Metilació i Acetilació a 10G i 1.2B

Alfred Cortés i Lucas Michel Todó

February 15, 2019

Contents

1 Introducció

Les dades representades en aquest set de gràfiques representen l'acetilació i meilació de de les mostres 10G i 1.2B del nostre estudi de Chip-Seq. El genoma sencer de les dues mostres s'ha partit en fragments de 200bp que s'on els que conformen la base de les dades representades. Cada fragment de 200bp porta associada informació respecte el gen a què correspon, l'estat d'acetilació i metilació i si correspon a una zona 3',5' o ORF.

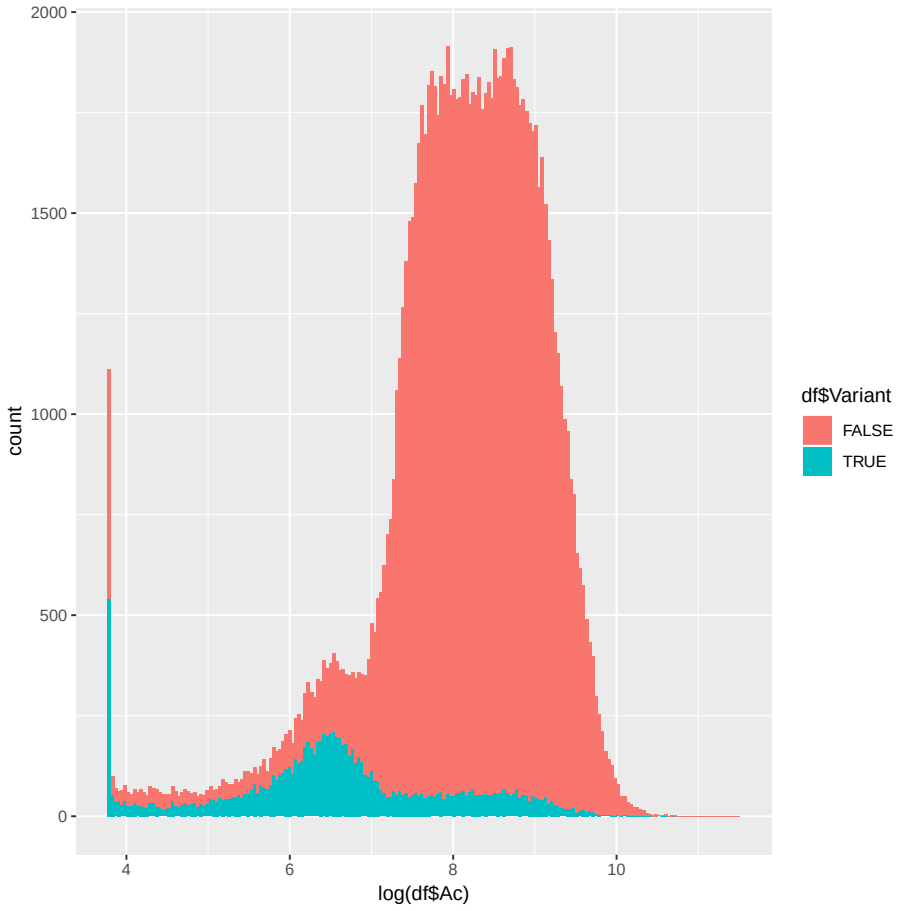
```
## Error in library(pscl):  there is no package called 'pscl'
## Error in library(XLConnect):  there is no package called 'XLConnect'
```

```
## Error in readWorksheetFromFile("/home/lucas/ISGlobal/Chip_Seq/Transcripció_CSV/3D7_Varian
:  could not find function "readWorksheetFromFile"
## Error in eval(expr, envir, enclos):  object 'trans_df' not found
## Error in eval(expr, envir, enclos):  object 'trans_df' not found
## Error in eval(expr, envir, enclos):  object 'trans_df' not found
## Error in ref$ID %in% noexprs:  object 'noexprs' not found
## Error in eval(expr, envir, enclos):  object 'noexprs_df' not found
```

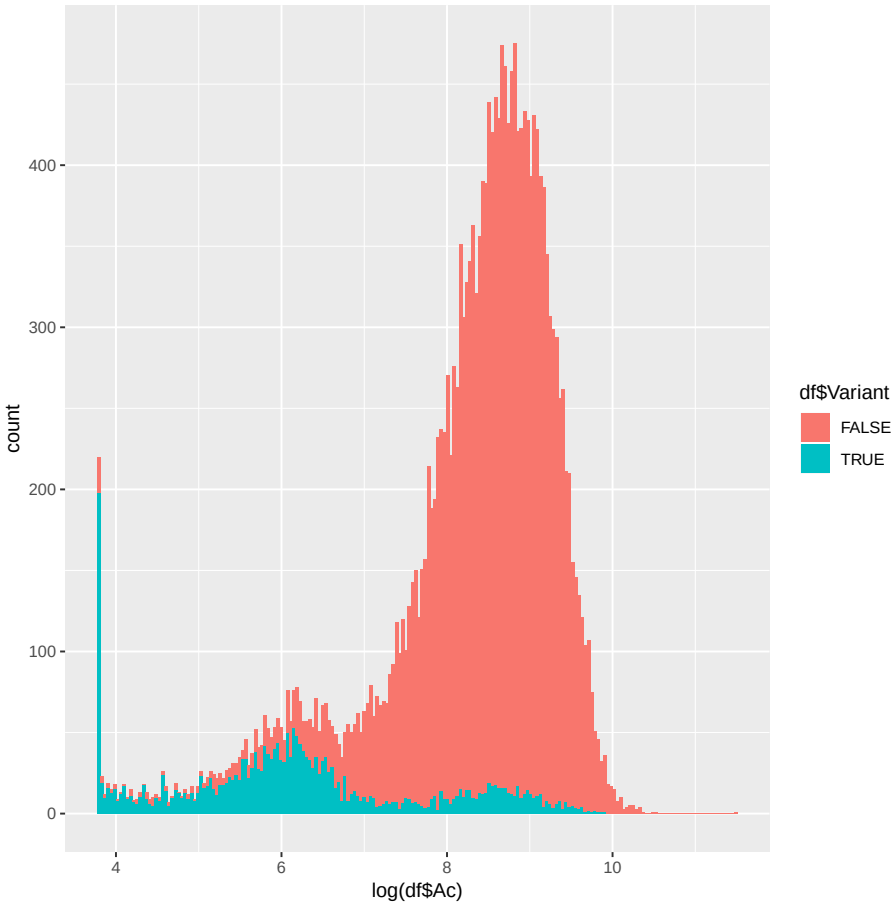
2 Histogrames

En primer lloc hem representat la distribució dels valors d'acetilació i de metilació als fragments (de 10G). (Archiu "Coverage_10G_200bp.csv", 116656 fragments corresponents al genoma sencer. Coverage directament del "bam" normalitzat per nombre de reads.)

