

Metilació i Acetilació a 10G i 1.2B

Alfred Cortés i Lucas Michel Todó

January 30, 2018

Contents

1	Introducció	2
2	Histogrames	3
2.1	log(Ac) All	3
2.2	log(Ac) 5'	4
2.3	log(Ac) 3'	5
2.4	log(Ac) ORF	6
2.5	log(Met) All	7
2.6	log(Met) 5'	8
2.7	log(Met) ORF	9
2.8	log(Ac) 3'	10
3	Acetilació vs Metilació	11
3.1	Transcripció i Metilació	12
3.2	Classificació segons Metilació	14
3.3	Classificació segons Metilació i estat Transcripcional	15
3.4	Gens diferencials	16
3.4.1	Gràfic de Gens diferencials: ORF	17
3.4.2	Gràfic de Gens diferencials: 5'	18
3.4.3	Gràfic de Gens diferencials: 3'	19
3.4.4	Gràfic de Gens diferencials: Només Var i Rifin	20
3.4.5	Gràfic de Gens diferencials: Excepte Var i Rifin	21
4	Model	22

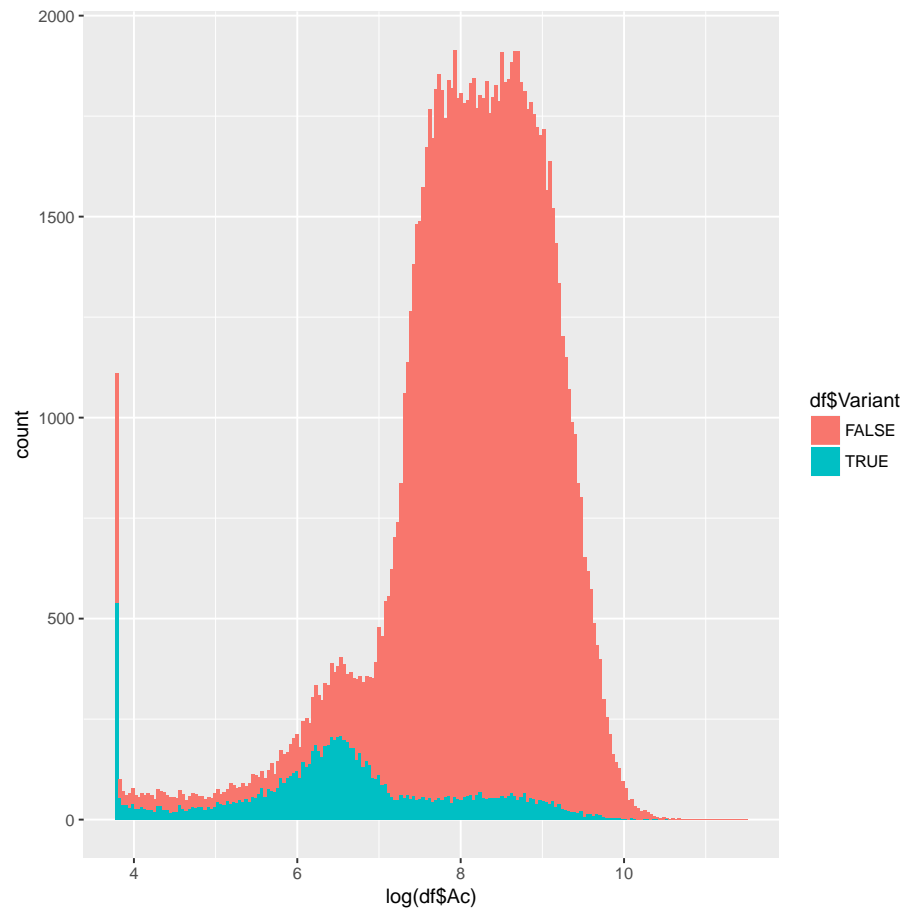
1 Introducció

Les dades representades en aquest set de gràfiques representen l'acetilació i metilació 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.

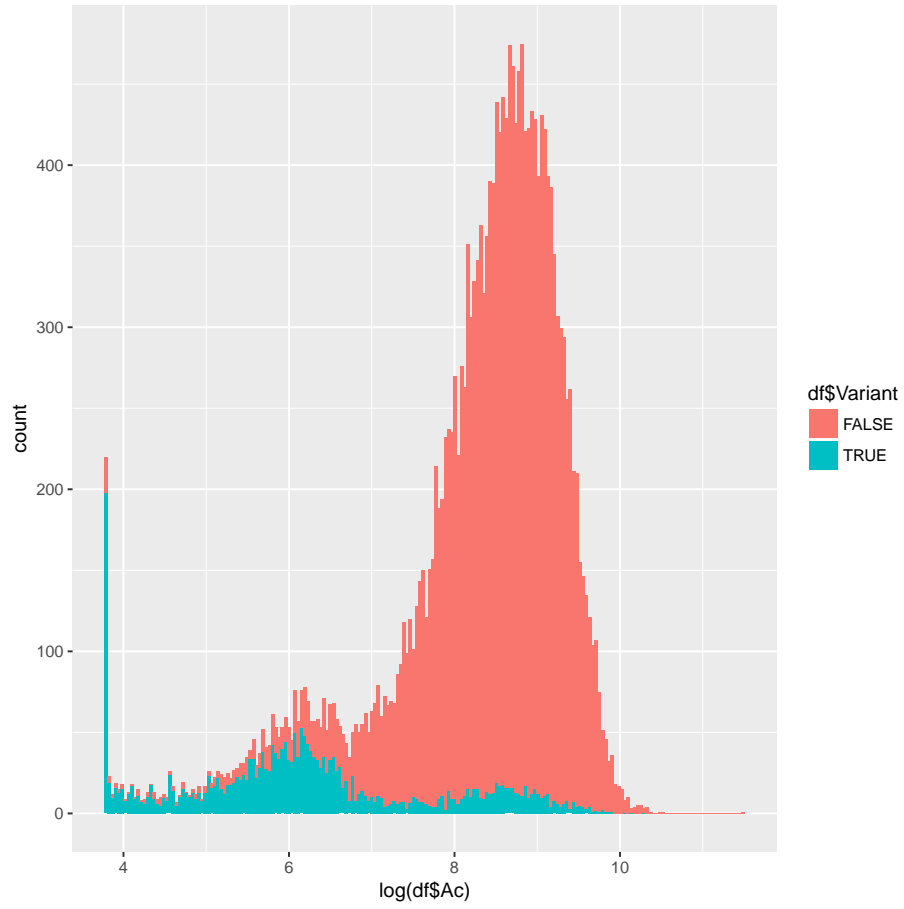
2 Histogrammes

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.)

