PlateletHW_Data Report

A Study of Platelet Data

Buranee Manoapan

2024-11-09

Coronary heart disease(CAD) is one of the disease that has a dramatically mortality and disability rate, the platelet aggregation rate(PAR) to gather the data and studies the enzymes encoded genes are PON1.192Q>R, CYP2C19*2, and CYP2C19*3 to observe the association of the polymorphisms and clopidogrel-resistance.This study has 211 IID of platelet aggregation data which indicate ADP (Adenosine diphoesphate), Resistance, Age, Sexand genotypes of CYP2C19*2 (rs4244285), CYP2C19*3 (rs4986893), and PON1.192Q>R (rs662) to perform statatistical analysis and association analysis. Basically, the platelet activity response can be use to measure by ADP due to it is a signaling molecule which can be studies involving antiplatelet drugs like clopidogrel. *R software* (version 2024.04.2+764) was aims to analyze and visualize the statistical analysis by plotting. Firstly, The data was import and read by *read_tsv function* and checking for an outliner by using IQR rule base on *quantile function.* write_tsv function was use to save the new clean PlateletHW file for further normalize the data. The data was taking an absolute *(abs function)* and log *(log function)* to get the positive integer, theoretically data, and avoiding zero. Furthermore, an association analysis were perform by linear regression, multiple linear regression, and logistic regression. The result from linear regression show statistically significant associations with ADP levels of rs4244285 and rs4986893 which can be observed through the expected value of the dependent variable, whereas rs662 didn't indicate any association with ADP levels. Multiple linear regression model suggest the impact of rs4244285 and rs4986893 on ADP levels as result shown as estimate, p-value., and multiple R-squared. Finally, logistic regression also confirmed that both SNPs located within CYP2C19 gene are strong predictors of resistance. The model supported that r662 nor does, age, and sex are non-significant predictors.

Import Raw Data

```
library(readr)
PlateletHW <- read_tsv("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.
```

```
PlateletHW
```

```
## # A tibble: 211 x 11
##      IID   ADP Resistance rs4244285 rs4986893 rs662   AGE   SEX `PON1.192Q>R`
##    <dbl> <dbl>      <dbl>     <dbl>     <dbl> <dbl> <dbl> <dbl> <chr>
##  1     1  1.60          0         0         0     1    70     1 A G
##  2     2  9.53          0         0         0     2    59     1 G G
##  3     3 12.8           0         0         0     1    69     1 A G
##  4     4 14.5           0         0         0     1    53     0 A G
##  5     5 18.3           0         0         0     2    44     0 G G
##  6     6 23.3           0         0         0     0    59     0 A A
##  7     7 32.9           0         0         0     2    76     0 G G
##  8     8 13.3           0         0         0     1    57     0 A G
##  9     9 33.6           0         1         0     1    57     1 A G
## 10    10 51.5           0         2         0     2    75     1 G G
## # i 201 more rows
## # i 2 more variables: `CYP2C19*2` <chr>, `CYP2C19*3` <chr>
```

```
dim(PlateletHW)
```

```
## [1] 211  11
```

Check for Outliners

IQR rules indicate that ADP value aren't identified as outliners.

```
quantiles <- quantile(PlateletHW$ADP, probs = c(0.25, 0.75))
iqr <- quantiles[2] - quantiles[1]
upper_limit <- quantiles[2] + 1.5 * iqr
lower_limit <- quantiles[1] - 1.5 * iqr
outliers <- PlateletHW$ADP[PlateletHW$ADP < lower_limit | PlateletHW$ADP > upper_limit]
print(outliers)
```

```
## numeric(0)
```

Clean data

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
PlateletHW_clean <- PlateletHW %>%filter(ADP >= lower_limit & ADP <= upper_limit)
summary(PlateletHW_clean$ADP)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -8.721  15.281  27.200  40.853  74.810 103.053
```

```
write_tsv(PlateletHW_clean,"clean_data/PlateletHW_clean.tsv")
head(PlateletHW_clean)
```

```
## # A tibble: 6 x 11
##     IID   ADP Resistance rs4244285 rs4986893 rs662   AGE   SEX `PON1.192Q>R`
##   <dbl> <dbl>      <dbl>     <dbl>     <dbl> <dbl> <dbl> <dbl> <chr>
## 1     1  1.60          0         0         0     1    70     1 A G
## 2     2  9.53          0         0         0     2    59     1 G G
## 3     3 12.8           0         0         0     1    69     1 A G
## 4     4 14.5           0         0         0     1    53     0 A G
## 5     5 18.3           0         0         0     2    44     0 G G
## 6     6 23.3           0         0         0     0    59     0 A A
## # i 2 more variables: `CYP2C19*2` <chr>, `CYP2C19*3` <chr>
```

Data Normalization

To normalize the ADP taking log transformation to get the positive integer and avoiding 0 and save the file.

