Phase 2 - HackBio Data Contest

Tasneem Juzer

This document contains exploratory data analysis results and their codes for the task given for phase 2 of the HackBio Data Contest.

Decoding and Understanding the dataset

```
#Read the data
sample <- read.table("sample.tsv", header=TRUE)

# Gives us the basic statistical info. Based on the summary statistics, median,
# minimum, maximum and mean were found for all numerical columns especially
# for MAS, Sample Size and P value.

summary(sample)
```

```
##     SNPID               RSID                CHR              POS
##  Length:25000       Length:25000       Min.   : 1.000   Min.   :    67365
##  Class :character   Class :character   1st Qu.: 4.000   1st Qu.: 31819538
##  Mode  :character   Mode  :character   Median : 8.000   Median : 71068010
##                                        Mean   : 8.571   Mean   : 79495862
##                                        3rd Qu.:13.000   3rd Qu.:115694815
##                                        Max.   :22.000   Max.   :249222450
##
##  EFFECT_ALLELE      OTHER_ALLELE       EFFECT_ALLELE_FREQ      BETA
##  Length:25000       Length:25000       Min.   :0.0000     Min.   :-1.53806
##  Class :character   Class :character   1st Qu.:0.0919     1st Qu.:-0.00539
##  Mode  :character   Mode  :character   Median :0.2670     Median :-0.00005
##                                        Mean   :0.3389     Mean   : 0.00030
##                                        3rd Qu.:0.5470     3rd Qu.: 0.00515
##                                        Max.   :1.0000     Max.   : 1.93485
##                                                           NA's   :194
##        SE                 P                 N              ANCESTRY
##  Min.   :0.00104   Min.   :0.0000   Min.   :    482   Length:25000
##  1st Qu.:0.00358   1st Qu.:0.1163   1st Qu.:  46408   Class :character
##  Median :0.00654   Median :0.3729   Median : 100692   Mode  :character
##  Mean   :0.01802   Mean   :0.4087   Mean   : 374821
##  3rd Qu.:0.00944   3rd Qu.:0.6742   3rd Qu.: 264725
##  Max.   :1.07000   Max.   :0.9999   Max.   :1597374
##  NA's   :194       NA's   :194
```

```
# unique super populations present (Ancestrors)
print(superpopulations <- unique(sample$ANCESTRY))
```

```
## [1] "AFRICAN"    "EUROPEAN"   "SOUTH_ASIA" "EAST_ASIA"  "HISPANIC"
```

```
#find how many chromosomes are present
unique(sample$CHR)
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22
```

```
# find if NA values present in our sample
print(colSums(is.na(sample)))
```

```
##              SNPID               RSID                CHR                POS
##                  0                  0                  0                  0
##      EFFECT_ALLELE      OTHER_ALLELE EFFECT_ALLELE_FREQ               BETA
##                  0                  0                  0                194
##                 SE                  P                  N           ANCESTRY
##                194                194                  0                  0
```

Number of significant SNPs in the total population and even across individual populations

For answering the first question given in the task, "How many SNPs are significant (p-value $<$ 0.01) for variability in height (MAF $>$ 0.01) in all the super populations" , the following code and plots conveys the answer.

```
data_color <- c("red", "blue", "green", "yellow", "purple")

# Filter data-  p-value < 0.01 and MAF > 0.01 for each super population
significant_snps <- data.frame()

for (sp in superpopulations) {
  subset_data <- sample[sample$ANCESTRY == sp, ]

  # Filter SNPs based on p-value and MAF
  filtered_snps <- subset_data[subset_data$ P < 0.01 &
                                 subset_data$EFFECT_ALLELE_FREQ > 0.01, ]

  # merge by rows
  significant_snps <- rbind(significant_snps, filtered_snps)
}

# Count the number of significant SNPs in all super populations
num_of_significant_snps <- nrow(significant_snps)
num_of_snps <- nrow(sample)

#or

length(significant_snps$SNPID)
```

```
## [1] 2253
```

```
#Answer for the 1st question given:
cat("The number of SNPs that are significant (p-value < 0.01) for
    variability in height (MAF > 0.01) in all the super populations are "
    ,num_of_significant_snps, "\n")
```

```
## The number of SNPs that are significant (p-value < 0.01) for
##     variability in height (MAF > 0.01) in all the super populations are  2253
```

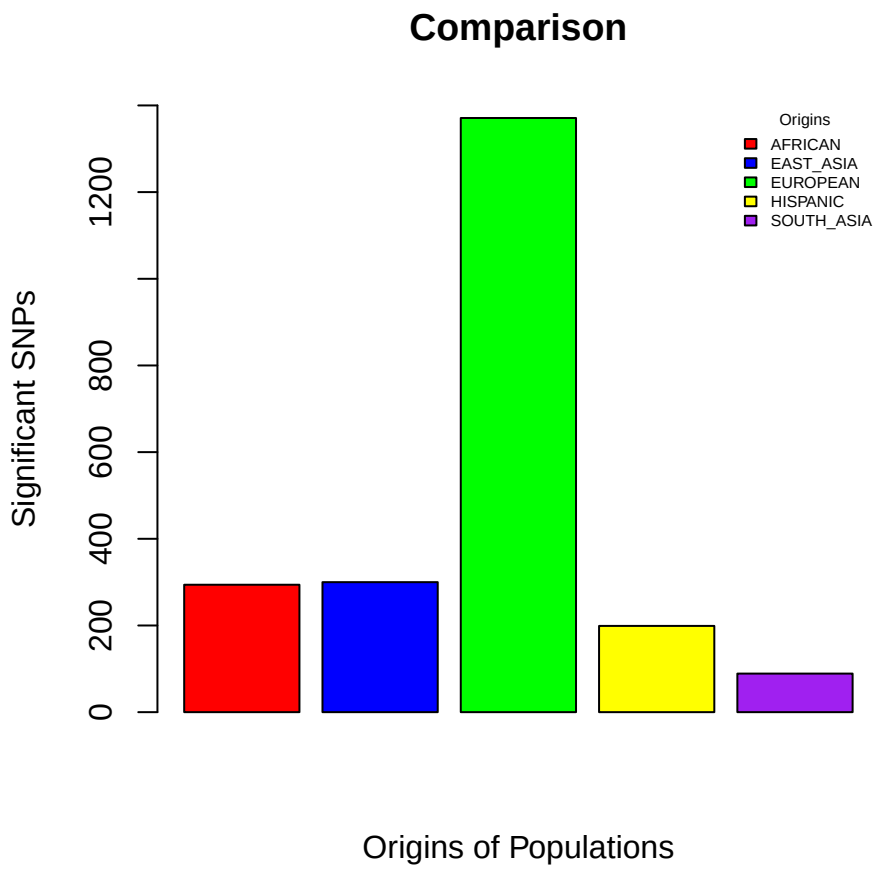
```
cat("Total number of significant SNPs vs Normal: ", num_of_significant_snps, ":" , num_of_snps,"\n")
```

```
## Total number of significant SNPs vs Normal:  2253 : 25000
```

```
# No. of significant SNPs per population:
print( significant_snps_per_super_population <- table(significant_snps$ANCESTRY))
```

```
##
##    AFRICAN  EAST_ASIA   EUROPEAN   HISPANIC SOUTH_ASIA
##        294        300       1371        199         89
```

```
#visual representation of the above data
pop<- c("AFRICAN" , "EAST_ASIA",  "EUROPEAN"  , "HISPANIC", "SOUTH_ASIA")

par(mfrow = c (1,1))
barplot(significant_snps_per_super_population, col = c("red", "blue", "green", "yellow", "purple"),
        xlab = "Origins of Populations",
        ylab = "Significant SNPs", xaxt = "n", ylim = c(0,1400),
        main = "Comparison")
legend("topright", legend = pop , fill = c("red", "blue", "green", "yellow", "purple"),
       title = "Origins", xjust = 1, yjust = 0, bty = "n", cex = 0.5)
```



Answer - The total number of significant SNPs are 2253

with Europe having the highest - 1371, and South Asia having the lowest - 89.

