Study Case 2

Miguel Sandim, Paula Fortuna, Vanessa Schmoll

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

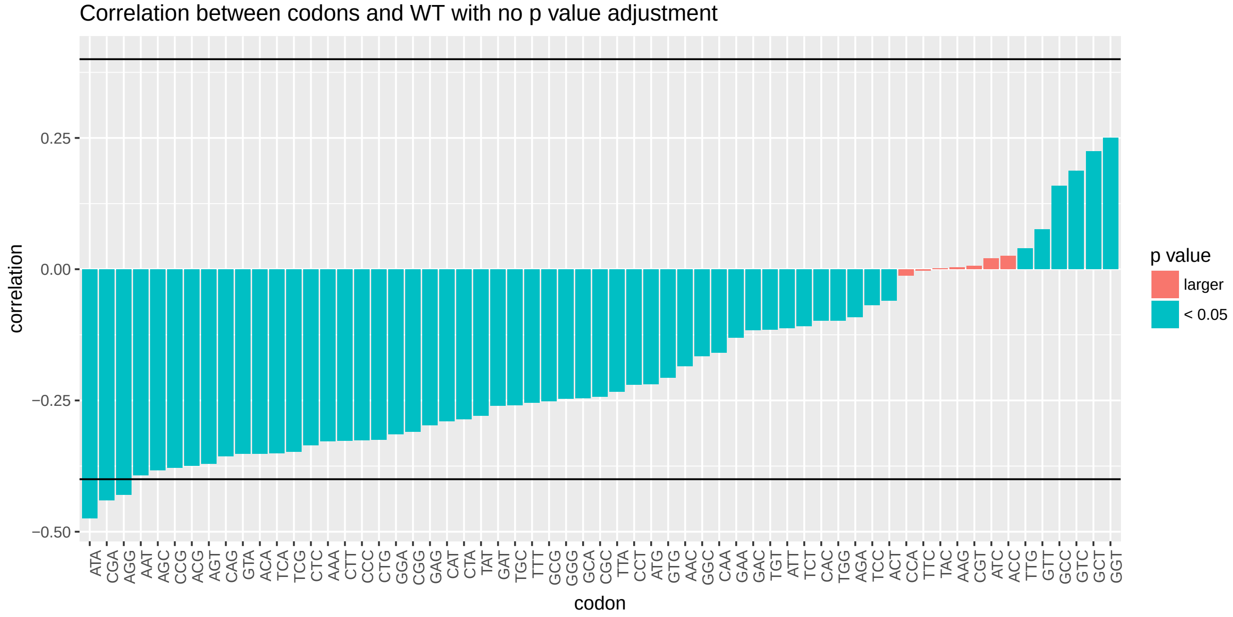
1.1) Association between codons and the half-life of the WT

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

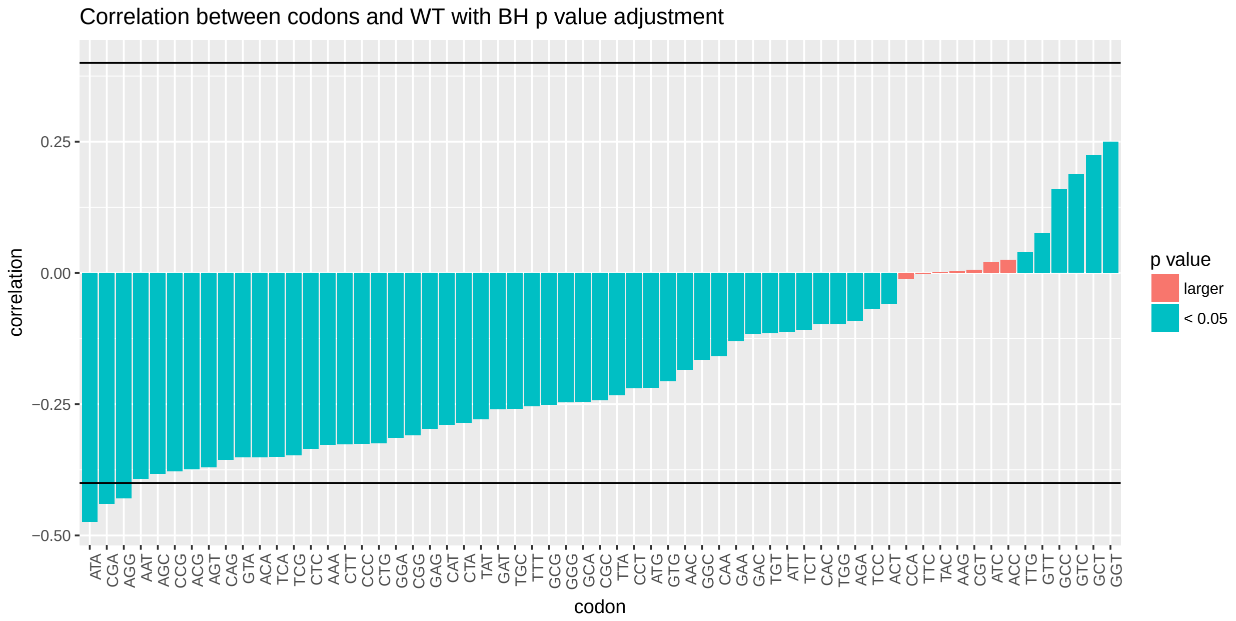
```
# 1 data preparation
dt <- readRDS("case_study_dt.rds")

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)
```

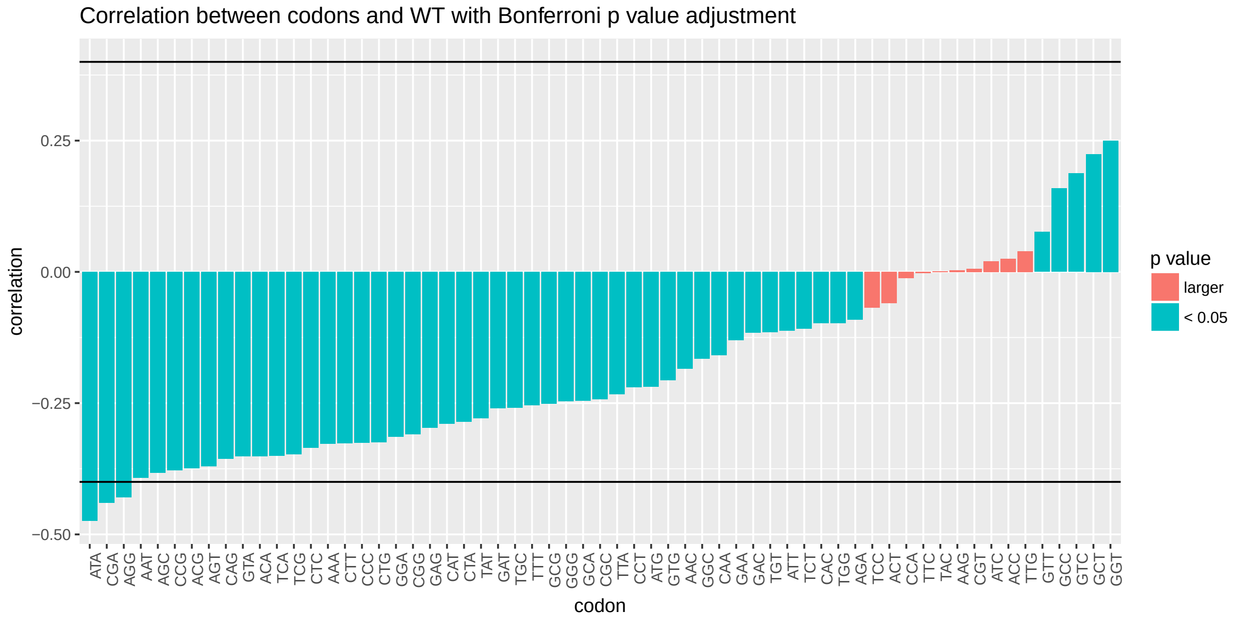
```
##
## Attaching package: 'psych'

## 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)
```