```
PlateletHW$ADP[PlateletHW$ADP == 0] <- mean(PlateletHW$ADP[PlateletHW$ADP > 0], na.rm = TRUE)
PlateletHW <- PlateletHW %>% mutate(ADP_log = log(ADP + 1))
```

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `ADP_log = log(ADP + 1)`.
## Caused by warning in `log()`:
## ! NaNs produced
```

```
PlateletHW$ADP_log[is.na(PlateletHW$ADP_log)] <- mean(PlateletHW$ADP_log, na.rm = TRUE)
PlateletHW$ADP <- PlateletHW$ADP_log
PlateletHW$ADP_log <- NULL
PlateletHW <- PlateletHW[, c("IID", "ADP", setdiff(names(PlateletHW_clean), c("IID", "ADP")))]
summary(PlateletHW$ADP)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -1.536   2.816   3.361   3.448   4.328   4.645
```

```
dim(PlateletHW)
```

```
## [1] 211  11
```

Analyses the data

Test for Associations: Linear regression model for each SNP

The association of each SNP has tested to observe the relationship between each SNP and ADP levels, a statistically significant of rs4244285 (figure 1.1), rs4986893 (figure 1.2), and rs662 (figure 1.3) can be viusulize through the plot of linear regression.

```
model_rs4244285 <- lm(ADP ~ rs4244285, data = PlateletHW)
model_rs4986893 <- lm(ADP ~ rs4986893, data = PlateletHW)
model_rs662 <- lm(ADP ~ rs662, data = PlateletHW)

summary(model_rs4244285)
```

```
##
## Call:
## lm(formula = ADP ~ rs4244285, data = PlateletHW)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7860 -0.5243 -0.0366  0.7537  1.3618
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.24959    0.07577  42.886  < 2e-16 ***
## rs4244285    0.36806    0.09172   4.013 8.36e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8326 on 209 degrees of freedom
## Multiple R-squared:  0.07153,    Adjusted R-squared:  0.06709
## F-statistic:  16.1 on 1 and 209 DF,  p-value: 8.358e-05
```

```
summary(model_rs4986893)
```

```
##
## Call:
## lm(formula = ADP ~ rs4986893, data = PlateletHW)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9439 -0.6095 -0.0470  0.8295  1.2374
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.40754    0.06058  56.251   <2e-16 ***
## rs4986893    0.61646    0.23517   2.621   0.0094 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8502 on 209 degrees of freedom
## Multiple R-squared:  0.03183,    Adjusted R-squared:  0.0272
## F-statistic: 6.871 on 1 and 209 DF,  p-value: 0.009404
```

```r
summary(model_rs662)
```

