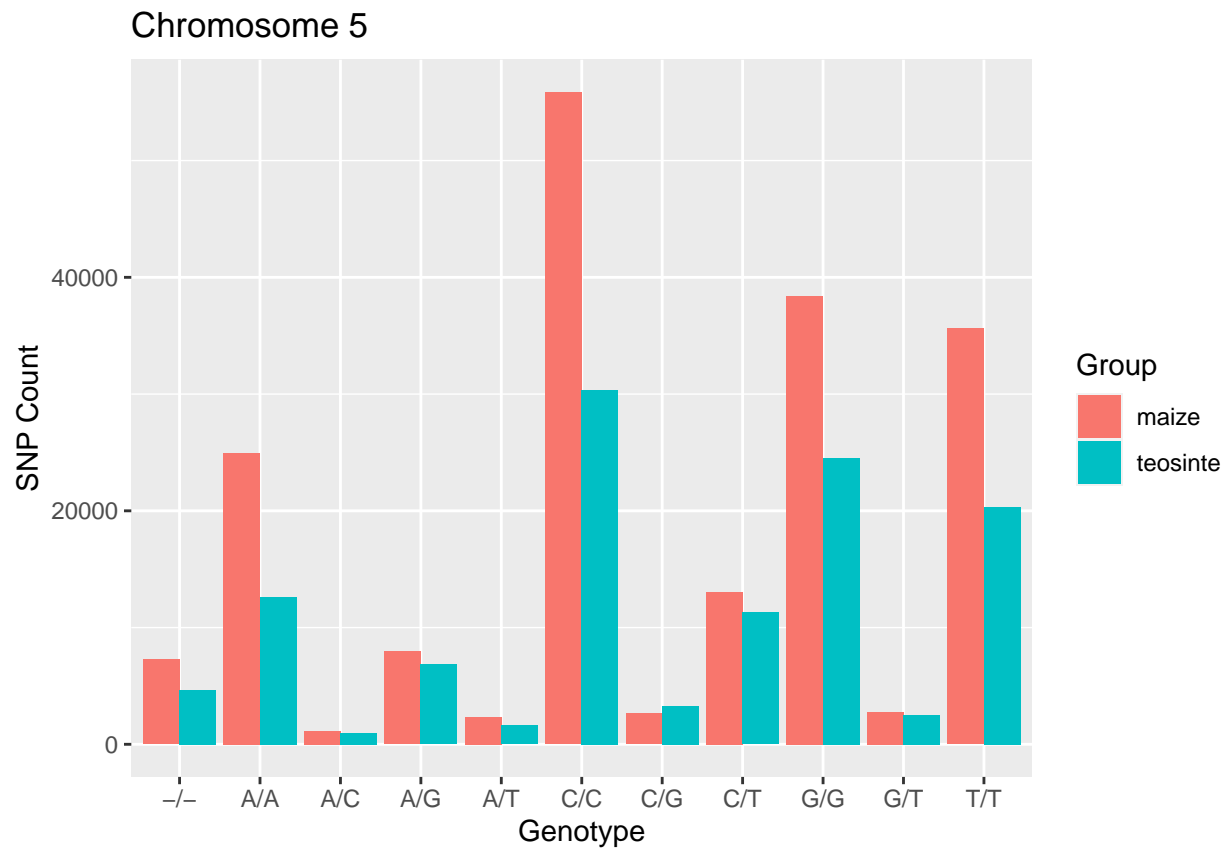
```

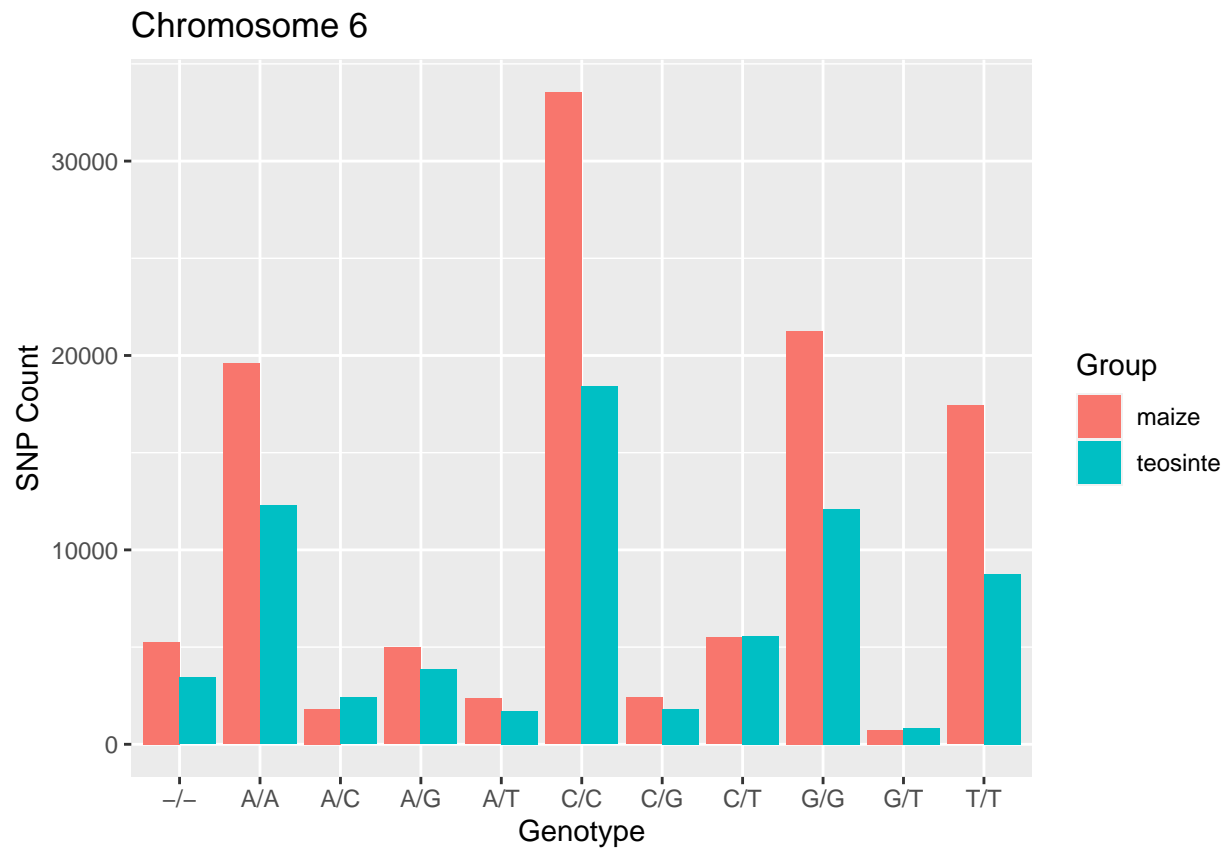


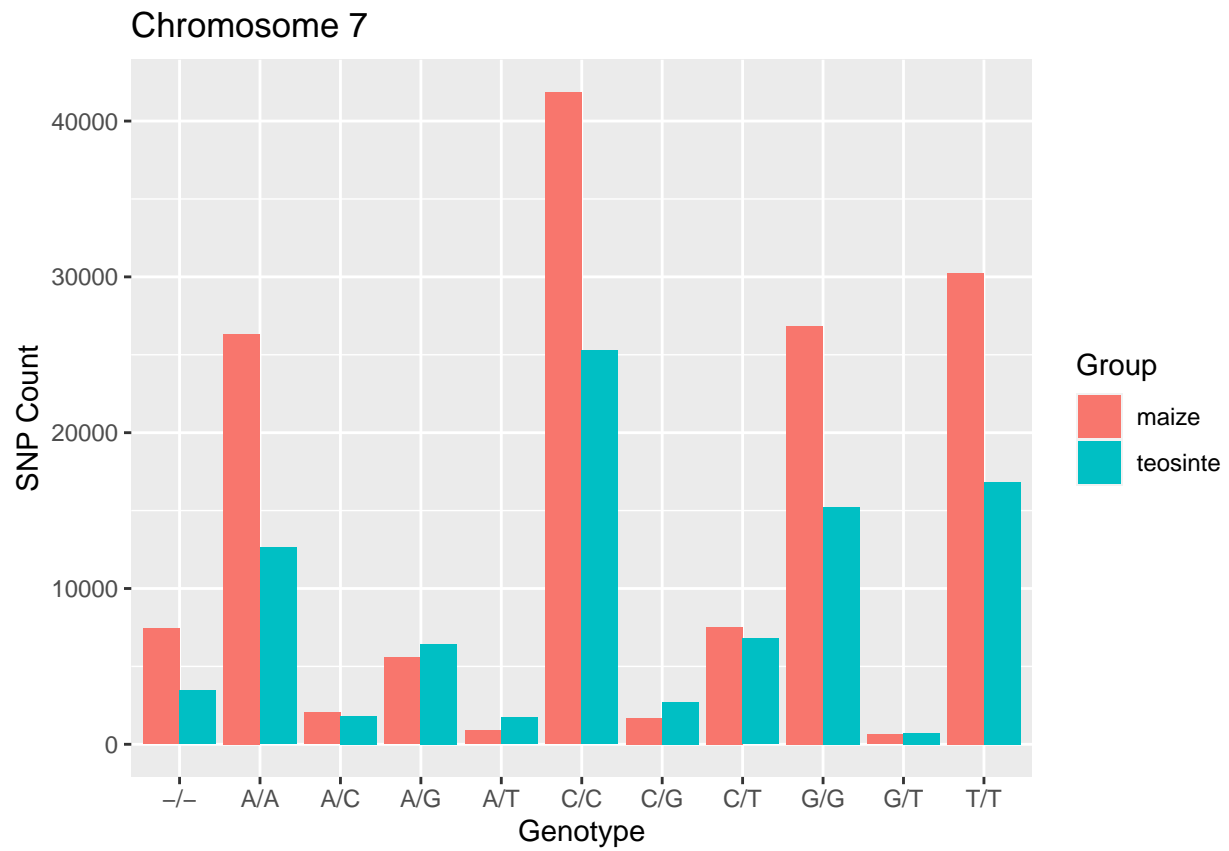