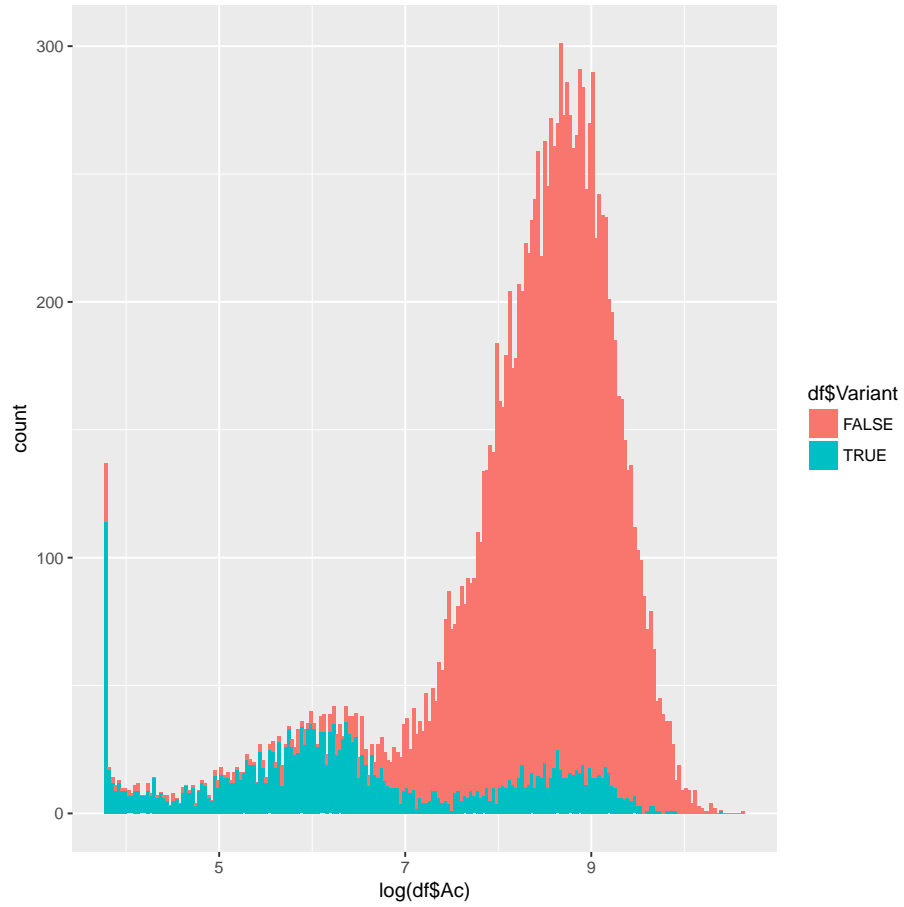
2.1 $\log(\text{Ac})$ All



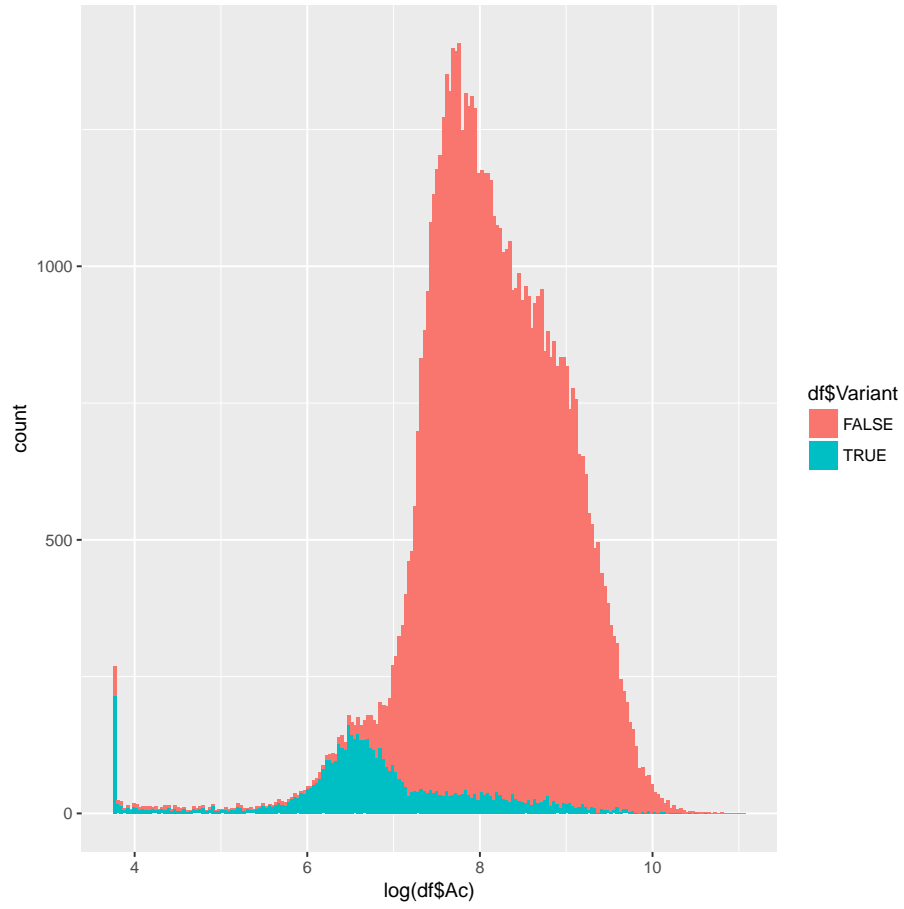
2.2 $\log(\text{Ac})$ 5'