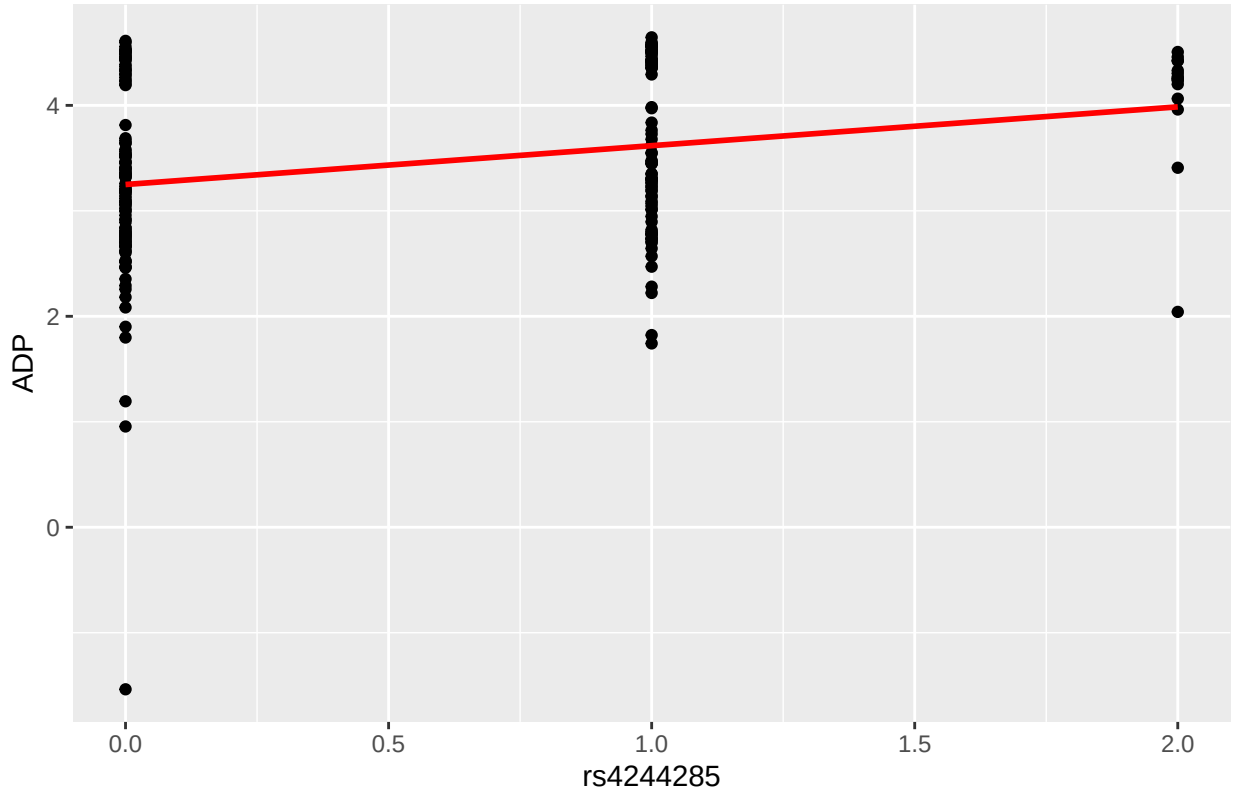
```
##
## Call:
## lm(formula = ADP ~ rs662, data = PlateletHW)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9681 -0.6294 -0.0749  0.8904  1.1921
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.38928    0.13741  24.665   <2e-16 ***
## rs662        0.04246    0.08891   0.478    0.633
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8636 on 209 degrees of freedom
## Multiple R-squared:  0.00109,    Adjusted R-squared:  -0.00369
## F-statistic: 0.228 on 1 and 209 DF,  p-value: 0.6335
```

The summarize of model in each SNPs can be visualize through ggplot.

```r
library(ggplot2)
ggplot(PlateletHW, aes(x = rs4244285, y = ADP)) + geom_point() + geom_smooth(method = "lm", se = FALSE,
```

