PROJECT 2

Code

GOODNESS NWOKEBU

1.0: Load the data

Hide

```
#Sample code to import the dataset in R
yengoHeight <- "https://raw.githubusercontent.com/HackBio-Internship/public_datasets/main/R/datasets/Contests/humanGeneticVariationsSamples.tsv"
yengoHeight <- read.table(yengoHeight)
head(yengoHeight)
```

	SNPID <chr>	RSID <chr>	C.. <int>	POS <int>	EFFECT_ALL... <chr>	OTHER_ALL... <chr>	EFFECT_ALLEI
15885	1:32296525:C:T	rs4949473	1	32296525	C	T	
23949	1:49121231:A:G	rs319993	1	49121231	A	G	
73516	1:171155103:C:T	rs6657314	1	171155103	T	C	
77457	1:179183766:C:T	rs2816213	1	179183766	C	T	
15298	1:31139078:A:C	rs1983822	1	31139078	C	A	
67183	1:161014649:A:G	rs1556259	1	161014649	G	A	

6 rows | 1-8 of 12 columns

Hide

```
class(yengoHeight)
```

```
[1] "data.frame"
```

Hide

```
##Data Preprocessing

summary(yengoHeight)
```

```
   SNPID              RSID              CHR                POS
Length:24806       Length:24806       Length:24806       Min.   :    67365
Class :character   Class :character   Class :character   1st Qu.: 31782304
Mode  :character   Mode  :character   Mode  :character   Median : 71128448
                                                         Mean   : 79520031
                                                         3rd Qu.:115715089
                                                         Max.   :249222450
EFFECT_ALLELE      OTHER_ALLELE       EFFECT_ALLELE_FREQ      BETA
Length:24806       Length:24806       Min.   :0.000017   Min.   :-1.5380600
Class :character   Class :character   1st Qu.:0.095425   1st Qu.:-0.0053944
Mode  :character   Mode  :character   Median :0.270000   Median :-0.0000528
                                      Mean   :0.341546   Mean   : 0.0002959
                                      3rd Qu.:0.549750   3rd Qu.: 0.0051544
                                      Max.   :1.000000   Max.   : 1.9348500
      SE                P                N             ANCESTRY
Min.   :0.00104   Min.   :0.0000   Min.   :    482   Length:24806
1st Qu.:0.00358   1st Qu.:0.1163   1st Qu.:  53717   Class :character
Median :0.00654   Median :0.3729   Median : 100692   Mode  :character
Mean   :0.01802   Mean   :0.4087   Mean   : 377682
3rd Qu.:0.00944   3rd Qu.:0.6742   3rd Qu.: 264725
Max.   :1.07000   Max.   :0.9999   Max.   :1597374
    CHRI
Length:24806
Class :character
Mode  :character
```