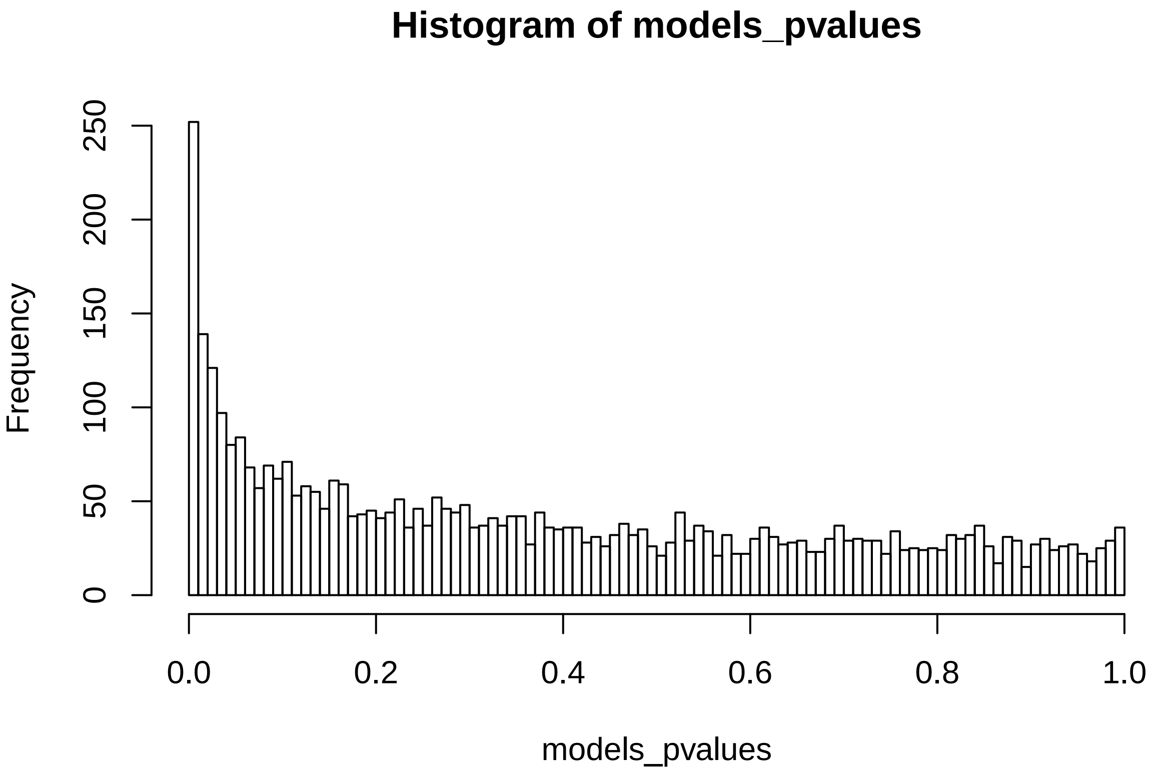
```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
```

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

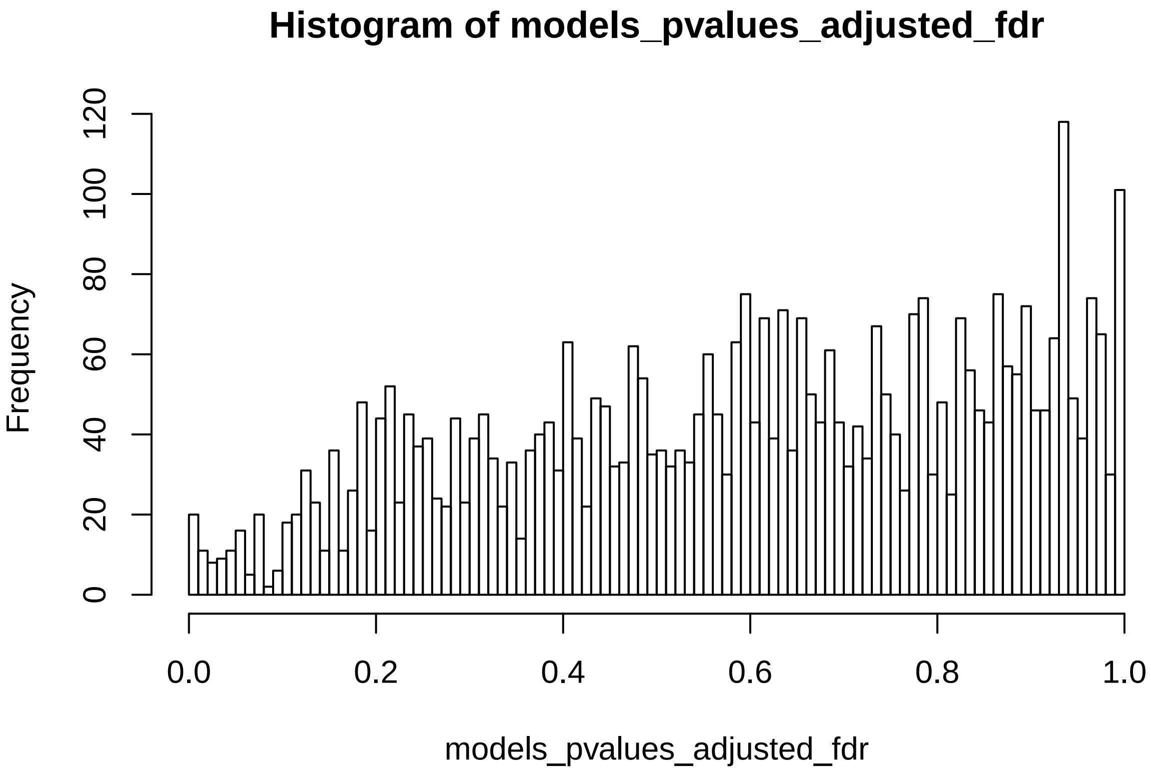
```
library(tidyr)
library(data.table)
```

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

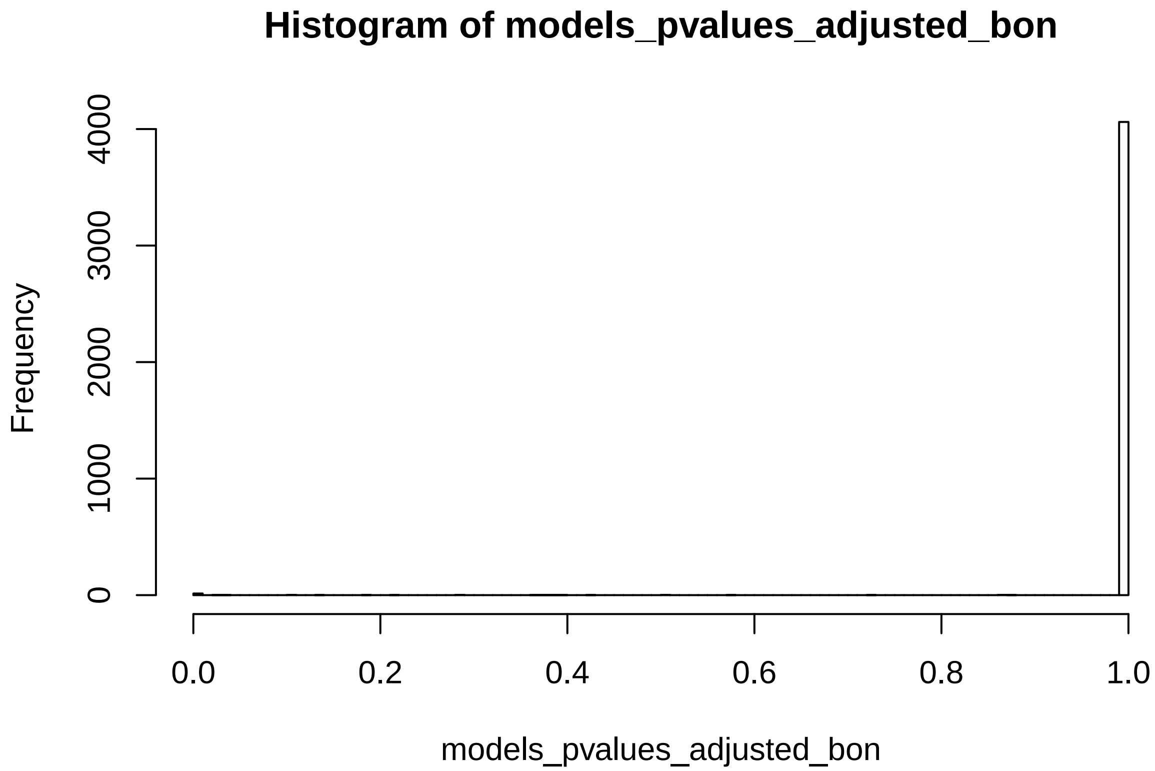
```
library(jsonlite)
library(corrplot)

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)
```



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



```r
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
```