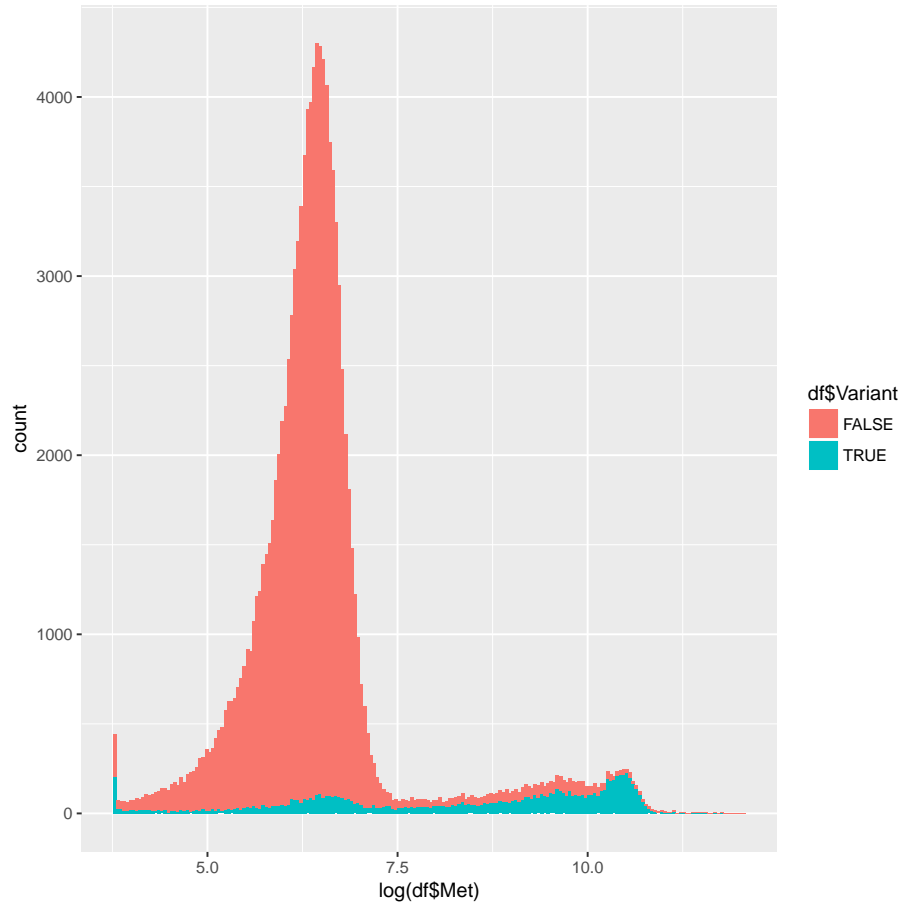
2.3 $\log(\text{Ac})$ 3'



