R_Assignment

Yu-Ru Chen

3/16/2021

Data inspection

Load the 2 files into genotype and pos data frame

```
library(tidyverse)
```

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

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.0     v dplyr   1.0.5
## v tidyr   1.1.3     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()
```

```
genotype <- as.data.frame(read.table("fang_et_al_genotypes.txt", sep="\t",header=TRUE))
pos <- as.data.frame(read.table("snp_position.txt", sep="\t",header=TRUE))
```

SNP genotypes data

`fang_et_al_genotypes.txt` is assigned to be the genotype data frame.

```
dim(genotype)
```

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

```
sapply(genotype, class)[1:6]
```

```
##   Sample_ID      JG_OTU       Group    abph1.20    abph1.22       ae1.3
## "character" "character" "character" "character" "character" "character"
```

```
#str(genotype)
#glimpse(genotype)
colnames(genotype)[1:6]
```

```
## [1] "Sample_ID" "JG_OTU"    "Group"     "abph1.20"  "abph1.22"  "ae1.3"
```

```
genotype[1:6,1:6]
```

```
##   Sample_ID   JG_OTU Group abph1.20 abph1.22 ae1.3
## 1     SL-15 T-aust-1 TRIPS      ?/?      ?/?   T/T
## 2     SL-16 T-aust-2 TRIPS      ?/?      ?/?   T/T
## 3     SL-11 T-brav-1 TRIPS      ?/?      ?/?   T/T
## 4     SL-12 T-brav-2 TRIPS      ?/?      ?/?   T/T
## 5     SL-18   T-cund TRIPS      ?/?      ?/?   T/T
## 6      SL-2 T-dact-1 TRIPS      ?/?      ?/?   T/T
```

```
genotype %>%
  group_by(Group) %>%
  count()
```

```
## # A tibble: 16 x 2
## # Groups:   Group [16]
##    Group     n
##    <chr> <int>
##  1 TRIPS    22
##  2 ZDIPL    15
##  3 ZLUXR    17
##  4 ZMHUE    10
##  5 ZMMIL   290
##  6 ZMMLR  1256
##  7 ZMMMR    27
##  8 ZMPBA   900
##  9 ZMPIL    41
## 10 ZMPJA    34
## 11 ZMXCH    75
## 12 ZMXCP    69
## 13 ZMXIL     6
## 14 ZMXNO     7
## 15 ZMXNT     4
## 16 ZPERR     9
```

SNP markers information

`snp_position.txt` is assigned to be the pos dataframe.

```
dim(pos)
```

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

```r
sapply(pos, class)[1:6]
```

```
##         SNP_ID  cdv_marker_id     Chromosome       Position        alt_pos
##    "character"      "integer"    "character"    "character"    "character"
## mult_positions
##    "character"
```

```r
str(pos)
```

```
## ’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 ...
```

```r
glimpse(pos)
```

```
## Rows: 983
## Columns: 15
## $ SNP_ID               <chr> "abph1.20", "abph1.22", "ae1.3", "ae1.4", "ae1.5"~
## $ cdv_marker_id        <int> 5976, 5978, 6605, 6606, 6607, 5982, 3463, 3466, 5~
## $ Chromosome           <chr> "2", "2", "5", "5", "5", "1", "3", "3", "4", "4",~
## $ Position             <chr> "27403404", "27403892", "167889790", "167889682",~
## $ alt_pos              <chr> "", "", "", "", "", "", "", "", "", "", "", "", "~
## $ mult_positions       <chr> "", "", "", "", "", "", "", "", "", "", "", "", "~
## $ amplicon             <chr> "abph1", "abph1", "ae1", "ae1", "ae1", "an1", "ba~
## $ cdv_map_feature.name <chr> "AB042260", "AB042260", "ae1", "ae1", "ae1", "an1~
## $ gene                 <chr> "abph1", "abph1", "ae1", "ae1", "ae1", "an1", "ba~
## $ candidate.random     <chr> "candidate", "candidate", "candidate", "candidate~
## $ Genaissance_daa_id   <int> 8393, 8394, 8395, 8396, 8397, 8398, 8399, 8400, 8~
## $ Sequenom_daa_id      <int> 10474, 10475, 10477, 10478, 10479, 10481, 10482, ~
## $ count_amplicons      <int> 1, 0, 1, 0, 0, 1, 1, 0, 1, 0, 0, 1, 0, 1, 1, 0, 1~
## $ count_cmf            <int> 1, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 1, 1, 0, 1~
## $ count_gene           <int> 1, 0, 1, 0, 0, 1, 1, 0, 1, 0, 0, 1, 0, 1, 1, 0, 1~
```

