Assignment3_6737933

Chutiwan Pankram

2024-11-11

The dataset PlateletHW.tsv provides essential data for analyzing platelet aggregation levels and clopidogrel resistance. The key attributes include ADP, representing the ADP-induced platelet aggregation level, and Resistant, a binary indicator where 1 denotes clopidogrel resistance and 0 indicates non-resistance. Genotype information for three specific SNPs—rs4244285 (CYP2C192), rs4986893 (CYP2C193), and rs662 (PON1 192Q>R)—is coded using an additive genetic model. Additional demographic data includes Age (in years) and Sex, coded as 0 for male and 1 for female. This dataset supports studies on genetic associations with platelet aggregation and resistance to clopidogrel, allowing for detailed examination of genetic and demographic factors in platelet response.

1.Data Cleaning

To convert ADP to a positive value, use the abs() function to convert the value to absolute.

Data characteristics: ADP values represent the level of platelet aggregation stimulated by ADP, which should normally be positive. Since aggregation should not be negative, the negative values found in the original data are likely errors in data recording. Negative values may result in incorrect statistical analysis or graphing, especially if we have to use log transformation, which does not accept negative values. Transforming to positive values will help to bring the data into the appropriate range for analysis.

```
options(repos = c(CRAN = "https://cran.rstudio.com/"))

install.packages("tidyverse")
```

```
## Installing package into 'C:/Users/chuti/AppData/Local/R/win-library/4.4'
## (as 'lib' is unspecified)

## package 'tidyverse' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\chuti\AppData\Local\Temp\RtmpagxFxt\downloaded_packages
```

```
install.packages("car")
```

```
## Installing package into 'C:/Users/chuti/AppData/Local/R/win-library/4.4'
## (as 'lib' is unspecified)

## package 'car' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\chuti\AppData\Local\Temp\RtmpagxFxt\downloaded_packages
```

```r
install.packages("e1071")
```

```
## Installing package into 'C:/Users/chuti/AppData/Local/R/win-library/4.4'
## (as 'lib' is unspecified)

## package 'e1071' successfully unpacked and MD5 sums checked

## Warning: cannot remove prior installation of package 'e1071'

## Warning in file.copy(savedcopy, lib, recursive = TRUE): problem copying
## C:\Users\chuti\AppData\Local\R\win-library\4.4\00LOCK\e1071\libs\x64\e1071.dll
## to C:\Users\chuti\AppData\Local\R\win-library\4.4\e1071\libs\x64\e1071.dll:
## Permission denied

## Warning: restored 'e1071'

##
## The downloaded binary packages are in
##  C:\Users\chuti\AppData\Local\Temp\RtmpagxFxt\downloaded_packages
```

```r
install.packages("data.table")
```

```
## Installing package into 'C:/Users/chuti/AppData/Local/R/win-library/4.4'
## (as 'lib' is unspecified)

## package 'data.table' successfully unpacked and MD5 sums checked

## Warning: cannot remove prior installation of package 'data.table'

## Warning in file.copy(savedcopy, lib, recursive = TRUE): problem copying
## C:\Users\chuti\AppData\Local\R\win-library\4.4\00LOCK\data.table\libs\x64\data_table.dll
## to
## C:\Users\chuti\AppData\Local\R\win-library\4.4\data.table\libs\x64\data_table.dll:
## Permission denied

## Warning: restored 'data.table'

##
## The downloaded binary packages are in
##  C:\Users\chuti\AppData\Local\Temp\RtmpagxFxt\downloaded_packages
```

```r
library(readr)
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.4.2

## Warning: package 'ggplot2' was built under R version 4.4.2
```

```
## Warning: package 'forcats' was built under R version 4.4.2
```

```
## Warning: package 'lubridate' was built under R version 4.4.2
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v purrr     1.0.2
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
data <- read_tsv("C:/Users/chuti/OneDrive/Documents/GitHub/Assignment3_6737933/raw_data/PlateletHW.tsv")
```

```
## Rows: 211 Columns: 11
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (3): PON1.192Q>R, CYP2C19*2, CYP2C19*3
## dbl (8): IID, ADP, Resistance, rs4244285, rs4986893, rs662, AGE, SEX
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
data_clean <- data %>%
  mutate(ADP_abs = abs(ADP))

data_clean$ADP_abs <- unlist(data_clean$ADP_abs)
data_clean$ADP <- NULL
names(data_clean)[names(data_clean) == "ADP_abs"] <- "ADP"
data_clean <- data_clean[, c("IID", "ADP", setdiff(names(data_clean), c("IID", "ADP")))]

write_tsv(data_clean, "C:/Users/chuti/OneDrive/Documents/GitHub/Assignment3_6737933/clean_data/Platelet
```

2.Visualizing ADP and Drug Resistance

After cleaning the data, we want to see the trend in the relationship between ADP and drug resistance. Specifically, we want to determine whether higher ADP levels are associated with increased drug resistance. To do this, we create a scatterplot, which allows us to check whether our data exhibits a linear relationship—a key condition if we intend to apply statistical analysis methods like linear regression in the next steps.

The scatterplot also helps us identify outliers (if any) that may still be present in the cleaned data, which can sometimes be difficult to detect by merely inspecting the numerical data. This check reduces the risk of drawing inaccurate conclusions due to abnormal data points.

