Acetilation

Lucas Michel Todó

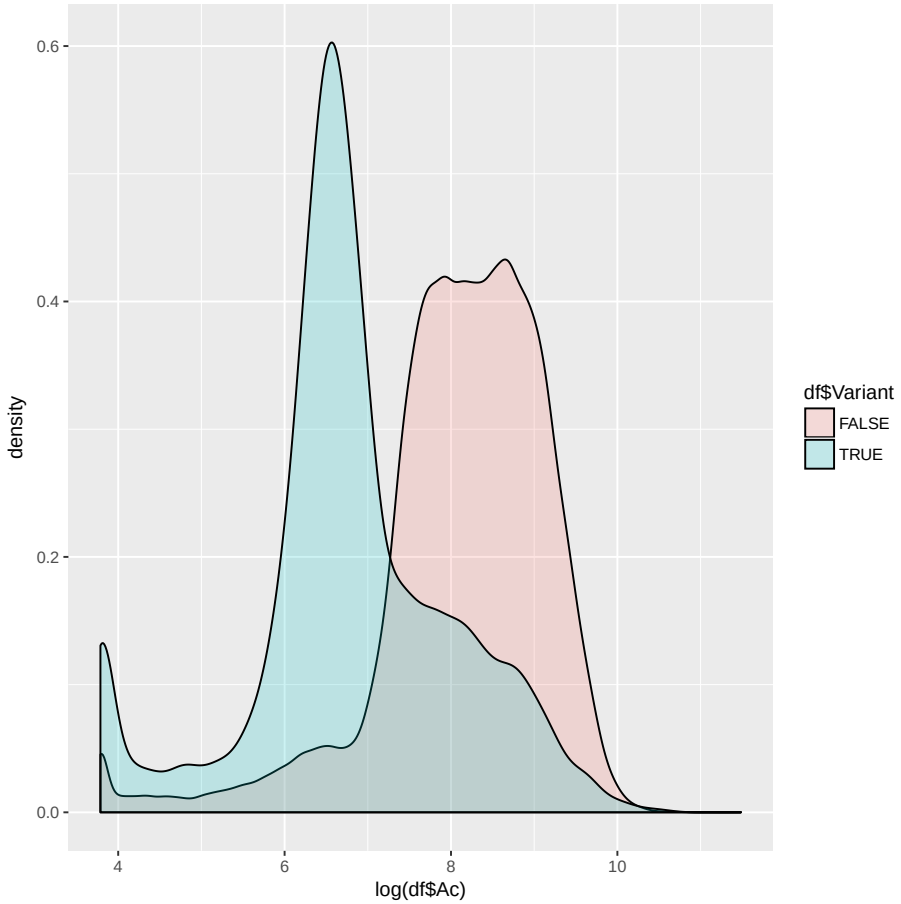
December 20, 2017

Contents

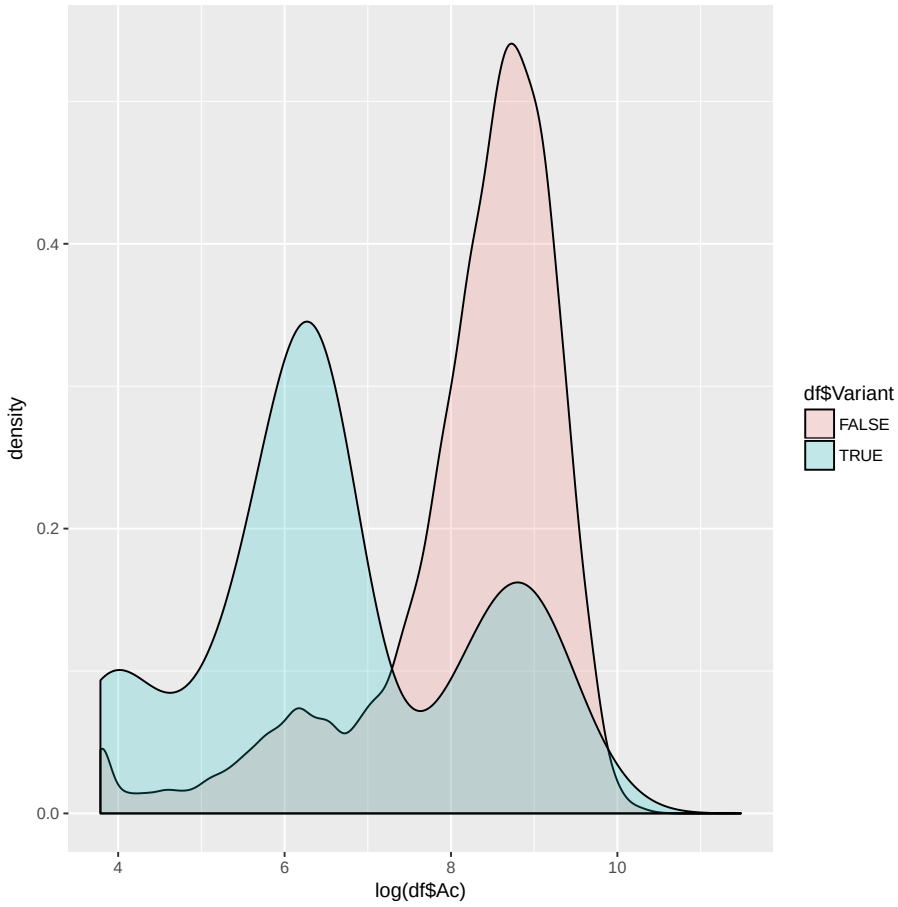
```
##
## Attaching package:  'dplyr'
## The following objects are masked from 'package:stats':
##
##    filter, lag
## The following objects are masked from 'package:base':
##
##    intersect, setdiff, setequal, union
```