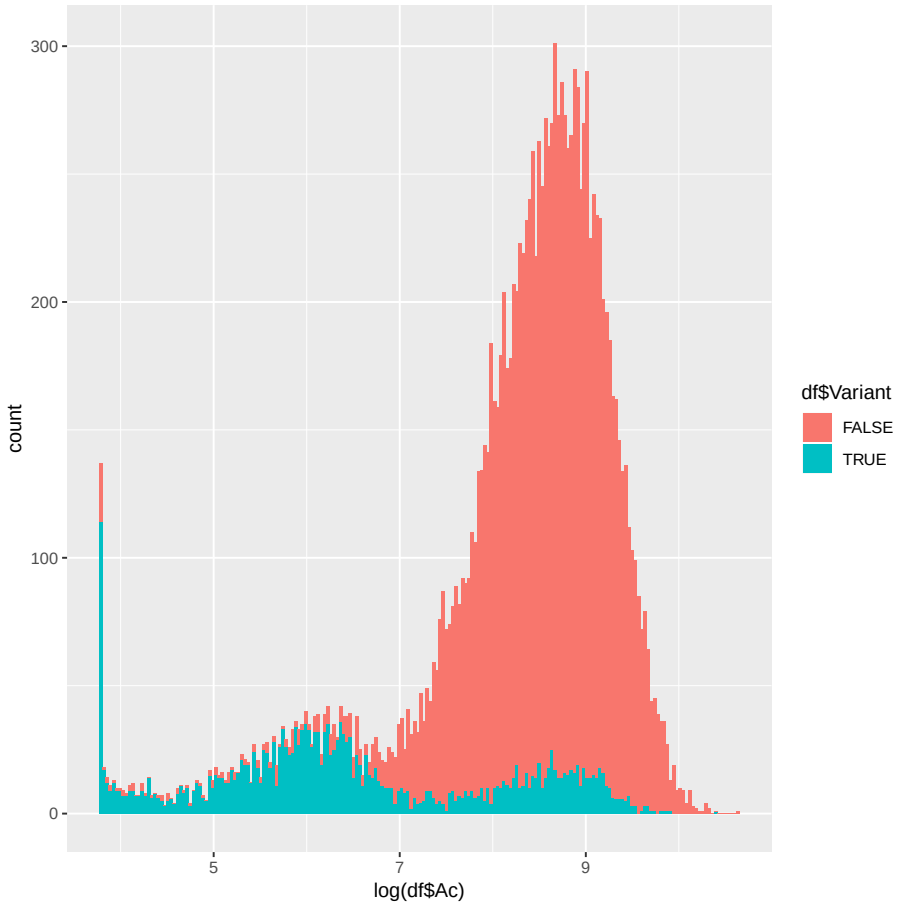
2.1 log(Ac) All


2000
1500
1000
500
0
count
4
6
8
10
log(df$Ac)
df$Variant
FALSE
TRUE

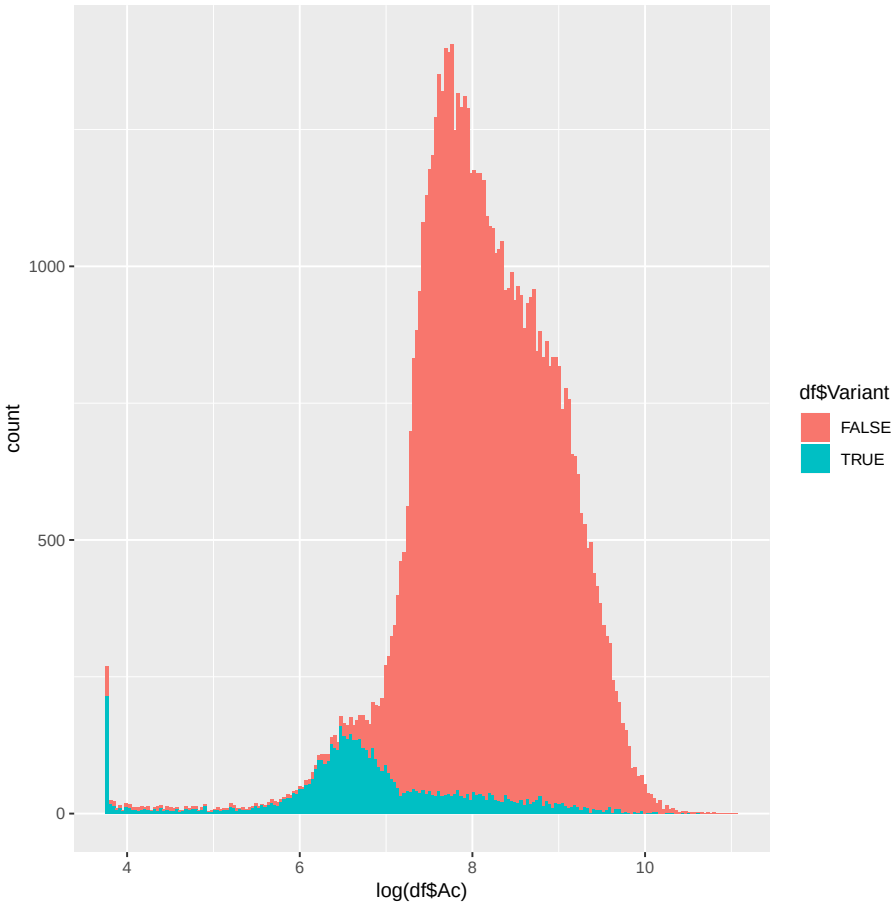
2.2 log(Ac) 5'