1.1 Data Cleaning

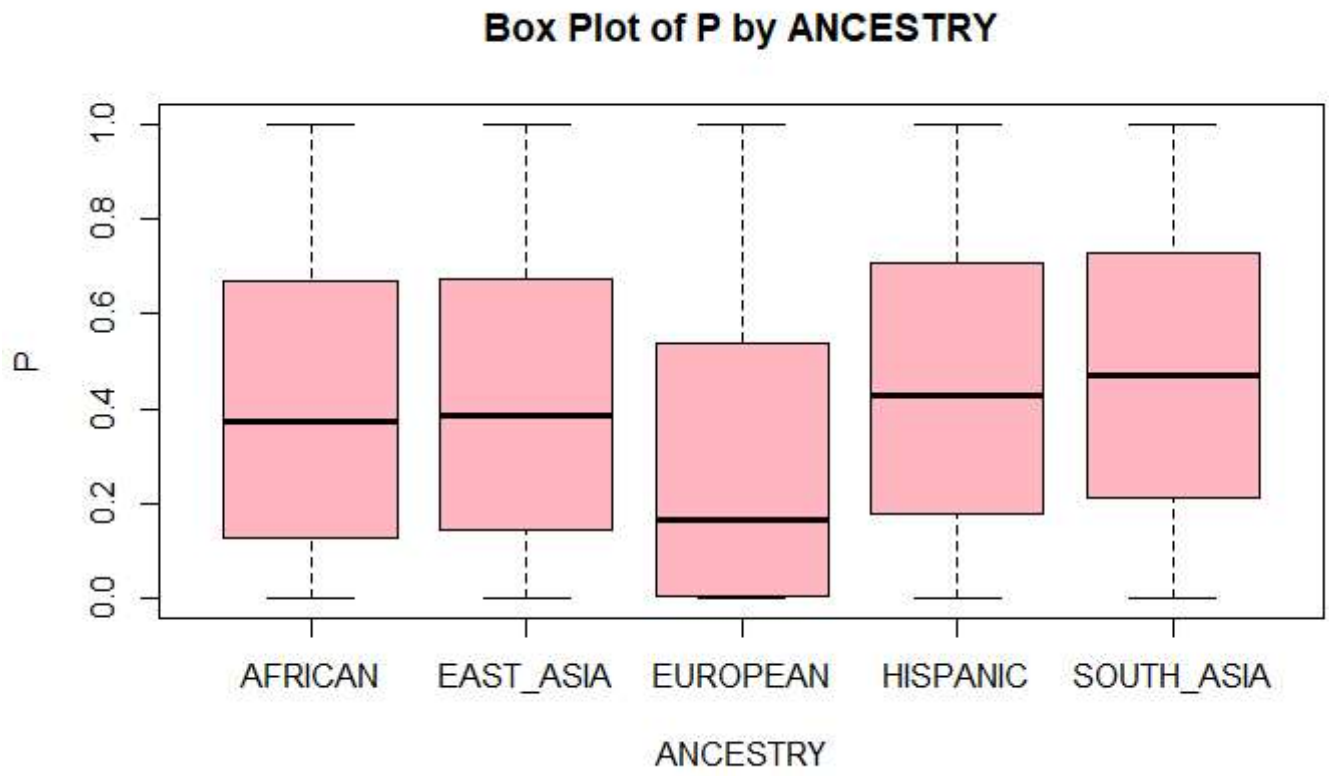
Hide

```
# Removing Null set
yengoHeight <- na.omit(yengoHeight)
#inappropriate data types
yengoHeight$CHR <- as.character(yengoHeight$CHR)
```

1.2 Data Visualisation

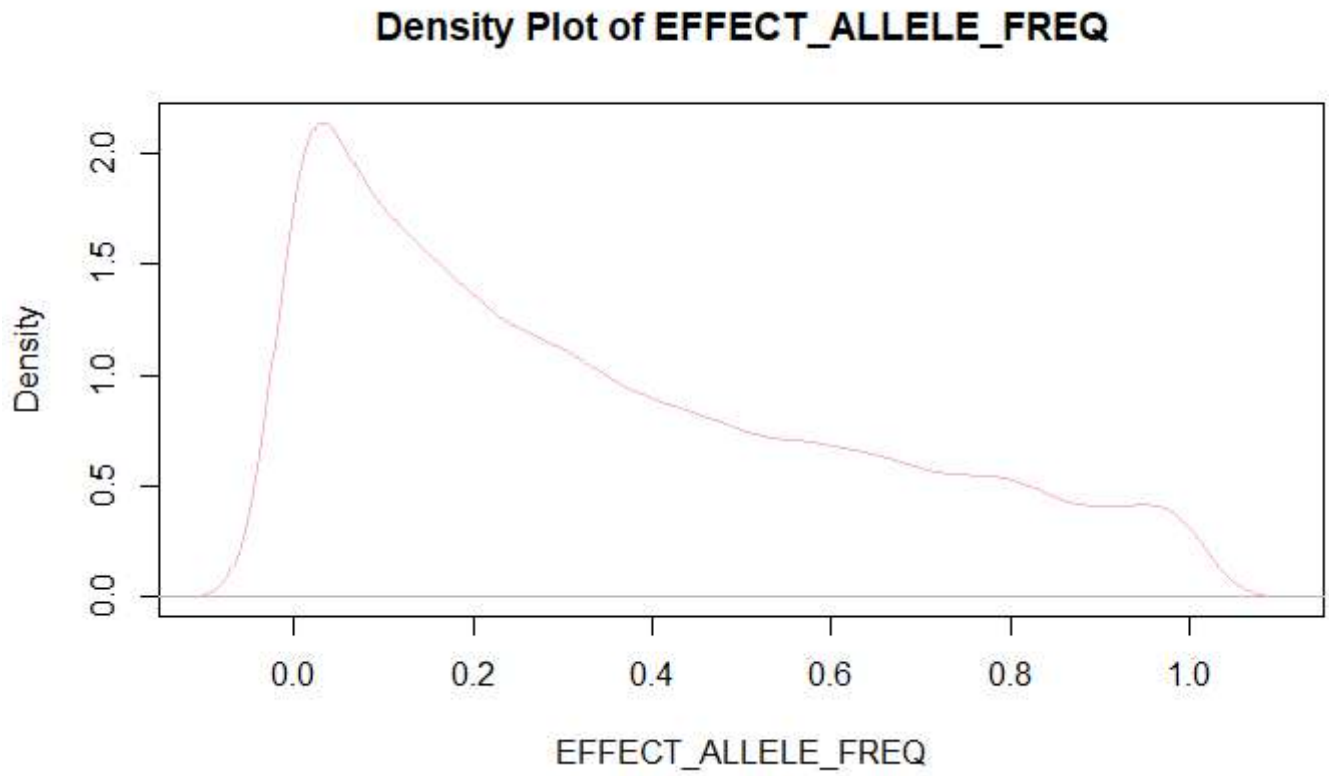
Hide

```
boxplot(P ~ ANCESTRY, data = yengoHeight, col = 'lightpink',
        main = "Box Plot of P by ANCESTRY", xlab = "ANCESTRY", ylab = "P")
```