```r
colnames(pos)[1:6]
```

```
## [1] "SNP_ID"         "cdv_marker_id"  "Chromosome"     "Position"
## [5] "alt_pos"        "mult_positions"
```

```
pos[1:6,1:6]
```

```
##     SNP_ID cdv_marker_id Chromosome  Position alt_pos mult_positions
## 1 abph1.20          5976          2  27403404
## 2 abph1.22          5978          2  27403892
## 3    ae1.3          6605          5 167889790
## 4    ae1.4          6606          5 167889682
## 5    ae1.5          6607          5 167889821
## 6    an1.4          5982          1 240498509
```

```
pos[pos == "unknown"] <- NA
pos[pos == "multiple"] <- NA

pos %>%
  group_by(Chromosome) %>%
  summarise(Max=max(Position, na.rm = T), Min=min(Position, na.rm = T),    Number=length(Position))
```

```
## # A tibble: 11 x 4
##    Chromosome Max        Min         Number
##    <chr>      <chr>      <chr>        <int>
##  1 1          "95897171" "10069039"     155
##  2 10         "96216463" "10432605"      53
##  3 2          "69623323" "10429605"     127
##  4 3          "95541392" "106631676"    107
##  5 4          "78946482" "103665461"     91
##  6 5          "945545"   "100227859"    122
##  7 6          "98507715" "113705211"     76
##  8 7          "43948320" "104898448"     97
##  9 8          "83913342" "115257234"     62
## 10 9          "94285743" "104237516"     60
## 11 <NA>       ""         ""              33
```

Data processing

Subset the pos data frame to keep the SNP_ID, Chr, and Pos columns

> adjust the characters of the column and remove the `unknown` and `multiple` which are regarded as NAs out

```
posred <- pos %>%
  select(SNP_ID, Chromosome, Position) %>%
  mutate(Chromosome=as.numeric(Chromosome),
         Position=as.numeric(Position))%>%
  filter_all(all_vars(. != "NA"))

str(posred)
```

```
## 'data.frame':    939 obs. of  3 variables:
##  $ SNP_ID    : chr  "abph1.20" "abph1.22" "ae1.3" "ae1.4" ...
##  $ Chromosome: num  2 2 5 5 5 1 3 3 4 4 ...
##  $ Position  : num  2.74e+07 2.74e+07 1.68e+08 1.68e+08 1.68e+08 ...
```

Subset the genotype data into maize and teosinte datasets

```
maize <- genotype[which(genotype$Group=="ZMMIL" | genotype$Group =="ZMMLR" | genotype$Group == "ZMMMR")
teosinte <-genotype[which(genotype$Group=="ZMPBA" | genotype$Group =="ZMPIL" | genotype$Group == "ZMPJA
```

Formatting the maize file to merge the redPos and maize file

```
maize <- maize[,c(-2,-3)]
maize[1:6,1:6] ## have a look
```

```
##      Sample_ID abph1.20 abph1.22 ae1.3 ae1.4 ae1.5
## 1210 ZDP_0752a      C/G      A/A   T/T   G/G   C/C
## 1211 ZDP_0793a      C/G      A/A   T/T   G/G   C/T
## 1212 ZDP_0612a      C/C      A/A   T/T   G/G   C/C
## 1213 ZDP_0602a      C/G      A/A   G/T   A/G   C/T
## 1214 ZDP_0581a      C/C      A/A   T/T   G/G   C/T
## 1215 ZDP_0552a      C/G      A/A   T/T   G/G   C/T
```