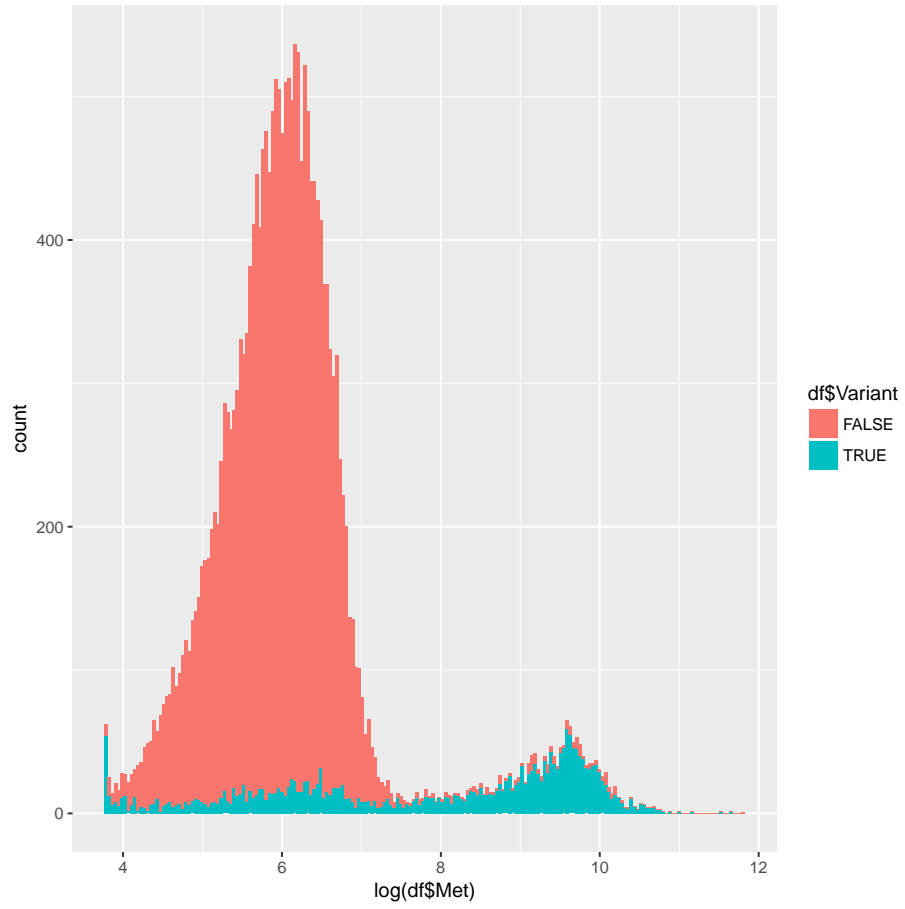
2.4 $\log(\text{Ac})$ ORF



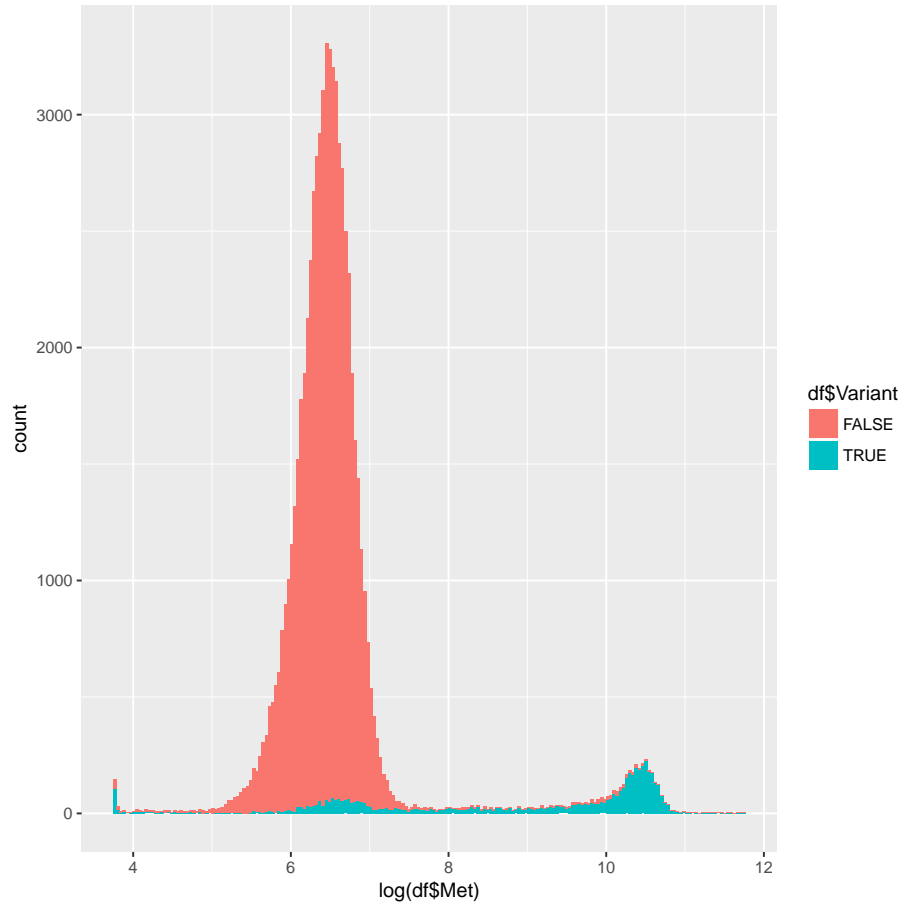
2.5 log(Met) All



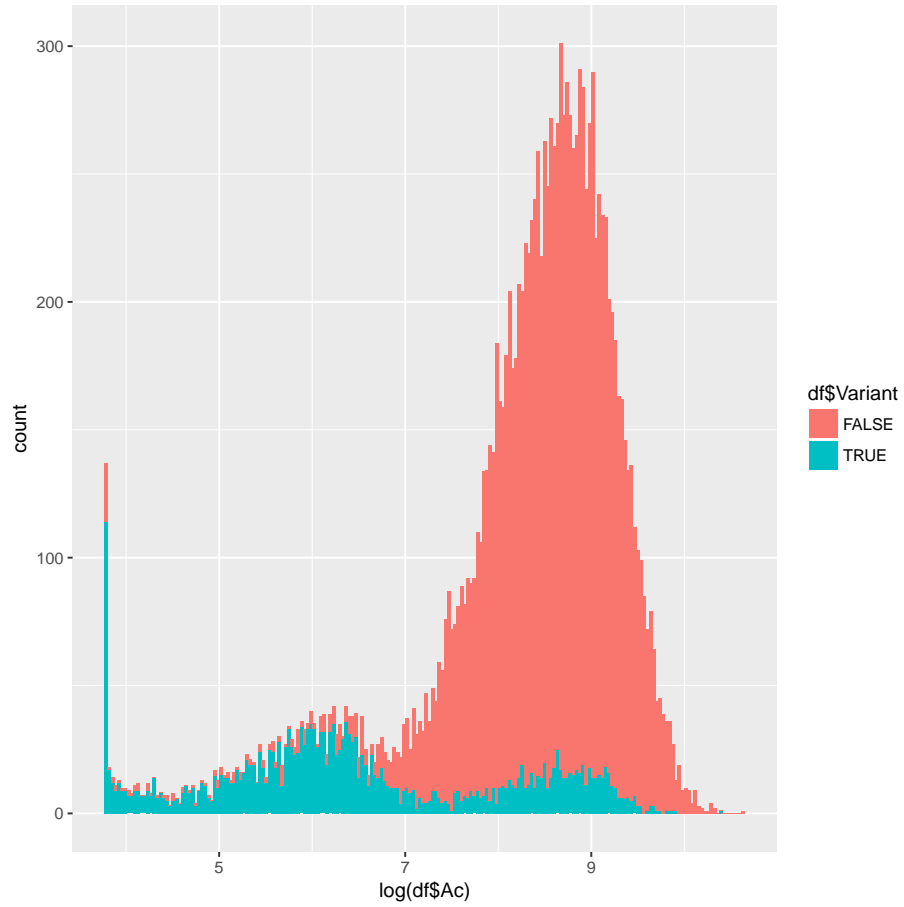
2.6 $\log(\text{Met})$ 5'



2.7 log(Met) ORF



2.8 $\log(\text{Ac})$ 3'



