GLM training

Compiled: May 15, 2022

```
library(ggplot2)
library(reshape2)
library(ggpubr)
library(dplyr)
```

Read Feature Matrix file

```
df <- read.csv("results/2022/Apr/GM12878_smartseq2.all.feature_matrix.tab.proc.csv", sep = "\t",
    header = T)
head(df)
```

SM	STROND	REP_COUNT	REP_COORDS	INDEL_LEN	variant_type	variant_id	error	AC	DP	RC	
sample	.	3	chr1:[631860,631861)/.	0	SNV	chr1:631860-631861:G>A	0.011	2	6	4	0.33
sample	.	3	chr1:[631861,631862)/.	0	SNV	chr1:631861-631862:G>A	0.011	6	6	0	1.00
sample	.	1	chr1:[1014227,1014228)/.	0	SNV	chr1:1014227-1014228:G>A	0.011	3	4	1	0.75
sample	.	1	chr1:[1254995,1254996)/.	0	SNV	chr1:1254995-1254996:C>T	0.011	2	4	2	0.50
sample	.	2	chr1:[1311887,1311888)/.	0	SNV	chr1:1311887-1311888:T>G	0.011	2	3	1	0.66
sample	.	2	chr1:[1318755,1318756)/.	0	SNV	chr1:1318755-1318756:G>A	0.011	2	2	0	1.00

```
dim(df)
```

```
## [1] 52511    23
```

```
table(df$LABEL)
```

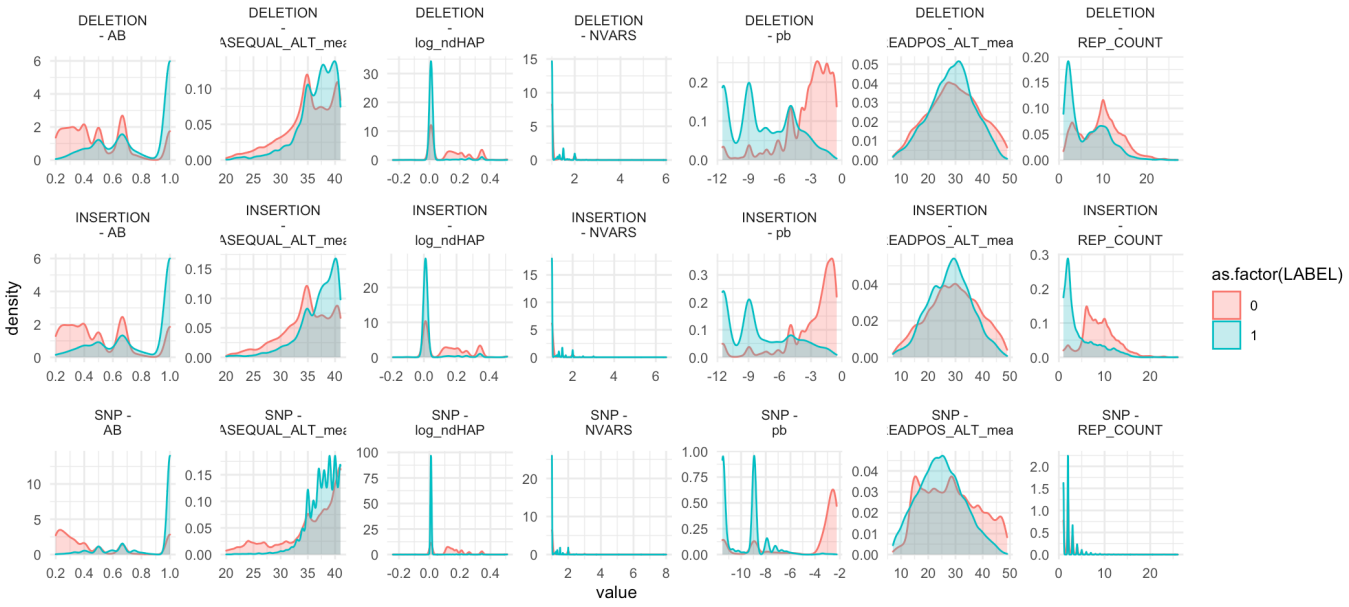
```
##
##     0     1
## 30301 22210
```