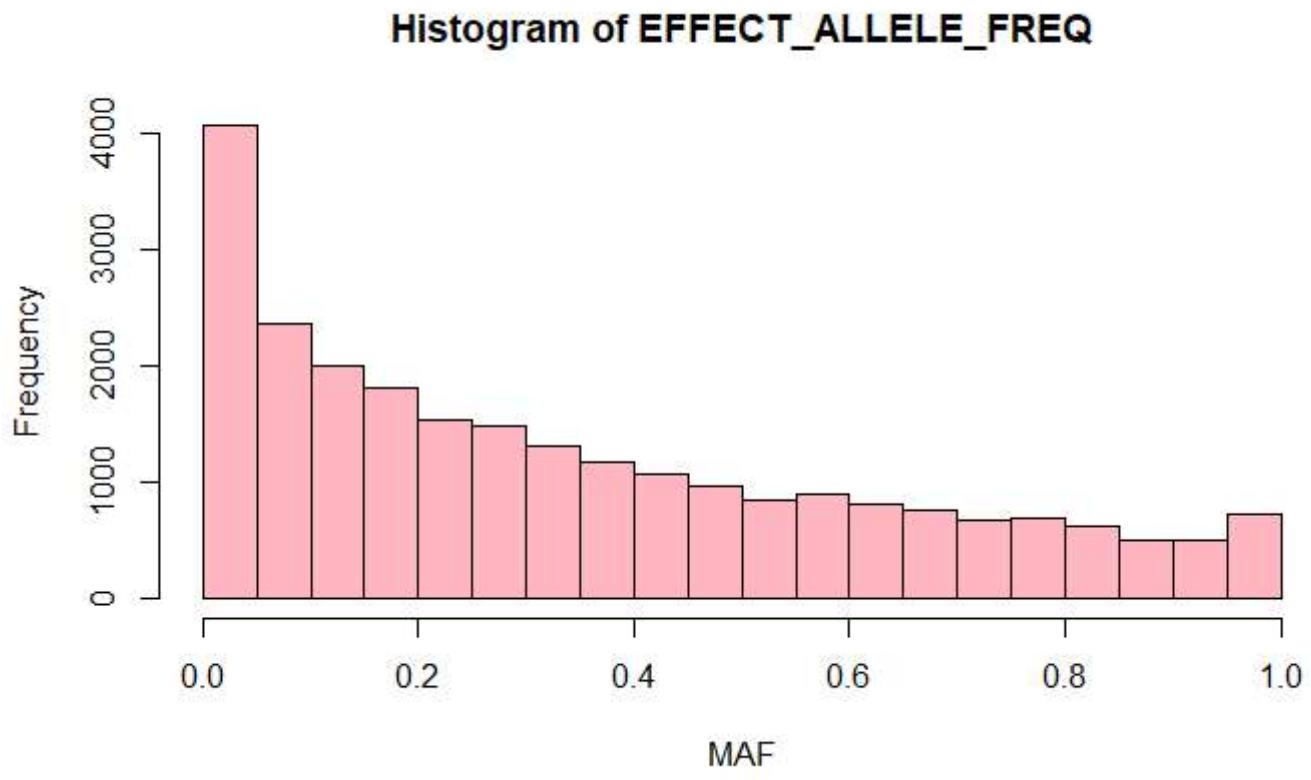
Hide

```
plot(density(yengoHeight$EFFECT_ALLELE_FREQ), col = "lightpink", main = "Density Plot of EFFECT_
ALLELE_FREQ", xlab = "EFFECT_ALLELE_FREQ")
```