1 Density Plots

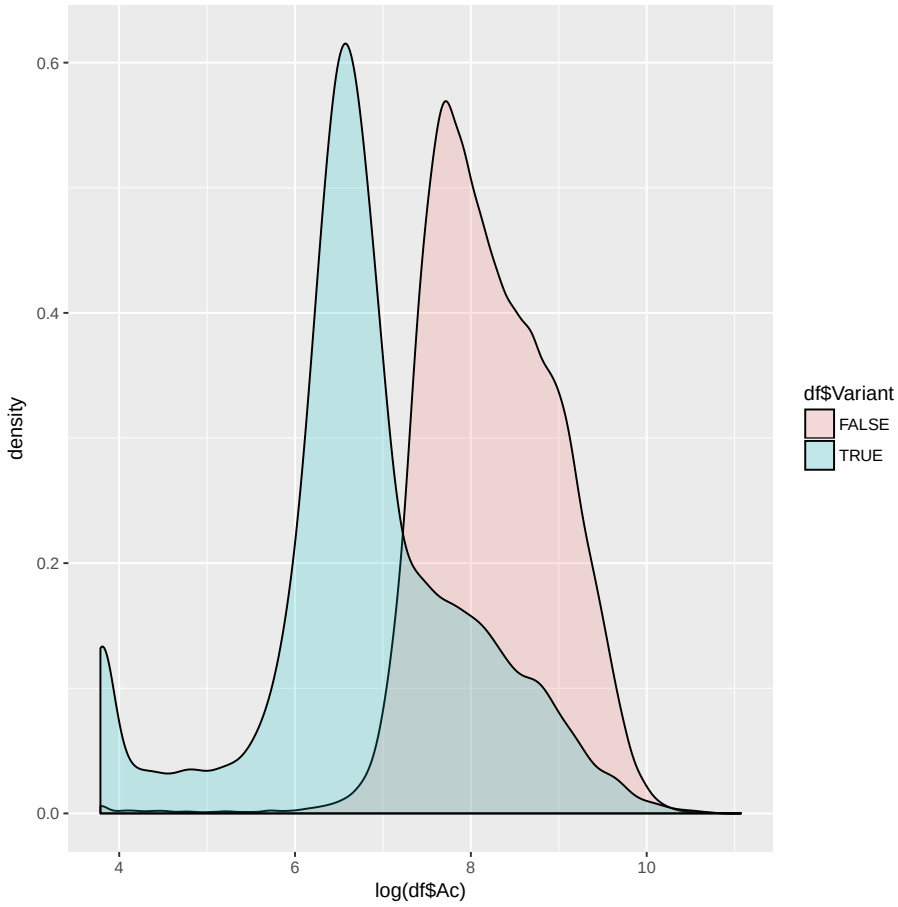
1.1 log(Ac) All



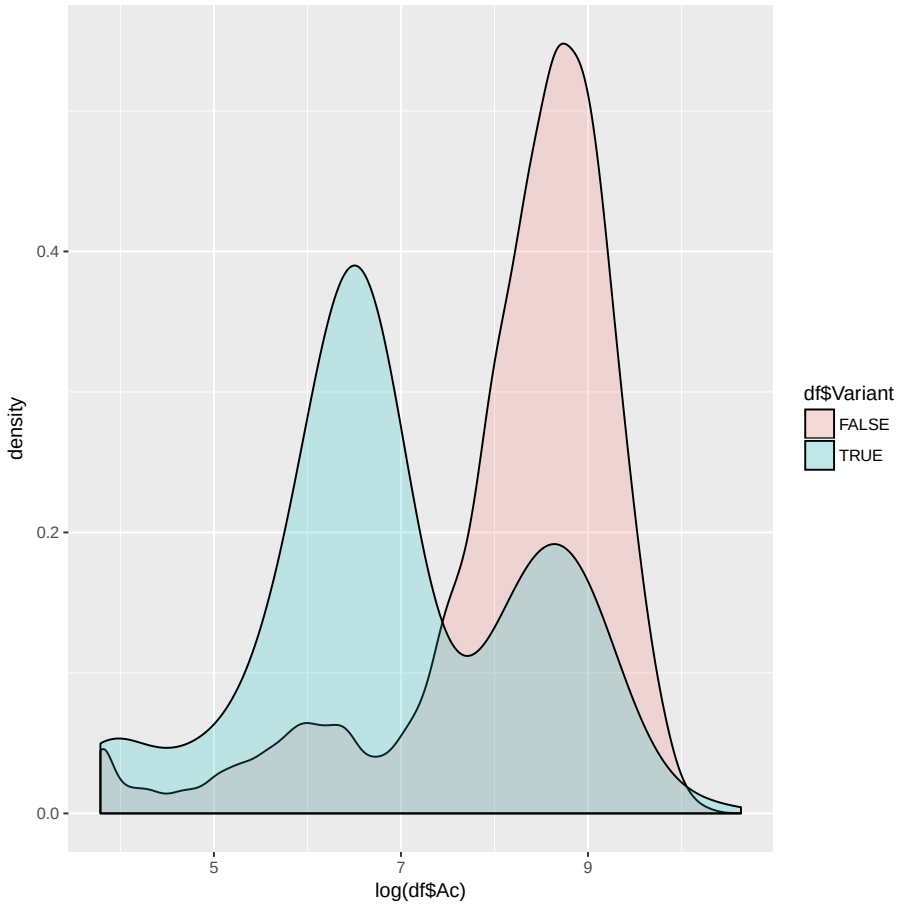
1.2 log(Ac) 5'



1.3 log(Ac) ORF



1.4 log(Ac) 3'




```
table(cov$Variant)

##
##  FALSE   TRUE
## 110292   6365

#df <- cov
df <- rbind(sample_n(cov[cov$Variant == FALSE,], 6365), cov[cov$Variant == TRUE,])

train_idx <- rownames(sample_n(df, 8486))
train_df <- df[rownames(df) %in% train_idx,]
test_df <- df[!rownames(df) %in% train_idx,]

model <- glm(Variant ~ Ac+Met+Type, family=binomial(link='logit'),data=train_df)
summary(model)
```