Question 2

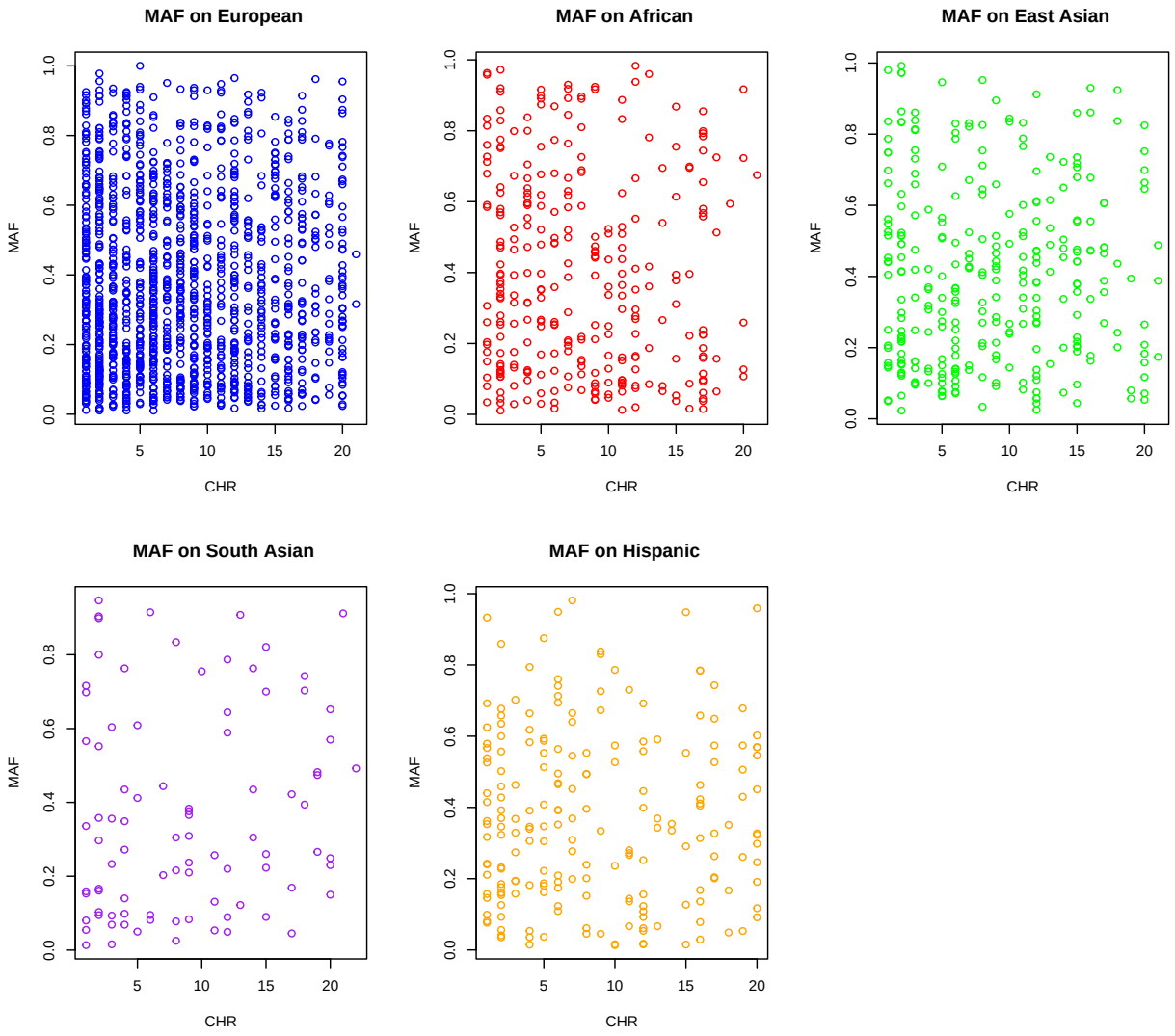
Europeans genetic variability can/cannot be found in other super populations

```
#better understanding of the new subset of data
summary(significant_snps)
```

```
##     SNPID              RSID               CHR              POS
##  Length:2253        Length:2253        Min.   : 1.000   Min.   :   123233
##  Class :character   Class :character   1st Qu.: 3.000   1st Qu.: 34383788
##  Mode  :character   Mode  :character   Median : 7.000   Median : 70051966
##                                        Mean   : 8.295   Mean   : 79655840
##                                        3rd Qu.:12.000   3rd Qu.:114226535
##                                        Max.   :22.000   Max.   :247010734
##  EFFECT_ALLELE      OTHER_ALLELE       EFFECT_ALLELE_FREQ      BETA
##  Length:2253        Length:2253        Min.   :0.0105     Min.   :-0.397104
```

```
##  Class :character   Class :character   1st Qu.:0.1590     1st Qu.:-0.010464
##  Mode  :character   Mode  :character   Median :0.3350     Median :-0.003257
##                                        Mean   :0.3799     Mean   :-0.001079
##                                        3rd Qu.:0.5720     3rd Qu.: 0.009352
##                                        Max.   :1.0000     Max.   : 0.081540
##        SE                  P                   N             ANCESTRY
##  Min.   :0.001040   Min.   :0.000e+00   Min.   :  21912   Length:2253
##  1st Qu.:0.001170   1st Qu.:1.210e-07   1st Qu.: 107061   Class :character
##  Median :0.001790   Median :2.433e-04   Median :1573790   Mode  :character
##  Mean   :0.003465   Mean   :1.840e-03   Mean   :1007854
##  3rd Qu.:0.004810   3rd Qu.:2.841e-03   3rd Qu.:1593869
##  Max.   :0.137000   Max.   :9.997e-03   Max.   :1597373
```

```r
#subsetting based on populations in the unfiltered dataset
african <- subset(sample, ANCESTRY == "AFRICAN")
european <- subset(sample, ANCESTRY == "EUROPEAN")
south_asia <- subset(sample, ANCESTRY == "SOUTH_ASIA")
east_asia <- subset(sample, ANCESTRY == "EAST_ASIA")
hispanic <- subset(sample, ANCESTRY == "HISPANIC")

#subsetting based on populations in the filtered dataset
sign_snps_african <- subset(significant_snps, ANCESTRY == "AFRICAN")
sign_snps_european <- subset(significant_snps, ANCESTRY == "EUROPEAN")
sign_snps_south_asia <- subset(significant_snps, ANCESTRY == "SOUTH_ASIA")
sign_snps_east_asia <- subset(significant_snps, ANCESTRY == "EAST_ASIA")
sign_snps_hispanic <- subset(significant_snps, ANCESTRY == "HISPANIC")

# remove europe alone
others_sign_snp <- significant_snps [significant_snps$ANCESTRY != "EUROPEAN", ]

# since european pop has highest number of filtered significant data (in human height),
# these plots shows how much MAF affects in all chromosomes across all the given populations.

par(mfrow = c (2,3))

plot( sign_snps_european$CHR, sign_snps_european$EFFECT_ALLELE_FREQ, type = "p",
      col = "blue",xlab = "CHR", ylab = "MAF", main = " MAF on European")
plot( sign_snps_african$CHR, sign_snps_african$EFFECT_ALLELE_FREQ, type = "p",
      col = "red",xlab = "CHR", ylab = "MAF", main = " MAF on African")
plot( sign_snps_east_asia$CHR, sign_snps_east_asia$EFFECT_ALLELE_FREQ, type = "p",
      col = "green",xlab = "CHR", ylab = "MAF", main = " MAF on East Asian")
plot( sign_snps_south_asia$CHR, sign_snps_south_asia$EFFECT_ALLELE_FREQ, type = "p",
      col = "purple", xlab = "CHR", ylab = "MAF", main = " MAF on South Asian")
plot( sign_snps_hispanic$CHR, sign_snps_hispanic$EFFECT_ALLELE_FREQ, type = "p",
      col = "orange", xlab = "CHR", ylab = "MAF", main = " MAF on Hispanic")
```