```
maize <- t(maize)
maize <- cbind(rownames(maize),maize)
rownames(maize) <- NULL
colnames(maize) <- maize[1,]
maize <- maize[-1,]
maize <- as.data.frame(maize)
colnames(maize)[1] <- "SNP_ID"
maizewp <- merge(posred, maize, by = "SNP_ID")
maizewp <- maizewp %>% arrange(Chromosome,Position)
## maize genotypes with SNP position information
```

Formatting the teosinte file to merge the redPos and maize file

```
teosinte <- teosinte[,c(-2,-3)]
teosinte <- t(teosinte)
teosinte <- cbind(rownames(teosinte),teosinte)
rownames(teosinte) <- NULL
colnames(teosinte) <- teosinte[1,]
teosinte <- teosinte[-1,]
teosinte <- as.data.frame(teosinte)
colnames(teosinte)[1] <- "SNP_ID"
teosintewp <- merge(posred, teosinte, by = "SNP_ID")
teosintewp <- teosintewp %>% arrange(Chromosome,Position)
```

Splitting the maize data into different files by chromosomes and SNP positions.

```
chr <- 1:10
for (i in chr) {
  files_inc <- maizewp[maizewp$Chromosome == i,]
  files_inc[files_inc == "?/?"] <- "?"
  if (i < 10) { write.table(files_inc, file = paste("Maize_Chr0",i,"_increase.txt",sep=""),row.names = F
  else {write.table(files_inc, file = paste("Maize_Chr",i,"_increase.txt",sep=""),row.names = FALSE, sep

  files_dec <- maizewp[maizewp$Chromosome == i,]
  files_dec[files_dec == "?/?"] <- "-"
  files_dec <- files_dec %>% arrange(desc(Chromosome),desc(Position))
  if (i < 10) { write.table(files_dec, file = paste("Maize_Chr0",i,"_decrease.txt",sep=""),row.names = F
  else {write.table(files_dec, file = paste("Maize_Chr",i,"_decrease.txt",sep=""),row.names = FALSE, sep
}
```

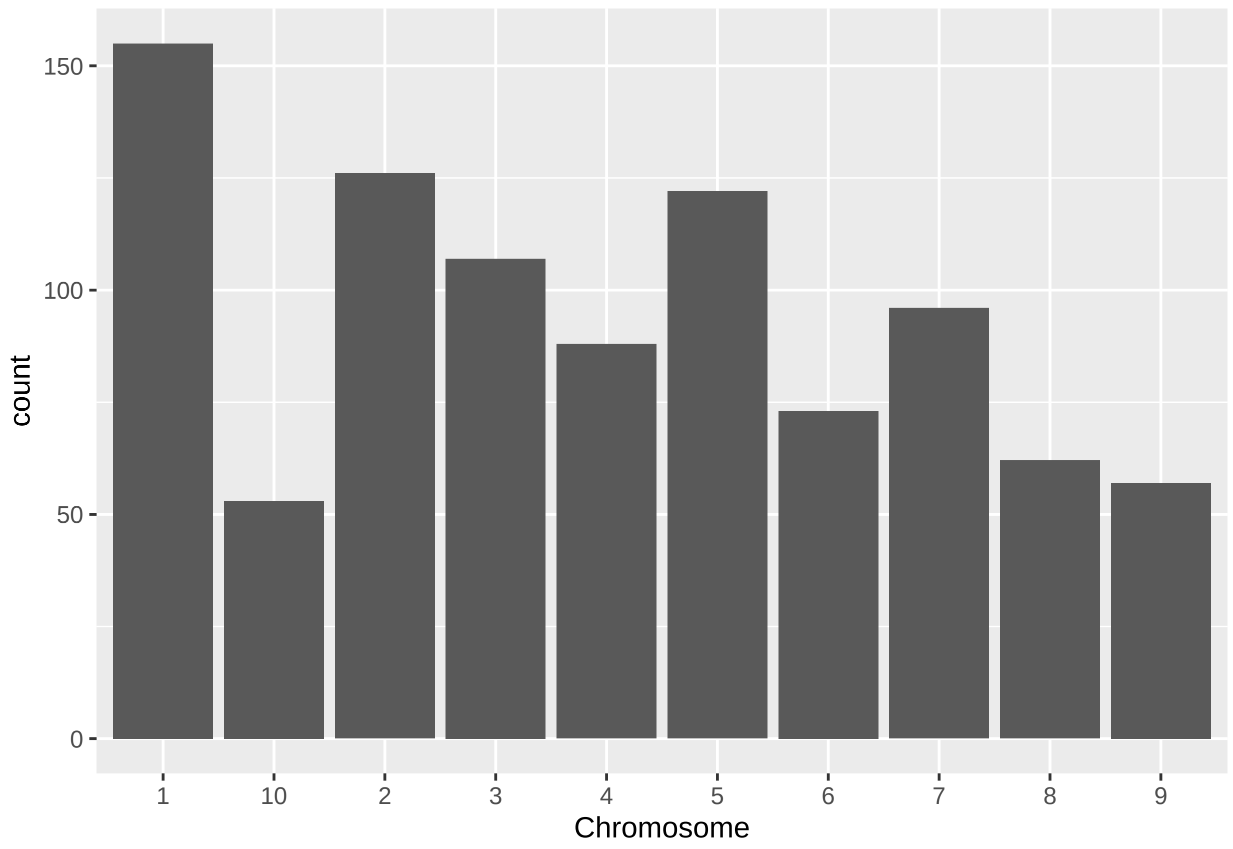
Splitting the teosinte data into different files by chromosomes and SNP positions.