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.4.2
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.4.2
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
scatterplot(Resistance ~ ADP, data=data_clean, reg.line
            = lm, smooth=FALSE)
```

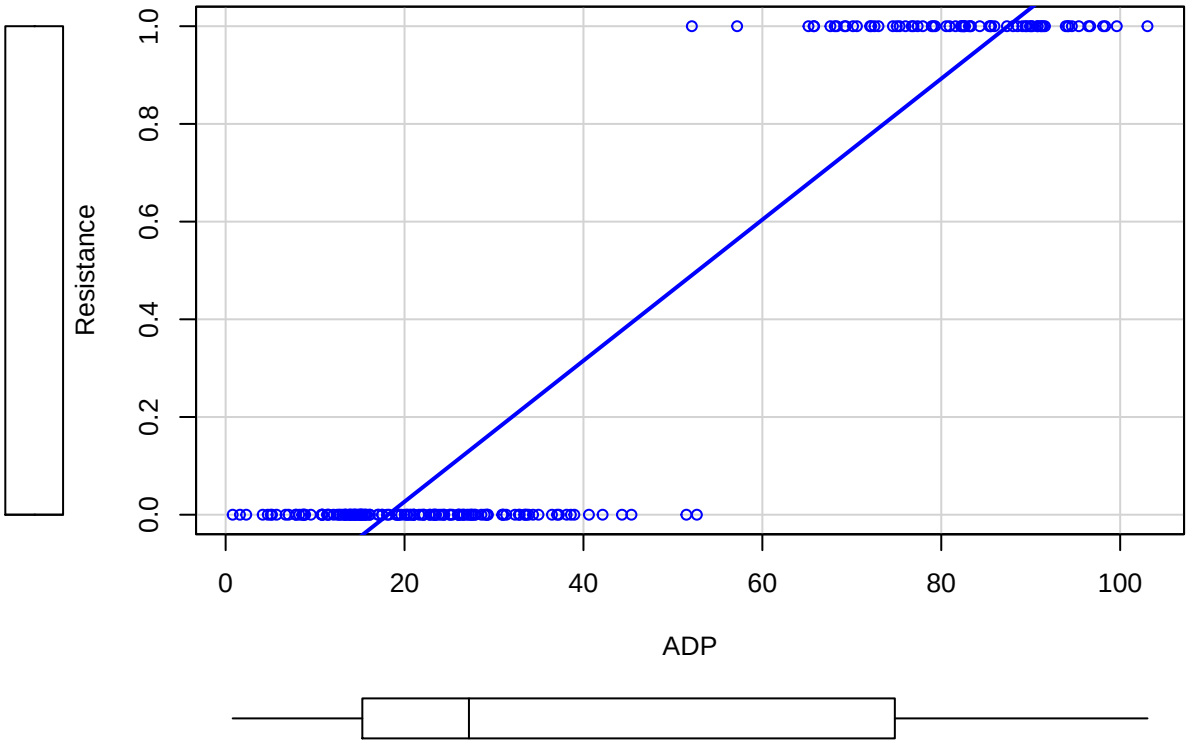
```
## Warning in plot.window(...): "reg.line" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "reg.line" is not a graphical parameter
```

```
## Warning in axis(side = side, at = at, labels = labels, ...): "reg.line" is not
## a graphical parameter
## Warning in axis(side = side, at = at, labels = labels, ...): "reg.line" is not
## a graphical parameter
```

```
## Warning in box(...): "reg.line" is not a graphical parameter
```

```
## Warning in title(...): "reg.line" is not a graphical parameter
```



The graph shows the relationship between resistance values and ADP (Adenosine Diphosphate) values. In the scatter plot above, most of the resistance values are scattered at very low or near-zero levels, especially in the range where ADP values are low. This may reflect various factors that keep the resistance values low in this range. For example, individuals with low ADP levels tend to show less drug resistance, as ADP plays a role in stimulating platelet function, which could be related to resistance to various treatments.

3.ADP Statistical Test

check the skewness and outliers using the IQR method.

```
library(tidyverse)
library(ggplot2)
library(e1071)
```

```
## Warning: package 'e1071' was built under R version 4.4.2
```

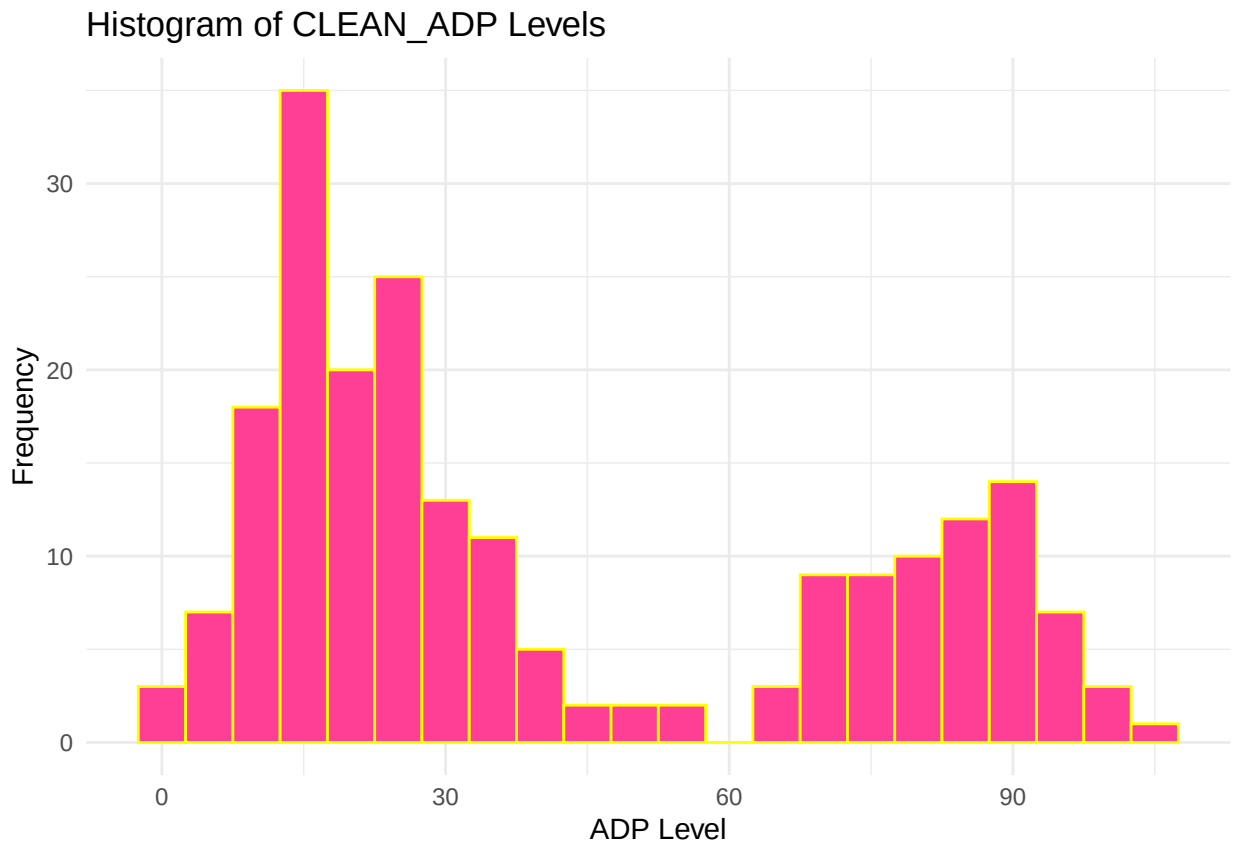
```
library(data.table)
```