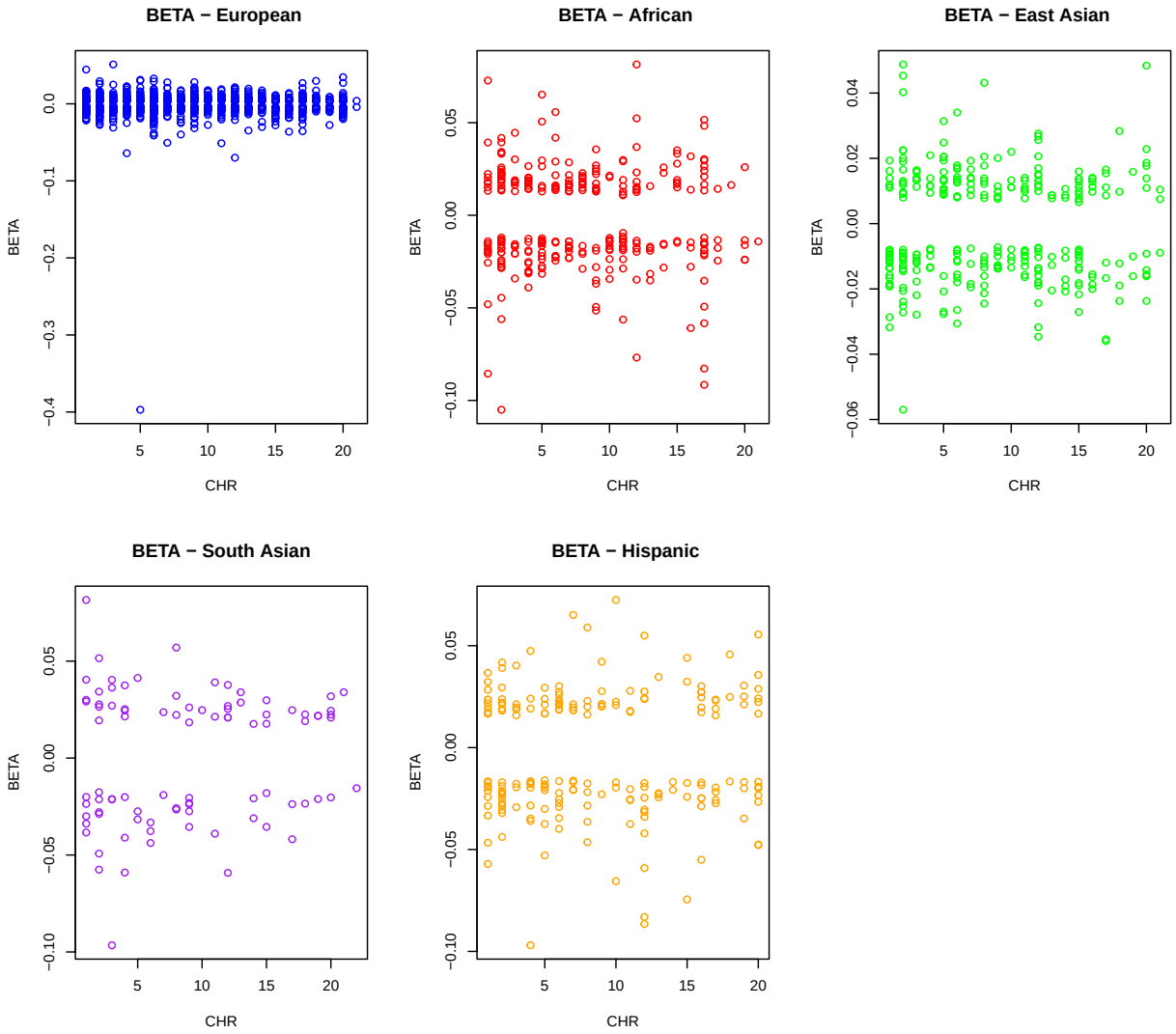


The above graph depicts that European population have the highest spread of MAFs and South Asian populations have the lowest spread of MAFS across all the Chromosomes, majority surpassing the MAF median value of 0.3350.

Relationship between BETA values and Chromosomes across all populations

```
par(mfrow = c (2,3))

plot( sign_snps_european$CHR, sign_snps_european$BETA, type = "p", col = "blue",
      xlab = "CHR", ylab = "BETA", main = " BETA - European")
plot( sign_snps_african$CHR, sign_snps_african$BETA, type = "p", col = "red",
      xlab = "CHR", ylab = "BETA", main = " BETA - African")
plot( sign_snps_east_asia$CHR, sign_snps_east_asia$BETA, type = "p", col = "green",
      xlab = "CHR", ylab = "BETA", main = " BETA - East Asian")
```

```r
plot( sign_snps_south_asia$CHR, sign_snps_south_asia$BETA, type = "p", col = "purple",
      xlab = "CHR", ylab = "BETA", main = "BETA - South Asian")
plot( sign_snps_hispanic$CHR, sign_snps_hispanic$BETA, type = "p", col = "orange",
      xlab = "CHR", ylab = "BETA", main = " BETA - Hispanic")
```



Performing statistical analysis:

Null hypothesis - Europeans' genetic variability can be found in other super populations

- no significant difference

Alternative hypothesis - Europeans' genetic variability cannot be found in other super populations

- significant difference

```
#Performing t test
# chosen alpha value 5%, i.e) 0.05%

t.test(sign_snps_european$EFFECT_ALLELE_FREQ, others_sign_snp$EFFECT_ALLELE_FREQ,
       paired = F, var.equal =F )
```

```
##
##  Welch Two Sample t-test
##
## data:  sign_snps_european$EFFECT_ALLELE_FREQ and others_sign_snp$EFFECT_ALLELE_FREQ
## t = -1.7762, df = 1826.2, p-value = 0.07587
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.04149409  0.00205483
## sample estimates:
## mean of x mean of y
## 0.3721338 0.3918534
```

```
t.test(sign_snps_european$BETA, others_sign_snp$BETA, paired = F, var.equal =F )
```

```
##
##  Welch Two Sample t-test
##
## data:  sign_snps_european$BETA and others_sign_snp$BETA
## t = 0.74064, df = 1263.5, p-value = 0.4591
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.001171384  0.002592225
## sample estimates:
##     mean of x     mean of y
## -0.0008011157 -0.0015115366
```

```
#BETA divided by SE - Effect Size

ES_E <- sign_snps_european$BETA / sign_snps_european$SE
ES_O <- others_sign_snp$BETA / others_sign_snp$SE
t.test(ES_E, ES_O, paired = F, var.equal =F )
```

```
##
##  Welch Two Sample t-test
```

```
##
## data:  ES_E and ES_O
## t = -0.94908, df = 2168, p-value = 0.3427
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.6825874  0.2373659
## sample estimates:
##  mean of x  mean of y
## -0.3766562 -0.1540455
```

```
#in all the above t tests, p-values are greater than alpha, which results
# in the acceptance of the null hypothesis indicating that
# Europeans' genetic variability can be found in other super populations

#t- values are also lesser than 1 in both, indicating difference between
# them is not statistically significant, hence no significant
# difference between the samples.
```

```
# ANOVA Tests
aov(BETA ~ ANCESTRY, data = significant_snps)
```

```
## Call:
##    aov(formula = BETA ~ ANCESTRY, data = significant_snps)
##
## Terms:
##                   ANCESTRY Residuals
## Sum of Squares  0.0014951 0.8932062
## Deg. of Freedom         4      2248
##
## Residual standard error: 0.01993323
## Estimated effects may be unbalanced
```

```
ANOVA_BETA <- aov(BETA ~ ANCESTRY, data = significant_snps)
summary(ANOVA_BETA)
```

```
##               Df Sum Sq   Mean Sq F value Pr(>F)
## ANCESTRY       4 0.0015 0.0003738   0.941  0.439
## Residuals   2248 0.8932 0.0003973
```