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

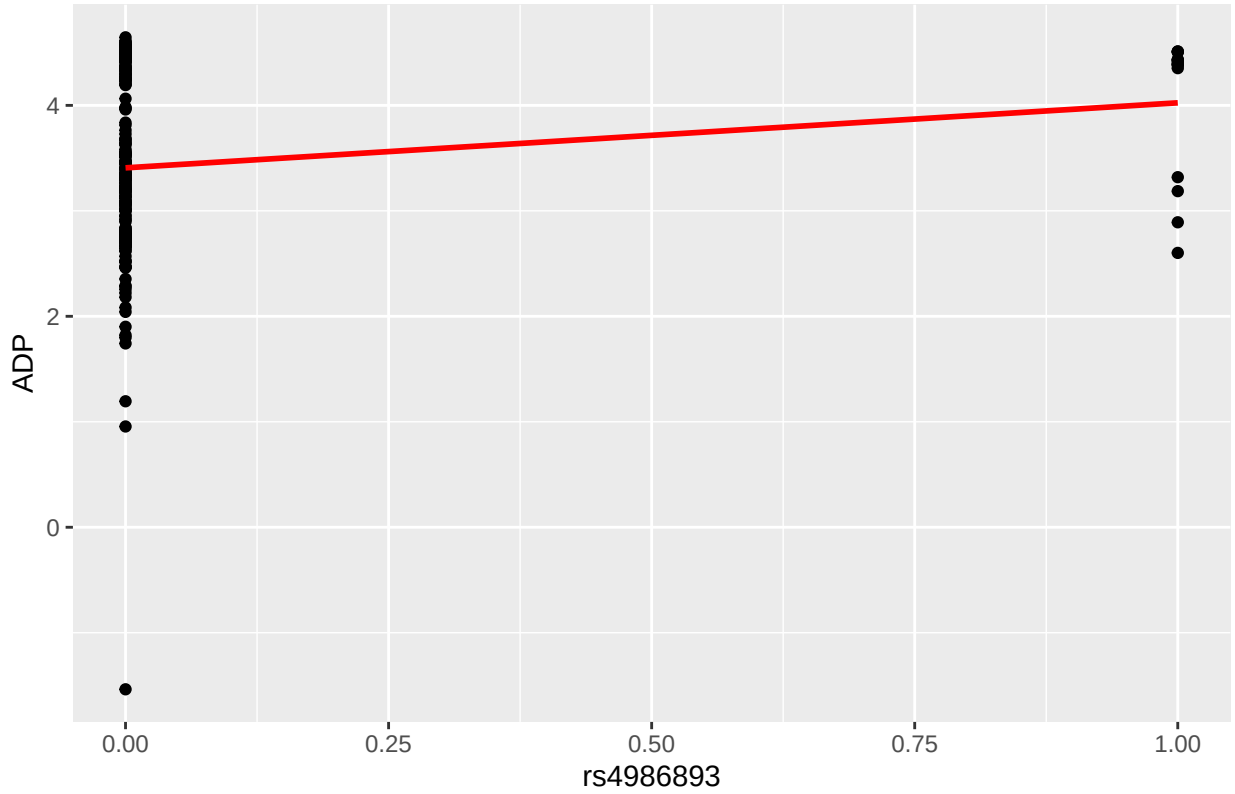
Figure 1.1 Linear Regression of ADP on rs4244285




```
ggplot(PlateletHW, aes(x = rs4986893, y = ADP)) +geom_point() + geom_smooth(method = "lm", se = FALSE, c
```

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

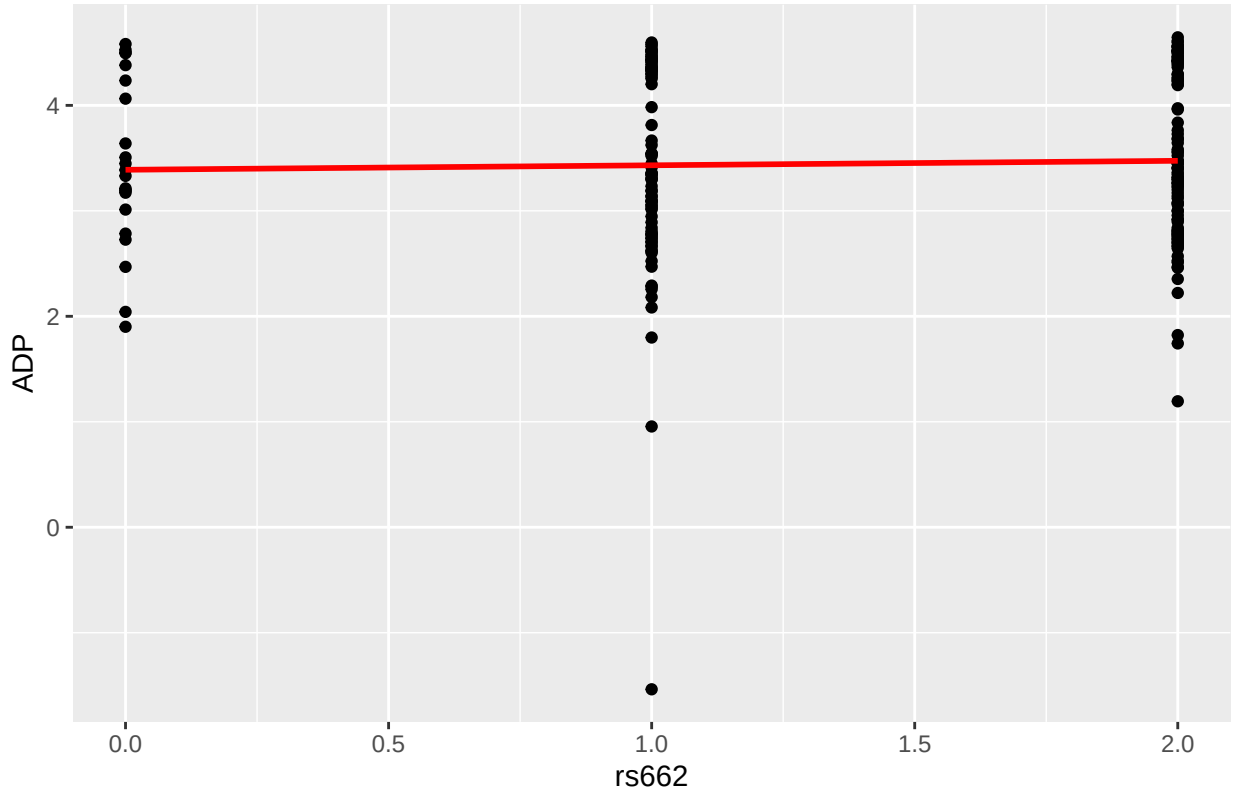
Figure 1.2 Linear Regression of ADP on rs4986893




```
ggplot(PlateletHW, aes(x = rs662, y = ADP)) +geom_point() + geom_smooth(method = "lm", se = FALSE, colo
```