```
## Warning: package 'data.table' was built under R version 4.4.2
```

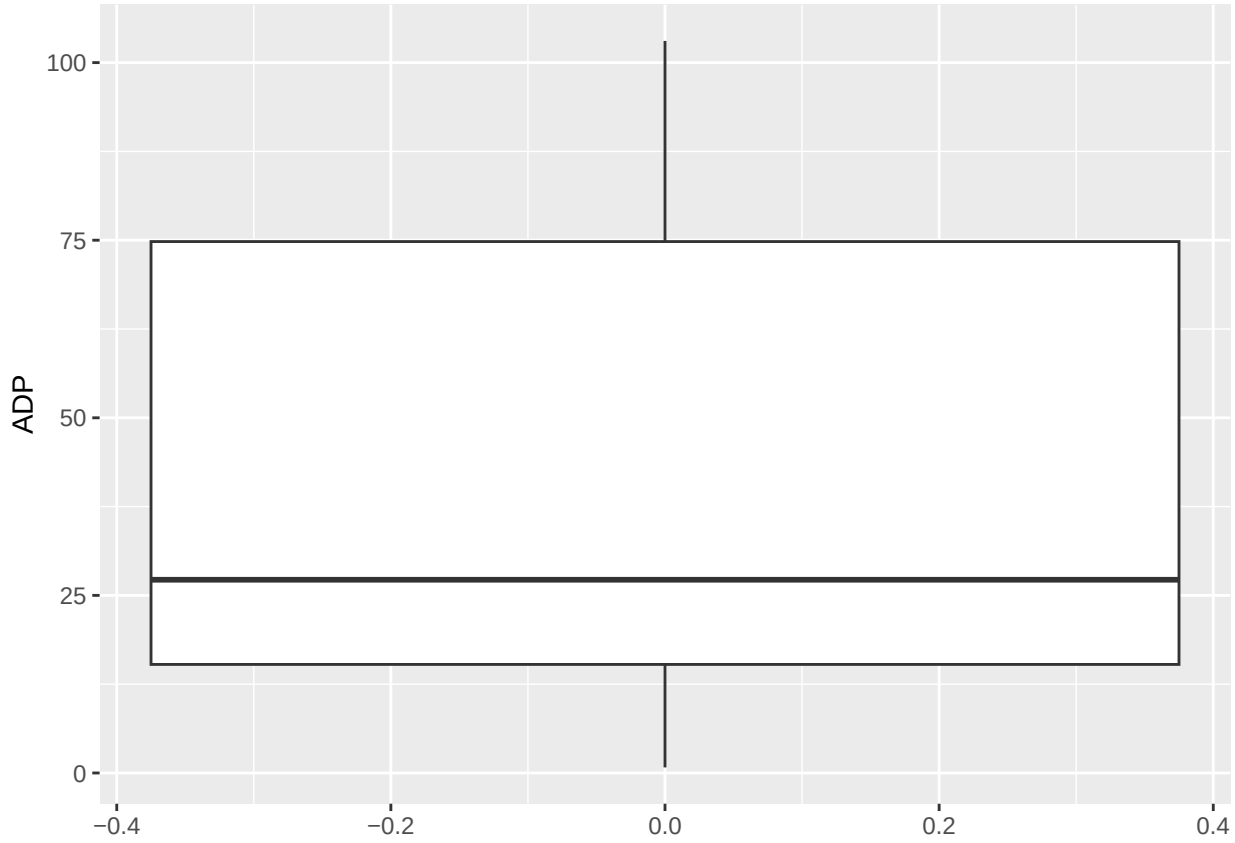
```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:lubridate':
##
##     hour, isoweek, mday, minute, month, quarter, second, wday, week,
##     yday, year

## The following objects are masked from 'package:dplyr':
##
##     between, first, last

## The following object is masked from 'package:purrr':
##
##     transpose
```

```r
# Checking clean data histogram
ggplot(data_clean, aes(x =ADP)) +
  geom_histogram(binwidth = 5, fill = "violetred1", color = "yellow") +
  theme_minimal() +
  labs(title = "Histogram of CLEAN_ADP Levels", x = "ADP Level", y = "Frequency")
```