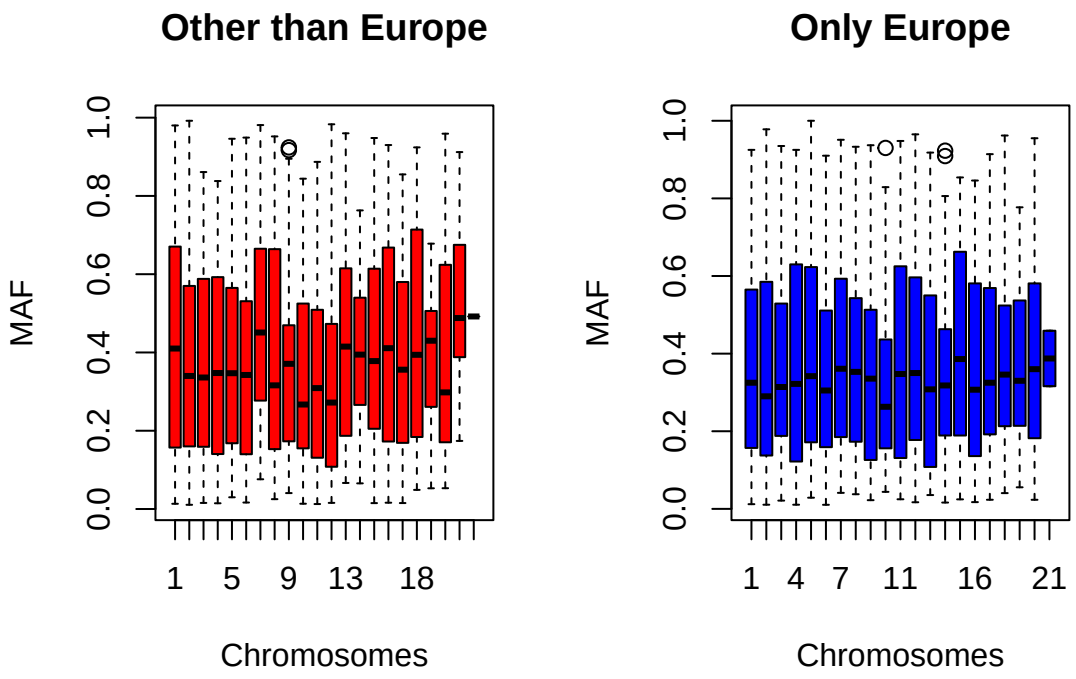
```
aov(EFFECT_ALLELE_FREQ ~ ANCESTRY, data = significant_snps)
```

```
## Call:
##    aov(formula = EFFECT_ALLELE_FREQ ~ ANCESTRY, data = significant_snps)
##
## Terms:
##                   ANCESTRY Residuals
## Sum of Squares    0.36848 146.27928
## Deg. of Freedom         4      2248
##
## Residual standard error: 0.2550899
## Estimated effects may be unbalanced
```

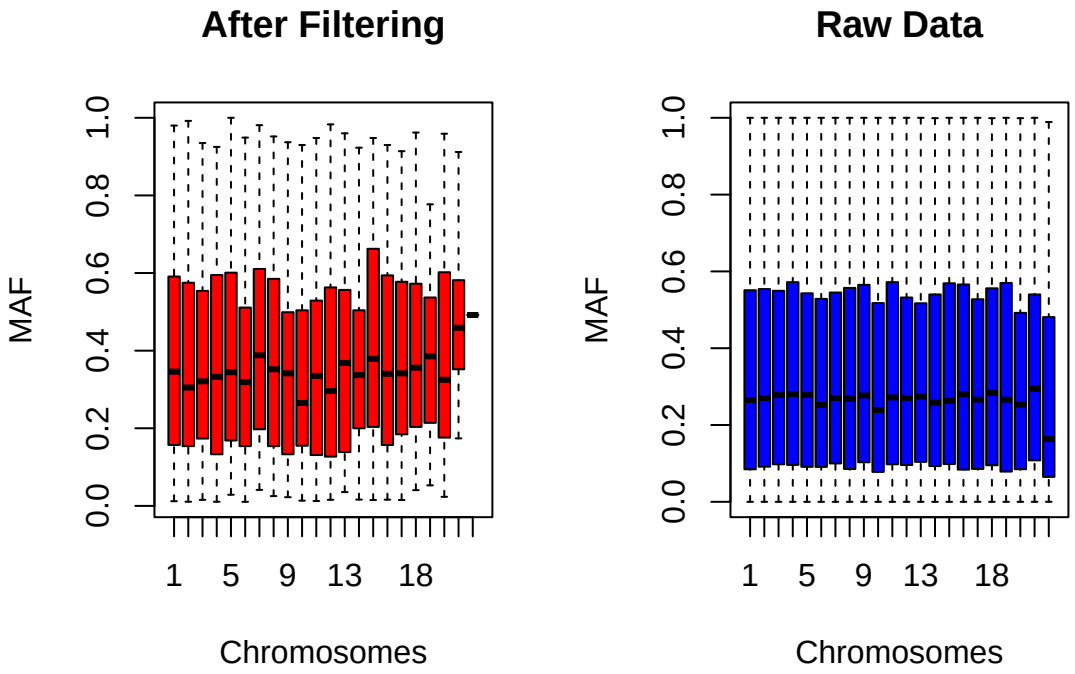
```r
ANOVA_MAF <- aov(EFFECT_ALLELE_FREQ ~ ANCESTRY, data = significant_snps)
summary(ANOVA_MAF)
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## ANCESTRY        4   0.37 0.09212   1.416  0.226
## Residuals    2248 146.28 0.06507
```