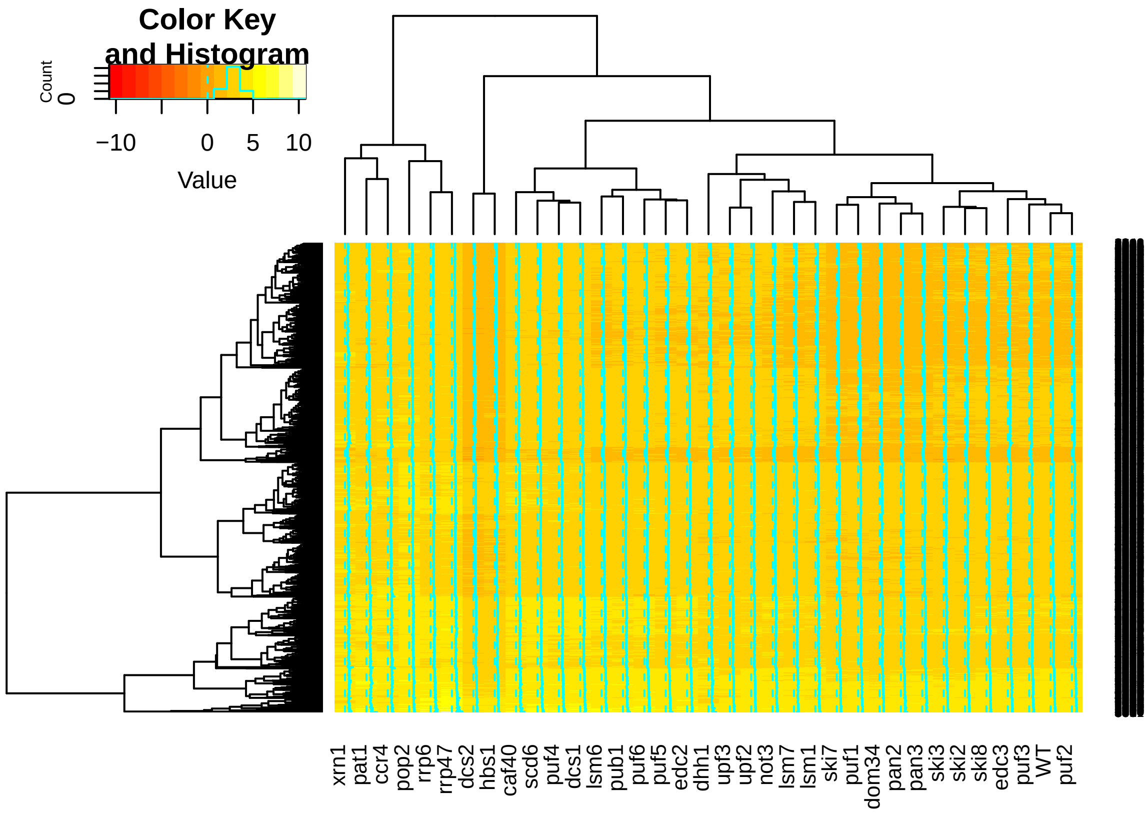
1.3) Some extra-work on the significance of the obtained results:

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

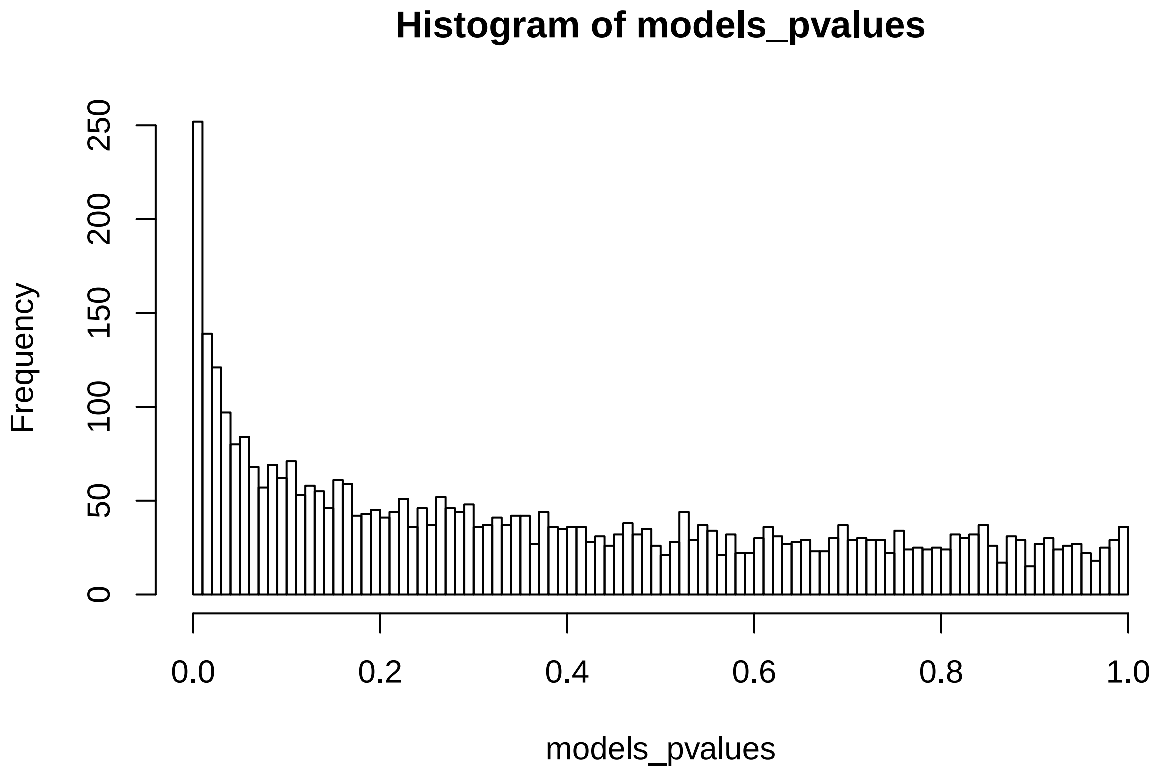
```
##
## Attaching package: 'magrittr'

## The following object is masked from 'package:tidyr':
##
##     extract
```

```
library(ggplot2)

library(dplyr)
library(tidyr)

info <- jsonlite::fromJSON("case_study_info.json")
dt <- readRDS("case_study_dt.rds")
```

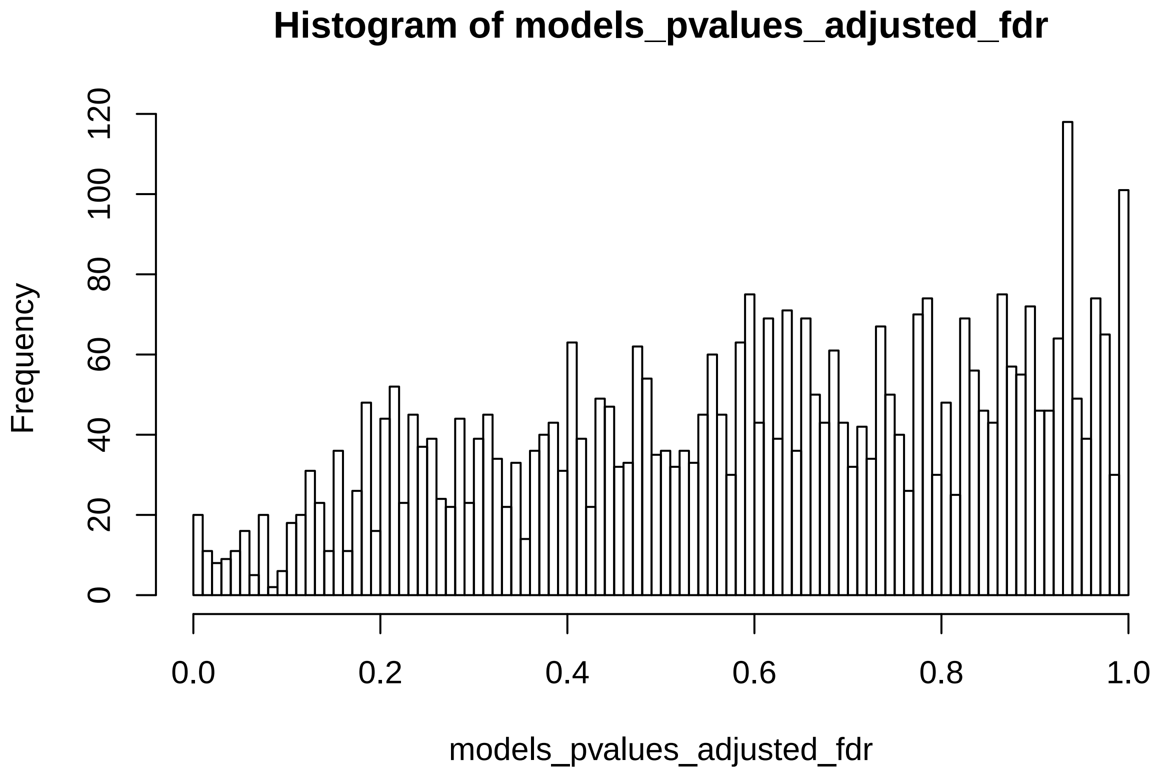
```
utr3_6mer <- as.data.table(readRDS("case_study_utr3_6mer.rds"))

# heatmap strains
gplots::heatmap.2(as.matrix(dt[,info$strains,with=F]))
```