```r
# Check a boxplot graph
ggplot(data_clean,aes(y= ADP)) + geom_boxplot()
```




```
#Check skewness
clean_skewness_value <- skewness(data_clean$ADP) #  right skewness
# Identify outliers using IQR method
Q1 <- quantile(data_clean$ADP, 0.25)
Q3 <- quantile(data_clean$ADP, 0.75)
IQR_value <- IQR(data_clean$ADP)
lower_bound <- Q1 - 1.5 * IQR_value
upper_bound <- Q3 + 1.5 * IQR_value
outliers_iqr <- data_clean %>%
  filter(ADP < lower_bound | ADP > upper_bound)   # There is no outliers that less than Q1 - 1.5 * IQR

cat("Number of outliers by IQR method:", nrow(outliers_iqr), "\n")
```

```
## Number of outliers by IQR method: 0
```

Number of outliers found using the IQR method is zero, it means there are no outliers according to the IQR method, which is consistent with a slightly right-skewed distribution of ADP values.

4.Linear regression

This step adjusts the data to make it more normally distributed. Generally, when the ADP values are right-skewed or contain outliers in one direction, applying a logarithmic transformation can help reduce the skewness and make the distribution more normal.

```r
data_clean$ADP_log <- log(data_clean$ADP)
```

```r
liner_logA <- lm(ADP_log ~ rs4244285, data = data_clean)
liner_logB <- lm(ADP_log ~ rs4986893, data = data_clean)
linear_logC <- lm(ADP_log ~ rs662, data = data_clean)
```

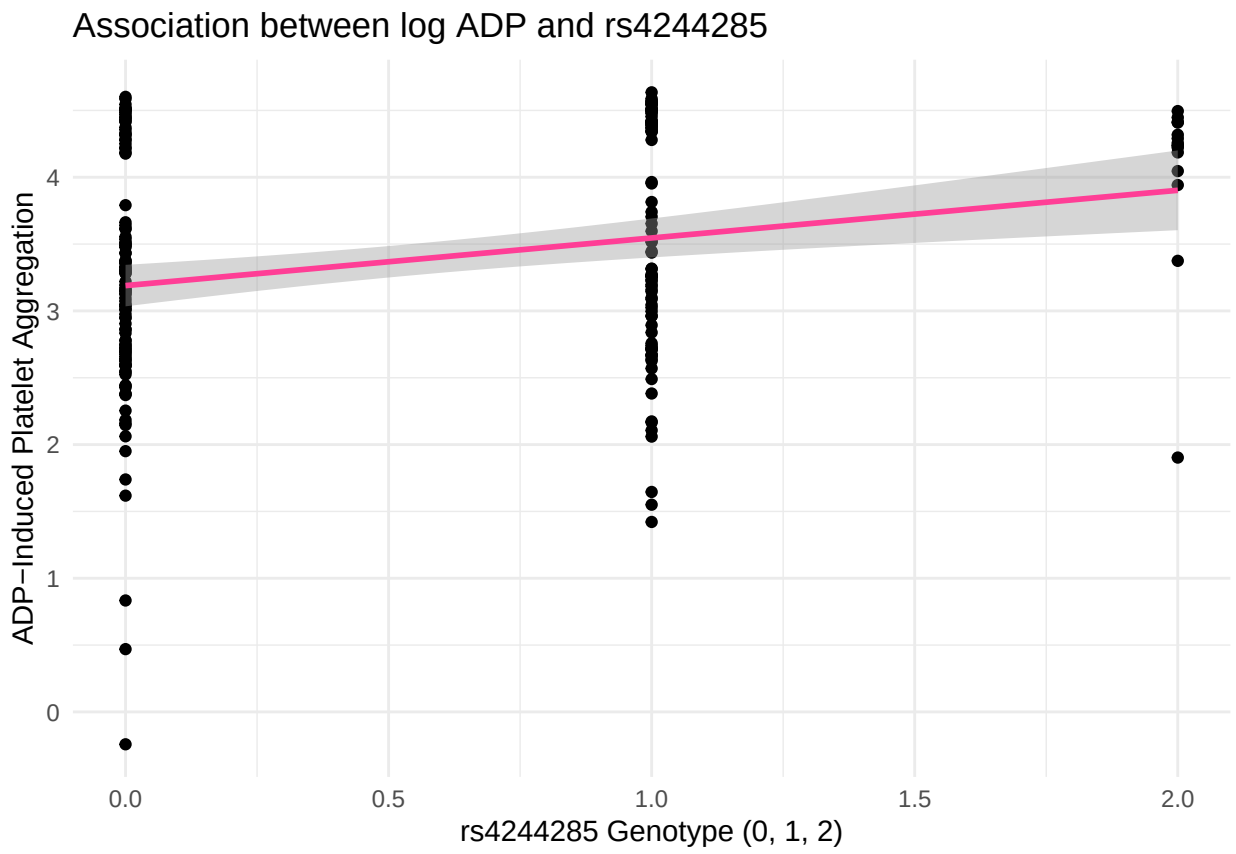
Furthermore, examining the relationship between SNPs and ADP using Linear Regression allows us to assess whether each SNP affects ADP levels by comparing the p-value and R-squared from the analysis. Additionally, using a Scatterplot and QQ plot helps verify whether the data obtained from the linear regression follows a normal distribution.