```
chr <- 1:10
for (i in chr) {
  files_inc <- teosintewp[teosintewp$Chromosome == i,]
  files_inc[files_inc == "?/?"] <- "?"
  if (i < 10) { write.table(files_inc, file = paste("Teosinte_Chr0",i,"_increase.txt",sep=""),row.names
  else {write.table(files_inc, file = paste("Teosinte_Chr",i,"_increase.txt",sep=""),row.names = FALSE,

  files_dec <- teosintewp[teosintewp$Chromosome == i,]
  files_dec[files_dec == "?/?"] <- "-"
  files_dec <- files_dec %>% arrange(desc(Chromosome),desc(Position))
  if (i < 10) { write.table(files_dec, file = paste("Teosinte_Chr0",i,"_decrease.txt",sep=""),row.names
  else {write.table(files_dec, file = paste("Teosinte_Chr",i,"_decrease.txt",sep=""),row.names = FALSE,
}
```

SNPs per chromosome

```
pos %>%
  select(SNP_ID, Chromosome, Position) %>%
  drop_na() %>%
  ggplot()+
  geom_bar(aes(x=Chromosome))
```



Missing data and amount of heterozygosity

```
genotype2 <- genotype[,-2]
genotype2[1:6,1:6]
```

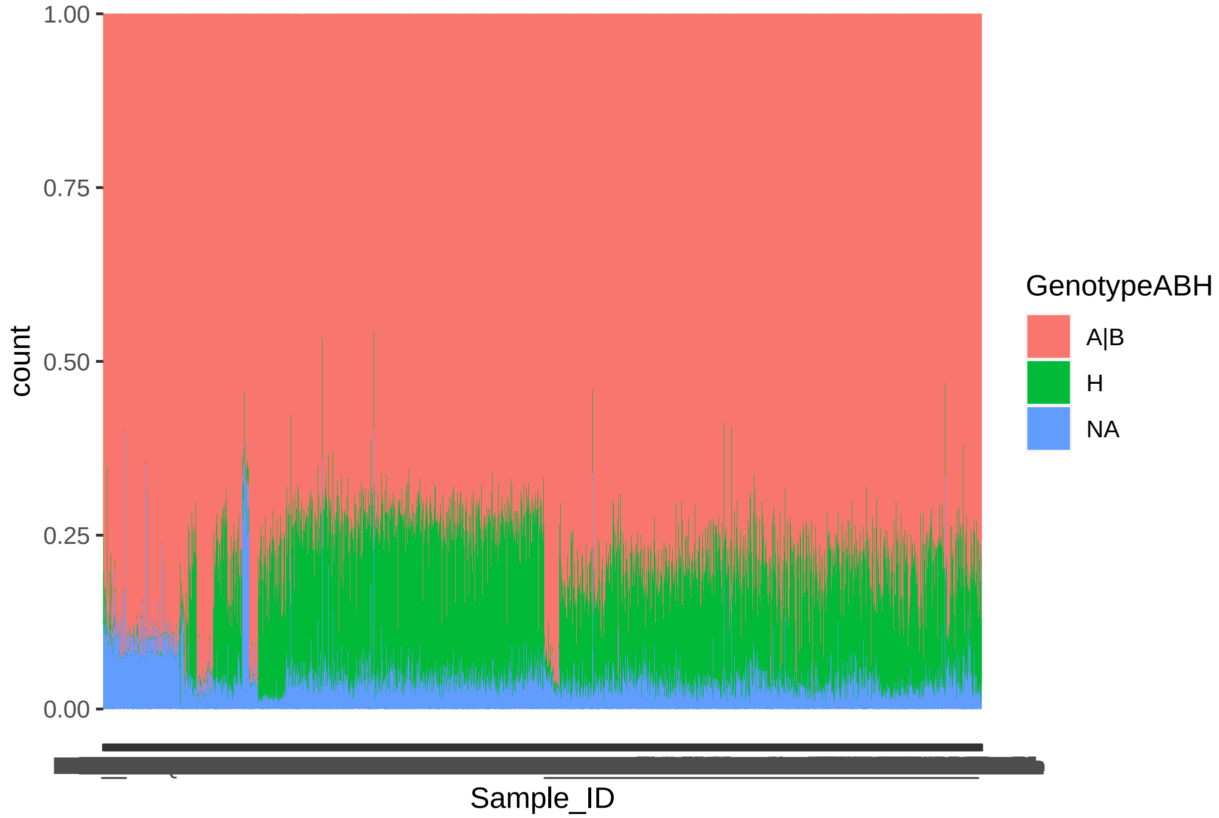
```
##   Sample_ID Group abph1.20 abph1.22 ae1.3 ae1.4
## 1     SL-15 TRIPS      ?/?      ?/?   T/T   G/G
## 2     SL-16 TRIPS      ?/?      ?/?   T/T   ?/?
## 3     SL-11 TRIPS      ?/?      ?/?   T/T   G/G
## 4     SL-12 TRIPS      ?/?      ?/?   T/T   G/G
## 5     SL-18 TRIPS      ?/?      ?/?   T/T   G/G
## 6      SL-2 TRIPS      ?/?      ?/?   T/T   G/G
```