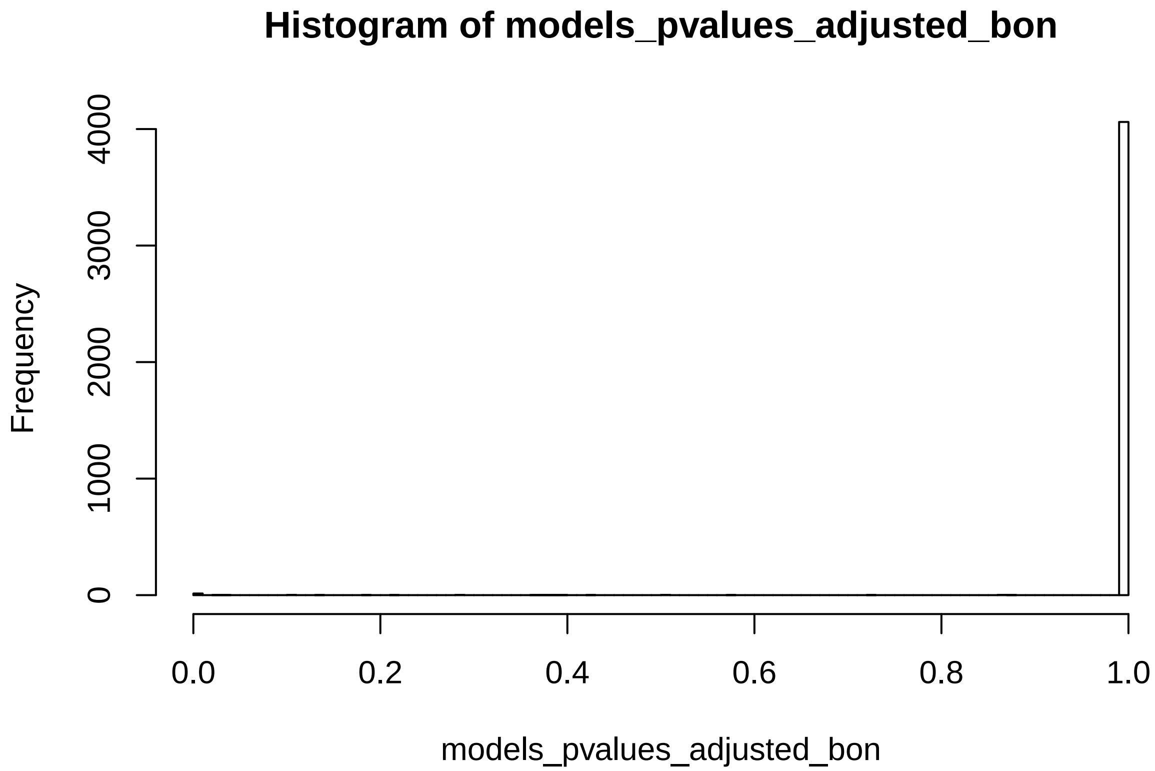
```
###start: code from yesterday
# linear model with only WT as Covariate
models_pvalues <- sapply(utr3_6mer, function(x,dt)
{
  sum <- summary(lm(x ~ dt))$coefficients[8]
},dt=dt$WT)

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)
```



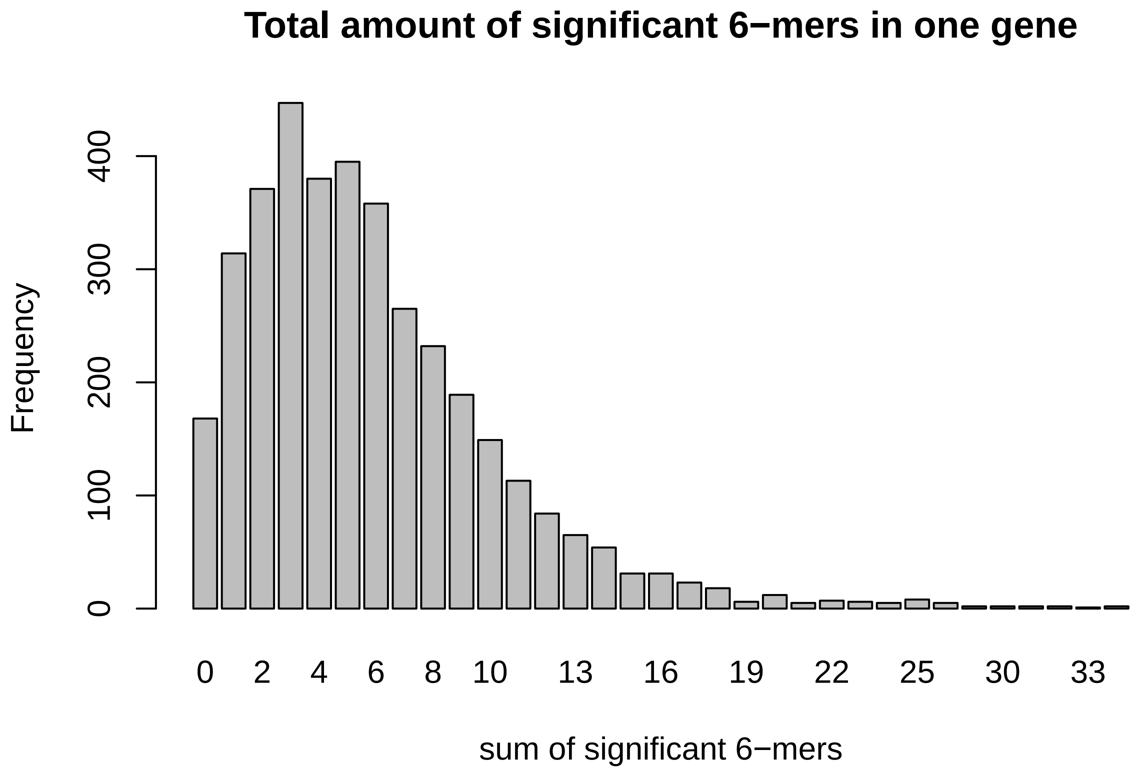
```
hist(models_pvalues_adjusted_bon, breaks=100)
```



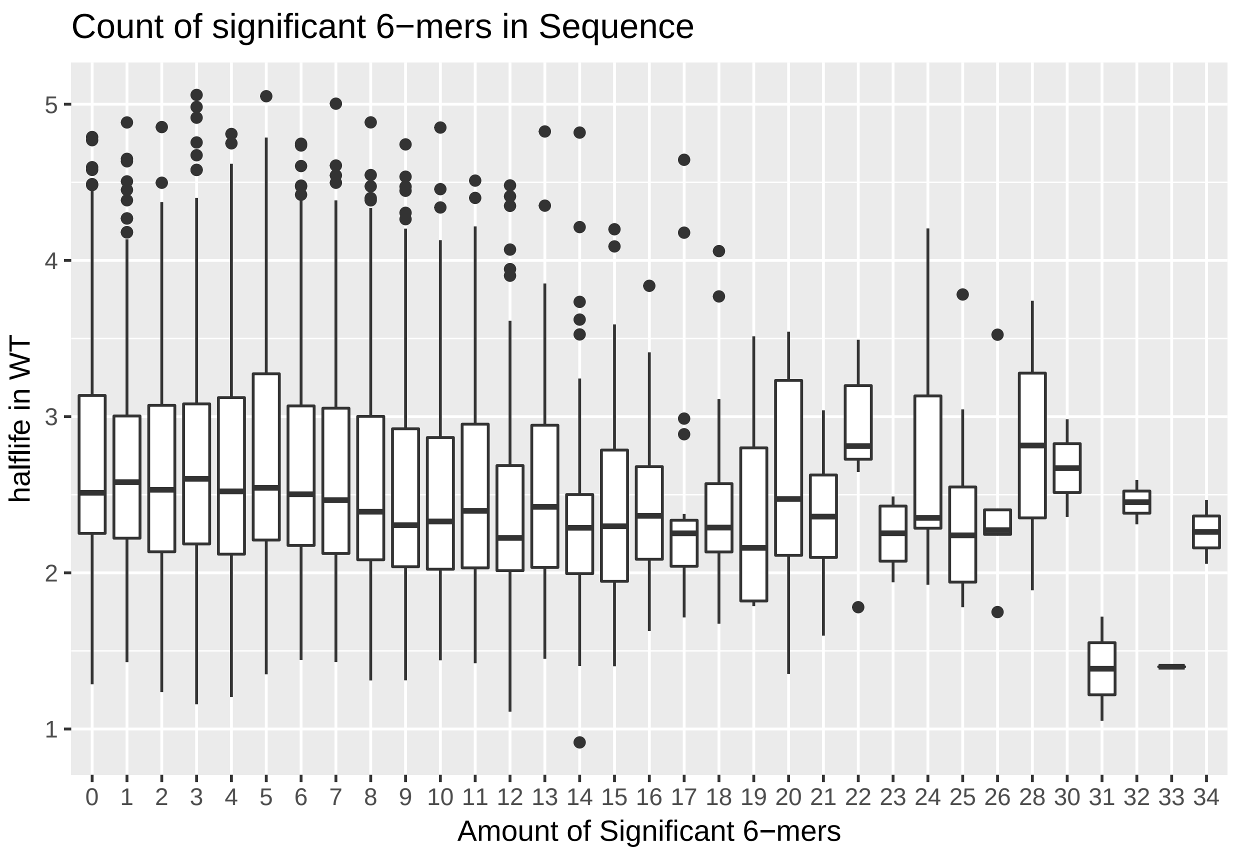
```
sum(models_pvalues_adjusted_bon <= 0.05)
```

