Study Case 2

1) Which features significantly associate with mRNA half-life?

1.1) Association between codons and the half-life of the WT

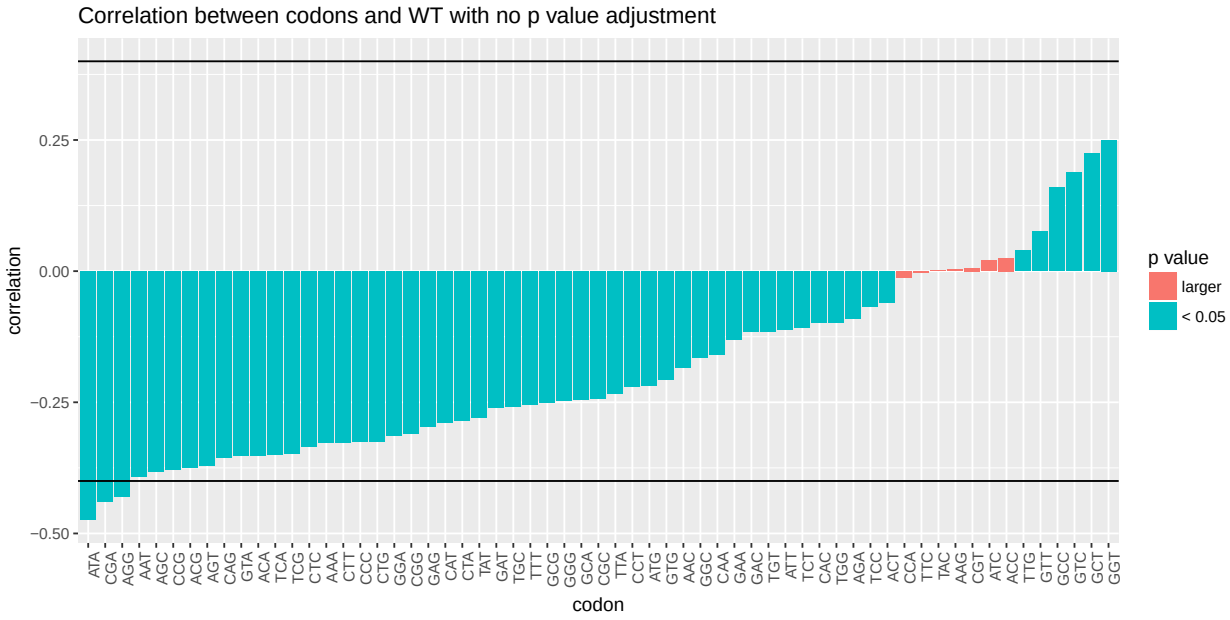
```
library(ggplot2)
set.seed(2)
```

```
# 1 data preparation
dt <- readRDS("case_study_dt.rds")

dt$CDS_seq <- NULL
dt$genename <- NULL
dt$UTR3_seq <- NULL

# 2 general correlations matrix

#3 only WT correlation
subset <- dt[,-c(1:34)]

#4 function to draw chart

correlation.chart <- function(adjust_str, title) {

  library("psych")
  res <- corr.test(subset, adjust = adjust_str)
  l_cor <- res$r[1,]
  l_p_values <- res$p[1,]

  df_cor <- data.frame(codon = names(l_cor), correlation = l_cor, p_value = l_p_values)
  df_cor$p_value_bool <- df_cor$p_value < 0.05
  df_cor <- df_cor[-1,]
  #sort by r value


  # Very basic bar graph

  ggplot(data=df_cor, aes(x=reorder(codon,correlation), y=correlation, fill=p_value_bool)) +
    geom_bar(stat="identity") +
    geom_hline(yintercept=0.40) +
    geom_hline(yintercept=-0.40) +
    ggtitle(title) +
    scale_fill_discrete(name="p value",
                        breaks=c("FALSE", "TRUE"),
                        labels=c("larger", "< 0.05")) +
    labs(x = "codon") +
    theme(axis.text.x = element_text(angle = 90, hjust = 1))

}

# no adjustment
```