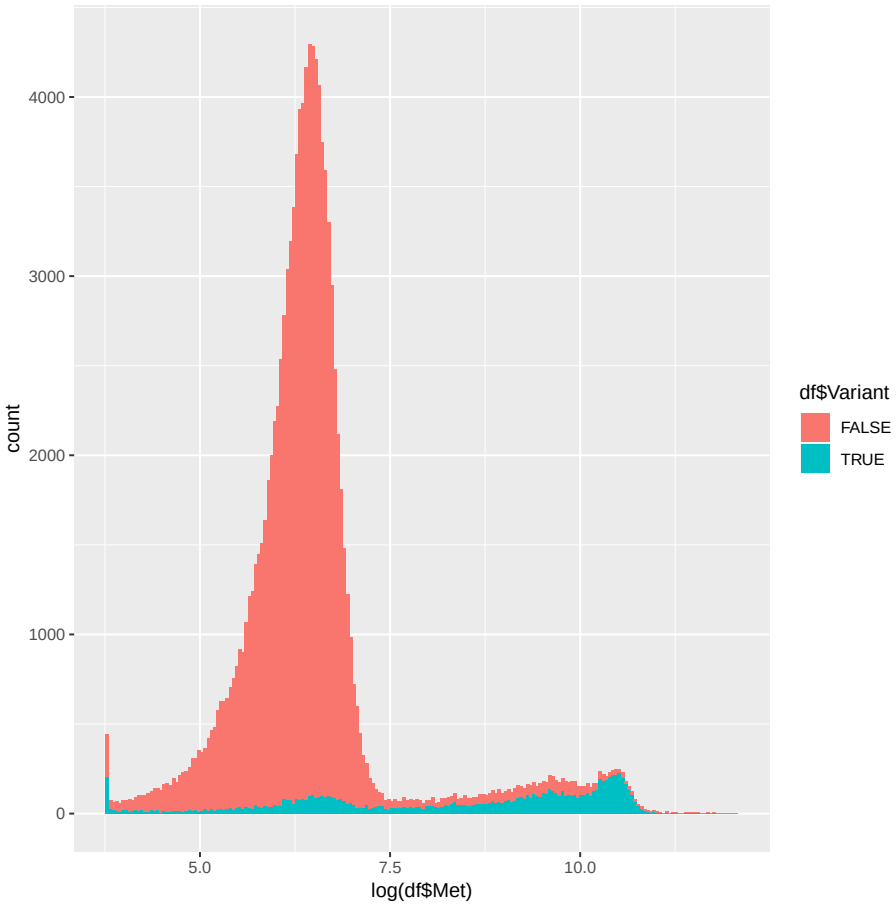
2.3 log(Ac) 3'