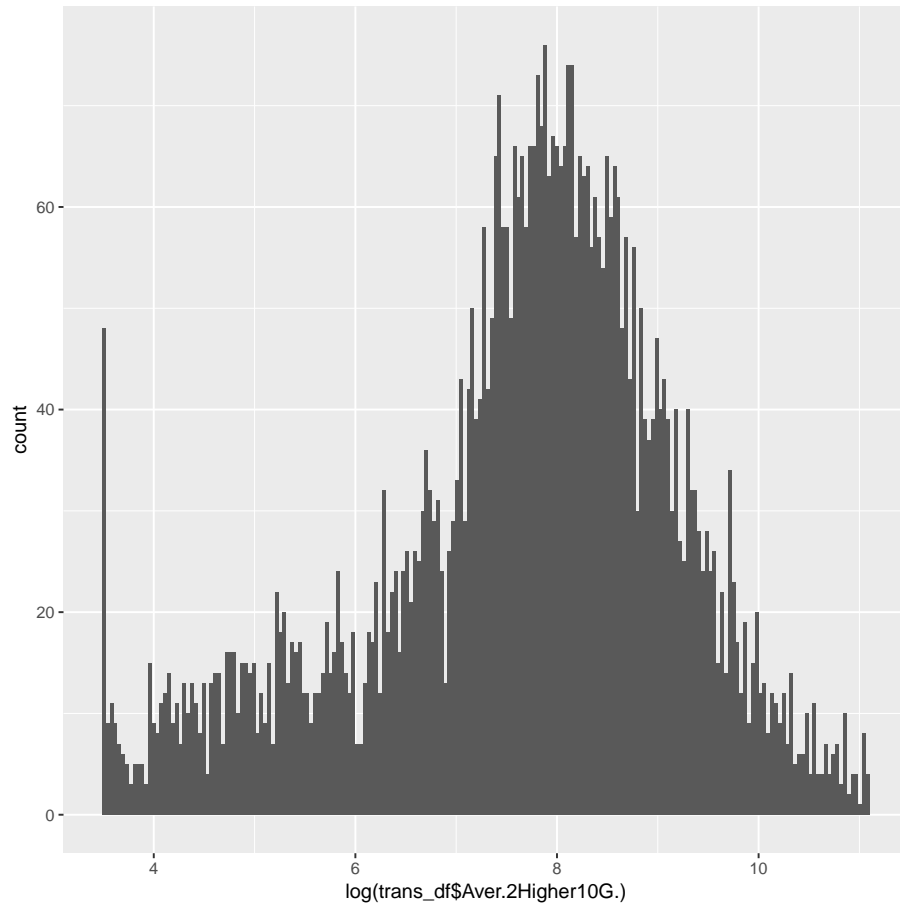
3 Acetilació vs Metilació

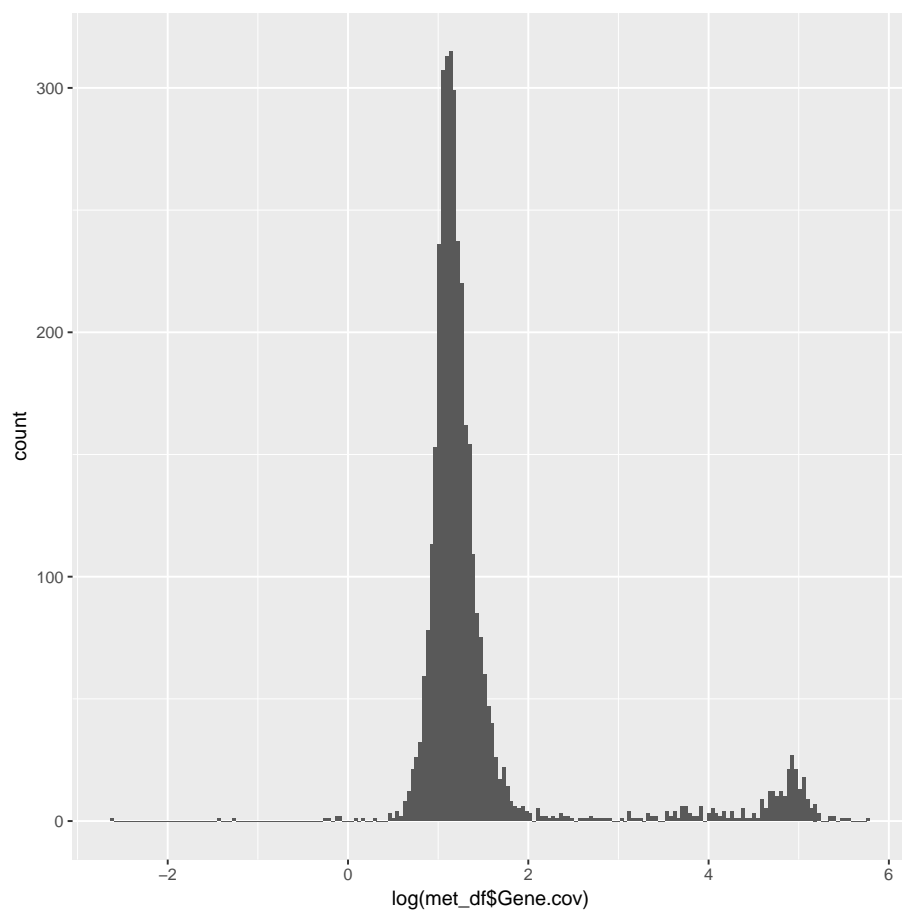
A continuació hem volgut dilucidar si la combinació de l'estat de metilació i acetilació dels fragments ens permet diferenciar entre gens variants/no-variants i entre els gens variants actius/incatius.

Hem classificat els gens com a variants utilitzant una llista obtinguda 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).