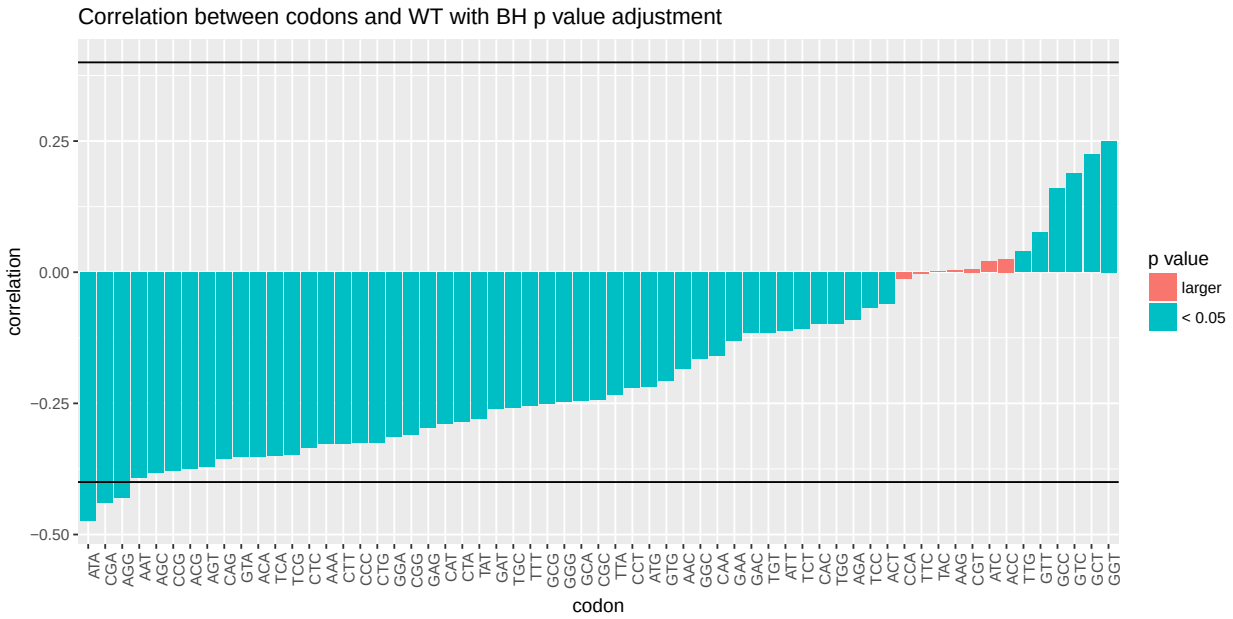
```
adjust_str <- "none"
title <- "Correlation between codons and WT with no p value adjustment"
correlation.chart(adjust_str,title)
```

```
##
## Attaching package: 'psych'
```

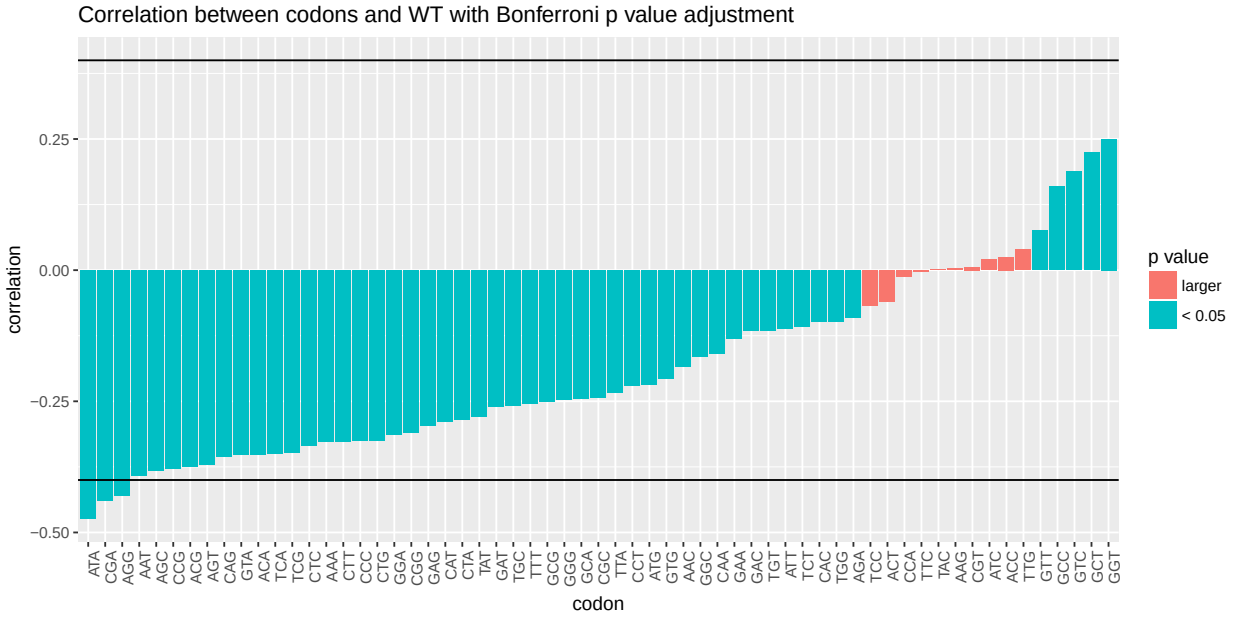
```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```