```
## create a function to detect the SNP genotypes
ABH <- function(x) {
  if ( x == "A/A" | x == "C/C" | x == "G/G" | x == "T/T") {
    return("A|B")
  }
  else if (x == "?/?") {
    return("NA")
  }
  else {return("H")}
}
```

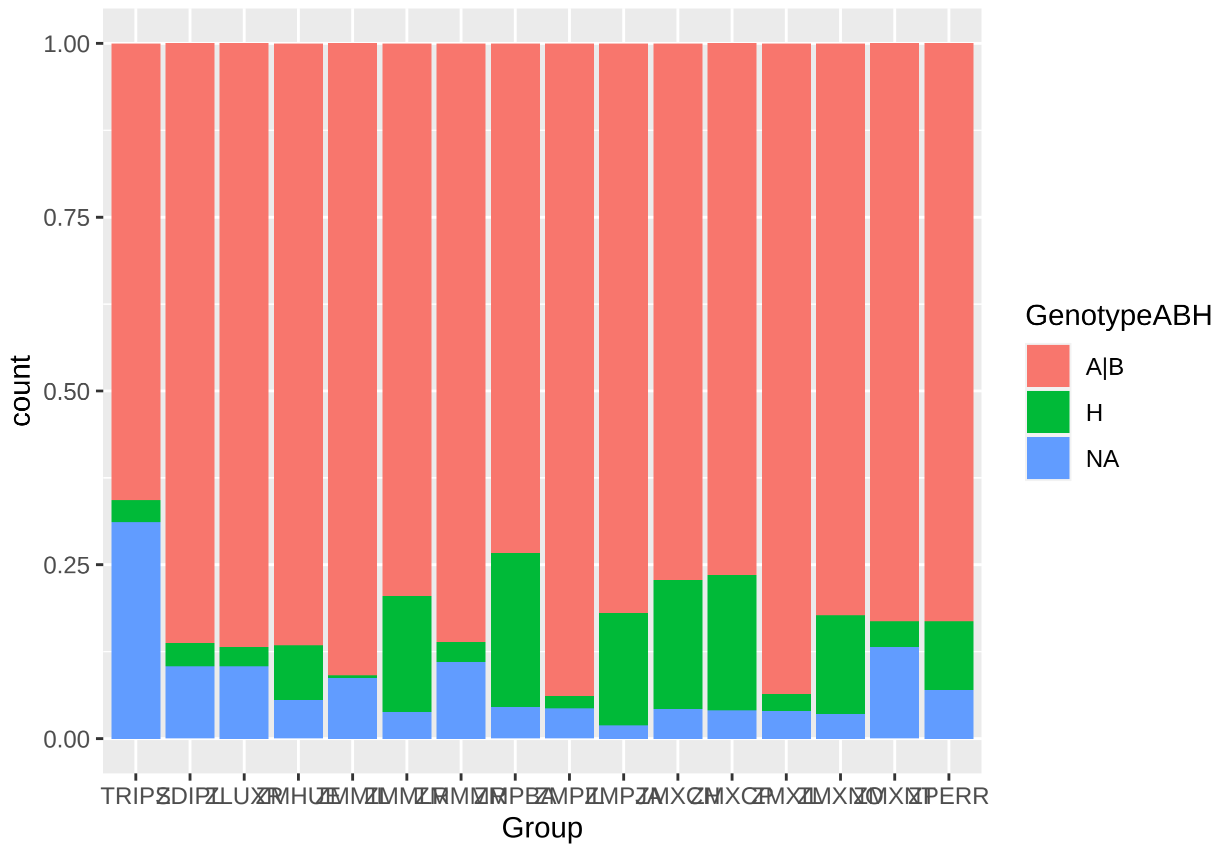
```r
ABH_V <- Vectorize(ABH) ## make the function be a vectorized function

genotype3 <- genotype2 %>%
  pivot_longer(3:last_col(), names_to = "SNP", values_to = "Genotype") %>%
  mutate( GenotypeABH = ABH_V(Genotype))

ggplot(genotype3)+
  geom_bar(aes(x=Sample_ID, fill=GenotypeABH), position = "fill")
```