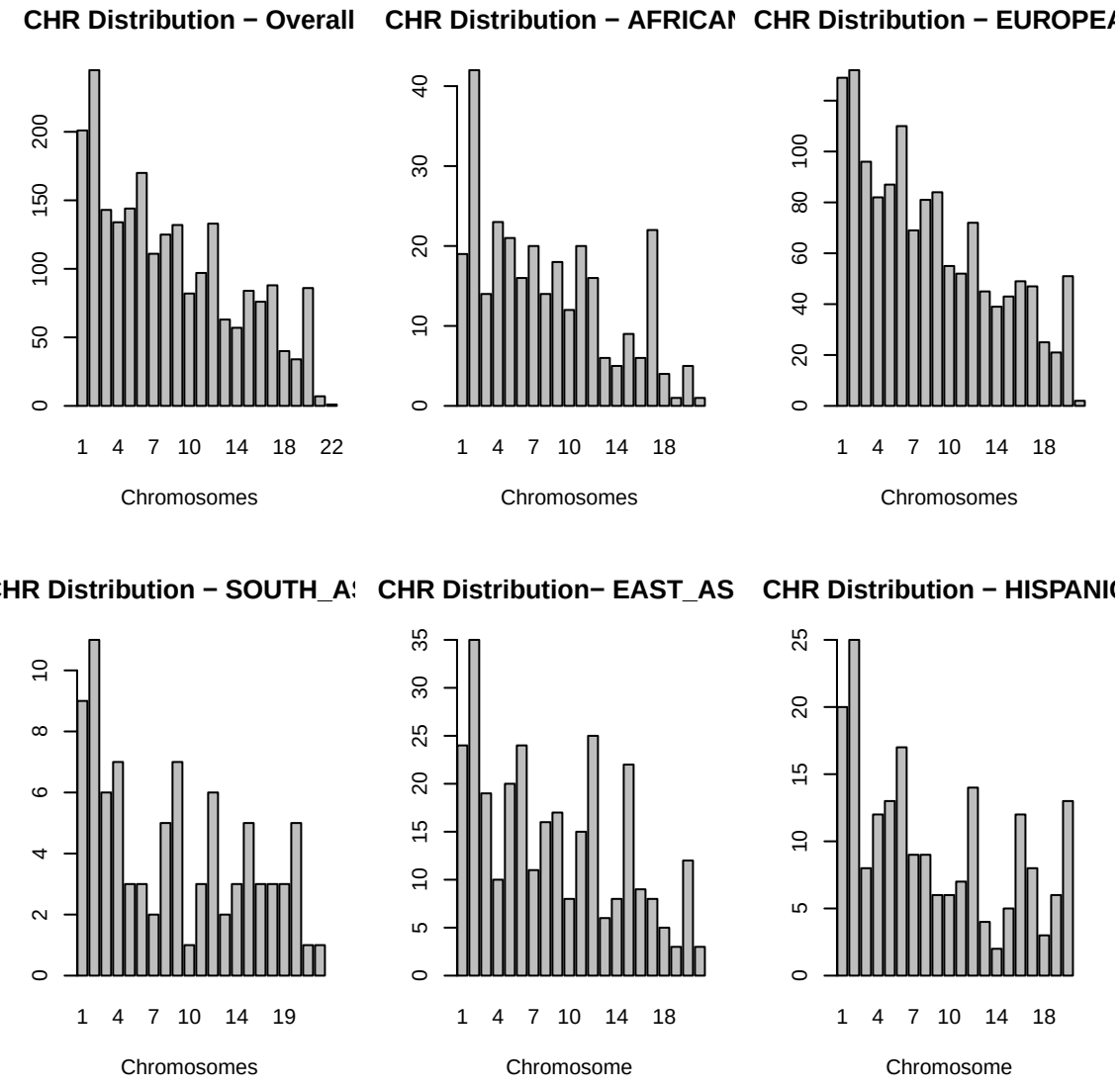
```r
par(mfrow = c (1,2))
boxplot(others_sign_snp$EFFECT_ALLELE_FREQ ~ others_sign_snp$CHR , col = "red",
        main = "Other than Europe", ylab = "MAF", xlab = "Chromosomes")
boxplot(sign_snps_european$EFFECT_ALLELE_FREQ ~ sign_snps_european$CHR , col = "blue",
        main = "Only Europe", ylab = "MAF", xlab = "Chromosomes")
```




```r
# normal data MAF vs significant data MAF-how much filtering  matters

par(mfrow = c (1,2))
boxplot(significant_snps$EFFECT_ALLELE_FREQ ~ significant_snps$CHR , col = "red",
        main = "After Filtering", ylab = "MAF", xlab = "Chromosomes")
boxplot(sample$EFFECT_ALLELE_FREQ ~ sample$CHR , col = "blue",
        ylab = "MAF", xlab = "Chromosomes", main = "Raw Data" )
```




```
# chr distribution in significant dataset
par(mfrow = c (2,3))
barplot(table(significant_snps$CHR), main = "CHR Distribution - Overall",
        xlab = "Chromosomes", xlim = c(1,22))
barplot(table(sign_snps_african$CHR), main = "CHR Distribution - AFRICAN ",
        xlab = "Chromosomes", xlim = c(1,22))
barplot(table(sign_snps_european$CHR), main = "CHR Distribution - EUROPEAN",
        xlab = "Chromosomes", xlim = c(1,22))
barplot(table(sign_snps_south_asia$CHR), main = "CHR Distribution - SOUTH_ASIA",
        xlab = "Chromosomes", xlim = c(1,22))
barplot(table(sign_snps_east_asia$CHR), main = "CHR Distribution- EAST_ASIA",
        xlab = "Chromosome", xlim = c(1,22))
barplot(table(sign_snps_hispanic$CHR), main = "CHR Distribution - HISPANIC ",
        xlab = "Chromosome", xlim = c(1,22))
```



The plot generated shows that maximum number of data distribution is in the 2nd chromosome. The total range in y-axis is 200, out of which 100 is by European Population.

Higher value of MAF is related to Genetic Variability

the following code proves that European population has more Minor allele frequency in their data having SNPs with MAFs value greater than the median value

```r
#get the median value from here
summary(sign_snps_european$EFFECT_ALLELE_FREQ)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0105  0.1580  0.3220  0.3721  0.5635  1.0000
```

```r
#MAF greater than median in the european population
print( MAF_E <- table(sign_snps_european$EFFECT_ALLELE_FREQ > 0.2670 ))
```

```
##
## FALSE  TRUE
##   573   798
```

```r
#get the median value from here
summary(others_sign_snp$EFFECT_ALLELE_FREQ)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0108  0.1610  0.3560  0.3919  0.5847  0.9920
```

```r
#how many MAF greater than median in the european population
print( MAF_Others <- table(others_sign_snp$EFFECT_ALLELE_FREQ > 0.3560 ))
```

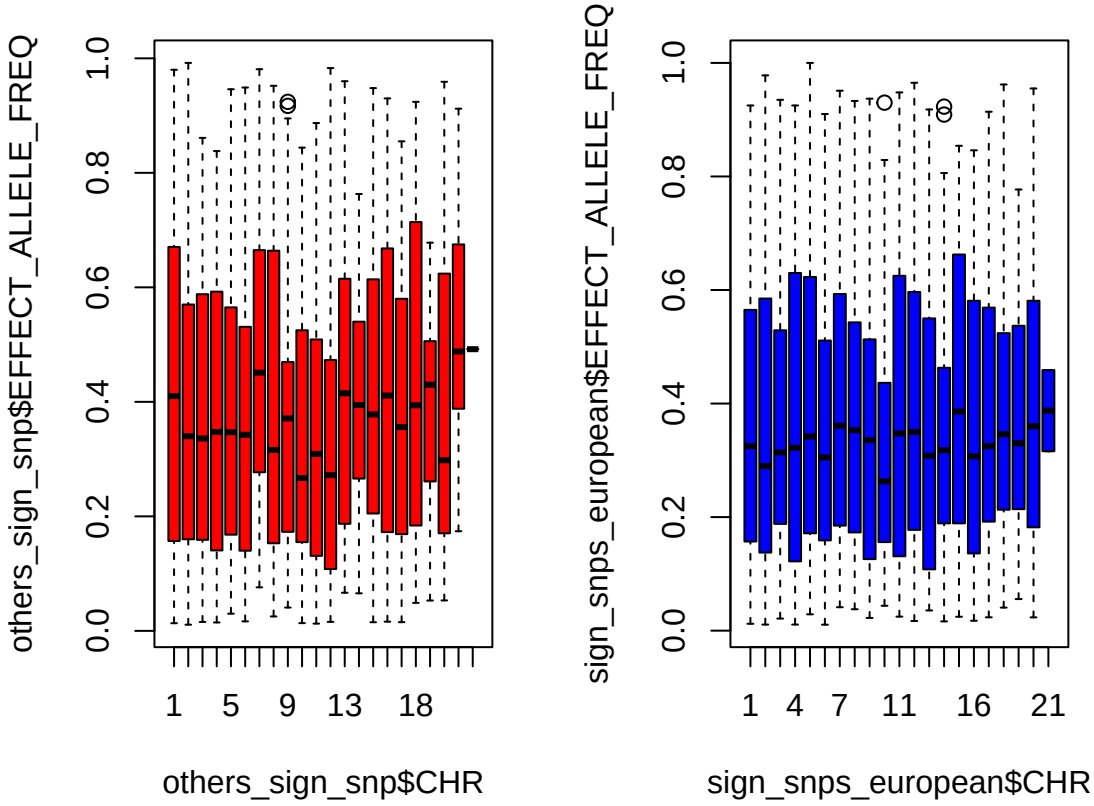
```
##
## FALSE  TRUE
##   442   440
```

```r
total_MAF_SP <- length(others_sign_snp$EFFECT_ALLELE_FREQ)

total_MAF_E <- length(sign_snps_european$EFFECT_ALLELE_FREQ)
```