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

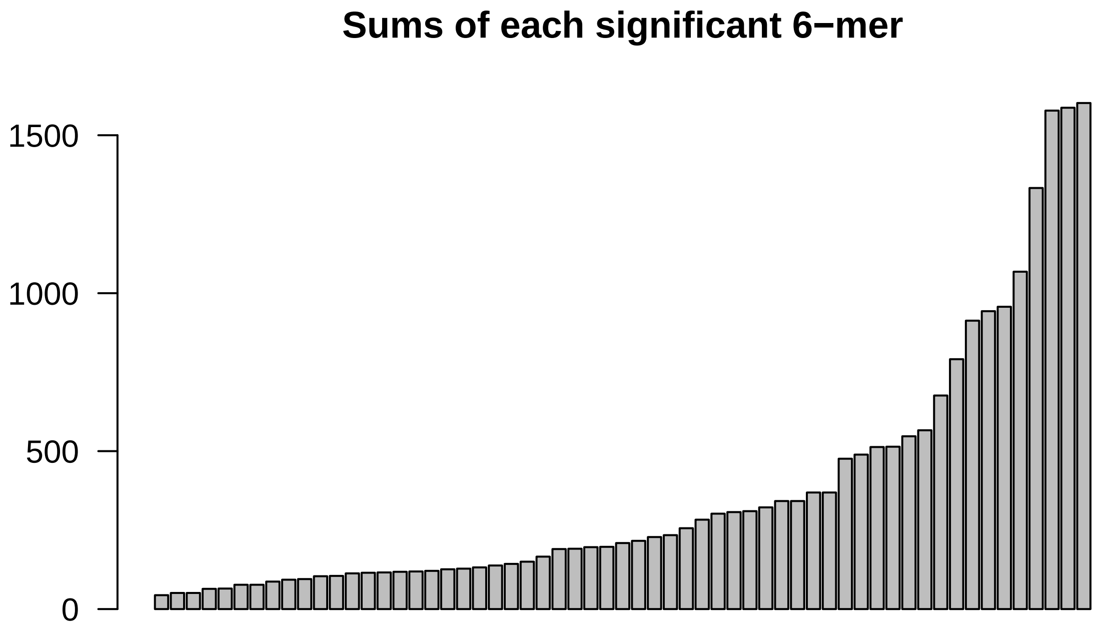
```
#### end: code from yesterday
## linear model with all strains
#all_models_pvalues <- sapply(utr3_6mer, function(x,dt)
#{
#  sum <- summary(lm(x ~ .,data = dt))$coefficients[(ncol(dt)+1)*4]
#},dt=dt[,c(info$strains, info$codons), with=F])

#all.pvals.fdr <- p.adjust(models_pvalues, "fdr")
#sum(all.pvals.fdr <= 0.05)
#identical(names(which(all.pvals.fdr <= 0.05)),names(which(models_pvalues_adjusted_fdr <= 0.05)))


### look at the half time
names_significant <- names(which(models_pvalues_adjusted_fdr <= 0.05))

utr3_significant <- utr3_6mer[, names_significant, with=FALSE]

utr3_significant$genename <- dt$genename

utr3_significant$sum <- apply(utr3_significant[, !"genename"], 1, sum)
barplot(table(utr3_significant$sum), main = "Total amount of significant 6-mers in one gene", xlab = "su
```



```
wt_dt_significant <- merge(utr3_significant, dt[, c("WT", "genename")], by="genename")

## boxplot of halflife in WT by amount of 6-mers in sequence
ggplot(wt_dt_significant,aes(factor(sum),WT))+
  geom_boxplot()+
  ggtitle("Count of significant 6-mers in Sequence")+
  labs(x = "Amount of Significant 6-mers", y = "halflife in WT")
```

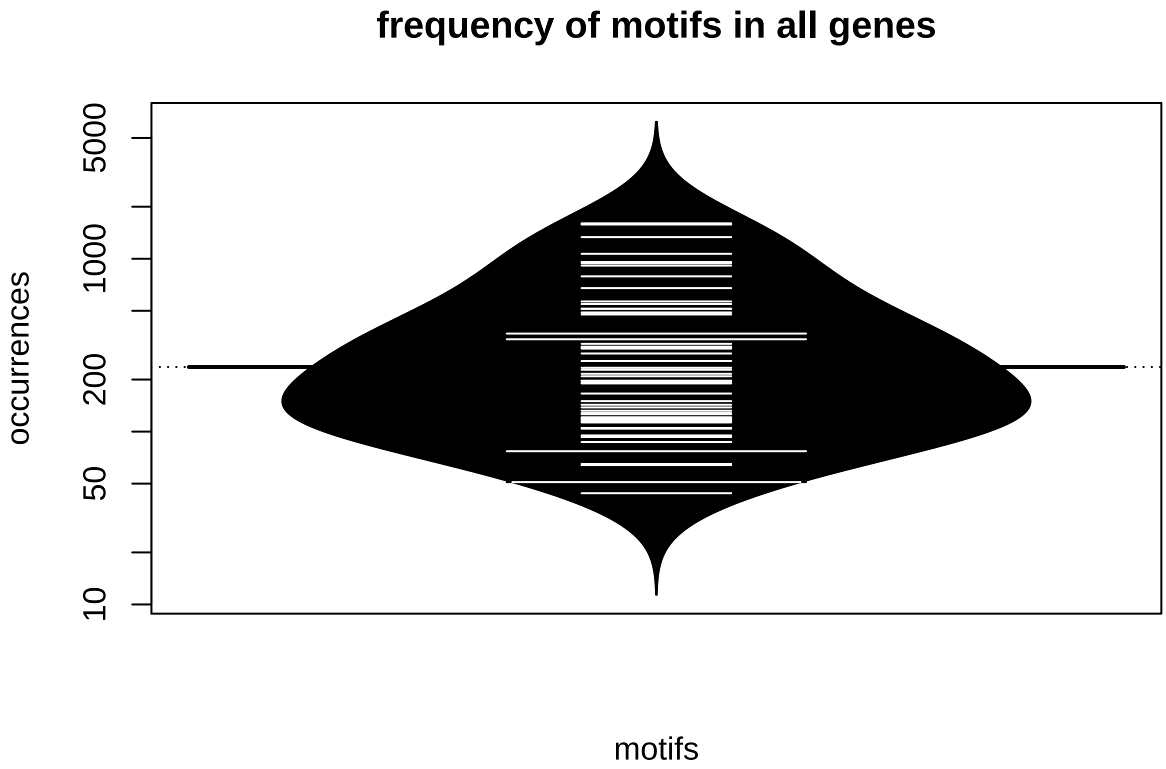


```
#wt_dt_significant[sum==33,]

# histogram of frequencies of the k-mers
barplot(sort(t(colSums(wt_dt_significant[ ,!c("genename","sum","WT")]))),main = "Sums of each significan
```