Feature distribution

```
df$log_ndHAP <- df$ndHAP + 0.01
df$VARTYPE <- ifelse(df$INDEL_LEN > 0, "INSERTION", "DELETION")
df$VARTYPE <- ifelse(df$INDEL_LEN == 0, "SNP", df$VARTYPE)

df1 <- df[c(c("log_ndHAP", "pb", "BASEQUAL_ALT_mean", "AB", "READPOS_ALT_mean", "NVARS", "REP_COUNT"),
    c("VARTYPE", "LABEL"))]
df0 <- melt(df1, id.vars <- c("VARTYPE", "LABEL"))

df0$varcomb = paste(df0$VARTYPE, df0$variable, sep = " - ")

g0 <- ggplot(df0) + geom_density(aes(x = value, fill = as.factor(LABEL), color = as.factor(LABEL)),
    alpha = 0.3, adjust = 0.75) + facet_wrap(varcomb ~ ., ncol = 7, scales = "free", labeller = label_wrap_gen(wi
dth = 8)) +
    theme_minimal() + theme(text = element_text(size = 10))

pdf("/Users/giovanni/hoffman_folder/micro_indel_project/FIGS/glm.training.pdf")
g0
dev.off()
```

```
## quartz_off_screen
##                 2
```

```
g0
```



GLM classification

```
df1 <- df[, c("VARTYPE", "LABEL", "QUAL_INSERTION", "QUAL_DELETION", "QUAL_SNP")]

df0 <- melt(df1, id.vars <- c("VARTYPE", "LABEL"))

df0$varcomb = paste0(df0$VARTYPE, df0$variable)

df0 = df0[df0$varcomb %in% c("SNPQUAL_SNP", "INSERTIONQUAL_INSERTION", "DELETIONQUAL_DELETION"),
    ]

df0$LABEL <- as.factor(df0$LABEL)

head(df0)
```

	VARTYPE	LABEL	variable	value	varcomb
12	INSERTION	0	QUAL_INSERTION	0.5531741	INSERTIONQUAL_INSERTION
21	INSERTION	0	QUAL_INSERTION	0.6077618	INSERTIONQUAL_INSERTION
26	INSERTION	1	QUAL_INSERTION	14.0299927	INSERTIONQUAL_INSERTION
38	INSERTION	0	QUAL_INSERTION	4.3243731	INSERTIONQUAL_INSERTION
42	INSERTION	0	QUAL_INSERTION	0.5865084	INSERTIONQUAL_INSERTION
53	INSERTION	1	QUAL_INSERTION	10.1986870	INSERTIONQUAL_INSERTION

```
label_df <- df0 %>%
    group_by(variable, LABEL) %>%
    summarise(n = n())
label_df$x <- c(5, 10, 5, 10, 5, 10)
head(label_df)
```

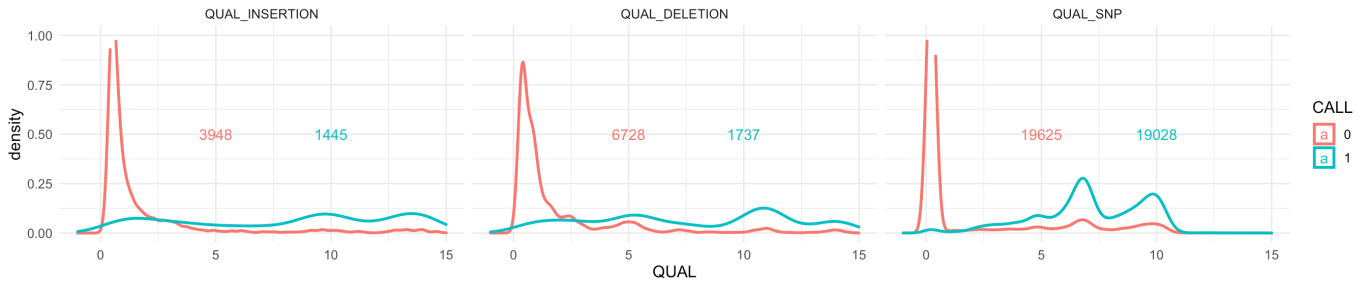
variable	LABEL	n	x
QUAL_INSERTION	0	3948	5
QUAL_INSERTION	1	1445	10
QUAL_DELETION	0	6728	5

variable	LABEL	n	x
QUAL_DELETION	1	1737	10
QUAL_SNP	0	19625	5
QUAL_SNP	1	19028	10