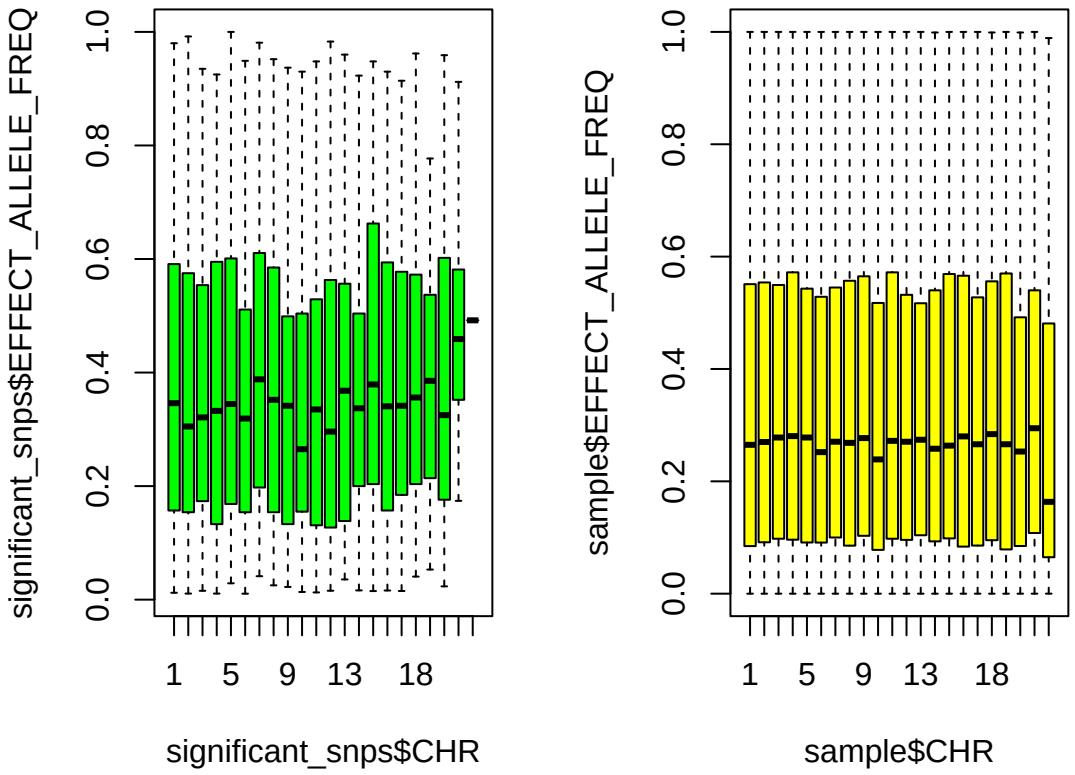
MAFs value greater than thier median value were found to be

higher in the European population

```r
# since european pop has highest number of MAF, this shows how much they are significant
par(mfrow = c (1,2))
boxplot(others_sign_snp$EFFECT_ALLELE_FREQ ~ others_sign_snp$CHR , col = "red")
boxplot(sign_snps_european$EFFECT_ALLELE_FREQ ~ sign_snps_european$CHR , col = "blue")
```




```
# normal data MAF vs significant data MAF
par(mfrow = c (1,2))
boxplot(significant_snps$EFFECT_ALLELE_FREQ ~ significant_snps$CHR , col = "green")
boxplot(sample$EFFECT_ALLELE_FREQ ~ sample$CHR , col = "yellow")
```



###understanding both unfiltered and filtered datasets:###

```
#Unfiltered: Original (sample)
#to understand the data (mean, min, max, median)
summary(sample$N)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     482   46408  100692  374821  264725 1597374
```

```
summary(sample$P)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  0.1163  0.3729  0.4087  0.6742  0.9999     194
```

```
summary(sample$POS)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##     67365  31819538  71068010  79495862 115694815 249222450
```

```r
summary(sample$BETA)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
## -1.53806 -0.00539 -0.00005  0.00030  0.00515  1.93485      194
```

```r
summary(sample$SE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## 0.00104 0.00358 0.00654 0.01802 0.00944 1.07000     194
```

```r
#to get the count of each
table(sample$EFFECT_ALLELE)
```

```
##
##    A    C    G    T
## 6410 6245 6113 6232
```

```r
table(sample$OTHER_ALLELE)
```

```
##
##    A    C    G    T
## 6004 6438 6471 6087
```

```r
table(sample$CHR)
```

```
##
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16
## 2084 2185 1632 1630 1407 1675 1384 1470 1330 1455 1230 1232  953  893  746  870
##   17   18   19   20   21   22
##  687  657  447  745  206   82
```

```r
table(sample$ANCESTRY)
```

```
##
##    AFRICAN  EAST_ASIA   EUROPEAN   HISPANIC SOUTH_ASIA
##       5000       5000       5000       5000       5000
```

Sample Size Distribution

```r
# sample size distribution among different origins
par(mfrow = c (2,3))
boxplot(sample$N, notch = TRUE, main = "Plot for sample size- overall")
boxplot(african$N, notch = TRUE, main = "Plot for sample size in Africa")
```