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

Figure 1.3 Linear Regression of ADP on rs662



Multiple Linear Regression

The relationship between the dependent variable as `ADP` and independent variable including `rs4244285`, `rs4986893`, `rs662`, `AGE` and `SEX` are show the significant predictors, insignificant predictors, and model fit.

```
model <- lm(ADP ~ rs4244285 + rs4986893 + rs662 + AGE + SEX, data = PlateletHW)
summary(model)
```

```
##
## Call:
## lm(formula = ADP ~ rs4244285 + rs4986893 + rs662 + AGE + SEX,
##     data = PlateletHW)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7215 -0.5318  0.0256  0.5549  1.3840
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.585918   0.363363   9.869  < 2e-16 ***
## rs4244285    0.368951   0.090701   4.068 6.77e-05 ***
## rs4986893    0.602403   0.228080   2.641   0.0089 **
## rs662        0.047532   0.085338   0.557   0.5781
## AGE         -0.006594   0.005374  -1.227   0.2212
```

```
## SEX         -0.019154   0.128480  -0.149   0.8816
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8226 on 205 degrees of freedom
## Multiple R-squared:  0.1111, Adjusted R-squared:  0.08944
## F-statistic: 5.125 on 5 and 205 DF,  p-value: 0.0001883
```

Test an association for Clopidogrel Resistance (Binary Outcome)

The performance of logistic regression model with all SNPs, age, and sex is use to confirm the relationship of the factors that influence ADP of the dataset. The model interpret that significant predictors are rs4244285 (CYP2C192) and rs4986893 (CYP2C193) due to p-value<0.05 whereas rs662 (PON1.192Q>R) is not show strong association with CR.

```
resistance_model <- glm(Resistance ~ rs4244285 + rs4986893 + rs662 + AGE + SEX, data = PlateletHW, famil
summary(resistance_model)
```

```
##
## Call:
## glm(formula = Resistance ~ rs4244285 + rs4986893 + rs662 + AGE +
##     SEX, family = binomial, data = PlateletHW)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.31636    0.99317   0.319  0.75007
## rs4244285    0.99647    0.25089   3.972 7.13e-05 ***
## rs4986893    1.85788    0.63550   2.923  0.00346 **
## rs662        0.01388    0.23940   0.058  0.95376
## AGE         -0.02697    0.01509  -1.787  0.07394 .
## SEX          0.02659    0.35993   0.074  0.94111
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 268.14  on 210  degrees of freedom
## Residual deviance: 239.11  on 205  degrees of freedom
## AIC: 251.11
##
## Number of Fisher Scoring iterations: 4
```

Note: The confusion matrix is perform to gather an understanding on the model that classify resistance case.

```
predicted_probs <- predict(model, type = "response")
predicted_classes <- ifelse(predicted_probs > 0.5, 1, 0)
actual_classes <- PlateletHW$Resistance
confusion_matrix <- table(Predicted = predicted_classes, Actual = actual_classes)
print(confusion_matrix)
```

```
##          Actual
```

```
## Predicted   0   1
##         1 141  70
```