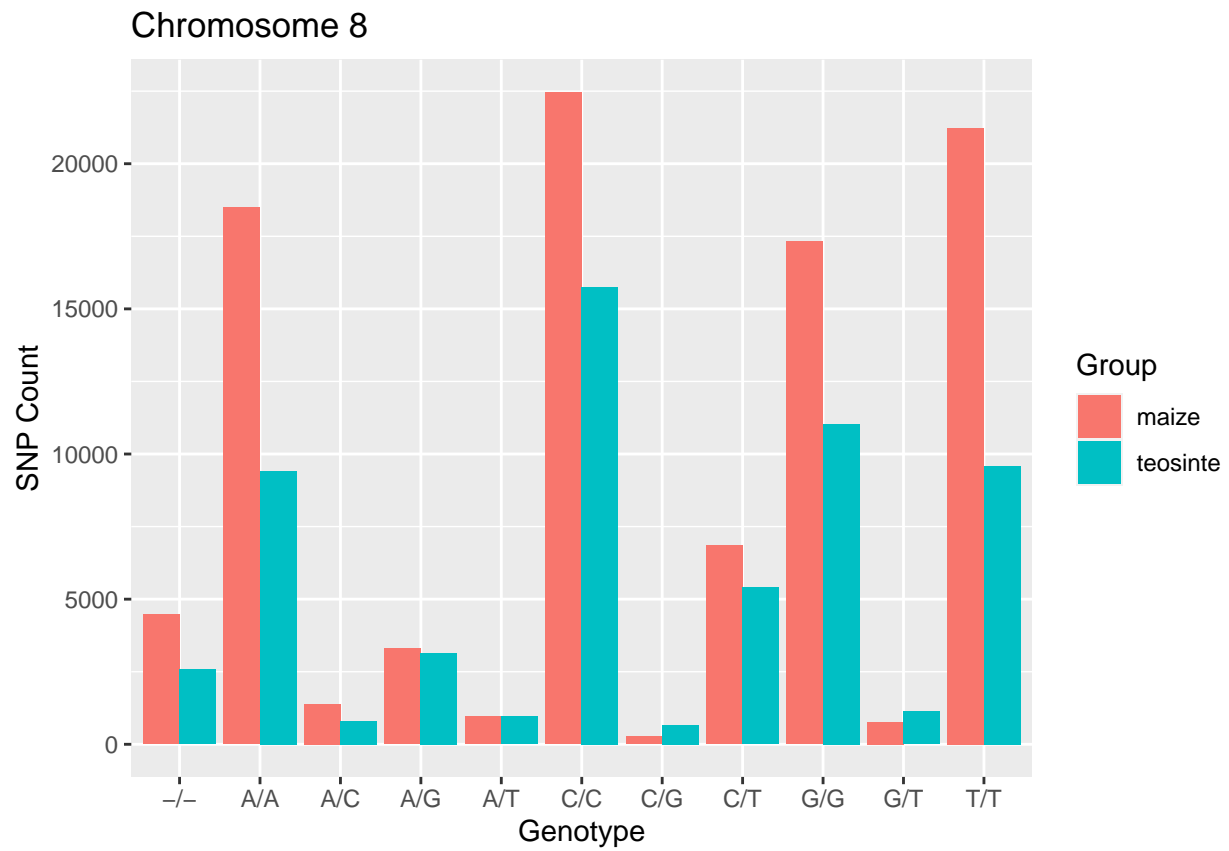


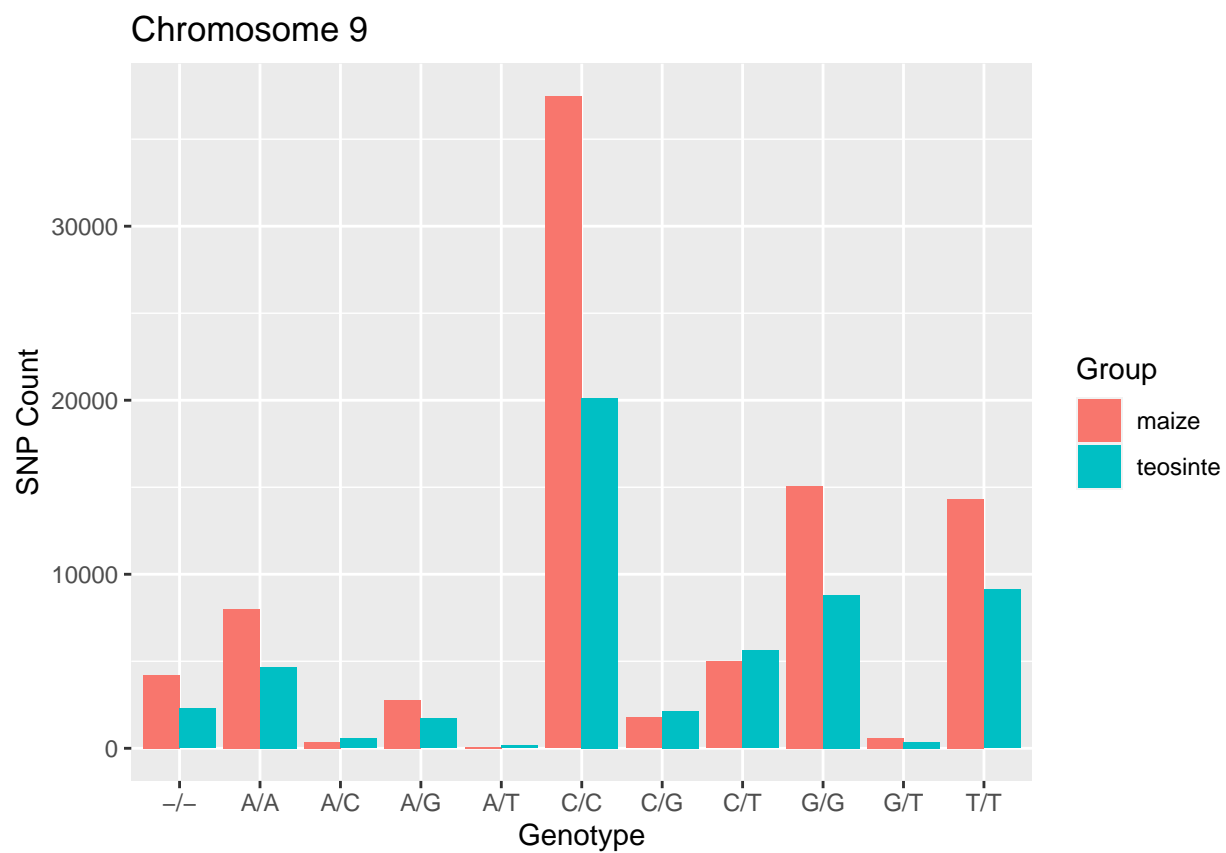


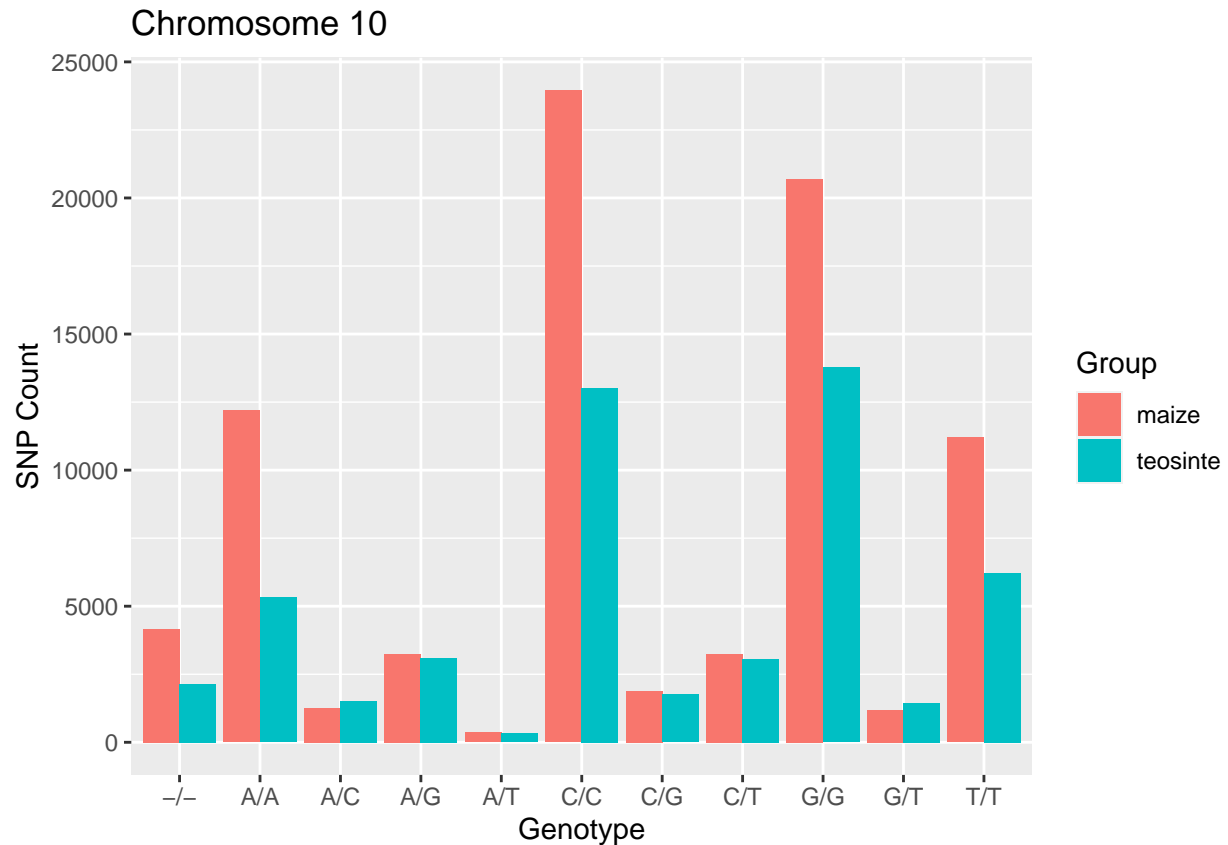