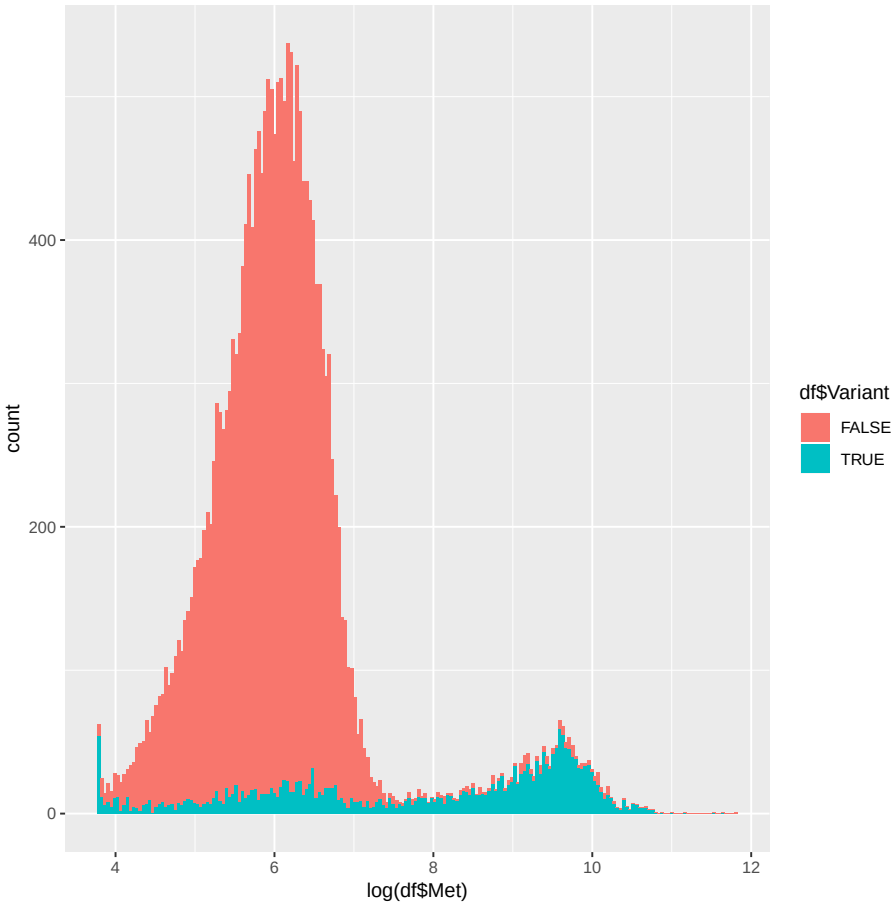
2.4 log(Ac) ORF



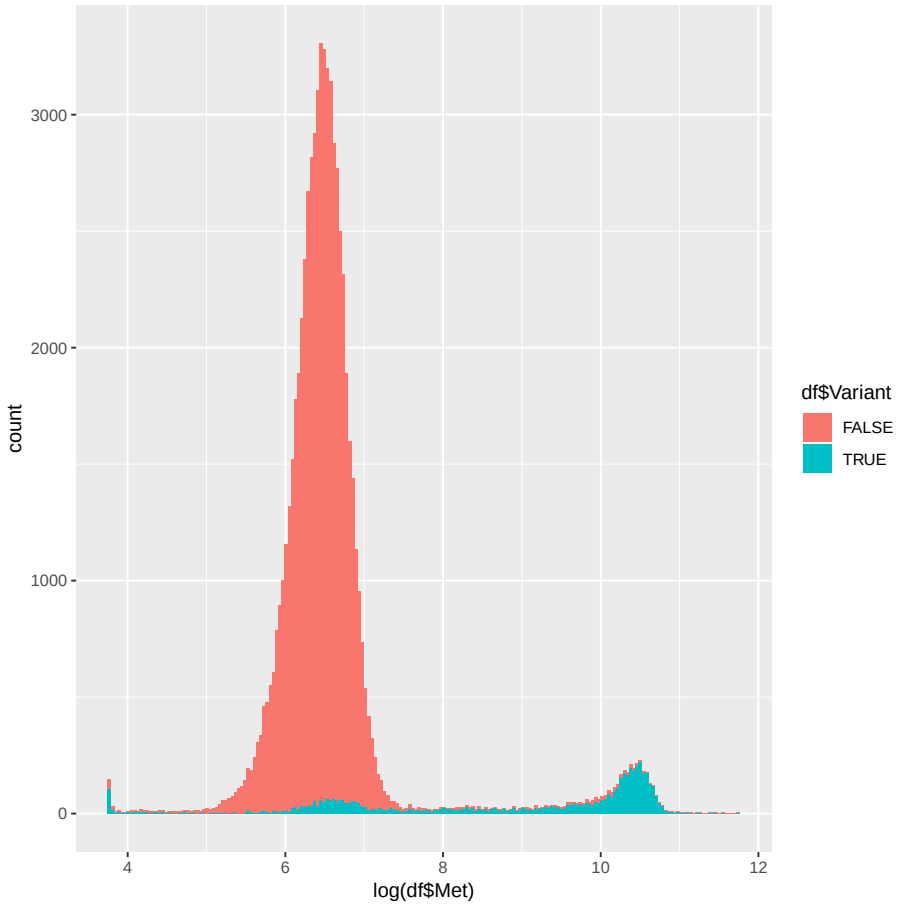
2.5 log(Met) All



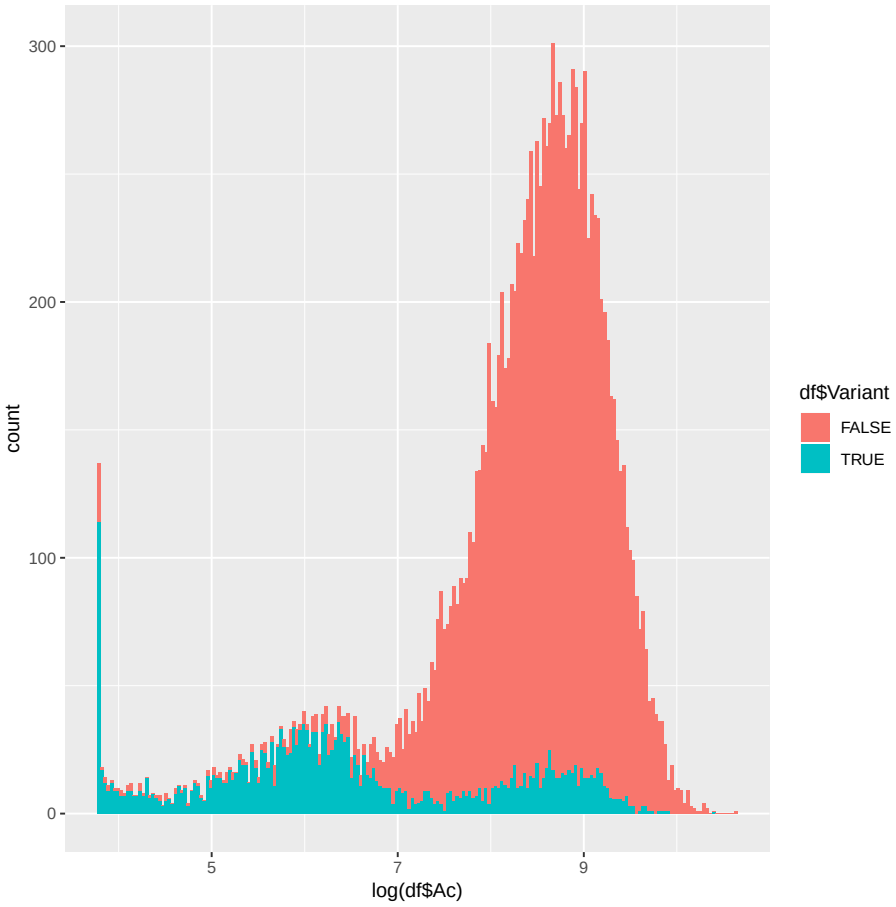
2.6 log(Met) 5'



2.7 log(Met) ORF



2.8 log(Ac) 3'



3 Acetilació vs Metilació

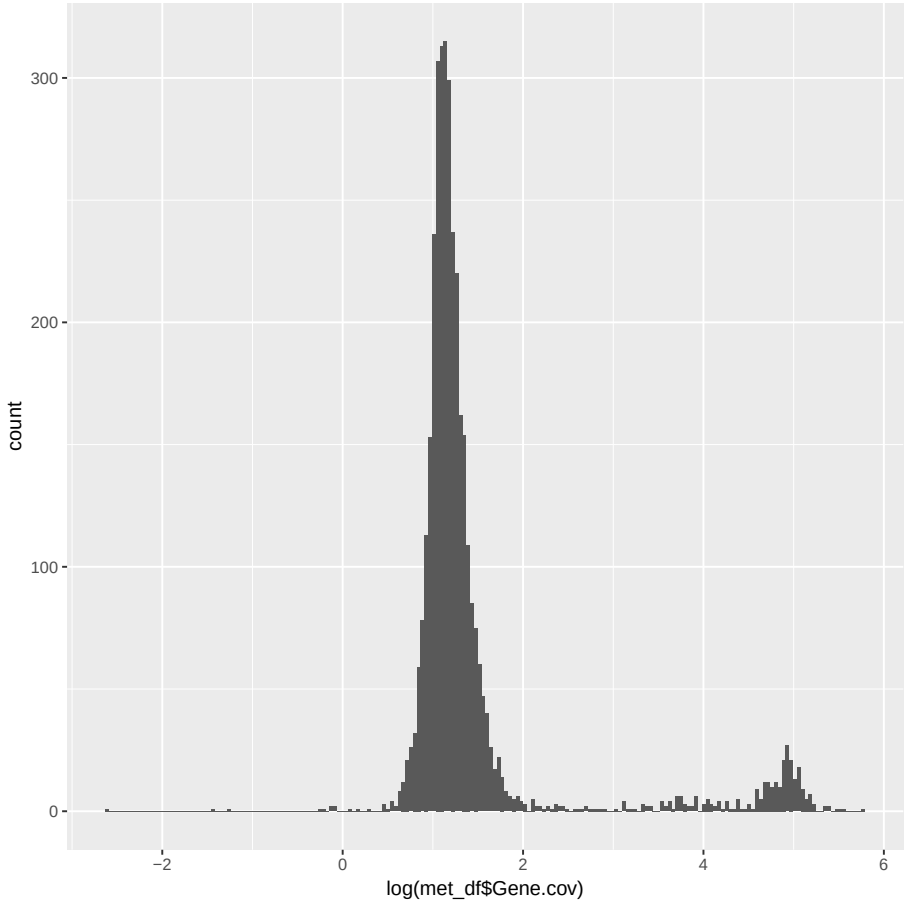
A continuació hem volgut dilucidar si la combinació de l'estat de metilació i acetilació dels fragments ens permet diferenciar entre gens variants/no-variants i entre els gens variants actius/incatius.