```r
library(ggplot2)
summary(liner_logA)
```

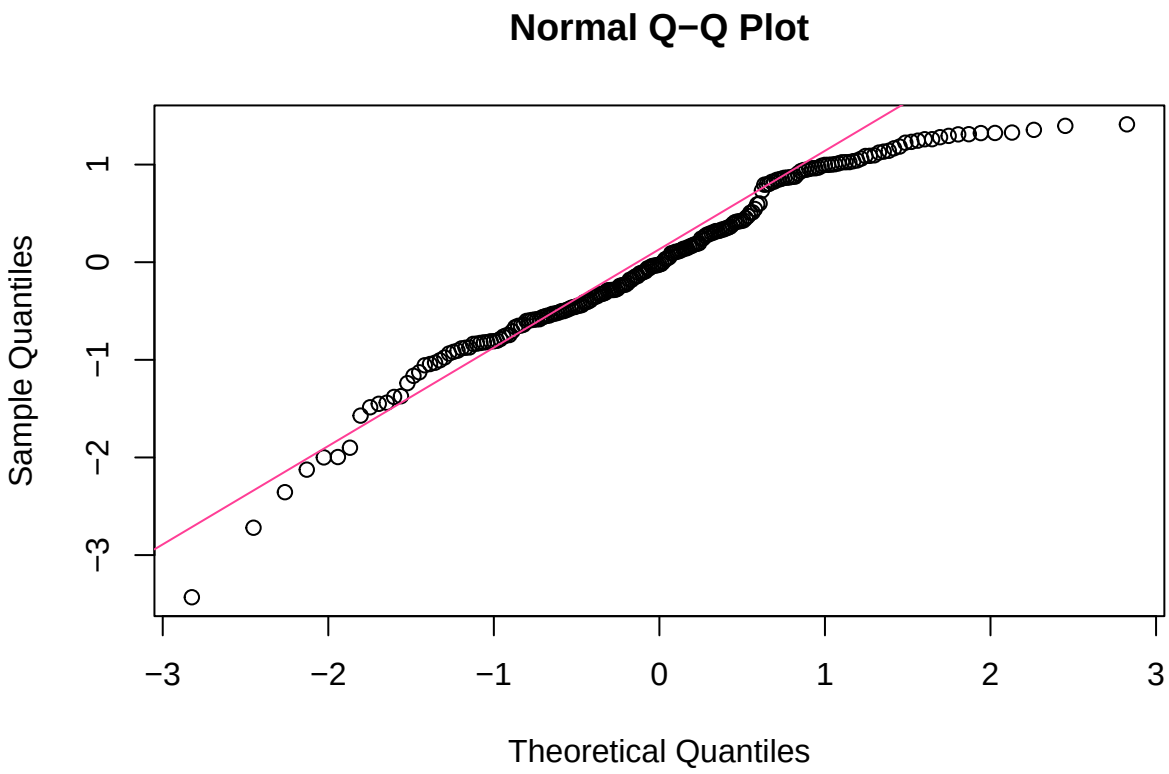
```
##
## Call:
## lm(formula = ADP_log ~ rs4244285, data = data_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4317 -0.5467 -0.0235  0.8128  1.4120
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.18940    0.07873   40.51  < 2e-16 ***
## rs4244285    0.35644    0.09530    3.74 0.000238 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8651 on 209 degrees of freedom
## Multiple R-squared:  0.06273,    Adjusted R-squared:  0.05825
## F-statistic: 13.99 on 1 and 209 DF,  p-value: 0.0002375
```

```r
ggplot(data_clean, aes(x = rs4244285, y = ADP_log)) +
  geom_point() +
  geom_smooth(method = "lm", color = "violetred1") +
  labs(title = "Association between log ADP and rs4244285",
       x = "rs4244285 Genotype (0, 1, 2)",
       y = "ADP-Induced Platelet Aggregation") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
qqnorm(liner_logA$residuals)
qqline(liner_logA$residuals, col = "violetred1")
```



The results of the linear regression analysis between ADP_log and rs4244285 indicate a statistically significant relationship between the two variables. The slope obtained from the analysis is 0.35644, meaning that when rs4244285 increases by 1 unit, the value of ADP_log increases on average by approximately 0.35644 units. However, despite the significance of this relationship, the R-squared value is only 6.27%, suggesting that rs4244285 may not be the primary factor affecting ADP_log. Other factors may have a greater impact.

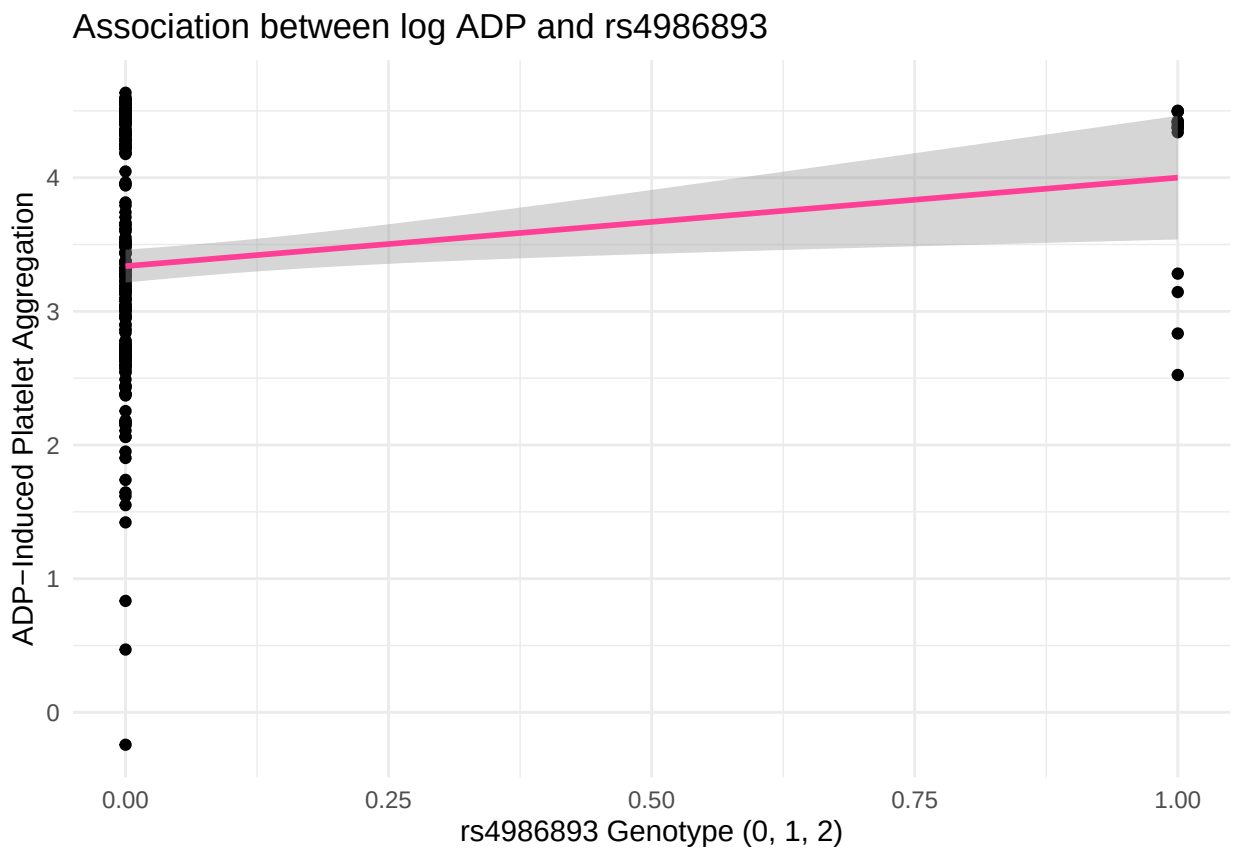
```
library(ggplot2)
summary(liner_logB)
```