Hide

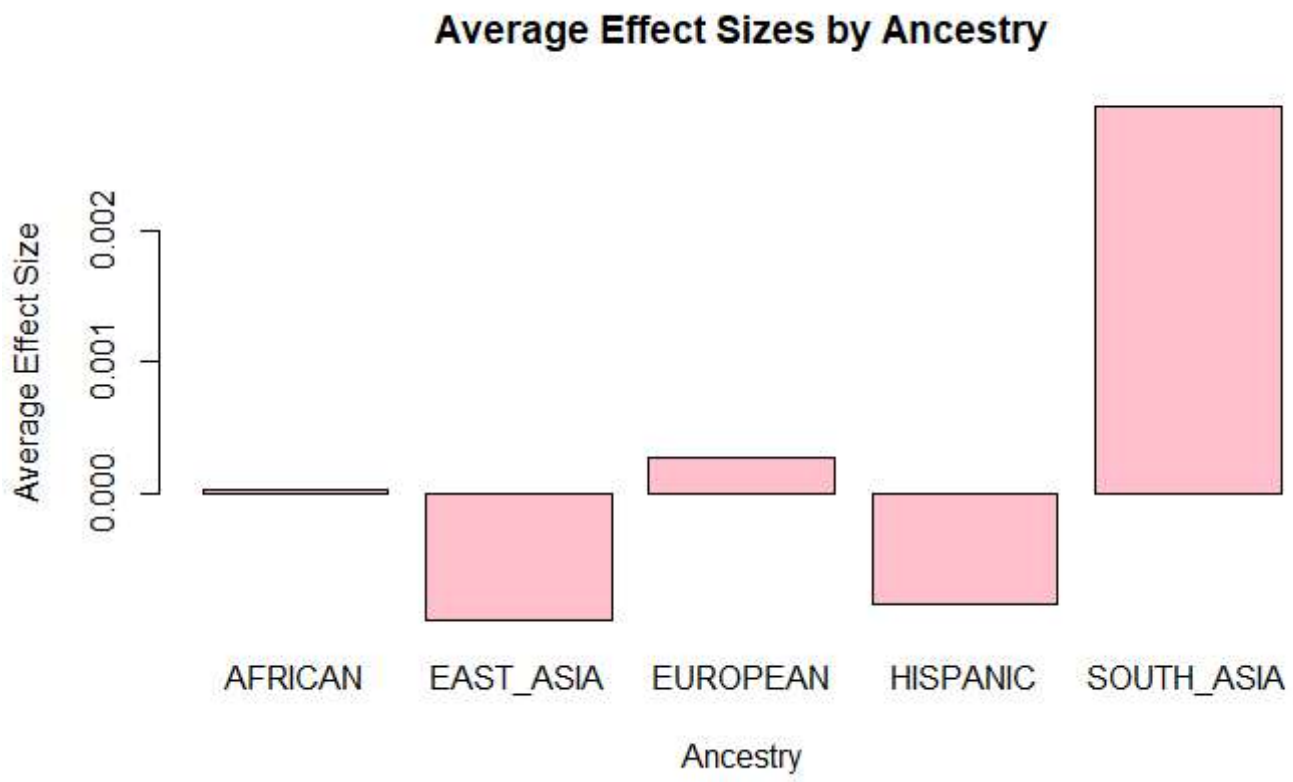
```
hist(yengoHeight$EFFECT_ALLELE_FREQ, xlab = 'MAF', ylab = 'Frequency', col = 'lightpink', main =
"Histogram of EFFECT_ALLELE_FREQ")
```