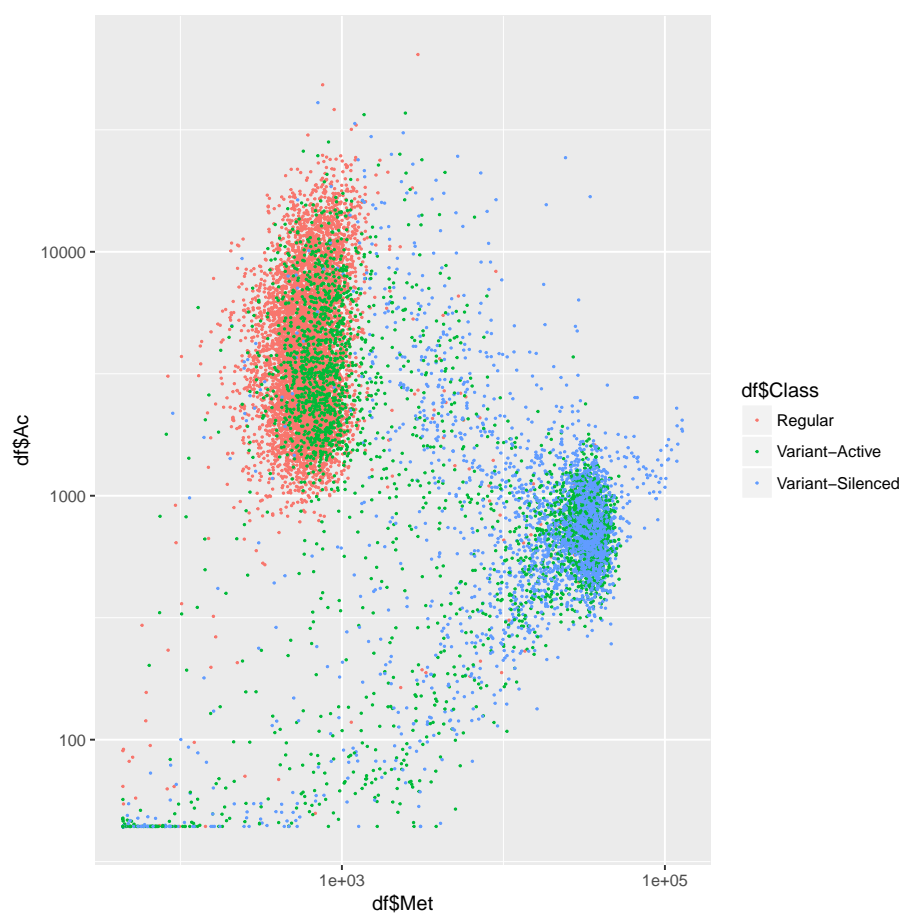




3.2 Classificació segons Metilació

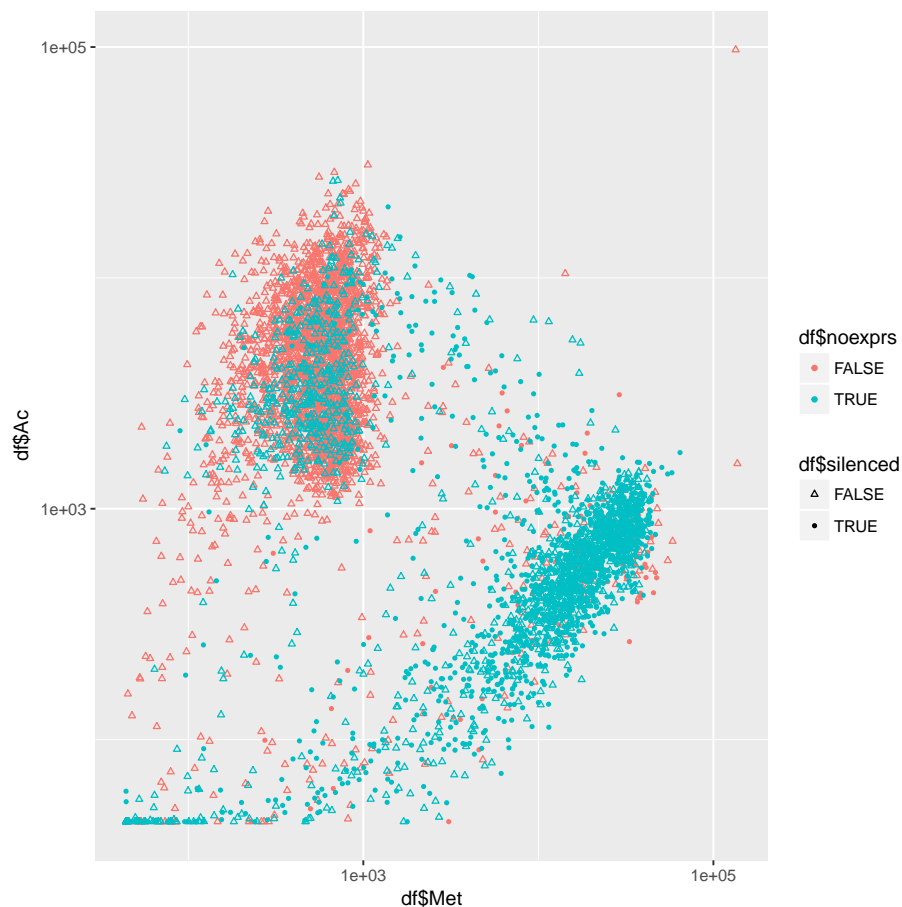
En un primer intent de diferenciar variants actius d'inactius hem creat el paremre "silenced" que hem considerat "TRUE" si el valor de metilació del fragment està per sobre d'un llindar (>3). (Els valors de metilació utilitzats corresponen a l'ORF de cada gen i s'han expandit a tots els fragments que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

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



3.3 Classificació segons Metilació i estat Transcripcional

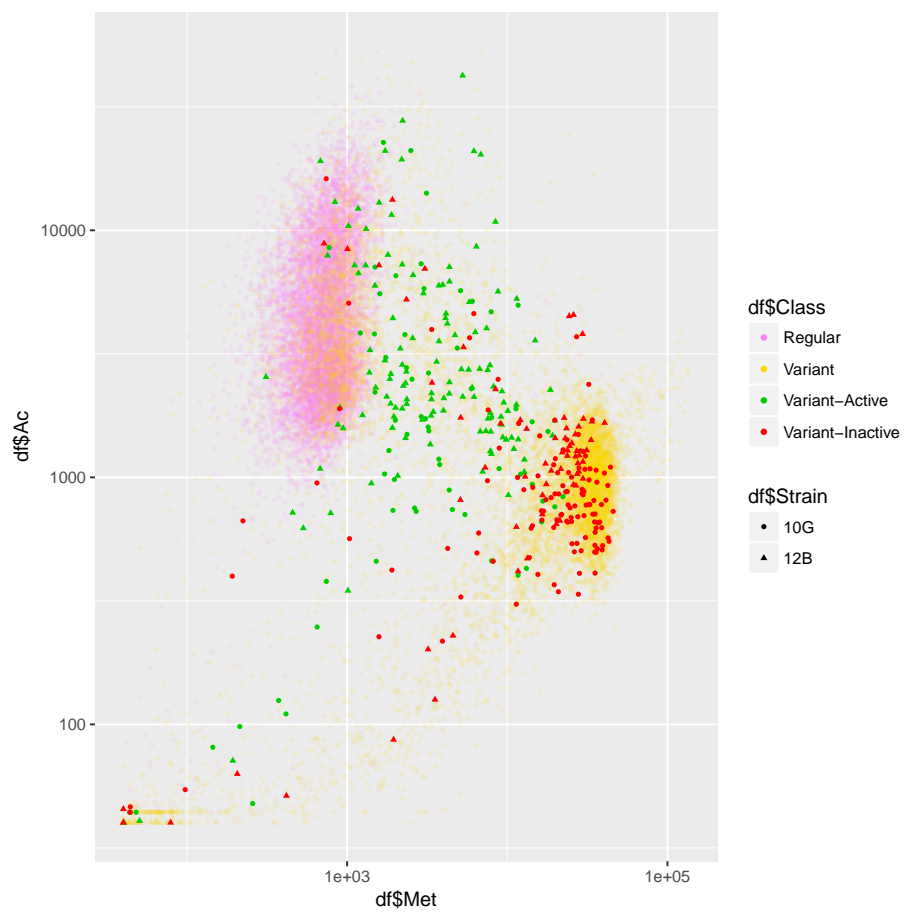
En un segon intent hem afegit la variable "noexprs" que hem considerat "TRUE" quan el valor d'expressió d'un fragment està per sota d'un llindar (<4) (Igual que en el cas anterior el valor de transcripció correspon a un gen i s'ha expandit a tots els fragments que en formen part.) (Dades de transcripció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