Hem classificat els gens com a variants utilitzant una llista obtiniguda en un estudi anterior. (La llista de gens variants es troba a "Gens_variants_extended.txt", 514 gens. Aquesta informació s'ha expandit a tots els fragments que formen part d'aquests gens, 5' i 3' inclosos.)

3.1 Transcripció i Metilació

Per a poder classificar els gens com a variants actius/inactius peimer hem mirat la distribució de les dades de transcripció i les de metilació. (amb l'idea de trobar llindars per a expressat/no expressat, metilat/no metilat).

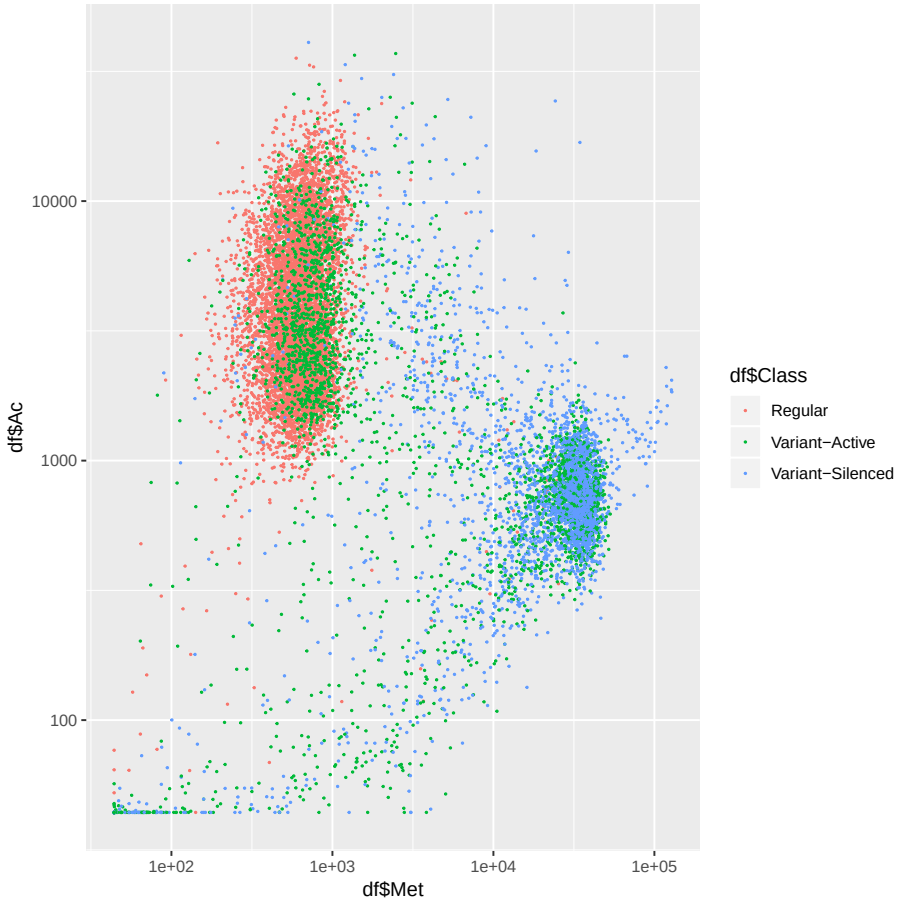
```
## Error in ggplot(trans_df, aes(log(trans_df$Aver.2Higher10G.))): object
'trans_df' not found
```


300
200
100
0
count
−2
0
2
4
6
log(met_df$Gene.cov)

3.2 Classificació segons Metilació

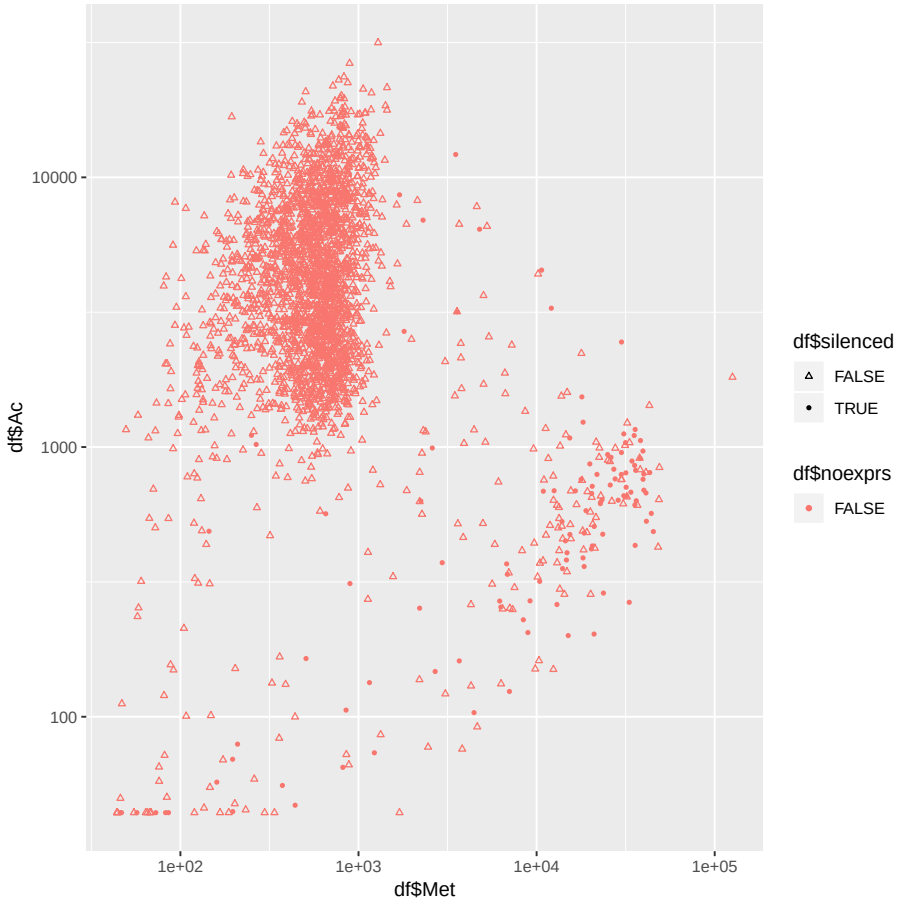
En un primer intent de diferenciar variants actius d'inactius hem creat el paremetre "silenced" que hem considerat "TRUE" si el valor de metilació del fragment està per sobre d'un llindar (>3). (Els valors de metilació utilitzats corresponen a l'ORF de cada gen i s'han expandit a tots els fragments que en formen part.)

```
##
##          Regular   Variant-Active Variant-Silenced
##           105873             5317             5467
```