Hide

```
# Calculate average effect sizes for each SNP, stratified by ancestry
average_effect_sizes <- aggregate(yengoHeight$BETA ~ yengoHeight$ANCESTRY, FUN = mean)

# Create a bar plot of average effect sizes by ancestry
barplot(average_effect_sizes[, 2], names.arg = average_effect_sizes[, 1], col = "pink",
        main = "Average Effect Sizes by Ancestry", xlab = "Ancestry", ylab = "Average Effect Siz
e")
```



Hide

```
NA
NA
```

1.3 Data Analysis

Question 1

How many SNPs are significant (p-value < 0.01) for variability in height (MAF > 0.01) in all the super populations.

Hide

```
# Step 1: Filter the data based on the conditions
filtered_snps <- yengoHeight[yengoHeight$P < 0.01 & yengoHeight$EFFECT_ALLELE_FREQ > 0.01, ]

# Step 2: Count the number of SNPs that satisfy the conditions in step1
num_significant_snps <- nrow(filtered_snps)


print(paste("The number of significant SNPs for variability in height (P-value < 0.01 and MAF > 0.01) in all super populations is:", num_significant_snps))
```

```
[1] "The number of significant SNPs for variability in height (P-value < 0.01 and MAF > 0.01) in all super populations is: 2253"
```

what five Chromosomes should we look out for significant SNPs ?

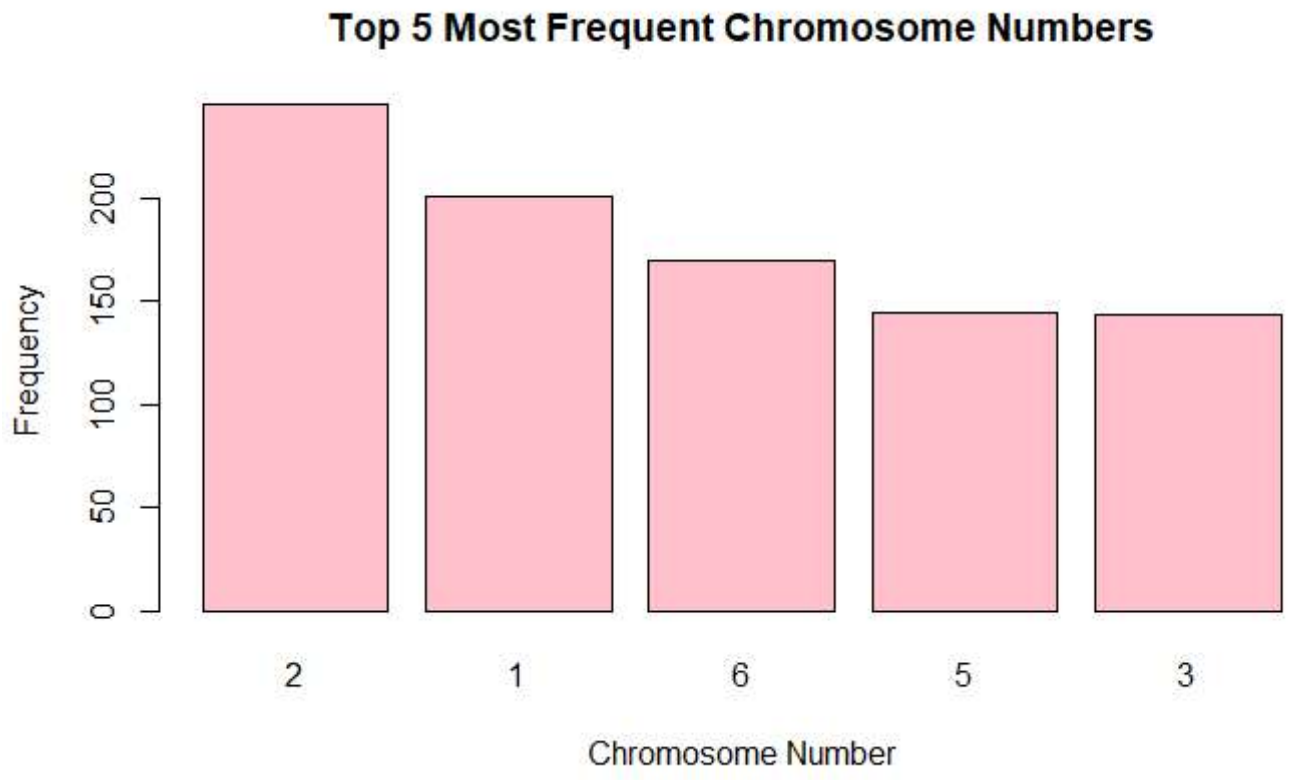
Hide

```
# Find the chromosome number that is most frequent
chromosome_counts <- table(filtered_snps$CHR)



# Sort the chromosome counts in descending order and select the top 5 values
top_chromosomes <- head(sort(chromosome_counts, decreasing = TRUE), 5)

# Create a bar plot for the top 5 most frequent chromosome numbers
barplot(top_chromosomes, main = "Top 5 Most Frequent Chromosome Numbers",
        xlab = "Chromosome Number", ylab = "Frequency", col = "pink")
```