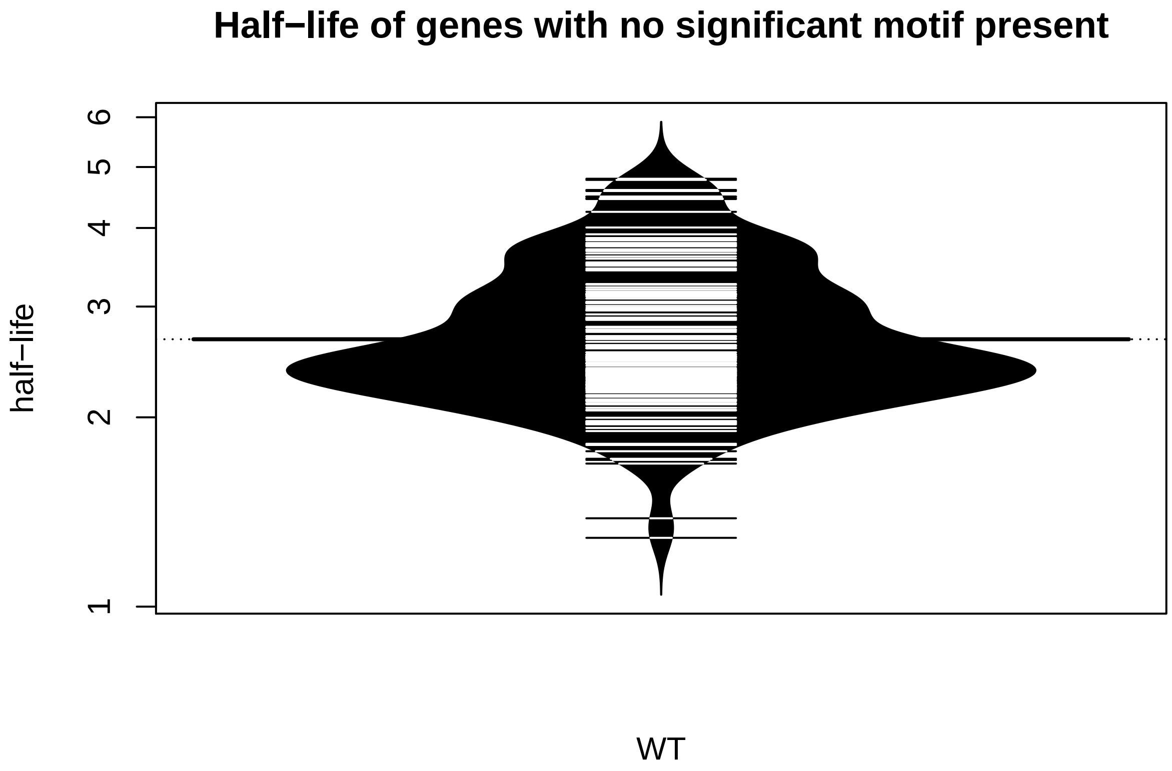
```
library(beanplot)
beanplot(colSums(wt_dt_significant[,!c("genename","sum","WT")]),main="frequency of motifs in all genes"
```

```
## log="y" selected
```

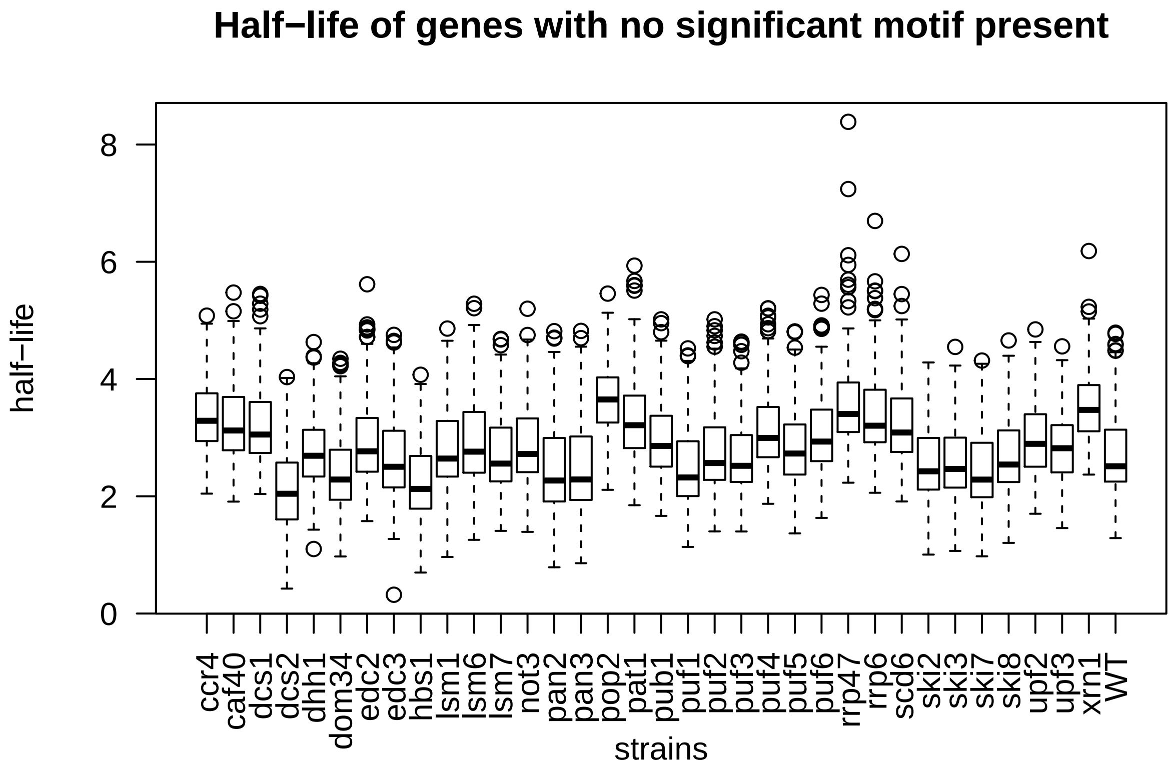


```
# TODO look at halflife where no significant motif is present
beanplot(wt_dt_significant[sum==0 , WT], main="Half-life of genes with no significant motif present",yla
```

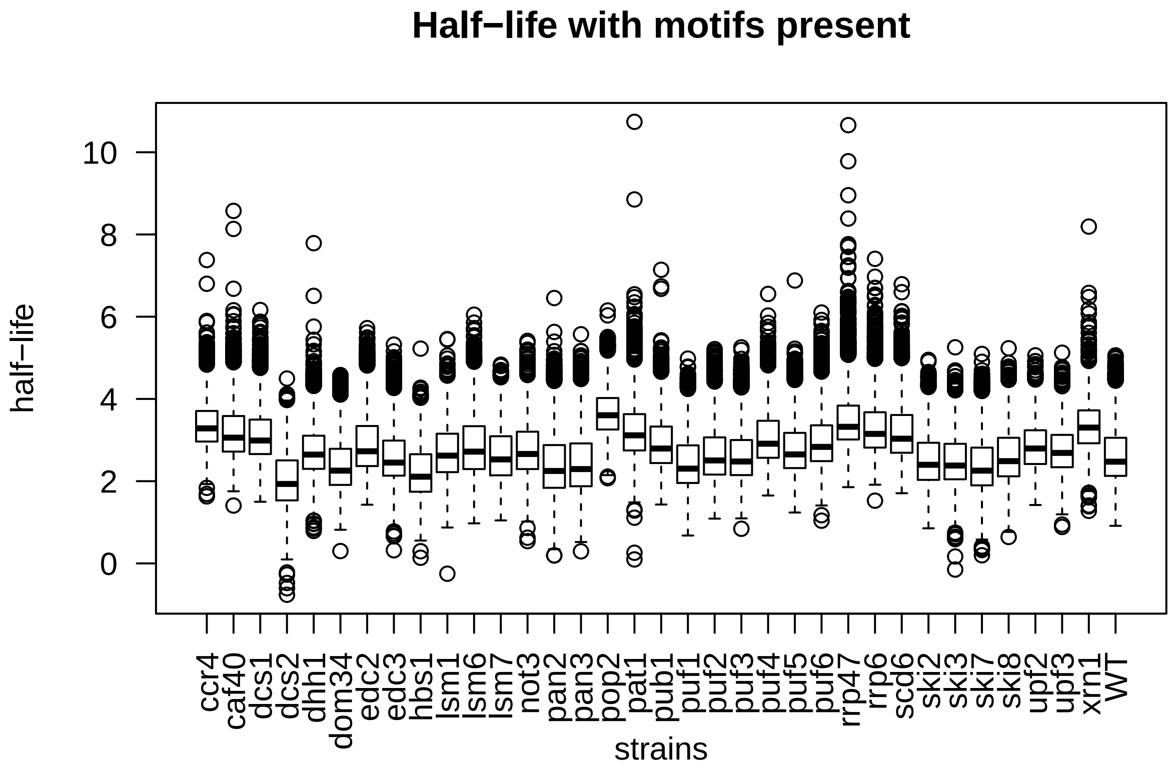
```
## log="y" selected
```



```
genes_nosig_motif <- wt_dt_significant[sum == 0 , genename]
halflifes_nosig_motif <- dt[genename %in% genes_nosig_motif,info$strains,with=F]
boxplot(halflifes_nosig_motif,main="Half-life of genes with no significant motif present",ylab="half-li
```