```
# BH adjustment
adjust_str <- "BH"
title <- "Correlation between codons and WT with BH p value adjustment"
correlation.chart(adjust_str,title)
```




```
# bonferroni adjustment
adjust_str <- "bonferroni"
title <- "Correlation between codons and WT with Bonferroni p value adjustment"
correlation.chart(adjust_str,title)
```



1.2) Association between 6-mers and the half-life of the WT

```
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
```

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

```
## ------------------------------------------------------------------------
```

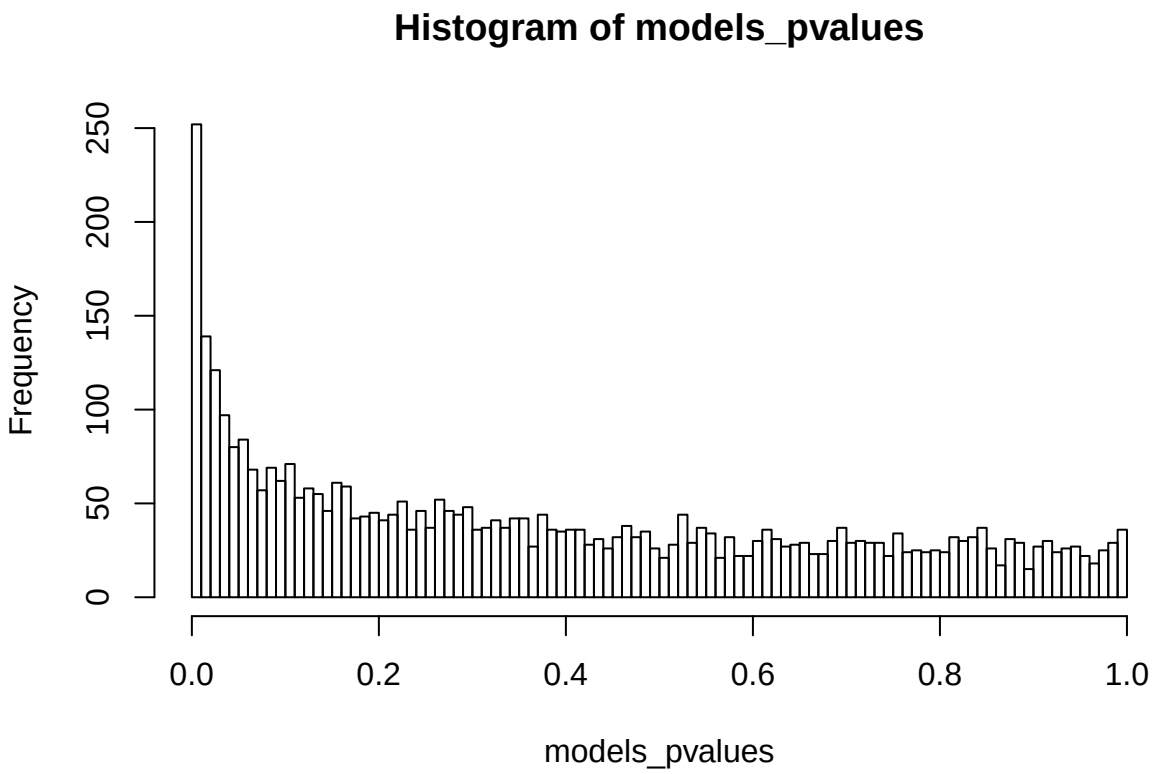
```
## data.table + dplyr code now lives in dtplyr.
## Please library(dtplyr)!
```