```
##
## Call:
## lm(formula = ADP_log ~ rs4986893, data = data_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5803 -0.6227 -0.0348  0.8844  1.2972
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.33804    0.06257  53.353  < 2e-16 ***
## rs4986893    0.66218    0.24289   2.726  0.00695 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8781 on 209 degrees of freedom
## Multiple R-squared:  0.03434,    Adjusted R-squared:  0.02972
```

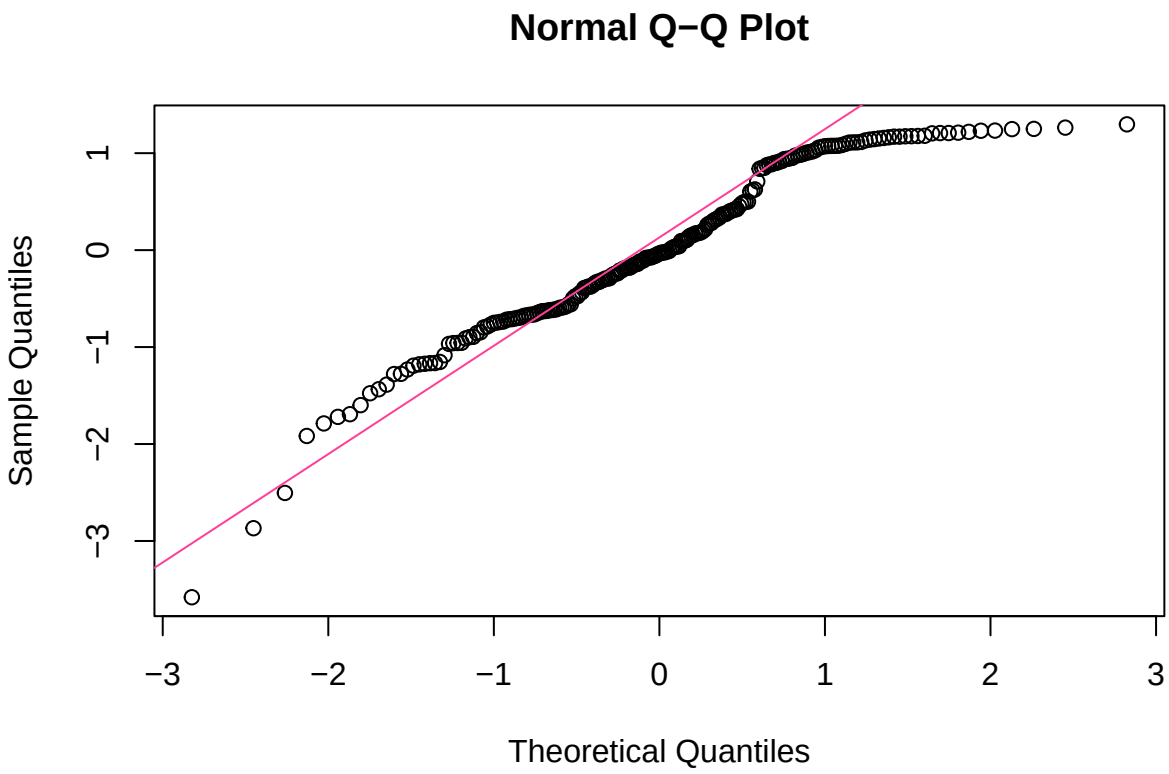
```
## F-statistic: 7.433 on 1 and 209 DF,  p-value: 0.006949
```