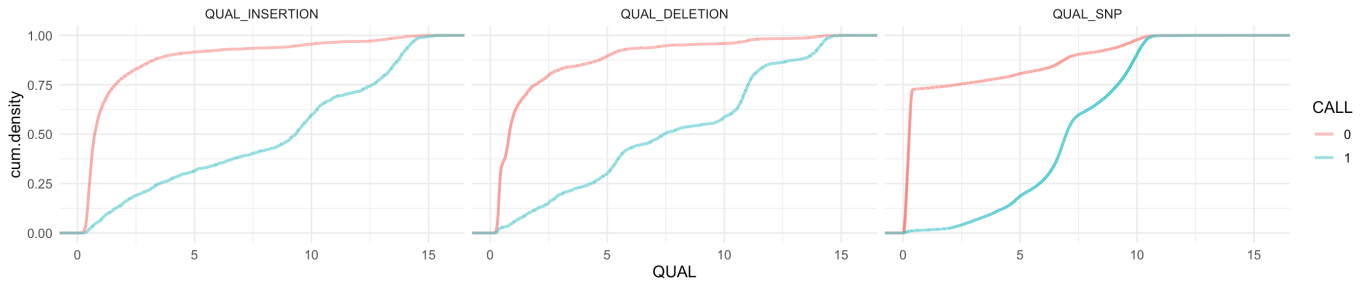
```
g1 <- ggplot(df0, aes(x = value)) + geom_density(aes(color = as.factor(LABEL)), alpha = 0.5, adjust = 0.95,
    size = 1) + labs(x = "QUAL", y = "density", color = "CALL") + geom_text(data = label_df, y = 0.5,
    aes(label = n, x = x, color = as.factor(LABEL))) + facet_grid(~variable) + scale_x_continuous(limits = c(-1,
    15)) + scale_y_continuous(limits = c(0, 1)) + theme_minimal() + theme(text = element_text(size = 12))

g2 <- ggplot(df0, aes(x = value)) + stat_ecdf(aes(color = as.factor(LABEL)), alpha = 0.5, adjust = 0.95,
    size = 1) + labs(x = "QUAL", y = "cum.density", color = "CALL") + facet_grid(~variable) + theme_minimal() +
    theme(text = element_text(size = 12))

pdf("/Users/giovanni/hoffman_folder/micro_indel_project/FIGS/glm.classification.pdf")
g1
g2
dev.off()
```

```
## quartz_off_screen
##                 2
```

```
g1
```




```
g2
```



Stutter noise analysis

```
head(df)
```

SM	STROND	REP_COUNT	REP_COORDS	INDEL_LEN	variant_type	variant_id	error	AC	DP	RC	
sample	.	3	chr1:[631860,631861)/.	0	SNV	chr1:631860-631861:G>A	0.011	2	6	4	0.33
sample	.	3	chr1:[631861,631862)/.	0	SNV	chr1:631861-631862:G>A	0.011	6	6	0	1.00
sample	.	1	chr1:[1014227,1014228)/.	0	SNV	chr1:1014227-1014228:G>A	0.011	3	4	1	0.75
sample	.	1	chr1:[1254995,1254996)/.	0	SNV	chr1:1254995-1254996:C>T	0.011	2	4	2	0.50
sample	.	2	chr1:[1311887,1311888)/.	0	SNV	chr1:1311887-1311888:T>G	0.011	2	3	1	0.66

SM	STROND	REP_COUNT	REP_COORDS	INDEL_LEN	variant_type	variant_id	error	AC	DP	RC	
sample	.	2	chr1:[1318755,1318756)/.	0	SNV	chr1:1318755-1318756:G>A	0.011	2	2	0	1.00