```r
ggplot(genotype3)+
  geom_bar(aes(x=Group, fill=GenotypeABH), position = "fill")
```



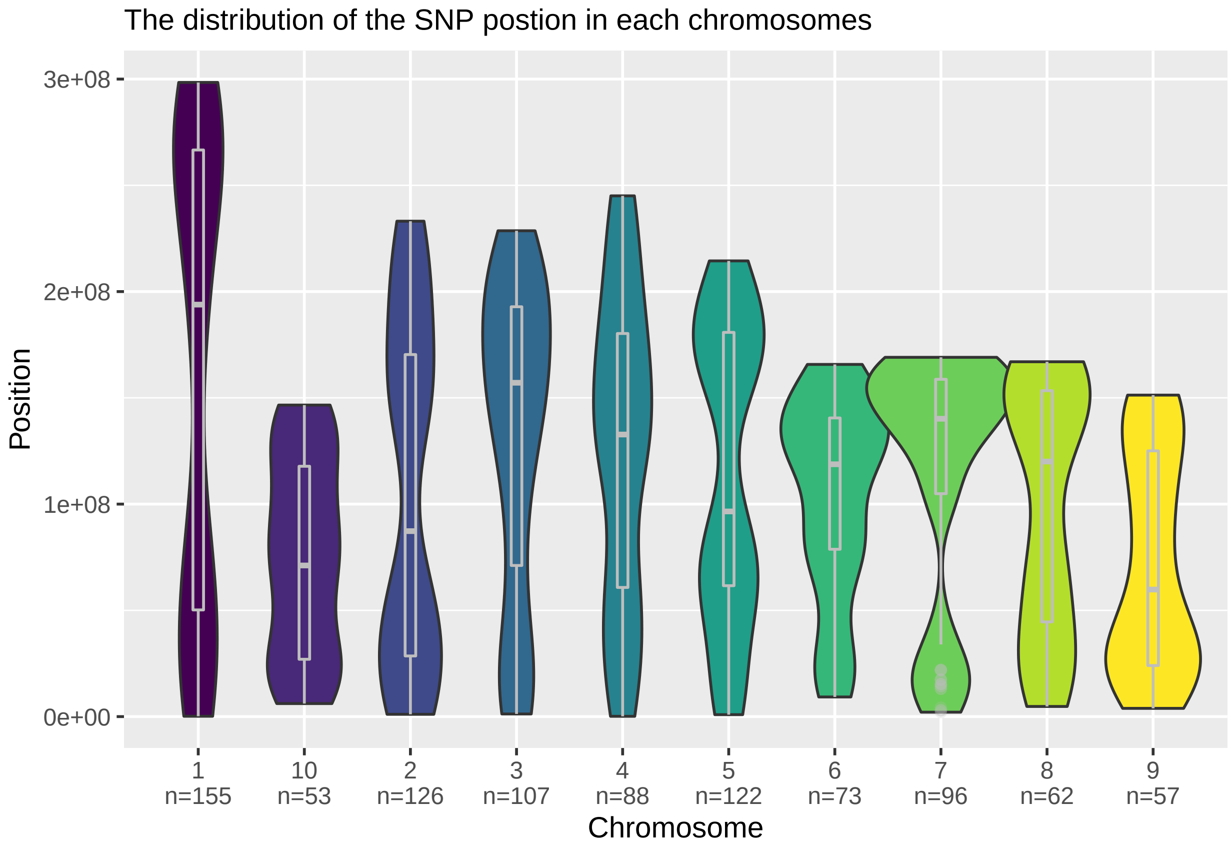
```
sample_size = posred %>% group_by(Chromosome) %>% summarize(num=n())

library(viridis)
```

```
## Loading required package: viridisLite
```

```
posred %>%
  left_join(sample_size) %>%
  mutate(myaxis = paste0(Chromosome, "\n", "n=", num)) %>%
  ggplot( aes(x=myaxis, y=Position, fill=as.character(Chromosome)))+
  geom_violin(width=1.4) +
  geom_boxplot(width=0.1, color="grey", alpha=0.2) +
  scale_fill_viridis(discrete = TRUE) +
    theme(
      legend.position="none",
      plot.title = element_text(size=11)
    ) +
    ggtitle("The distribution of the SNP postion in each chromosomes") +
    xlab("Chromosome")
```

```
## Joining, by = "Chromosome"
```


The distribution of the SNP postion in each chromosomes
Position
3e+08
2e+08
1e+08
0e+00
1
n=155
10
n=53
2
n=126
3
n=107
4
n=88
5
n=122
6
n=73
7
n=96
8
n=62
9
n=57
Chromosome