3.3 Classificació segons Metilació i estat Transcripcional

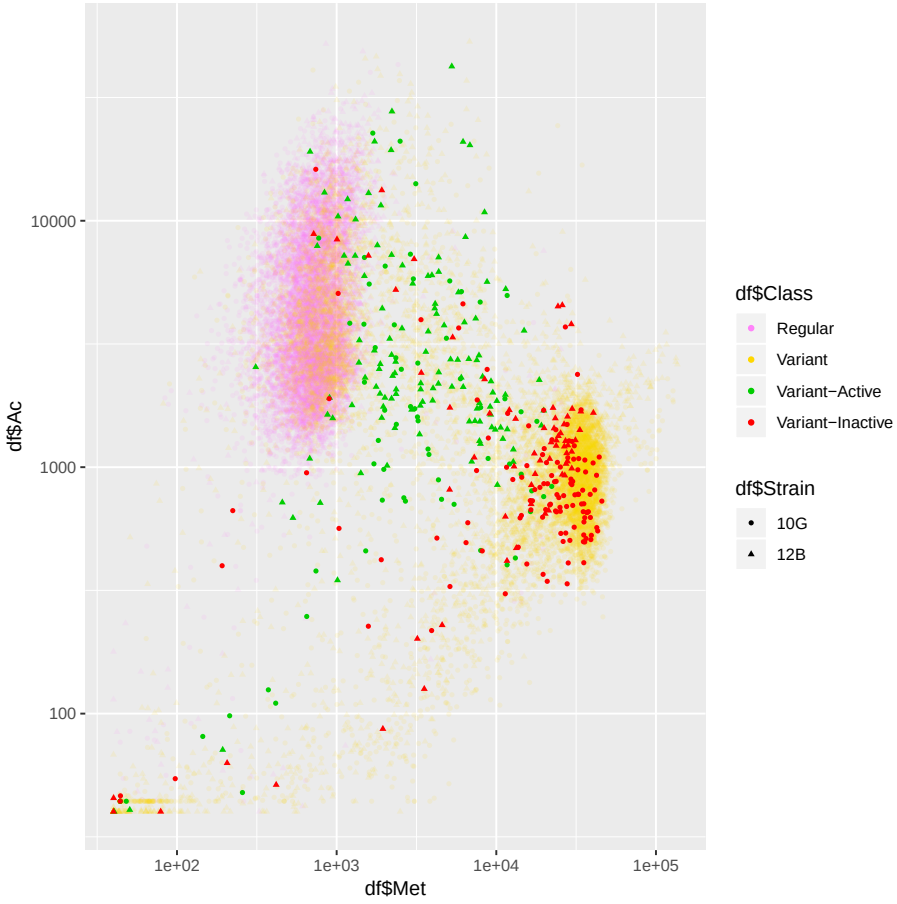
En un segon intent hem afegit la variable "noexprs" que hem considerat "TRUE" quan el valor d'expressió d'un fragment està per sota d'un llindar (<4)(Igual que en el cas anterior el valor de transcripció correspon a un gen i s'ha expandit a tots els fragments que en formen part.)(Dades de trascripció a l'arxiu "3D7_Variantome_AllData_withGam.xls" fulla 1 columnes 1 i 26).