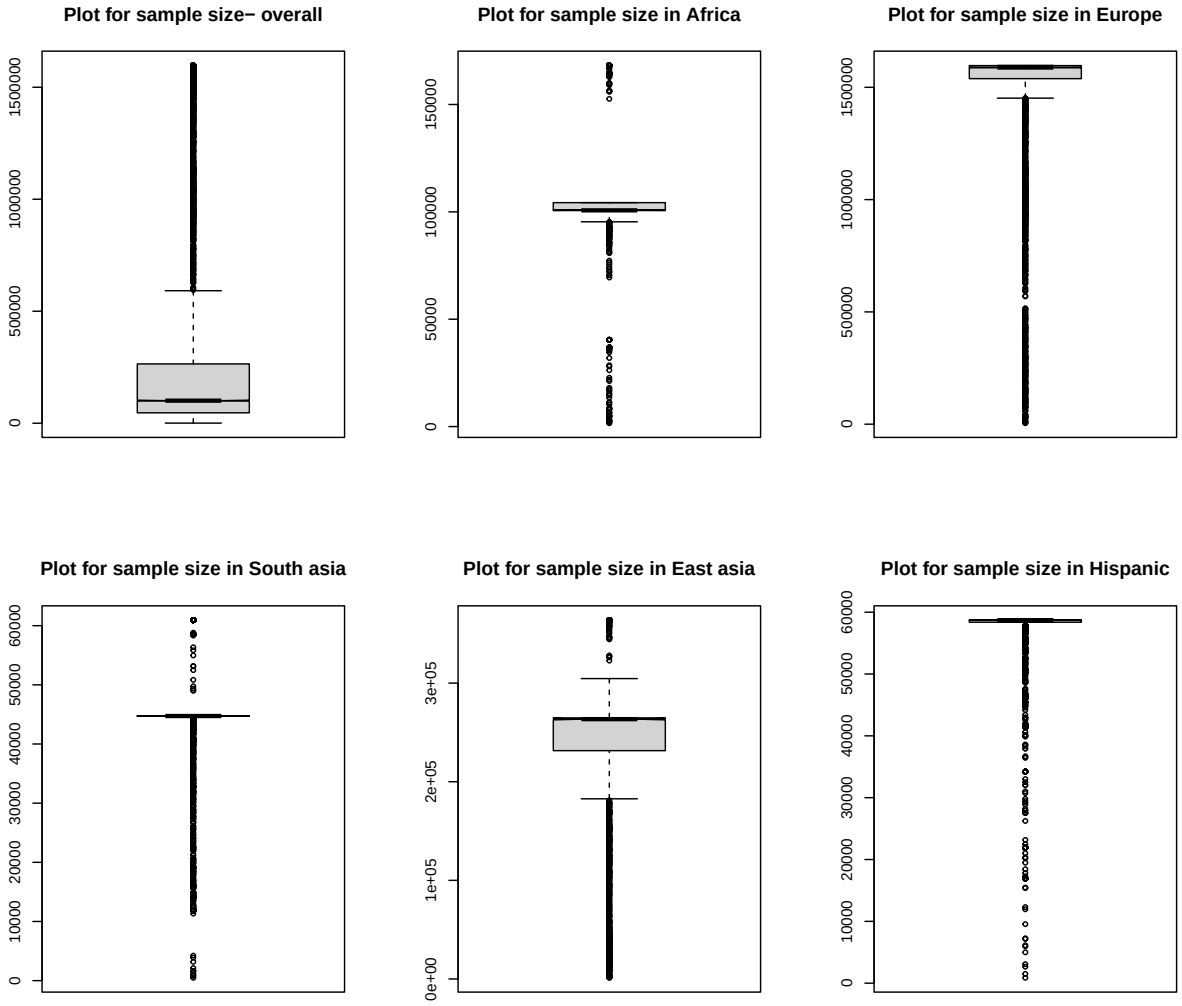
```
## Warning in (function (z, notch = FALSE, width = NULL, varwidth = FALSE, : some
## notches went outside hinges ('box'): maybe set notch=FALSE
```

```
boxplot(european$N, notch = TRUE, main = "Plot for sample size in Europe")
boxplot(south_asia$N, notch = TRUE, main = "Plot for sample size in South asia")
```

```
## Warning in (function (z, notch = FALSE, width = NULL, varwidth = FALSE, : some
## notches went outside hinges ('box'): maybe set notch=FALSE
```

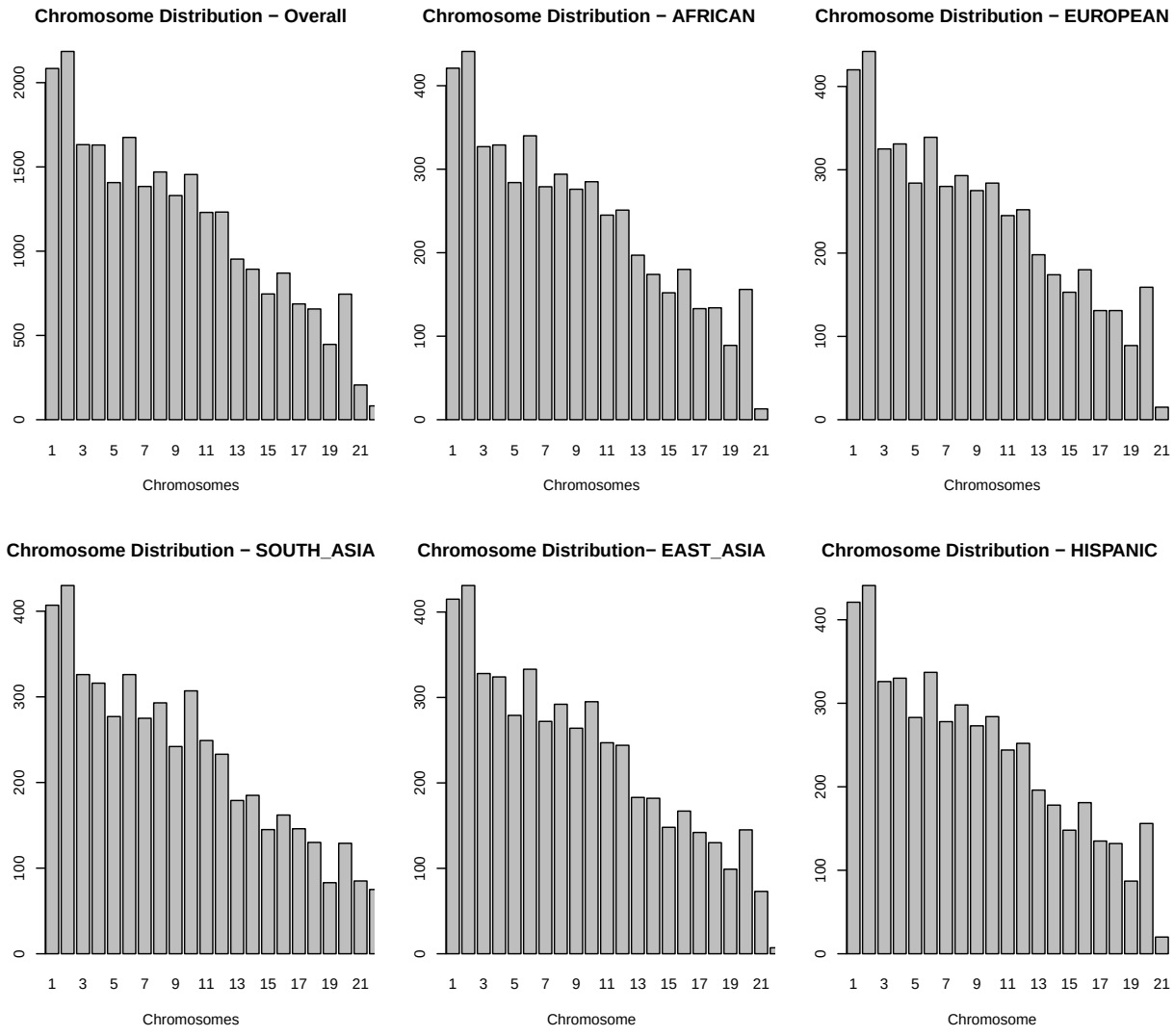
```
boxplot(east_asia$N, notch = TRUE, main = "Plot for sample size in East asia")
boxplot(hispanic$N,notch = TRUE, main = "Plot for sample size in Hispanic")
```

```
## Warning in (function (z, notch = FALSE, width = NULL, varwidth = FALSE, : some
## notches went outside hinges ('box'): maybe set notch=FALSE
```