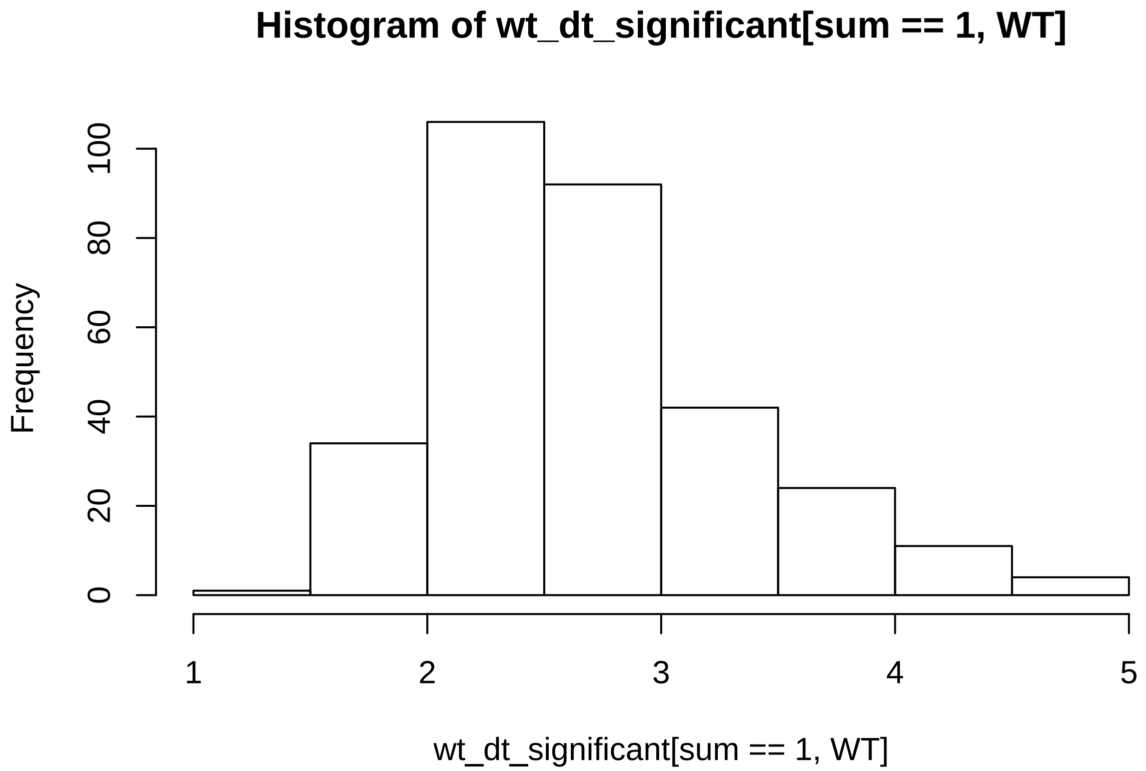
```
halflifes_sig_motifs <- dt[genename %in% utr3_significant$genename, info$strains,with=F]
boxplot(halflifes_sig_motifs,main="Half-life with motifs present",ylab="half-life",xlab="strains",las=2)
```



```
# boxplot of halflife for every significant marker -> melt table
melted_motifs <- melt(wt_dt_significant[,!"sum"],id.vars = c("genename","WT"),measure.vars=names(wt_dt_s
names(melted_motifs)[2]<- "halflife"
melted_motifs <- melted_motifs[freq>0,]
head(melted_motifs)
```

```
##     genename halflife  motif freq
## 1:   YBR018C 2.425463 AACGGA    1
## 2:   YBR057C 2.161163 AACGGA    1
## 3:   YBR146W 2.492733 AACGGA    1
## 4:   YCR015C 2.158706 AACGGA    1
## 5:   YDL045C 2.335400 AACGGA    1
## 6: YDL045W-A 2.261548 AACGGA    1
```

```
## look at genes with only one significant 6-mer that have a high half-life and check if they are also 
hist(wt_dt_significant[sum==1,WT])
```



```
single.sig.motifs.high.halflife.genename <- wt_dt_significant[sum==1&WT>4,genename]
single.sig.motifs.low.halflife.genename <- wt_dt_significant[sum==1&WT<2,genename]

## try to match all motifs that occur together
motifs.pergene <- sapply(unique(melted_motifs$genename), function(gene,dt){
  dt[genename == gene,motif]
}, dt = melted_motifs)

#length(motifs.pergene[c(single.sig.motifs.high.halflife.genename)])
sig.motifs.high.halflife <- as.character(unlist(motifs.pergene[c(single.sig.motifs.high.halflife.genenam
table(sig.motifs.high.halflife)
```

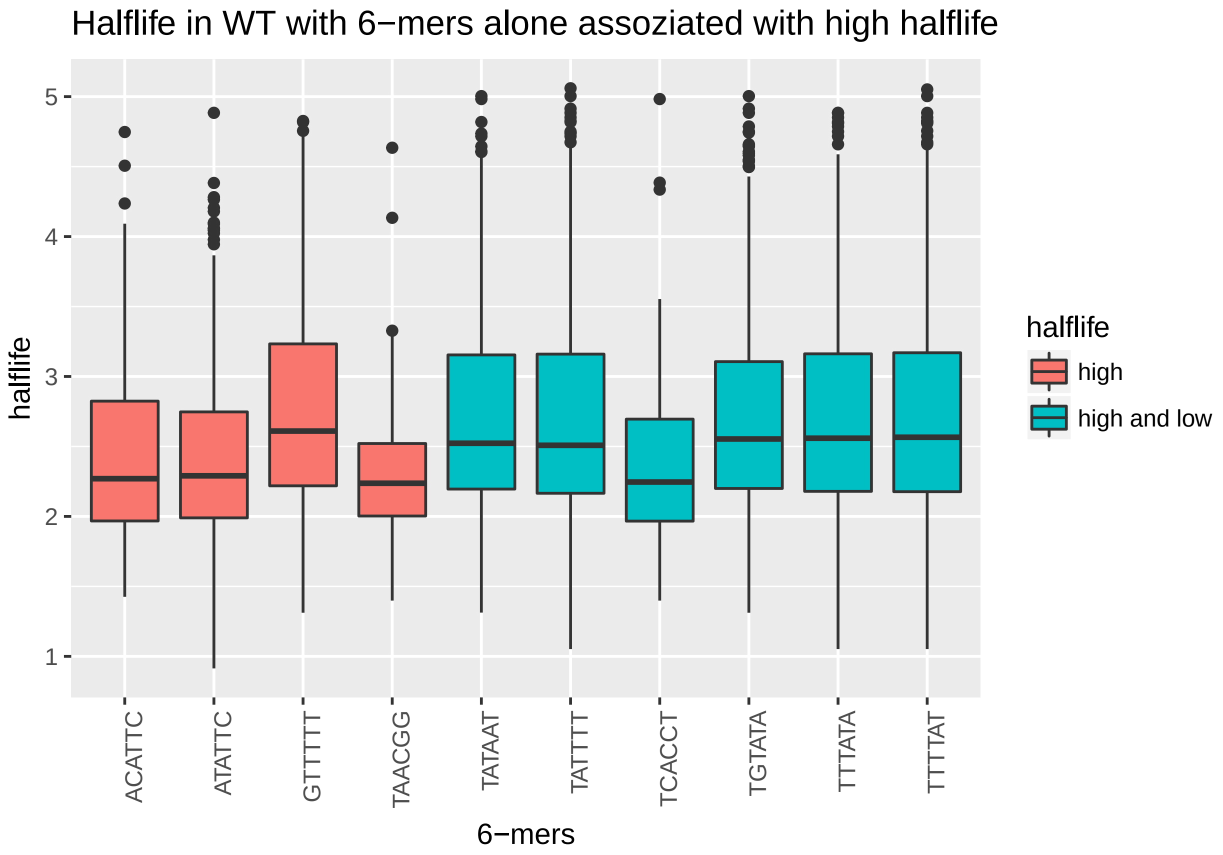
```
## sig.motifs.high.halflife
## ACATTC ATATTC GTTTTT TAACGG TATAAT TATTTT TCACCT TGTATA TTTATA TTTTAT
##      1      1      1      1      2      1      1      4      2      1
```

```
sig.motifs.low.halflife <- as.character(unlist(motifs.pergene[c(single.sig.motifs.low.halflife.genename
table(sig.motifs.low.halflife)
```

```
## sig.motifs.low.halflife
## AATATT ACTGCA ATTCAC CAAATT CATATT CATTTC CGCATA GCATTT TAGCAT TATAAT
##      2      1      2      2      2      1      1      1      1      4
## TATTTT TCACCT TGACCA TGTAAA TGTATA TTCGGA TTTATA TTTGCA TTTTAT TTTTTA
##      1      1      2      4      2      1      3      1      2      1
```

```
in.both <- intersect(sig.motifs.high.halflife, sig.motifs.low.halflife)
```