```
## ------------------------------------------------------------------------
```

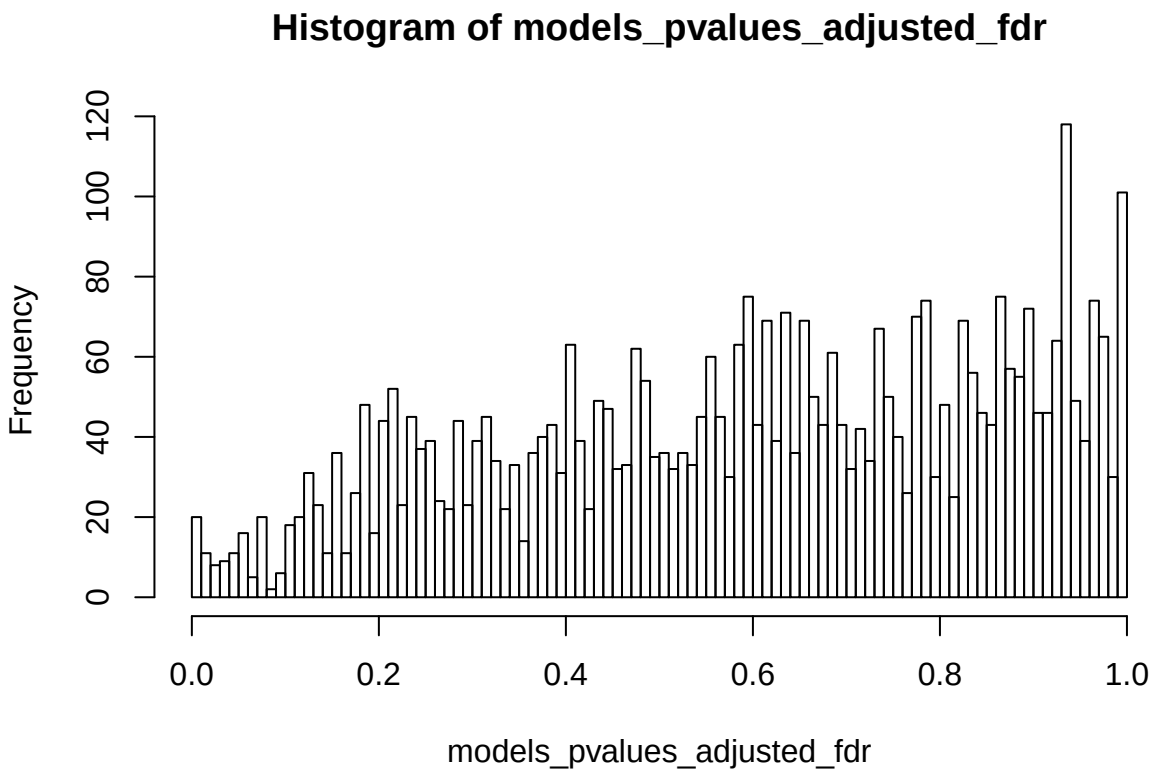
```
##
## Attaching package: 'data.table'
```

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

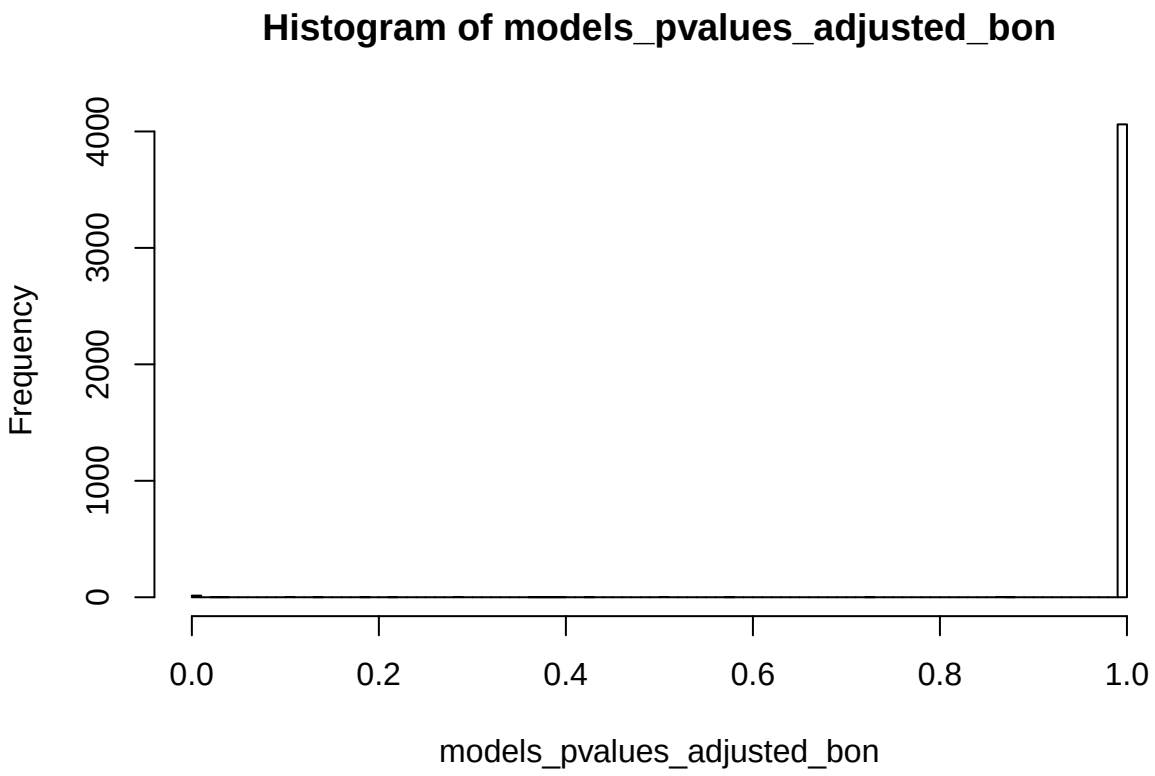
```
library(jsonlite)
library(corrplot)