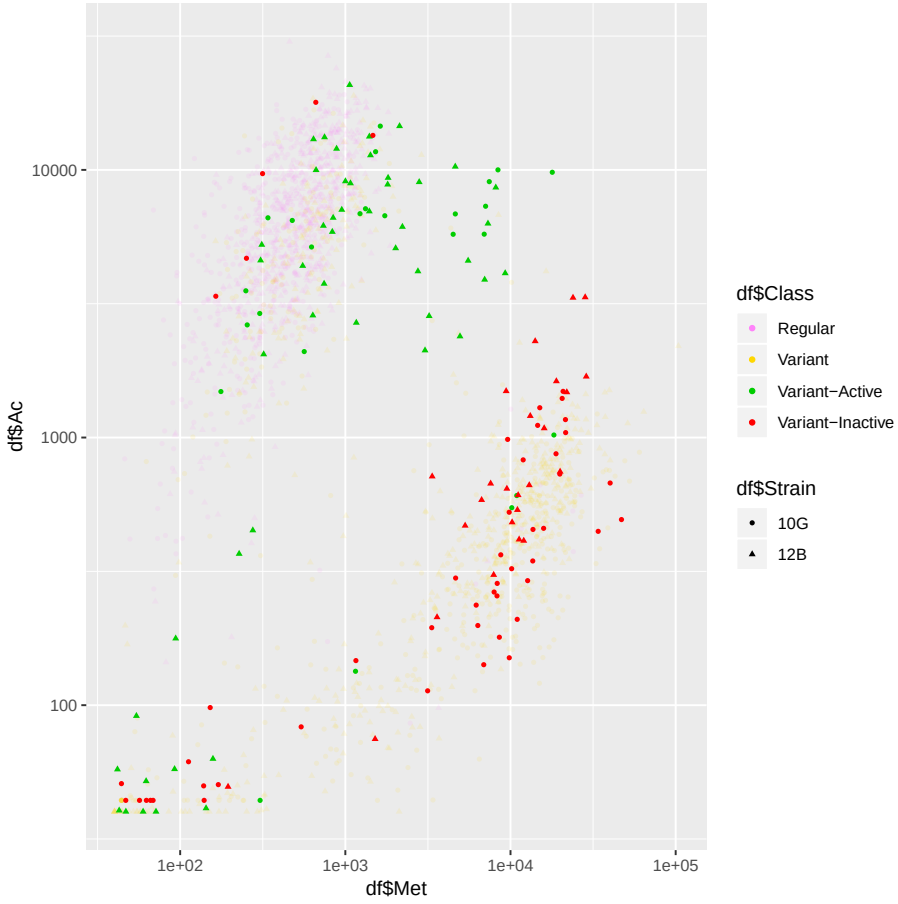
3.4 Gens diferencials

Finalment, vist que els anterior mètodes no enc classificaven satisfactòriament els gens com a variants actius o inactius, hem decidit centrar-nos en aquells gens que s'expressen diferencialment entre 10G i 1.2B. Els gens variants que estàn sobreexpressats en una soca respecte una altra els hem classificat com a actius i viceversa. Al fons del gràfic hem afegit la resta de gens (tots aquells que no tenen una expressió diferencial.) El que hi ha representat al gràfic són fragments de 200bp (tota la informació respectiva a gens s'ha traslladat als fragments que els representen). (Gens diferencials i nivells d'expressió a "Trans2.csv", 30 gens a la llista dels quals només usem els 20 amb majors diferències d'expressió.)

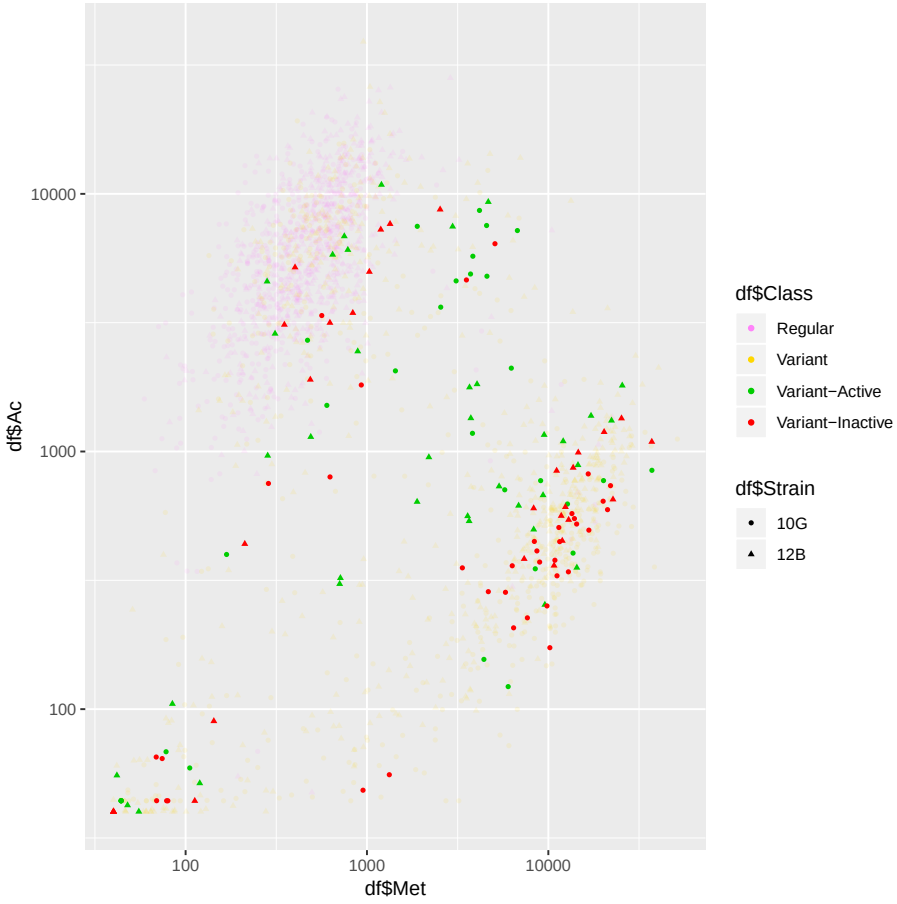
3.4.1 Gràfic de Gens diferencials: ORF



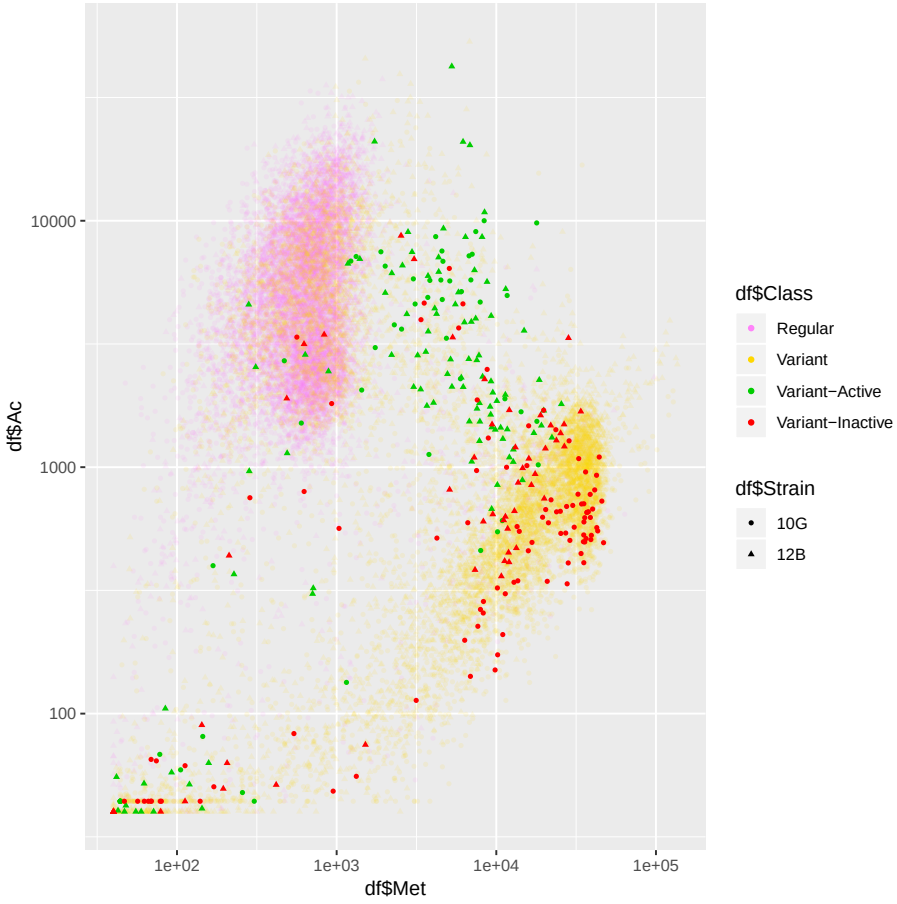
3.4.2 Gràfic de Gens diferencials: 5'