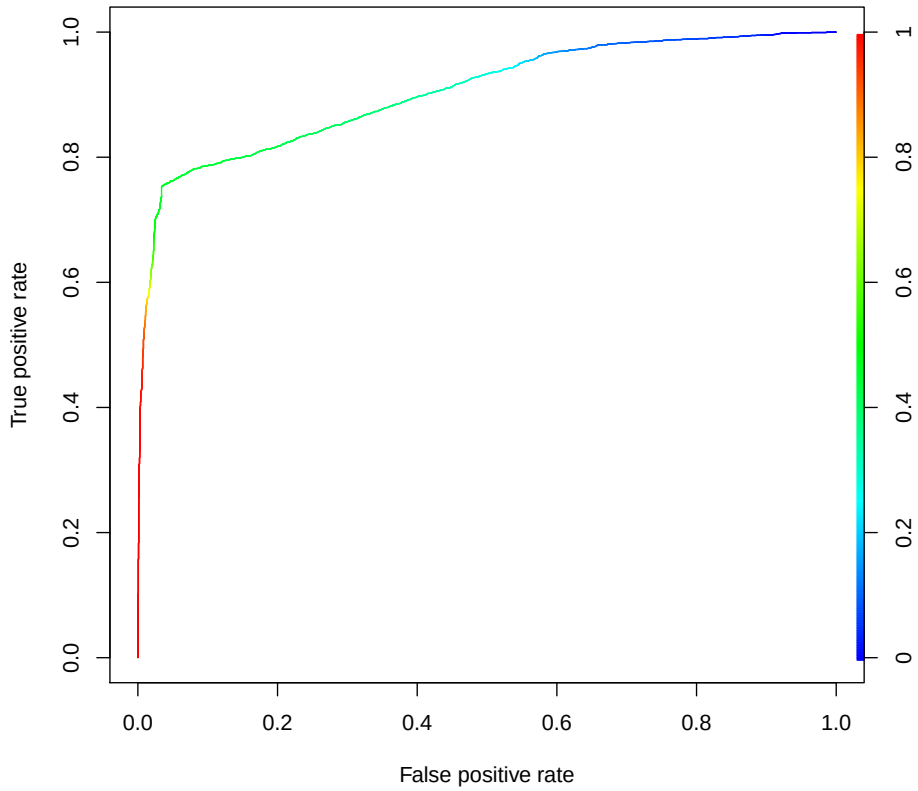
```
##
## Call:
## glm(formula = Variant ~ Ac + Met + Type, family = binomial(link = "logit"),
##     data = train_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.9780  -0.8216  -0.0517   0.3802   3.5533
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.519e+00  1.214e-01 -12.513  < 2e-16 ***
## Ac          -1.103e-04  9.176e-06 -12.018  < 2e-16 ***
## Met          1.873e-04  7.302e-06  25.651  < 2e-16 ***
## Type5prima  -5.411e-01  1.572e-01  -3.443 0.000575 ***
## TypeORF      1.375e+00  1.167e-01  11.783  < 2e-16 ***
## Typeother   -3.570e+00  3.848e-01  -9.278  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11764.1  on 8485  degrees of freedom
## Residual deviance:  6670.4  on 8480  degrees of freedom
## AIC: 6682.4
##
## Number of Fisher Scoring iterations: 7

fitted.results <- predict(model, test_df, type='response')
table(test_df$Variant, fitted.results > 0.5)

##
##         FALSE TRUE
##   FALSE  2071   48
##   TRUE    689 1436

fitted.results <- ifelse(fitted.results > 0.5,1,0)
misClasificError <- mean(fitted.results != test_df$Variant)
print(paste('Accuracy',1-misClasificError))

## [1] "Accuracy 0.826343072573044"

test_df["Pred"] <- fitted.results
test_df["all_0"] <- 0
misClasificError2 <- mean(test_df$all_0 != test_df$Variant)
print(paste('Accuracy of null model',1-misClasificError2))
```

```
## [1] "Accuracy of null model 0.499293119698398"
```

```
library(ROCR)
```

```
## Loading required package:  gplots
##
## Attaching package:  'gplots'
## The following object is masked from 'package:stats':
##
##     lowess
```

```
predict <- predict(model, type = 'response')
ROCRpred <- prediction(predict, train_df$Variant)
ROCRperf <- performance(ROCRpred, 'tpr','fpr')
plot(ROCRperf, colorize = TRUE, text.adj = c(-0.2,1.7))
```



```r
false_pos_idx <- rownames(test_df[test_df$Pred == 1 & test_df$Variant == FALSE,])
false_pos <- cov[rownames(cov) %in% false_pos_idx,]
table(false_pos$Type)

##
## 3prima 5prima    ORF  other
##     14     19     10      5

false_neg_idx <- rownames(test_df[test_df$Pred == 0 & test_df$Variant == TRUE,])
false_neg <- cov[rownames(cov) %in% false_neg_idx,]
table(false_neg$Type)

##
## 3prima 5prima    ORF  other
##     51     60    572      6
```