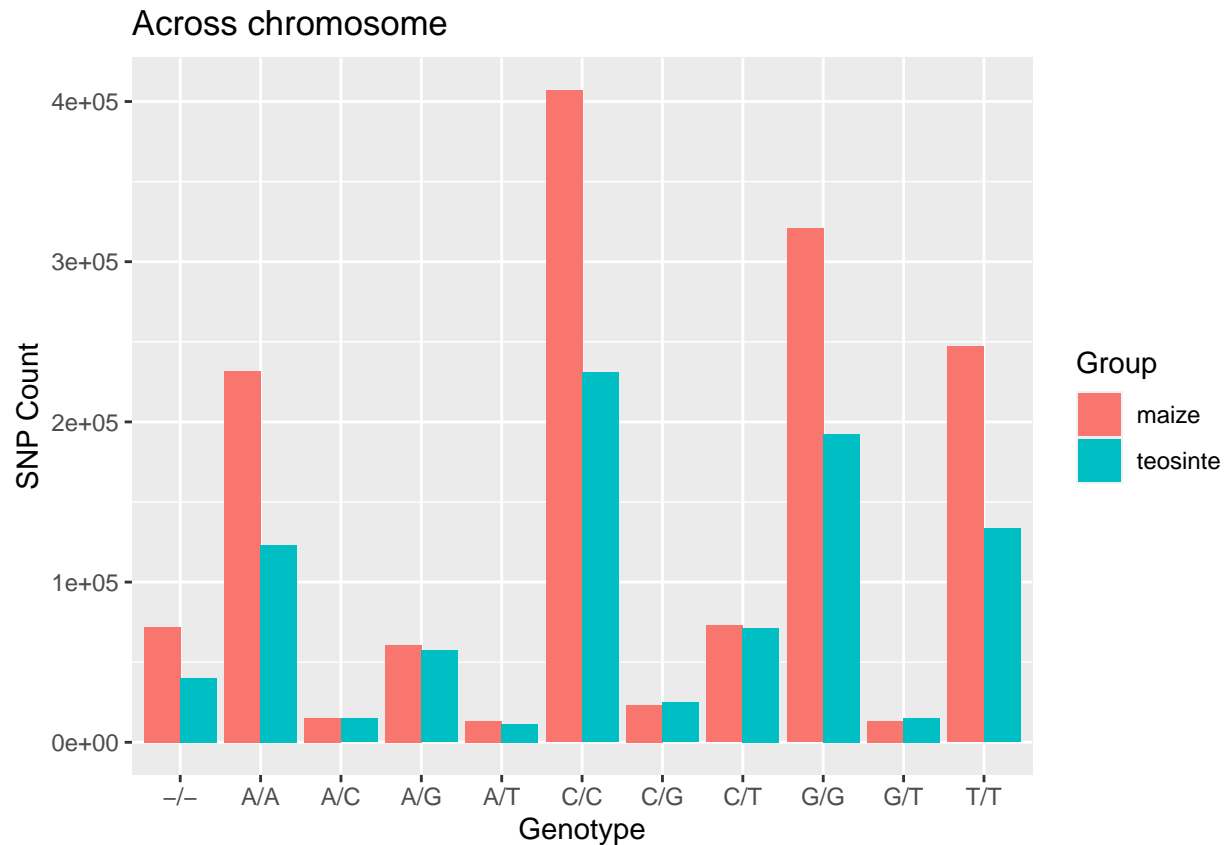








```
## Across chromosome
maize <- rbind(md1, md2, md3, md4, md5, md6, md7, md8, md9, md10)
teosinte <- rbind(td1, td2, td3, td4, td5, td6, td7, td8, td9, td10)
# Find the frequency count of genotype by groups maize and teosinte
m.freq <- maize[,4:ncol(maize)] %>% unlist() %>% table() %>% as.data.frame()
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
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 = "%")

```