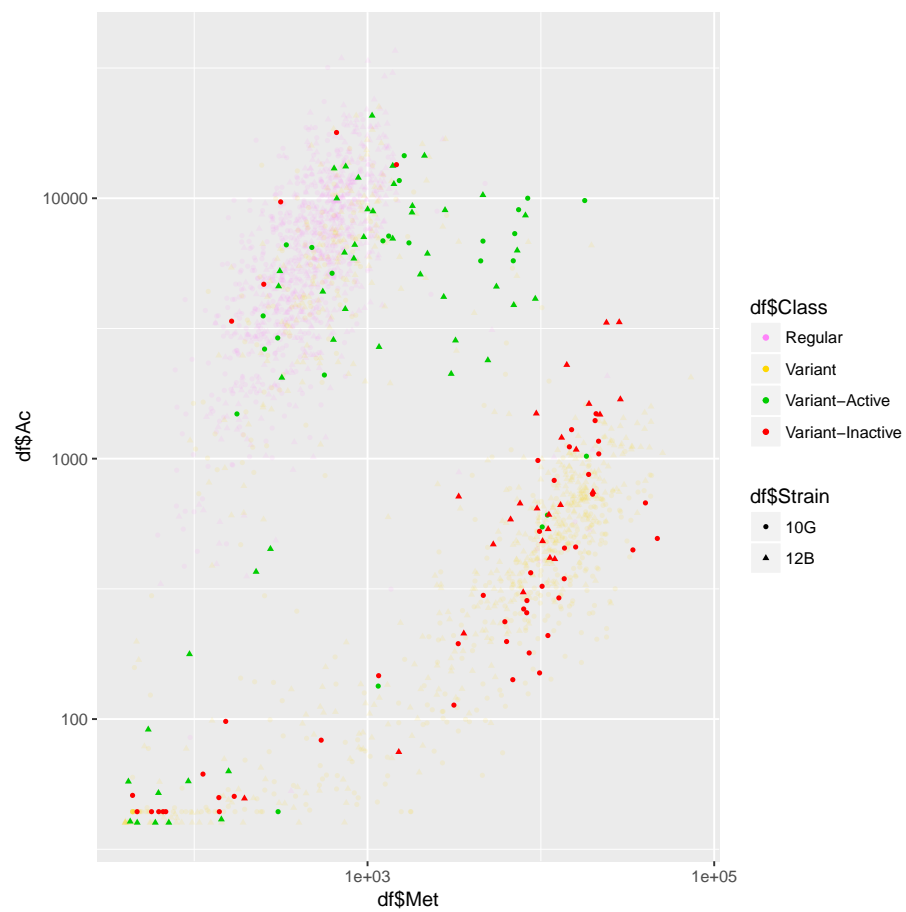
3.4 Gens diferencials

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.)