```r
ggplot(data_clean, aes(x = rs4986893 , y = ADP_log)) +
  geom_point() +
  geom_smooth(method = "lm", color = "violetred1") +
  labs(title = "Association between log ADP and rs4986893",
       x = "rs4986893 Genotype (0, 1, 2)",
       y = "ADP-Induced Platelet Aggregation") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```r
qqnorm(liner_logB$residuals)
qqline(liner_logB$residuals, col = "violetred1")
```



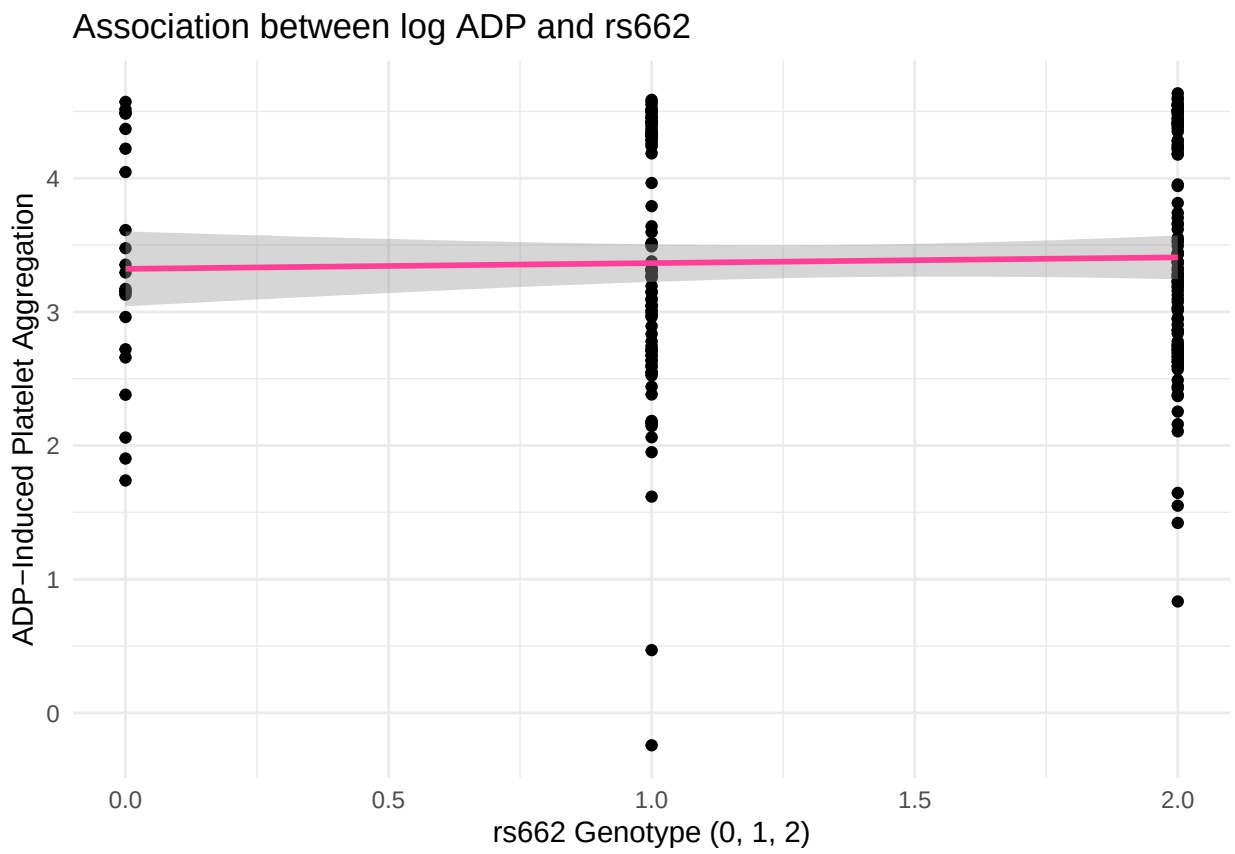
The results of the linear analysis between rs4986893 and ADP_log show that rs4986893 is statistically significantly associated with ADP_log (p-value = 0.00695). When rs4986893 increases by 1 unit, ADP_log increases by 0.66218 units. However, the R-squared value is 0.03434, indicating that rs4986893 explains only a small portion of the variance in ADP_log. Despite the statistical significance, this model explains very little of the variance in ADP_log.

```
summary(linear_logC)
```

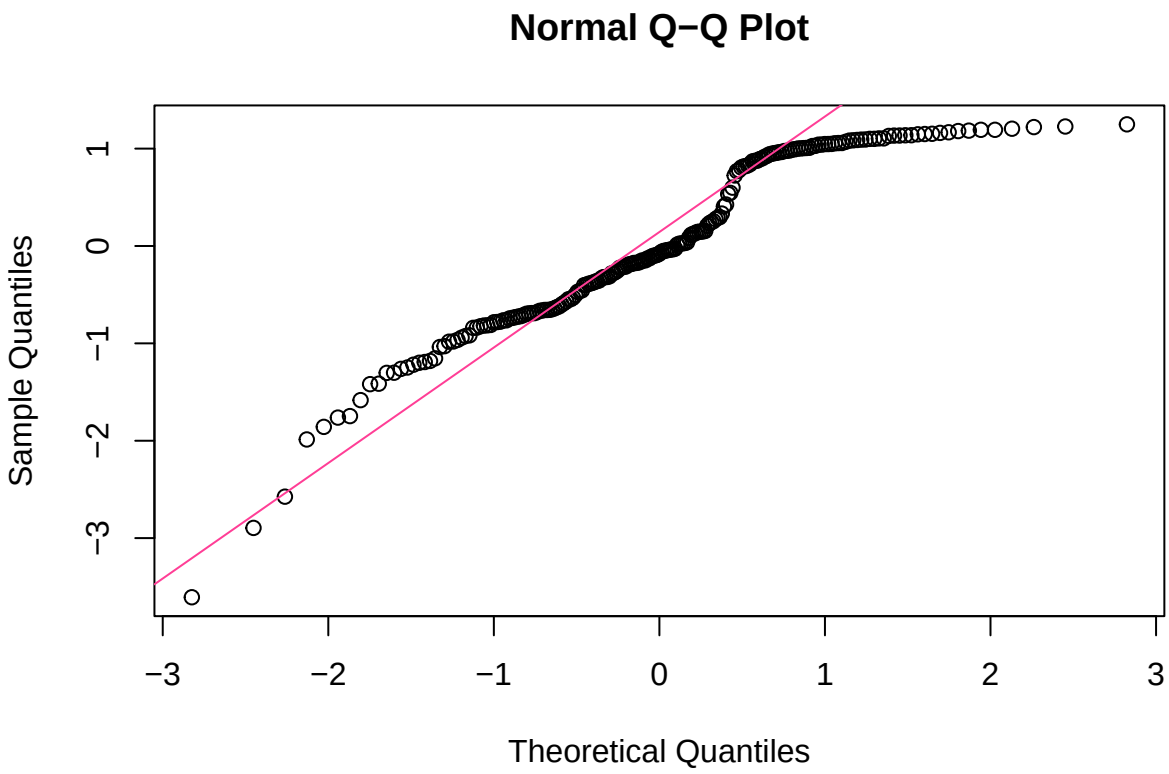
```
##
## Call:
## lm(formula = ADP_log ~ rs662, data = data_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6073 -0.6565 -0.0787  0.9437  1.2492
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.32192    0.14211  23.376   <2e-16 ***
## rs662        0.04310    0.09195   0.469     0.64
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8932 on 209 degrees of freedom
## Multiple R-squared:  0.00105,    Adjusted R-squared:  -0.003729
## F-statistic: 0.2197 on 1 and 209 DF,  p-value: 0.6397
```