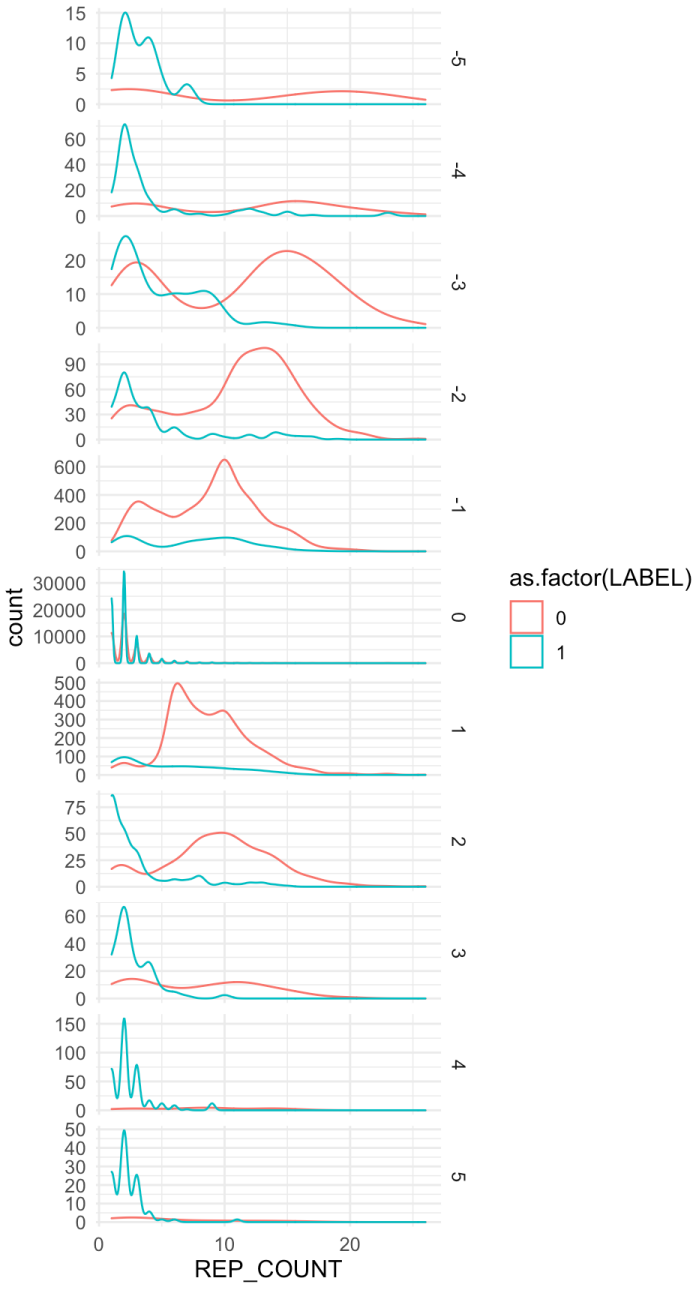
```
df0 <- df[, c("LABEL", "AB", "BASEQUAL_ALT_mean", "INDEL_LEN", "REP_COUNT")]
df0$LABEL <- as.factor(df0$LABEL)
df1 <- df0 %>%
    group_by(INDEL_LEN, REP_COUNT, LABEL) %>%
    summarise(n = n())
df1 <- df1 %>%
    dcast(INDEL_LEN + REP_COUNT ~ LABEL)
df1[is.na(df1)] <- 0

df1$`0` = df1$`0` + 1
df1$`1` = df1$`1` + 1
df1$n = df1$`0` + df1$`1`
df1 <- df1[df1$n > 50, ]
df1$FP_rate = log10(df1$`0`/(df1$n))
df1$vt <- df1$INDEL_LEN > 0


g1 <- ggplot(df0[abs(df0$INDEL_LEN) <= 5, ], aes(x = REP_COUNT)) + facet_grid(INDEL_LEN ~ ., scales = "free") +
    geom_density(aes(color = as.factor(LABEL), y = ..count..)) + theme_minimal() + theme(text = element_text(size
= 12))

g2 <- ggplot(df1, aes(x = REP_COUNT, y = -FP_rate)) + geom_smooth(method = "glm") + geom_line(aes(group = as.fact
or(INDEL_LEN),
    size = 3, color = as.factor(INDEL_LEN))) + geom_point(aes(size = n, color = as.factor(INDEL_LEN))) +
    facet_grid(vt ~ .) + scale_colour_brewer(palette = "RdYlBu") + theme_minimal() + theme(text = element_text(si
ze = 12))

m <- lm(df1$FP_rate ~ df1$REP_COUNT + abs(df1$INDEL_LEN))
coef(m)
```

```
##        (Intercept)      df1$REP_COUNT abs(df1$INDEL_LEN)
##        -0.32384522         0.02887569        -0.08001282
```