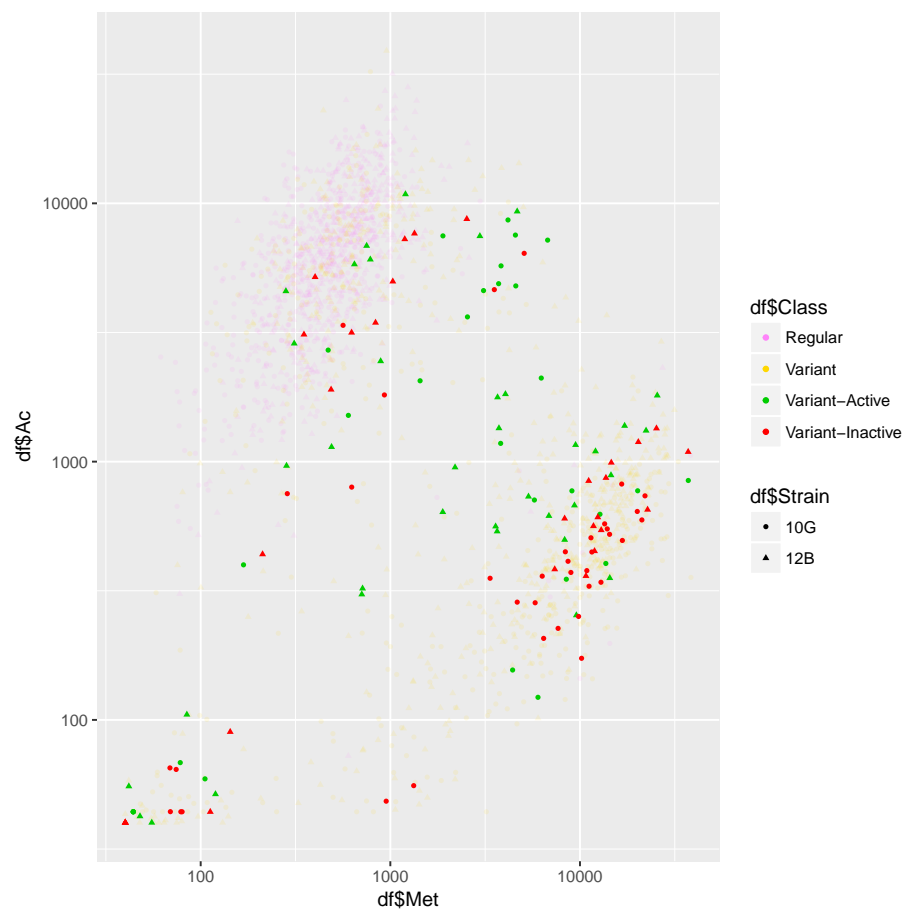
3.4.1 Gràfic de Gens diferencials: ORF



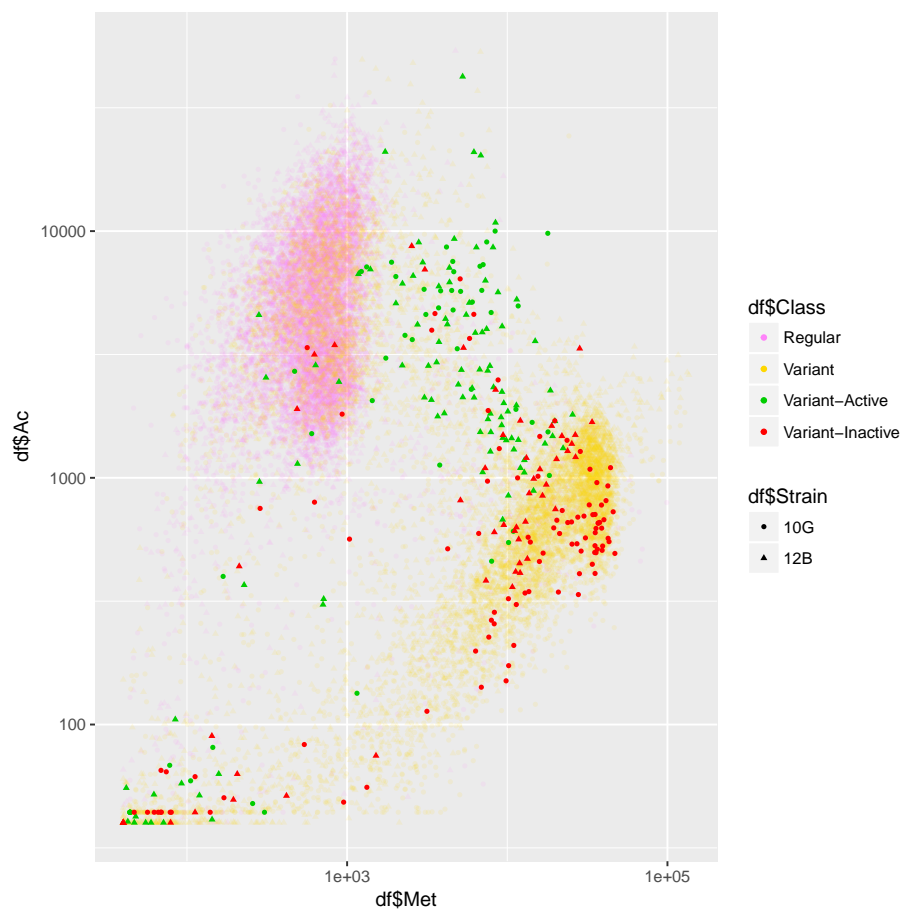
3.4.2 Gràfic de Gens diferencials: 5'



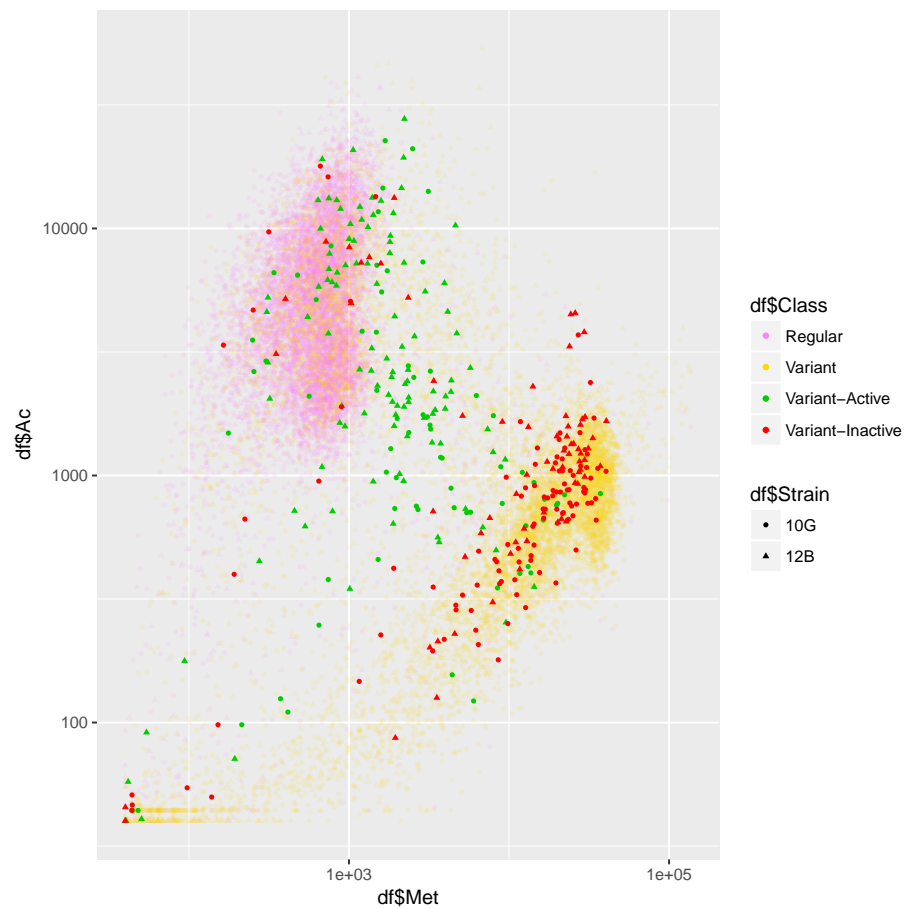
3.4.3 Gràfic de Gens diferencials: 3'



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 podem 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      8477      5352.6
## 2      8479      6066.9 -2   -714.26 < 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
## -5.1670  -0.7336   0.0001   0.0879   2.9123
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.052e-01  1.009e-01   4.017 5.90e-05 ***
## Ac          -1.201e-04  9.478e-06 -12.674 < 2e-16 ***
## Met           6.071e-04  5.022e-05  12.089 < 2e-16 ***
## Type5prima  -3.264e-01  1.009e-01  -3.234  0.00122 **
## TypeORF     -1.069e+00  8.907e-02 -12.007 < 2e-16 ***
## Typeother   -5.454e+01  2.080e+02  -0.262  0.79310
## Start       -2.331e-07  4.533e-08  -5.143 2.71e-07 ***
## Stop                NA             NA      NA      NA
## silencedTRUE  4.856e+00  4.348e-01  11.168 < 2e-16 ***
## noexprsTRUE   1.841e+00  2.811e-01   6.549 5.79e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 11212.4  on 8485  degrees of freedom
## Residual deviance:  5352.6  on 8477  degrees of freedom
```

```

## AIC: 5370.6
##
## Number of Fisher Scoring iterations: 16
##
##      FALSE TRUE
## FALSE 3056 142
## TRUE  1185 4280
## [1] "Accuracy 0.846819808380469"
## [1] "Accuracy of null model 0.495209511716495"

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

##
## 3prima 5prima  ORF
##    85    23    34
##
## 3prima 5prima  ORF
##    202    301   682
##      llh      llhNull      G2      McFadden      r2ML      r2CU
## -2676.3084504 -5606.1770458 5859.7371909 0.5226144 0.4986838 0.6801418

```