dt <- readRDS("case_study_dt.rds")
utr3 <- as.data.table(readRDS("case_study_utr3_6mer.rds"))

json <- fromJSON("case_study_info.json")

merged <- data.table(dt, utr3)

models_pvalues <- sapply(utr3, function(x,dt)
{
  sum <- summary(lm(x ~ dt))$coefficients[8]
},dt=dt$WT)

models_pvalues_adjusted_fdr <- p.adjust(models_pvalues, "fdr")
models_pvalues_adjusted_bon <- p.adjust(models_pvalues, "bonferroni")
hist(models_pvalues, breaks=100)
```




```
hist(models_pvalues_adjusted_fdr, breaks=100)
```




```
hist(models_pvalues_adjusted_bon, breaks=100)
```




```
sum(models_pvalues_adjusted_fdr < 0.05)
```

```
## [1] 59
```

```
sum(models_pvalues_adjusted_bon < 0.05)
```

```
## [1] 16
```

3) Can we predict mRNA half-life from the given features, in wild-type and knock-outs? What are the relevant features?

We tried to predict the half-life of the WT strain by using the codons' frequencies. We then tried to analyze the coefficients of the features of the model and their significancy.

```
library(dplyr)
library(data.table)
library(caret)
```

```
## Loading required package: lattice
```

```
dt <- readRDS("case_study_dt.rds")
utr3 <- as.data.table(readRDS("case_study_utr3_6mer.rds"))

merged <- data.table(dt, utr3)

merged <- merged %>% select(WT, TTT:GGG)
```

```
train_index <- createDataPartition(merged$WT, p = .75, list = FALSE)
bh_tr <- merged[ train_index, ]
bh_te <- merged[-train_index, ]

lm_fit <- train(WT ~ .,
                data = merged,
                method = "lm")

bh_pred <- predict(lm_fit, bh_te)

coefficients <- as.data.frame(summary(lm_fit)$coefficients)
coefficients$codon <- rownames(coefficients)
coefficients <- coefficients[-1,]
names(coefficients) <- c("coefficient", "error", "tValue", "pValue", "codon")
coefficients$p_value_bool <- coefficients$pValue < 0.05
```

Results of the linear regression:

Residual standard error: 0.5141 on 3690 degrees of freedom

Multiple R-squared: 0.4748, Adjusted R-squared: 0.4661

F-statistic: 54.68 on 61 and 3690 DF, p-value: $<$ 2.2e-16

```
ggplot(data=coefficients, aes(x=reorder(codon,coefficient), y=coefficient, fill=p_value_bool)) +
    geom_bar(stat="identity") +
    geom_hline(yintercept=0.03) +
    geom_hline(yintercept=-0.03) +
    ggtitle("Coefficients of the features in the linear regression for WT half-life prediction") +
    scale_fill_discrete(name="p value",
                        breaks=c("FALSE", "TRUE"),
                        labels=c("larger", "< 0.05")) +
    labs(x = "codon") +
    theme(axis.text.x = element_text(angle = 90, hjust = 1))
```