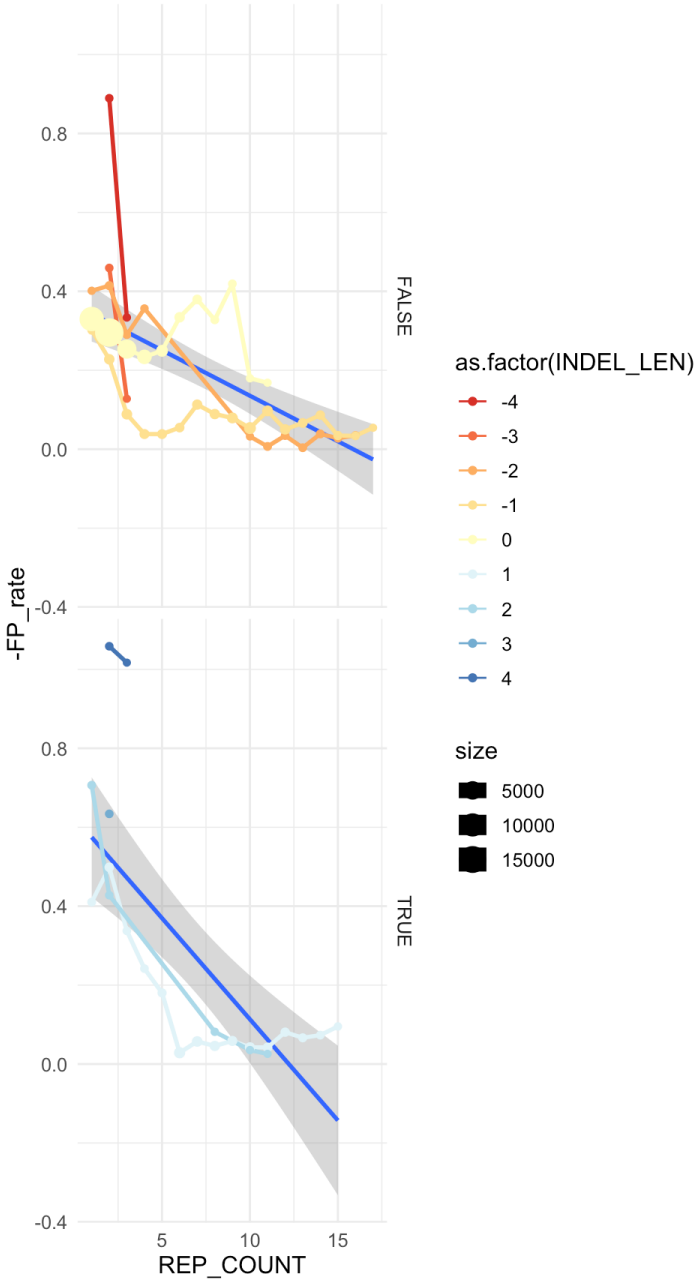
```
pdf("/Users/giovanni/hoffman_folder/micro_indel_project/FIGS/glm.stutter_noise.pdf")
g1
g2
dev.off()
```

```
## quartz_off_screen
##                 2
```

```
g1
```




```
g2
```


0.8
0.4
0.0
-0.4
FALSE
TRUE
-FP_rate
5
10
15
REP_COUNT
as.factor(INDEL_LEN)
-4
-3
-2
-1
0
1
2
3
4
size
5000
10000
15000