Hide

```
NA
NA
NA
```

Chromosomes 2, 1,6,5,3 is shown to be the top chromosomes that genetic variation occurs.

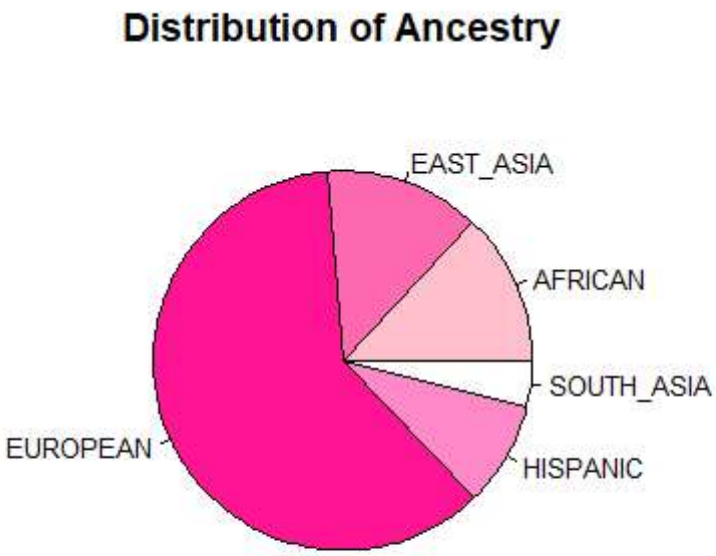
Which ancestry have the most signiicant genetic variability?

Hide

```
# Calculate the frequency of each unique value in the 'ANCESTRY' column
ancestry_counts <- table(filtered_snps$ANCESTRY)
most_common_ancestry <- names(ancestry_counts)[which.max(ancestry_counts)]



num_ancestries <- length(ancestry_counts)
custom_palette <- colorRampPalette(c("pink","deeppink", "white"))(num_ancestries)
pie(ancestry_counts, main = "Distribution of Ancestry", col = custom_palette, labels = names(anc
estry_counts), cex = 0.8)
```



Question 2

How much of Europeans genetic variability can/cannot be found in other super populations.

The question about the proportion of European genetic variability that can or cannot be found in other super populations raises points is related to population genetics and the significance of diversity in sequencing projects.

Defining a Null Hypothesis will help with that:

Null Hypothesis: The genetic variability observed in the European population is not significantly different from the genetic variability observed in other super populations.

To investigate this question, a Chi square statistical test will be used for the analysis.

Hide

```
european_data <- filtered_snps[yengoHeight$ANCESTRY == 'European', ]
european_variability <- sum(european_data$EFFECT_ALLELE_FREQ)

# Calculate the observed genetic variability in other super populations
other_data <- yengoHeight[yengoHeight$ANCESTRY != 'European', ]
other_variability <- sum(other_data$EFFECT_ALLELE_FREQ)

# Perform a chi-squared test
chisq_result <- chisq.test(c(european_variability, other_variability))
print(chisq_result)
```

```
    Chi-squared test for given probabilities

data:  c(european_variability, other_variability)
X-squared = 8472.4, df = 1, p-value < 2.2e-16
```

The extremely small p-value strongly suggests that there is a significant difference in genetic variability between the European population and other super populations. The large chi-squared value further supports this conclusion, indicating a substantial deviation from the expected values under the null hypothesis.

Hence we reject the Null Hypothesis