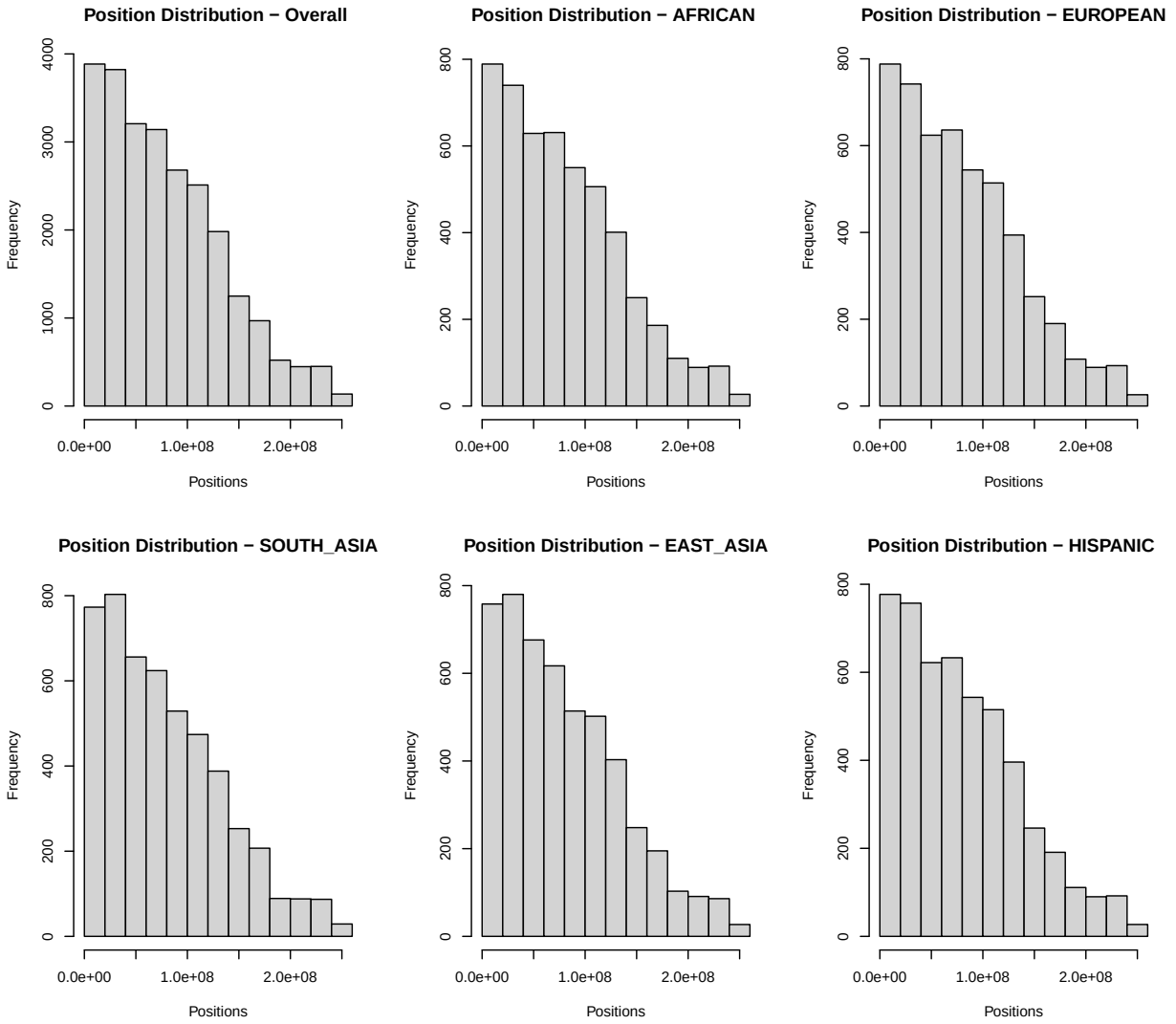
Chromosome distribution in the raw data

```r
#Chromosome distribution across different origins
par(mfrow = c (2,3))
barplot(table(sample$CHR), main = "Chromosome Distribution - Overall",
        xlab = "Chromosomes", xlim = c(1,22))
barplot(table(african$CHR), main = "Chromosome Distribution - AFRICAN ",
        xlab = "Chromosomes", xlim = c(1,22))
barplot(table(european$CHR), main = "Chromosome Distribution - EUROPEAN",
        xlab = "Chromosomes", xlim = c(1,22))
barplot(table(south_asia$CHR), main = "Chromosome Distribution - SOUTH_ASIA",
        xlab = "Chromosomes", xlim = c(1,22))
barplot(table(east_asia$CHR), main = "Chromosome Distribution- EAST_ASIA",
        xlab = "Chromosome", xlim = c(1,22))
barplot(table(hispanic$CHR), main = "Chromosome Distribution - HISPANIC ",
        xlab = "Chromosome", xlim = c(1,22))
```



Position distribution

```
# Position distribution across different origins
par(mfrow = c (2,3))
hist(sample$POS, main = "Position Distribution - Overall ", xlab = "Positions")
hist(african$POS, main = "Position Distribution - AFRICAN ", xlab = "Positions")
hist(european$POS, main = "Position Distribution - EUROPEAN", xlab = "Positions")
hist(south_asia$POS, main = "Position Distribution - SOUTH_ASIA", xlab = "Positions")
hist(east_asia$POS, main = "Position Distribution - EAST_ASIA", xlab = "Positions")
hist(hispanic$POS, main = "Position Distribution - HISPANIC ", xlab = "Positions")
```



the below code results in the creation of dataframe containing MAFs value greater than 0.95, hence indicating that the following SNPs are highly responsible in Genetic Variation (more common and has been present in the population for a longer time).

```
imp <- subset(significant_snps ,significant_snps$EFFECT_ALLELE_FREQ > 0.95 )
summary(imp)
```

```
##     SNPID               RSID                CHR              POS
##  Length:19          Length:19          Min.   : 1.000   Min.   :  3737495
##  Class :character   Class :character   1st Qu.: 2.000   1st Qu.: 40990906
##  Mode  :character   Mode  :character   Median : 5.000   Median : 64692041
##                                        Mean   : 7.211   Mean   : 92057998
##                                        3rd Qu.:12.000   3rd Qu.:135323988
##                                        Max.   :20.000   Max.   :227503125
##  EFFECT_ALLELE      OTHER_ALLELE       EFFECT_ALLELE_FREQ      BETA
##  Length:19          Length:19          Min.   :0.9510     Min.   :-0.39710
##  Class :character   Class :character   1st Qu.:0.9585     1st Qu.:-0.03348
##  Mode  :character   Mode  :character   Median :0.9650     Median : 0.00745
##                                        Mean   :0.9691     Mean   :-0.01870
##                                        3rd Qu.:0.9790     3rd Qu.: 0.02245
##                                        Max.   :1.0000     Max.   : 0.06510
##        SE                 P                   N             ANCESTRY
##  Min.   :0.00255   Min.   :4.100e-08   Min.   :  58371   Length:19
##  1st Qu.:0.00321   1st Qu.:2.221e-03   1st Qu.: 104293   Class :character
##  Median :0.00983   Median :3.520e-03   Median : 262846   Mode  :character
##  Mean   :0.01638   Mean   :3.897e-03   Mean   : 595950
##  3rd Qu.:0.01475   3rd Qu.:6.866e-03   3rd Qu.:1451252
##  Max.   :0.13700   Max.   :8.413e-03   Max.   :1597370
```

```
# a unique dataframe comtaining MAF values greater than 0.95
MAF_above_0.95 <- data.frame( Origin = imp$ANCESTRY,MAF = imp$EFFECT_ALLELE_FREQ,
imp = imp$SNPID , RSID = imp$RSID, BETA = imp$BETA, sample_size = imp$N )
```

Credits goes to HackBio for the datasets and organizing the data contest

Citation: Yengo, L., Vedantam, S., Marouli, E. et al. A saturated map of common genetic variants associated with human height. Nature 610, 704–712 (2022). https://doi.org/10.1038/s41586-022-05275-y

Thank you HackBio for the opportunity