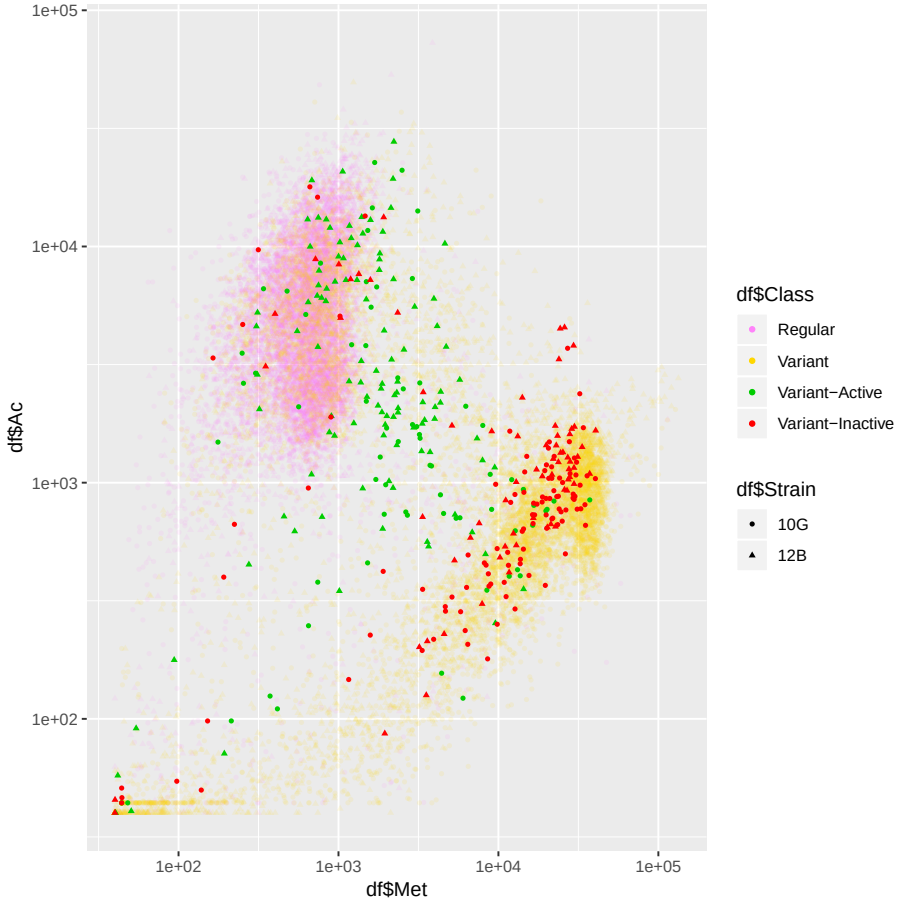
3.4.3 Gràfic de Gens diferencials: 3'



3.4.4 Gràfic de Gens diferencials: Només Var i Rifin



3.4.5 Gràfic de Gens diferencials: Excepte Var i Rifin



4 Model

Finalment hem volgut comprovar si amb les dades de metilació i acetilació podiem crear un model basat en regressió logística que fós capaç de classificar correctament gens variants i no variants.

```
##
##  FALSE   TRUE
## 105873  10784
## Analysis of Deviance Table
##
## Model 1: Variant ~ Ac + Met + Type + Start + Stop + silenced + noexprs
## Model 2: Variant ~ Ac + Met + Type + Start + Stop
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      8478     5697.8
## 2      8479     6339.3 -1  -641.59 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Call:
## glm(formula = Variant ~ Ac + Met + Type + Start + Stop + silenced +
##     noexprs, family = binomial(link = "logit"), data = train_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -7.5153  -0.7720   0.0073   0.1886   3.0433
##
## Coefficients: (2 not defined because of singularities)
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   8.063e-01  9.816e-02   8.214  < 2e-16 ***
## Ac           -1.297e-04  9.454e-06 -13.720  < 2e-16 ***
## Met           3.064e-04  2.234e-05  13.715  < 2e-16 ***
## Type5prima   -3.511e-01  9.901e-02  -3.546 0.000391 ***
## TypeORF      -1.102e+00  8.637e-02 -12.758  < 2e-16 ***
## Typeother    -3.843e+01  1.200e+03  -0.032 0.974462
## Start        -2.501e-07  4.152e-08  -6.025 1.69e-09 ***
## Stop                NA         NA      NA       NA
## silencedTRUE  4.148e+00  2.997e-01  13.840  < 2e-16 ***
## noexprsTRUE          NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11150.5  on 8485  degrees of freedom
## Residual deviance:  5697.8  on 8478  degrees of freedom
```

```
## AIC: 5713.8
##
## Number of Fisher Scoring iterations: 20
##
##         FALSE TRUE
##   FALSE  2965  291
##   TRUE   1046 4361
## [1] "Accuracy 0.845665473854323"
## [1] "Accuracy of null model 0.495786678979568"

## Error in library(ROCR): there is no package called 'ROCR'
## Error in prediction(predict, train_df$Variant):  could not find
function "prediction"
## Error in performance(ROCRpred, "tpr", "fpr"):  could not find function
"performance"
## Error in plot(ROCRperf, colorize = TRUE, text.adj = c(-0.2, 1.7)):
object 'ROCRperf' not found

##
## 3prima 5prima    ORF  other
##    170    104     16      1
##
## 3prima 5prima    ORF
##    138    197    711

## Error in pR2(model):  could not find function "pR2"
```