```r
ggplot(data_clean, aes(x = rs662 , y = ADP_log)) +
  geom_point() +
  geom_smooth(method = "lm", color = "violetred1") +
  labs(title = "Association between log ADP and rs662",
       x = "rs662 Genotype (0, 1, 2)",
       y = "ADP-Induced Platelet Aggregation") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```r
qqnorm(linear_logC$residuals)
qqline(linear_logC$residuals, col = "violetred1")
```



The results of the linear analysis between rs662 and ADP_log show that rs662 is not statistically significantly associated with ADP_log (p-value = 0.6397), which is greater than the significance level of 0.05. This means that rs662 cannot predict or explain the variance in ADP_log. The estimate for rs662 is 0.04310, indicating that an increase of 1 unit in rs662 does not result in a statistically significant change in ADP_log. The R-squared value (0.00105) suggests that this model explains only a small portion of the variance in ADP_log.

Logistic regression

```
snp_list <- c("rs4244285", "rs4986893", "rs662")

results_list <- list()

for (snp in snp_list) {
  model_sum <- lm(as.formula(paste("ADP_log ~ AGE + SEX +", snp)), data = data_clean)
  results_list[[snp]] <- summary(model_sum)
}
print(results_list[["rs4244285"]]) # significant only SNP
```

```
##
## Call:
## lm(formula = as.formula(paste("ADP_log ~ AGE + SEX +", snp)),
##     data = data_clean)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -3.4329 -0.5847  0.0234  0.7691  1.3790
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.642427   0.369440   9.859  < 2e-16 ***
## AGE         -0.006644   0.005629  -1.180 0.239184
## SEX         -0.047234   0.134027  -0.352 0.724883
## rs4244285    0.360830   0.095387   3.783 0.000203 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8652 on 207 degrees of freedom
## Multiple R-squared:  0.07152,    Adjusted R-squared:  0.05806
## F-statistic: 5.315 on 3 and 207 DF,  p-value: 0.001507
```

```
print(results_list[["rs4986893"]])# significant only SNP
```

```
##
## Call:
## lm(formula = as.formula(paste("ADP_log ~ AGE + SEX +", snp)),
##     data = data_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5707 -0.5926 -0.0460  0.8435  1.3307
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.787300   0.372706  10.162  < 2e-16 ***
## AGE         -0.006749   0.005720  -1.180  0.23941
## SEX         -0.006994   0.136411  -0.051  0.95916
## rs4986893    0.663689   0.243691   2.723  0.00701 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.879 on 207 degrees of freedom
## Multiple R-squared:  0.04167,    Adjusted R-squared:  0.02778
## F-statistic:     3 on 3 and 207 DF,  p-value: 0.03161
```

```
print(results_list[["rs662"]]) # not siginifcant all variables
```

```
##
## Call:
## lm(formula = as.formula(paste("ADP_log ~ AGE + SEX +", snp)),
##     data = data_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6018 -0.6330 -0.0758  0.9179  1.2493
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  3.758708   0.392094   9.586   <2e-16 ***
## AGE         -0.006572   0.005838  -1.126    0.262
## SEX         -0.024108   0.139273  -0.173    0.863
## rs662        0.047704   0.092747   0.514    0.608
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8941 on 207 degrees of freedom
## Multiple R-squared:  0.008599,   Adjusted R-squared:  -0.005769
## F-statistic: 0.5985 on 3 and 207 DF,  p-value: 0.6167
```

Summary

In this analysis of the PlateletHW.tsv dataset, we explored the relationship between ADP levels, clopidogrel resistance, and genetic factors (SNPs: rs4244285, rs4986893, rs662), as well as age and sex. We cleaned the data by converting negative ADP values to positive and checked the distribution, finding a slight right skew but no outliers.

We applied a logarithmic transformation to ADP to normalize the data and conducted Linear Regression with each SNP. The results showed that rs4244285 and rs4986893 were significantly associated with ADP_log, though their explanatory power was low (R-squared values of 6.27% and 3.43%, respectively). rs662 was not significantly associated with ADP_log. Adding age and sex to the models did not substantially change the results.

In conclusion, while some SNPs showed significance, their impact on platelet aggregation was minimal, indicating other factors may influence ADP levels and clopidogrel resistance more strongly.