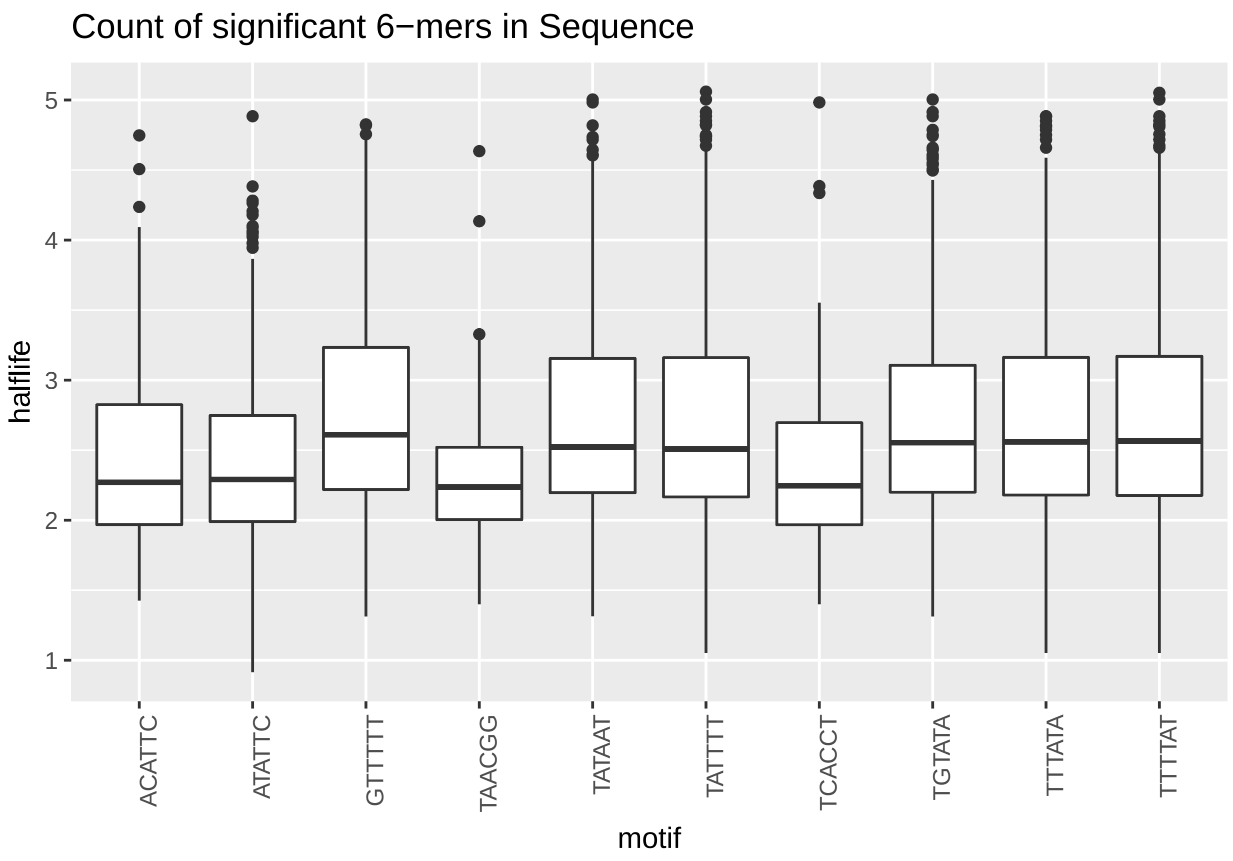
```
melted_motifs.high.halflife <- melted_motifs[motif %in% (unique(sig.motifs.high.halflife)),]

low.and.high <- melted_motifs.high.halflife[,motif] %in% in.both

ggplot(melted_motifs.high.halflife,aes(motif,halflife, fill=low.and.high))+
  geom_boxplot()+
  ggtitle("Halflife in WT with 6-mers alone assoziated with high halflife") +
  scale_fill_discrete(name="halflife",
                      breaks=c("FALSE", "TRUE"),
                      labels=c("high", "high and low")) +
  labs(x = "6-mers") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



```
# are those in genes with low half-lifes?


## get genes with different methylation status classification:
range(melted_motifs$halflife)
```

```
## [1] 0.9140725 5.0595045
```

```
ggplot(melted_motifs.high.halflife,aes(motif,halflife))+
  geom_boxplot()+
  ggtitle("Count of significant 6-mers in Sequence") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



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")
```

```r
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
```

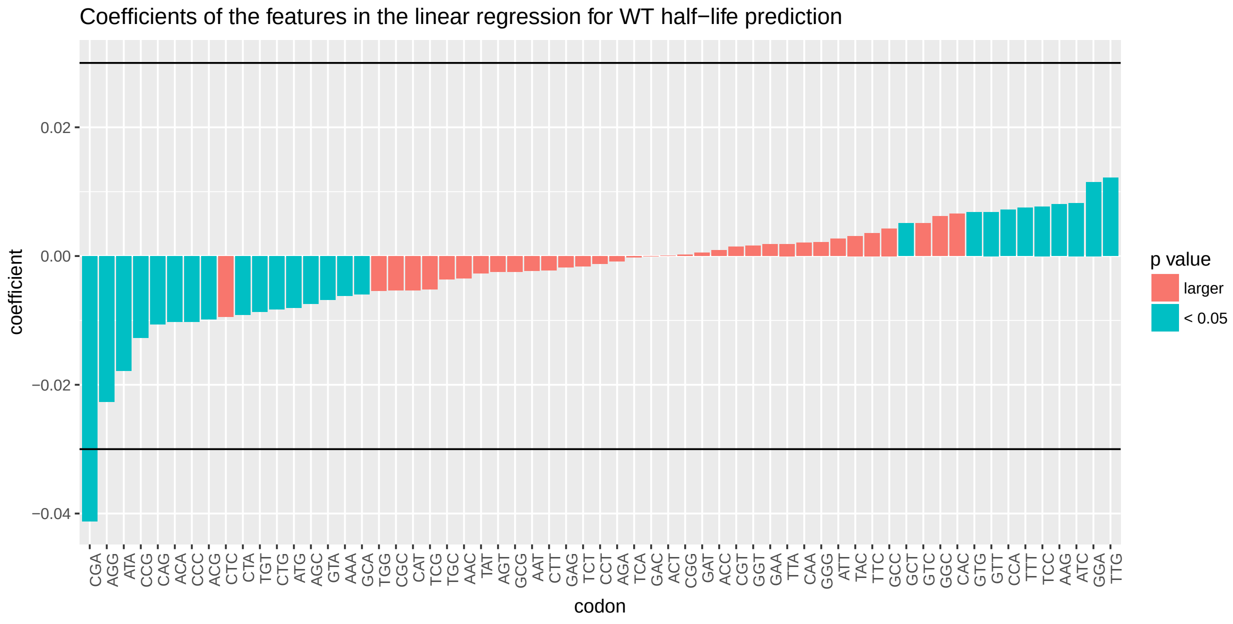
Results of the linear regression:

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

```r
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))
```


Coefficients of the features in the linear regression for WT half−life prediction
coefficient
CGA
AGG
ATA
CCG
CAG
ACA
CCC
ACG
CTC
CTA
TGT
CTG
ATG
AGC
GTA
AAA
GCA
TGG
CGC
CAT
TCG
TGC
AAC
TAT
AGT
GCG
AAT
CTT
GAG
TCT
CCT
AGA
TCA
GAC
ACT
CGG
GAT
ACC
CGT
GGT
GAA
TTA
CAA
GGG
ATT
TAC
TTC
GCC
GCT
GTC
GGC
CAC
GTG
GTT
CCA
TTT
TCC
AAG
ATC
GGA
TTG
codon
p value
larger
< 0.05